Supporting Exploratory Search Tasks Through Alternative Representations of Information

by

Bahareh Sarrafzadeh

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Computer Science

Waterloo, Ontario, Canada, 2020

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner: Rob Capra, Associate Professor
School of Information and Library Science,
University of North Carolina at Chapel Hill

Supervisor(s): Edward Lank, Professor
Cheriton School of Computer Science, University of Waterloo

Olga Vechtovoma, Associate Professor
Department of Management Sciences, University of Waterloo

Internal Member(s): Charles L. A. Clarke, Professor
Cheriton School of Computer Science, University of Waterloo

Edith Law, Assistant Professor
Cheriton School of Computer Science, University of Waterloo

Internal-External Member: Mark Smucker, Associate Professor
Department of Management Sciences, University of Waterloo

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

This dissertation includes first-authored peer-reviewed material that has appeared in conference and journal proceedings published by the Association for Computing Machinery (ACM). The ACM's policy on reuse of published materials in a dissertation is as follows:

> "*Authors can include partial or complete papers of their own (and no fee is expected) in a dissertation as long as citations and DOI pointers to the Versions of Record in the ACM Digital Library are included.*"

The following list serves as a declaration of the Versions of Record for works included in this dissertation:

Portions of Chapter 4:
Bahareh Sarrafzadeh, Olga Vechtomova, and Vlado Jokic. "*Exploring knowledge graphs for exploratory search*". In proceedings of the 5th Information Interaction in Context Symposium (IIiX 2014), Regensburg, Germany, 2014. ACM.
https://dl.acm.org/doi/abs/10.1145/2637002.2637019

Portions of Chapter 5:
Bahareh Sarrafzadeh, Alexandra Vtyurina, Edward Lank, and Olga Vechtomova. "*Knowledge graphs versus hierarchies: An analysis of user behaviours and perspectives in information seeking*". In proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval (CHIIR 2016), Carrboro, North Carolina, USA, 2016. ACM.
https://doi.org/10.1145/2854946.2854958

Portions of Chapter 6:
Bahareh Sarrafzadeh and Edward Lank. "*Improving Exploratory Search Experience through Hierarchical Knowledge Graphs*". In proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SigIR 2017), Shinjuku, Tokyo, Japan, 2017. ACM.
https://doi.org/10.1145/3077136.3080829

Portions of Chapter 7:
Bahareh Sarrafzadeh, Adam Roegiest, and Edward Lank. "*Cost of Errors in Exploratory Search Outcomes and Behaviors*". Submitted to the ACM Transactions on Information Systems (TOIS 2020), 2020. ACM.

Portions of Appendix A:
Bahareh Sarrafzadeh, Alex C. Williams, Olga Vechtomova, and Edward Lank. "*Fostering Crowdworker Reliability in Synthetic Labeling Tasks Via Qualitative Codebook Interaction Design*". In preparation for submission to the ACM CHI Conference on Human Factors in Computing Systems (CHI 2021), 2021. ACM.

Abstract

Information seeking is a fundamental component of many of the complex tasks presented to us, and is often conducted through interactions with automated search systems such as Web search engines. Indeed, the ubiquity of Web search engines makes information so readily available that people now often turn to the Web for all manners of information seeking needs. Furthermore, as the range of online information seeking tasks grows, more complex and open-ended search activities have been identified. One type of complex search activities that is of increasing interest to researchers is *exploratory search*, where the goal involves "learning" or "investigating", rather than simply "looking-up".

Given the massive increase in information availability and the use of online search for tasks beyond simply looking-up, researchers have noted that it becomes increasingly challenging for users to effectively leverage the available on-line information for complex and open-ended search activities. One of the main limitations of the current document retrieval paradigm offered by modern search engines is that it provides a ranked list of documents as a response to the searcher's query with no further support for locating and synthesizing relevant information. Therefore, the searcher is left to find and make sense of useful information in a massive information space that lacks any overview or conceptual organization.

This thesis explores the impact of alternative representations of search results on user behaviors and outcomes during exploratory search tasks. Our inquiry is inspired by the premise that exploratory search tasks require sensemaking, and that sensemaking involves constructing and interacting with representations of knowledge. As such, in order to provide the searchers with more support in performing exploratory activities, there is a need to move beyond the current document retrieval paradigm by extending the support for locating and externalizing semantic information from textual documents and by providing richer representations of the extracted information coupled with mechanisms for accessing and interacting with the information in ways that support exploration and sensemaking. This dissertation presents a series of discrete research endeavour to explore different aspects of *providing information* and *presenting this information* in ways that both extraction and assimilation of relevant information is supported.

We first address the problem of extracting information – that is more granular than documents – as a response to a user's query by developing a novel information extraction system to represent documents as a series of entity-relationship tuples. Next, through a series of designing and evaluating alternative representations of search results, we examine how this extracted information can be represented such that it extends the document-based search framework's support for exploratory search tasks. Finally, we assess the ecological validity of this research by exploring error-prone representations of search results and how they impact a searcher's ability to leverage our representations to perform exploratory search tasks.

Overall, this research contributes towards designing future search systems by providing insights into the efficacy of alternative representations of search results for supporting exploratory search activities, culminating in a novel hybrid representation called Hierarchical Knowledge Graphs (HKG). To this end we propose and develop a framework that enables a reliable investigation of the impact of different representations and how they are perceived and utilized by information seekers.

Acknowledgements

I'd like to thank my PhD supervisors, Edward Lank and Olga Vechtovoma, for their guidance, mentorship and encouragement during my PhD journey. I am thankful to Olga for providing opportunities in areas I have been long interested in: Natural Language Understanding and Generation of Ontologies. And I owe special thanks to Ed, who helped me transition to a new world where human perceptions and interactions with machines were as important as machines themselves, if not more. A world, which used to be very foreign to me but through exposure to the field of HCI and Ed's continuous support and guidance I believe it is not only my PhD focus that is broadened, it is also my own personal mindset and beliefs that have been transformed and in some ways I am now more open to move past my own biases and seek alternative viewpoints.

I'm also very fortunate to have had an outstanding committee. Charlie Clarke, Edith Law, Mark Smucker and Rob Capra. I'm grateful for their support and valuable feedback. They also contributed to having a pleasant and meaningful defence experience despite the special circumstances caused by the Covid-19 pandemic.

I've been incredibly fortunate to have worked with different researchers and mentors at Microsoft Research: Ahmed Hassan Awadallah, Milad Shokouhi, Susan Dumais, Ryen White, Michael Gamon and Sujay Jauhar. My internships at Microsoft Research allowed me to understand the meaning of impactful research and afforded me a privileged opportunity to pursue mixed-methods research on a scale not possible at a public institution. These collaborations helped me become confident about leveraging what I learned through my PhD research and apply it to new domains distant from Web Search. These opportunities have also paved the way for the next chapter of my life after PhD and I am very fortunate to have had this experience beyond my academic accomplishments.

I'd like to thank the members of the HCI Lab at the University of Waterloo, for helping create such a positive and supportive research community. I am immensely grateful for your friendship and encouragement. I would also like to thank my wonderful friend Maheedhar Kolla, for helping me navigate through the ambiguities of getting started with a PhD, pushing me to master the art of Information Visualization and making sure I wouldn't lose my courage after the many number of rejections I received throughout the years; and of course for all these years of friendship and great memories. As well, I am thankful to my research collaborators from both IR and HCI labs, Sasha Vtyurina, Adam Roegiest, Gaurav Baruah and Alex Williams for productive brainstorming sessions and your contributions to my research endeavors. Finally, there are many people who supported me during my PhD years, one way or another, in particular: Daniel Tunkelang for his support both leading to my first internship and hosting me during my second internship; Margaret Towell, Paula Roser and Angie Hewitt for their assistance with the PhD programs logistics and paperwork.

I owe my family a great deal of gratitude in supporting me and the decision to move to an entirely different country in pursuit of better opportunities and to be impactful and thrive in a new home.

And finally, my partner in crime, a true companion in all these years of living a life full of challenges, my best friend and soulmate, Saeed. I cannot imagine going through the thick and thin of the PhD without you by my side. I am so grateful we both went through this doctorate journey together, and during a global pandemic too, and we are now almost at the finish line.

In the end, I'd like to recognize the role of all local and franchise coffee shops where I spent unbelievable number of hours working on my papers and writing my dissertation, whether on Christmas Eve or the New Year's day, without them much of this thesis would not exist.

My PhD journey concluded during this unprecedented time of Covid-19 pandemic. It is still hard to believe how much every aspect of our lives has been affected by this global pandemic; the university closure, moving all classes, talks and PhD defences online, not being able to have anyone attend the defence in person including the committee members. The anxiety caused by the immense uncertainty around the logistics of a remote PhD defence and growing distress due to isolation made these last two months especially stressful. I was glad my defence proceeded as scheduled despite the pandemic, but I also felt a profound sense of loss. That I couldn't hug my supervisor and thank him for all his support in person; that I wasn't able to shake hands with all the committee members and thank them for being a part of this journey, and that I was unable to share this moment with those who mean a lot to me — my parents, my colleagues, and my best friends.

Yet, I am immensely grateful to live in a country which cares about the well-being of all of its residents, researchers who have devoted their lives to preventing and addressing pandemics like this one, and all those healthcare and essential workers, your dedication, commitment and courage deserve our deepest gratitude and admiration.

Dedication

This thesis is dedicated to Saeed, for his companionship, love and support. You are the best thing that has happened in my life.

To my beautiful mom, Niloofar, whose passion for science and the world of mathematics inspired me to become a researcher I am today and to my dad, Mohammad, a creative artist who taught me about the perception of beauties and the colors in this world. And finally my amazing sister, Mehrnaz for always being there for me.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

> *We shall not cease from exploration. And the end of all our exploring will be to arrive where we started and know the place for the first time.*
>
> – T.S. Eliot Little Gidding (1942, Part V)

We are living in a world that is increasingly more complex, which in turn has led to our continued dependence on information. The ubiquity of Web search engines makes information so readily available that people now often turn to the Web for all manners of information seeking needs. These information needs extend beyond simply looking up facts and include tasks such as acquiring new skills, learning about a domain, making informed decisions, solving complex problems, and other information intensive tasks. Some examples include understanding a medical problem and possible treatments for a loved one, acquiring a new technical skill to prepare for a job interview, or making sense of the literature in an unfamiliar domain. Information seeking is a fundamental component of many of the complex tasks presented to us, and is often conducted through interactions with automated search systems such as web search engines. As a result, alongside retrieval, the comprehension of information returned by these systems is a key part of decision making and action in a broad range of settings [355].

As the range of online information seeking tasks grows, more complex and open-ended search activities have been identified. One type of complex search activities that is of increasing interest to researchers is *exploratory search*, where the goal involves "learning" or "investigating", rather than simply "looking-up". In exploratory search, people seek to learn about a topic of interest or discover new information.

However, for searchers who are either unfamiliar with their problem domain, unfamiliar with the process needed to achieve their goal, or who lack a well-defined goal, there is a pressing need for assistance in search. Since the seminal work of pioneers in information seeking research (c.f. [63, 306, 556, 352, 597]) there have been many calls to move beyond the traditional turn-taking interaction model supported by major search engines, and toward support for human intelligence amplification and information use. Despite the widespread acknowledgement that – rather than just providing search results – search systems should provide effective browsing mechanisms [397, 257], to help users explore and examine the search results and to overcome uncertainty, there is still limited research into the design of alternative representations of search results that can help users explore, contrast, and learn, i.e. representations that best support exploratory search.

The recent SWIRL Workshop [129] emphasizes the need for continuing and ongoing research in supporting complex, evolving, and exploratory information seeking goals. This research requires advances in:

- algorithms to provide information;
- interfaces that represent this information;
- evaluation methods that support these goals.

My PhD thesis contributes towards designing future Exploratory Search Support (ESS) systems by providing insights into the efficacy of alternative representations of search results for supporting exploratory search activities, culminating in a novel hybrid representation called Hierarchical Knowledge Graphs (HKG). To this end we propose and develop a framework that enables a reliable investigation of the impact of different representations and how they are perceived and utilized by information seekers.

1.1 Thesis Statement

My thesis can be stated as follows:

> This thesis explores the impact of alternative representations of search results on user behavior during exploratory search tasks. Supporting complex and exploratory search tasks requires designing search systems that

> move beyond the current query-response paradigm in three main directions: (1) algorithms that move beyond document retrieval and provide information relevant to a user's query; (2) interfaces that move beyond a turn-taking interaction with a ranked list of documents and provide richer representations of the search results, as well as (3) mechanisms for accessing and interacting with them in ways that support exploration and sensemaking.

To elaborate, the retrieval of information, representation of this information, and interaction mechanisms that are supported by the search UI are the three pillars of modern search systems. For example, current search engines retrieve documents relevant to a user's query (i.e. retrieval of information units), represent the search results as a linear ranked list of documents (i.e. representation) and the search UI employs a simplified interaction model which allows the user to enter a few keywords in the search box to indicate their information need and then interact with the retrieved results by scrolling a ranked list and clicking on items that are perceived to be useful for satisfying their information need (i.e. interaction paradigm).

The research presented in this dissertation extends the current query-response paradigm by providing information at a more granular level than documents, i.e., through extracting semantic information from the content of retrieved documents; organizing this extracted information as non-linear, spatial representations that highlight salient concepts of the domain of interest and their relationships; and finally, designing interaction mechanisms that enable the searcher to explore the retrieved information and navigate through the space using two alternative navigation paradigms, i.e., Overview-First-Details-On-Demand [495] and Expand-from-Known [562].

We hypothesize that this extended framework enhances the information seekers' ability to acquire knowledge, and to investigate and make sense of the information space, essentially, to perform exploratory search tasks more effectively.

In order to accomplish the above tasks, this thesis specifically:

1. Explores the design of Information Extraction systems that can extract semantic information from a document as a series of entity-relationship tuples.

2. Assesses the potential benefit to exploratory search of the presentation of search results in the form of a knowledge graph.

3. Contrasts Knowledge Graphs (leveraging Expand-from-Known paradigm) with Hierarchies (leveraging Overview-first-Details-on-Demand paradigm) to understand the strengths and weaknesses of these alternative representations.

4. Proposes a novel representation of search results, a Hierarchical Knowledge Graph (HKG), that combines the benefits of both hierarchical and network representations into a single representation.

5. Assesses the impact of current error-prone IE algorithms on hierarchical knowledge graphs, demonstrating that they have good resiliency to typical error rates.

Figure 1.1 outlines the progression of this thesis around these five, primary research contributions. The following section provides context for these contributions, focusing specifically on the domain of exploratory search. We end this chapter by providing an overview of the research presented in this dissertation and how it is structured.

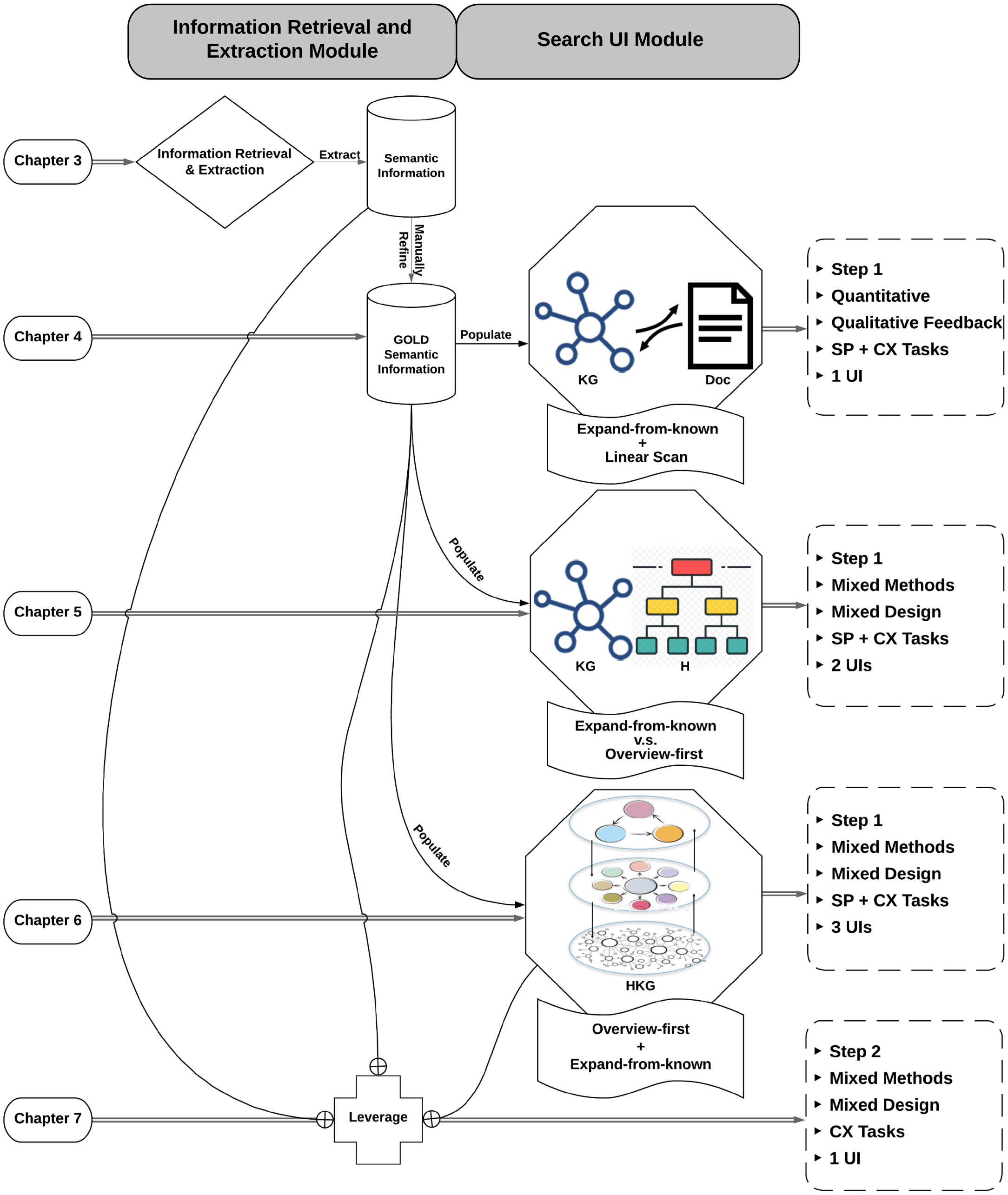


Figure 1.1: Progression of Thesis over the Chapters Based on the three main dimensions described in the Thesis Statement: Information Retrieval & Extraction, Representation and Interaction. Three main Representations of Search Results were developed throughout the thesis: Coupling of Documents with Knowledge Graphs (Doc $\Longleftrightarrow$ KG); Knowledge Graphs with Embedded Document Mappings; Hierarchies; and Hierarchical Knowledge Graphs (HKG). Our Evaluation Methodologies, depicted in dashed blocks, is based on the 2-step process described in Section 1.3, where Step 1 refers to assuming Perfect IE algorithms are in place and focus on the evaluation of representations and interaction mechanism only and in Step 2 we relax this simplifying assumption and contrast Error-prone and Perfect IE algorithms effects on Information Seeking Outcomes and Behaviors.

1.2 Supporting Exploratory Search

In order to motivate the need to go beyond the current query-response paradigm (as offered by all major search engines), we first need to define what exploratory search is and understand what support is needed for complex information seeking activities. The following subsections provide different pieces of a foundation, supported by past research, that is required in order to understand the motivation behind our effort for supporting exploratory search tasks.

1.2.1 What is Exploratory Search?

Search and exploration is a fundamental human activity that goes well beyond the digital domain. Humans are explorers by nature. This exploration is often derived from curiosity, i.e. as a means for bridging the gap between what we know and what we would like to know.

In defining exploratory search, researchers acknowledge that exploratory search covers a broader class of search activities than traditional IR and Interactive Information Retrieval (IIR), which targets query-document matching under the assumption that relevant information exists and that a well-formed query statement will retrieve it from the collection [597]. People engaged in exploratory searches are generally [605, 596, 352]: (1) unfamiliar with the domain of their goal (i.e., need to learn about the topic in order to understand how to achieve their goal); (2) unsure about the ways to achieve their goals (either the technology or the process); and/or even (3) unsure about their goals. Exploratory search is a specialization of information seeking, which describes the activity of attempting to obtain information through a combination of querying and collection browsing.

We adopt the following definition of exploratory search as *a type of information seeking and a type of sensemaking focused on the gathering and use of information to foster intellectual development.* [597]. This definition aligns well with work by Marchionini [352] who suggested that there exist two broad classes of exploratory search: learning and investigation.

In order to design systems that are effective in supporting exploratory search activities, we must first understand how searchers find specific facts, locate fragments of information or make sense of gathered information.

There are many theories in related domains that can collectively explain this exploratory behavior. Among them, information foraging [412], sensemaking [140] and berrypicking [44] are the most established theories that are essential to understanding how information seekers explore a domain and look for information. These related theories and different models of information seeking are reviewed in Chapter 2.

1.2.2 Going beyond Document Retrieval

When we examine, specifically, the idea of exploratory search as a type of information seeking to foster intellectual development, i.e. as learning or investigation tasks, one question we can ask is whether users' information needs really answered by a list of documents without any further processing? Consider a variety of everyday context that may trigger information seeking: planning a trip, starting one's own business or deciding which university is offering the best graduate program. In all of these scenarios solutions that help scope the information need to the most relevant subset of resources to consider, and improve the results presented in order to facilitate knowledge discovery and synthesis are needed.

The work in this thesis is not alone in questioning this simplistic assumption. Different models of information seeking [50, 305, 157, 158, 351, 110] indicate that the search process does not end with a set of retrieved documents. In fact, document retrieval is just the beginning of the information seeking journey. Past research [462] has shown that extracting relevant information from a set of documents and encoding them into structures or canonical forms required over 75% of the total time of the sensemaking process. Hence, any system support that mitigates some of these steps can significantly reduce the time and cognitive demands of information seekers during sensemaking and examination of search results. Chapter 2 reviews some of the relevant efforts that aimed at retrieving more focused answers, snippets or informative nuggets from the content of documents.

1.2.3 Why even a perfect search engine is not enough?

Web has become the default global repository for information. Perhaps the key technology that made the Web the default information finding medium is Search. Web search, enables users to find the information they want via the simplest of interaction paradigms: type some keywords into a box and get back an informative results list ranked in order of predicted relevance.

The success of the search engine enables us to do so much information discovery that it is difficult to imagine what we can do without it. Yet, in turn, successful paradigms can indeed constrain our ability to imagine other ways to ask questions or interact with the information space that might open up new and more powerful possibilities [611, 484]. That is, the prominence of Web Search based on the fundamental efficacy of keyword search makes it difficult to imagine *alternative search paradigms.*

It turns out, however, that this elegant paradigm is especially effective for the "1 minute search" [611], while many users have come to the Web for substantive research or exploration that can take hours or weeks.

Prior to the emergence of search engines, the Web was explored by links (i.e. hypertext). As recently as 2004, surfing the Web was still a common way for browsing the Web, following from link to link, from one site to another. Browsing was indeed the best practice for knowledge discovery as well as creating new knowledge. As the Web grew in size, it soon deemed to be beyond surfable, which in turn led to a shift towards "Searching by query". This growth itself is a result of the existence of search as it motivates publishing more and more content knowing that they would be accessible by keyword search.

Searching (by query) and Browsing are the two predominant paradigms for finding information on the Web. While they have shown to exhibit complementary advantages for information seeking, neither paradigm is adequate for complex information seeking scenarios if applied in isolation [397, 110]. Researchers have tried to support directed search (i.e. search by query) by attempting to build a "perfect" search engine: one that returns exactly what is sought given a fully specified information need. However, such a perfect search engine, despite perhaps impossible to flawlessly construct, might not be enough. There is a body of prior research that highlights information seeking scenarios that lend themselves to browsing as opposed to directly searching by query.

Exploration is a step by step process.

First and foremost, browsing is a natural way of exploration, information finding and knowledge discovery. Theories of information foraging [412] and berrypicking [44] describe the process of exploration as a step by step journey in the space, where the analysis of one step is needed to inform the next step to be taken [394, 351].

Browsing was the dominant way of looking for information and knowledge discovery before the advent of the Web and the search engines. While search engines drastically changed the way we perceive information finding, browsing based information seeking remains a viable way of exploration and information seeking.

Not all search types can be addressed by retrieving a target answer.

Search topics and search types have changed over the years, while the support from the search engines for query formulation and interaction with the search results haven't: Not only people still use few terms and few operators in their queries [514], but studies show that people typically view only the first few result pages [247], while more complex and exploratory queries often need multiple results. In fact, more complex queries require through examination of search results for extracting and integrating relevant information as well as sensemaking activities.

Even for known item search, *orienteering* is shown to be an effective, and often preferred, approach to locate what users are looking for, both in personal repositories (e.g. email or computers file system) as well as on the Web [535]. To elaborate, orienteering offers three main advantages over direct search or *teleporting*:

- Cognitive Ease: Orienteering can lessen cognitive burden during search activities by eliminating the need to articulate exactly what the searchers are looking for; narrowing the space they needed to explore.
- Sense of Location: the relatively small steps taken in orienteering can allow searchers to maintain a sense of where they were, helping them to feel in control, to know they are traversing in the right direction with the ability to backtrack, and to feel certain they have fully explored the space when unable find what they were looking for.
- Understanding the Answer: orienteering can also give searchers a context for their results they find [152, 331] and help them get a sense of how trustworthy those results are [594].

Not all information needs can be expressed as a well defined query

Even a perfect search engine would depend on a user's ability to formulate clear and well-specified queries. There are a variety of scenarios, however, that do not lend themselves to keyword search [397, 535, 351]:

- Searcher is not sure what they are looking for until the available options are presented;
- Some information needs are difficult to formulate as a clear query;

- Lack of appropriate terminology to formulate a query;
- Information can be gained along the way and not at the final step only (Accumulating results as opposed to landing at a final set of results);

Once we move towards more complex search tasks we encounter a variety of search scenarios that lend themselves to both search by query and search by navigation paradigms, whether due to the ambiguity of the information need or the need to accumulate results and improve the understanding of the space as opposed to landing at a final set of results. While current search engines are becoming more efficient at supporting search by query scenarios, we argue that they lack sufficient browsing support which is essential for examining the retrieved search results and sensemaking activities.

1.2.4 Supporting Browsing or Search by Navigation

We described the interplay between the two paradigms for finding information on the Web, that is search by query and search by navigation or browsing. While searching requires the user to translate their information need into queries, browsing accommodates the knowledge gap between what the user is able to communicate and what the system requires to find the desired information [50, 397]. This knowledge gap is more evident when the information need is more complex.

Supporting browsing and navigation is closely coupled with structuring the information into meaningful representations and groupings. Even beyond the Web and online information seeking activities, our understanding of the world is largely determined by our ability to organize information. We organize to understand, to explain and to control. The way we organize, label and relate information influences the way people comprehend that information [611, 452].

Going back to supporting exploratory search activities on the Web as the main focus of this thesis, we are inspired by the past research on stages of exploratory search and theoretical models that characterize the process of sense making, that is at the core of examining the search results and assimilation of the information into the information seeker's knowledge of the problem domain. In order to design systems that are effective in supporting exploratory search, we first must understand how users find specific facts, locate fragments of information and make sense of the collected information pieces.

Sensemaking can be viewed as a process of creating a representation of a large volume of information. These representations provide an overview of the underlying content, while creating a way for organizing and accessing the accumulated information. An important observation about many sensemaking methods is that they are *anytime algorithms* [464]. Anytime methods provide the best solution that they can find in given limited time. Given more resources, they continue to search for better solutions. Russell et al. call the common use of such tradeoffs in information processing tasks the "anytime principle".

Based on this principle, a reduction of cost (or increase in gain) associated with a step frees time to invest in other steps. We argue that, if search systems provide richer representations of the retrieved search results for information, that are better suited for investigation and knowledge synthesis, then more effort may be placed into other steps of the information gathering and sensemaking process (e.g., finding more pieces of relevant information or more in-depth analysis of acquired information). In other words, providing meaningful organizations of information and creating representations of search results can assist the sensemaking process in order to understand these results and synthesize new knowledge. As well, these organizations are essential for browsing since they accommodate the state of knowledge and can help users to get to the point to be able to formulate accurate queries.

However, the simplistic assumption that any automated assistance with generating representations of search results will improve searcher's performance, is not true. In fact, any search tool or information representation that does not match the searcher's perception of the information space or the requirements of the task at hand, can actually slow down the performance. For example, clustering documents in a collection is a fast and scalable methods that can provide some groupings for a set of retrieved documents. However, it is shown [215] that generated clusters often do not make sense to searchers and the searchers might need to begin a potentially long and complex negotiation with the system to correct the misalignments [463].

1.2.5 Exploring Alternative Representations of Information

In order to provide information seekers with more support in performing exploratory activities we have motivated the need to go beyond document retrieval and provide relevant information as well as the necessity to provide more support for browsing and search by navigation through utilizing richer representations of search results and more advanced interaction mechanisms. Wilson et al [611] emphasize the significance of research and evaluations of alternative approaches to data exploration for knowledge discovery and synthesis as the best preparation for the next generation of the Web. In fact, evaluation of representations of knowledge and mechanisms for accessing and manipulating this knowledge is an integral part of designing any search system that is effective in supporting exploratory search tasks. The search systems support information seeking by structuring knowledge and constraining access.

The way that knowledge is organized and made available affects the way that information seekers are able to access this knowledge and thus their information seeking performance [351]. To elaborate, all search systems offer specific features that define and constrain search. For example, a book provides a table of content that supports direct search for specific topics and encourages scanning and linear reading. Similarly, digital search systems can support linear, hierarchical or network structures for representing content and have the potential of providing alternative representations according to users' needs. Understanding ways to organize information that is meaningful to the searchers and can improve the outcome of exploratory information seeking is not a trivial task and requires extensive research to contrast alternative representations and their efficacy for supporting complex search tasks.

1.3 Research Overview

Our thesis statement stated in Section 1.1 highlights three elements of designing search systems that are essential for supporting users during exploratory search tasks: [$Information$], i.e., the units of information that are retrieved in response to users' information need, and [*Representation* and $Interaction$], i.e., the ways these information pieces are structured and presented to the user such that it can facilitate the process of exploration, sensemaking and obtaining knowledge. Essentially, the main research question we are investigating in this dissertation is "*how differences in the ways that information is represented impacts the users' behaviors and outcomes of exploratory search tasks?*". To this end, we propose algorithms that extract semantic information from the text of retrieved documents and develop specific representations of these results that can be leveraged by information seekers more effectively to perform learning and investigation tasks, as two main components of exploratory search.

In order to ensure the validity of our methodology to investigate our main research question, in Section 1.3.1 we propose a framework for all of our experiments that describes our efforts for extending the current query-response paradigm as well as the holistic evaluation approach we leverage to assess the efficacy of tools we developed.

1.3.1 Proposed Framework

We propose an experimental framework based on two main modules that we develop in order to extend the current query-response paradigm into a platform that we hypothesize is more effective in supporting exploratory search tasks. We broadly characterize these two modules as an (1) *Information* Retrieval Module and a (2) Search UI Module. Our characterization of these two modules are inline with how Marchionini [351] envisions *search system* as a source that retrieves and represents knowledge and provides tools and rules for accessing and using that knowledge. In our framework, the IR module requires both a document retrieval and ranking component as well as an Information Extraction (IE) component. As such, we refer to this module as Information Retrieval and Extraction (IRE) to distinguish it from the current Document Retrieval model employed by search engines that is more commonly referred to as Information Retrieval (IR). The Search UI, on the other hand, is an *interface* that provides rich representations of the information that is extracted by the IRE module and supports an interaction paradigm that is enabling and restricting access to this information.

A final aspect of developing this experimental framework is the decision regarding the choice of evaluation methodologies that enable a holistic assessment of the impact of each of these modules on searchers' behaviors as well as the outcomes of exploratory search tasks. Our proposed approach to evaluating the impact of different representations of the search results on the searchers' behaviors and outcomes is a two step process that separates the effect of the performance of the IRE module from the efficacy of the Search UI (i.e. the ways that the retrieved information is represented and the interaction mechanisms that are supported). In other words, we study representations independently from extraction methods that generate them.

This separation is valuable as it provides an opportunity to observe the impact of different representations of the search results on the outcome of complex information seeking scenarios independent of the accuracy of the algorithms that produce the underlying data. That is, assuming there is a perfect IRE module in place, which retrieves documents for a search query and extracts semantic information that is as accurate as the source content, we can solely focus on assessing how different ways of structuring and presenting this information impacts users exploratory behaviors and outcomes. As a result of the Step 1 of our evaluation strategy we can identify the representations of search results and interaction mechanisms that are effective in supporting users during exploratory search tasks. Next, the Step 2 of the evaluation process involves deploying a Search UI module that leverages our ideal version of the search results representation and then experiment with varying levels of IRE module's output accuracy along with the same search UI across experiments in order to study the impact of errors in the output of the IR module on the overall outcome of exploratory search tasks.

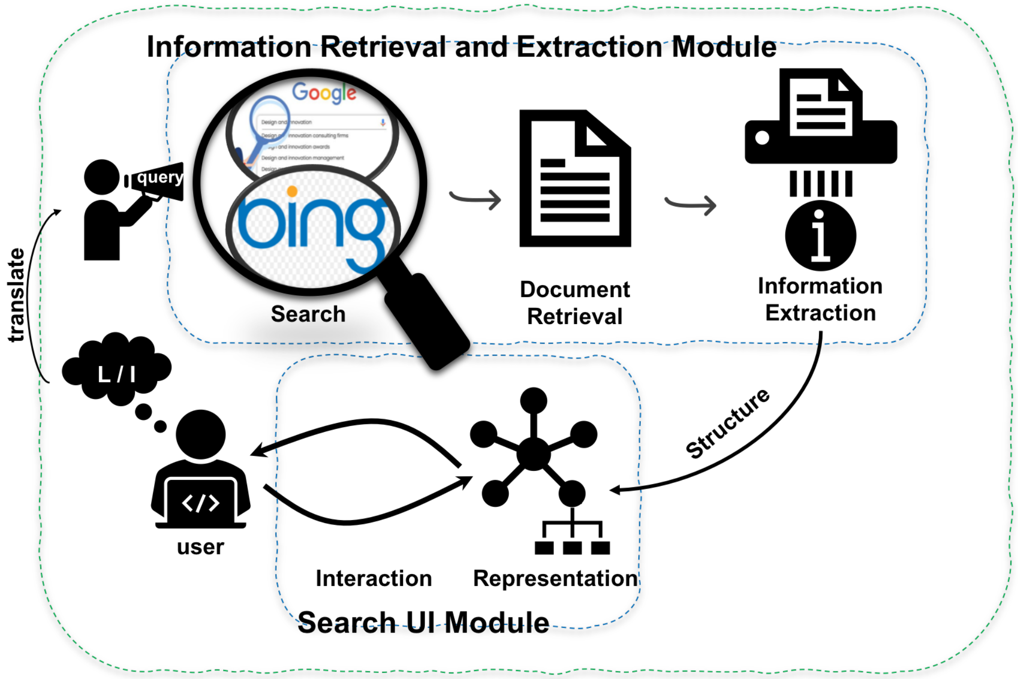


Figure 1.2: Our Proposed Exploratory Search Support Framework Consisting of Two Main Modules. This research is mainly concerned with information needs pertaining to exploratory search tasks with two main categories of Learn and Investigate (L/I).

In this dissertation Chapter 3 describes the development of our IRE module, while Chapters 4 and 5 elaborate on our investigation of alternative representations of the search results and their efficacy in supporting look up and exploratory search tasks given perfect Information Retrieval and Extraction algorithms. Inspired by the findings presented in the previous chapters, Chapter 6 introduces and describes a novel representation of search results, called Hierarchical Knowledge Graph (HKG), and a search UI that provides mechanisms for interaction. Finally, Chapter 7 expands on the second step of our evaluation method and investigates the impact of *imperfect* IE algorithms on the information seeking behaviors and outcomes.

1.3.2 Research Questions and Contributions

To explore the impact of alternative representations of search results on user behavior during exploratory search tasks, we stated that:

> *Supporting complex and exploratory search tasks requires designing search systems that move beyond the current query-response paradigm in three main directions: (1) algorithms that move beyond document retrieval and provide information relevant to a user's query; (2) interfaces that move beyond a turn-taking interaction with a ranked list of documents and provide richer representations of the search results, as well as (3) mechanisms for accessing and interacting with them in ways that support exploration and sensemaking.*

To address the three aspects of designing search systems for supporting exploratory search tasks we proposed a search framework with two main components (see Section 1.3.1): An information retrieval and extraction module and a search UI module that leverages richer representations of the extracted information and supports more advanced interaction models. Given this characterization of our proposed framework we started with two primary research questions as follows:

[RQ1]. How can semantic information be automatically externalized as a response to a searcher's query?

[RQ2]. How can this extracted semantic information be presented such that it extends the document-based search framework's support for information seeking activities?

As our first contribution, we extend the current Document Retrieval framework by developing an Open Information Extraction (IE) tool that can extract semantic information from textual content of a set of retrieved documents such that the extractions are tailored to a searcher's query. We detail our developed retrieval and extraction system in Chapter 3.

As our second contribution, in an attempt to address RQ2 we leverage knowledge graphs, as one canonical spatial structure for information representation supported by theories of learning and comprehension. We developed a search tool that couples retrieves documents with their corresponding knowledge graphs and enables an initial foray into how such a framework can be utilized for information seeking tasks of varying complexity. Based on an analysis of logged traces of 20 information seekers completing simple lookup and complex exploratory search tasks using our extended search framework we identified different information seeking patterns and their likelihood of resulting in finding relevant information.

Our initial study of the efficacy of the developed knowledge graph extension of retrieved documents resulted in somewhat counter-intuitive findings favoring these representations for supporting look-up and known-item tasks. These early results, reported in Chapter 4 gave rise to a series of additional research questions.

First, we explored the space of spatial representations of information beyond vectorial models [340] further and directly contrasted networks and hierarchies to understand how they can specifically assist the searchers in the exploration and making sense of the search results.

[RQ3]. How do alternative spatial representations of information, that externalize semantic information in documents' content, fare in presenting results for exploratory search tasks?

As our third contribution, in Chapter 5 we present the results of the first and only mixed methods evaluation and characterization of network and hierarchical representations of search results and provide novel insights into their comparative efficacy for supporting simple and exploratory search tasks. Our quantitative and qualitative findings broadened our understanding of strengths and weaknesses of each of these representations and highlighted the complementary nature of hierarchical structures and knowledge graphs as representations of search results. These results led naturally to a new inquiry that whether these representations can be combined such that the strengths of the underlying representations are preserved?

[RQ4]. How can we design a hybrid representation that can seamlessly combine alternative representations of search results while preserving their complementary strengths and minimizes their shortcomings?

We investigated this inquiry in two steps:

[RQ4-a]. How can we combine alternative representations of search results – specifically hierarchical and knowledge graph representations –

into a unified structure where local and global views of the data are co-visualized and seamless transitions between these views are enabled?

This inquiry in turn led to a series of design questions we discuss in Chapter 6:

1. How do we integrate network and hierarchical views into a single, seamless data structure?
2. How can both the global and the local view of a knowledge graph be co-visualized?
3. How can transitions between views be designed to maximize visualization stability?

[RQ4-b]. Given that we introduced "*Hierarchical Knowledge Graphs*" as a new structure that combines the hierarchical overviews with local context of knowledge graphs representations, a related question is whether this new structure preserves the strengths of the underlying representations?

As our forth contribution, we introduce and evaluate a novel representation of search results called Hierarchical Knowledge Graphs (HKG) and demonstrate that HKGs preserve many of the previously observed advantages of traditional knowledge graphs, i.e. fewer document views and reduced reading time. Alongside this, HKGs introduce an effective hierarchical representation into knowledge graphs that can offer both global and local views of the information space.

And finally our last research question involves an overall assessment of how Hierarchical Knowledge Graphs can work in the real world.

[RQ5]. Given that IE algorithms are not perfect in the real world, how resilient these HKGs are to these errors? And how do error-prone HKGs impact a searcher's ability to leverage these representations to perform exploratory search tasks?

To probe these questions, and as our fifth contribution, we adopt a tiered evaluation model that embeds IE in a task flow and evaluates IE performance in situ using a balanced mix of quantitative and qualitative (i.e., mixed) methods. Using this in situ evaluation model, we analyze the effect of precision and recall on the performance of hierarchical knowledge graphs for two different exploratory search tasks. While the quantitative data shows a limited effect of precision and recall on user performance and user effort, qualitative data provides evidence that the type of exploratory search task (e.g., learning versus investigating) can be impacted by precision and recall. The implications of these results and an analysis of other factors that more significantly impact exploratory search performance in our experimental tasks are discussed in Chapter 7.

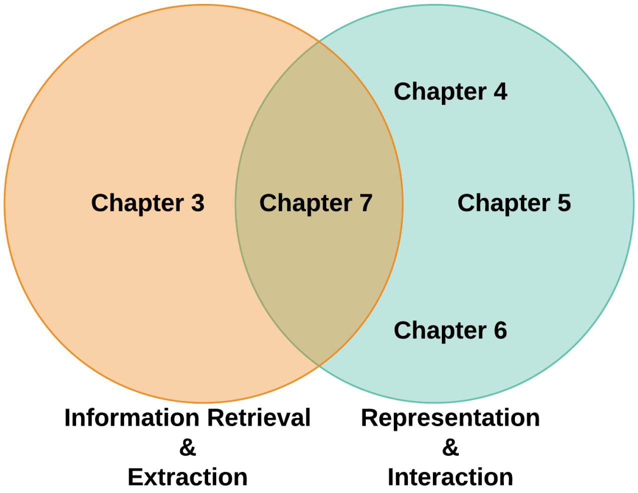

Chapter 3
Chapter 7
Chapter 4
Chapter 5
Chapter 6
Information Retrieval & Extraction
Representation & Interaction

Figure 1.3: Venn Diagram indicating the contribution of each chapter to addressing the two main components of our proposed framework.

1.4 Thesis Structure and Summary

This dissertation details the research and development efforts related to the main components of our proposed ESS framework (discussed in Section 1.3.1). In Chapter 2 we review related research areas that can broadly characterize exploratory search tasks, how systems can support these activities and how these systems can be evaluated. Chapter 3 describes our designed information extraction algorithm that is leveraged by our IR module in order to extend the current document retrieval paradigm. Chapters 4 to 6 involves developing and experimenting with alternative representations of search results (e.g. Textual Documents, Knowledge Graphs and Hierarchies) and a range of interaction capabilities that culminated in a Search UI that can enable a more effective exploration of search results. Finally, in Chapter 7 we extend the ecological validity of our findings by investigating how error prone representations of information can impact the ways searchers' behaviors and outcomes. Figure 1.3 illustrates this structure.

Chapter 2: Background. In this chapter we are covering the related work pertaining to different aspects of designing systems for supporting exploratory search. The work presented in this dissertation builds upon work in a number of related research areas including extensions to document retrieval, information extraction systems, taxonomies of Web search queries, seminal theories and models of information seeking and exploratory search. We also look at prior research in information organization and structuring and visualizing search results. We conclude the chapter with different aspects of evaluating exploratory search systems.

Chapter 3: Extending the Document Retrieval Paradigm. In this chapter we begin to develop our extension to the Query-Response paradigm and focus on the first element specified in our thesis statement, i.e., *providing information.* Motivated by the widespread acknowledgement that document retrieval is only the first step towards satisfying searchers' complex information seeking goals, we design information extraction (IE) algorithms that are effective in extracting semantic information given a set of retrieved documents. This chapter details our implementation of the designed (open) IE algorithm and elaborates on the development of the IR module in our proposed framework described in Section 1.3.1.

Chapter 4: Exploring a Knowledge Graph Extension. As a first step towards understanding the users behaviors and search outcomes given alternative representations of search results, we designed an exploratory study to look at how a conceptual representation of the semantic information extracted by our IR module can be leveraged to extend the textual representation of documents and enable new ways of exploring the search results.

The main idea behind this framework is based on combining knowledge graphs with document retrieval in order to provide a conceptual overview for the information space. Knowledge Graphs have been widely used to promote meaningful learning as well as browsing knowledge and navigation. However, there is limited insight into how these graphs can be utilized by searchers to aid with locating relevant information and making sense of them. The research described in Chapter 4 challenges the models that focus on either traditional document retrieval or the use of linked data for finding relevant information. The findings demonstrate that knowledge graphs and the coherent content of textual documents are both crucial for supporting users during their exploratory activities.

Overall, the results of this initial study highlighted the role of spatial representations in structuring the information that can help locate relevant information when they are coupled with textual content of documents. We also observed how different representations of the same underlying information can lead to different information seeking strategies and how the complexity of the search tasks can bias searchers towards utilizing different components of a representation (e.g. starting from documents text versus interacting with nodes in the knowledge graphs versus examining the edges in the graphs corresponding to semantic relationships).

Chapter 5: Alternative Representations of the Search Results. In exploratory search, how information is presented to the user and how the user interacts with the presented information heavily influence the user's success [354]. In fact, information seekers often express a desire for a user interface that organizes search results into meaningful groups to help make sense of the results, to infer relationships between concepts, and to help decide what to do next [213, 414, 397].

Given our observations in Chapter 4, regarding the ways that graphs of concepts and relationships, which are derived from documents, can be utilized by searchers in search tasks of varying complexity, in Chapter 5 we focused specifically on spatial representations of search results; Two spatial representations, knowledge graphs and hierarchies, were contrasted to provide a broader understanding of how different ways of structuring the search results can impact the information seeking behaviors and outcomes. More specifically, we were interested to identify the unique characteristics of each representation as they both externalize semantic information from textual content and structure them in a 2D graphical structure. To this end, Chapter 5 describes the first mixed methods study of multiple representations of search space and identifying their relative strengths and weaknesses in supporting look-up and exploratory information seeking tasks. The findings of this work culminated in designing a novel representation of search results, deemed Hierarchical Knowledge Graphs [475], that enables the user to engage in two alternative navigation paradigms: they can exploit overview layers to explore the collection at a higher level followed by targeted immersion in the detailed view (See Chapter 6).

Chapter 6: Putting It All Together: Hierarchical Knowledge Graphs. In information retrieval and information visualization, hierarchies are a common tool to structure information into topics or facets, and network visualizations such as knowledge graphs link related concepts within a domain. Given the complementary benefits of hierarchies and network structures to support exploratory browsing of search results (described in Chapter 5), in this chapter, we explore a multi-layer extension to knowledge graphs, hierarchical knowledge graphs (HKGs), that combines hierarchical and network visualizations into a unified data representation.

Through interaction logs, we show that HKGs preserve the benefits of single-layer knowledge graphs at conveying domain knowledge while incorporating the sensemaking advantages of hierarchies for knowledge seeking tasks. Specially, this chapter describes our algorithm to construct these visualizations, analyzes interaction logs to quantitatively demonstrate performance parity with networks and performance advantages over hierarchies, and synthesizes data from interaction logs, interviews, and thinkalouds on a testbed data set to demonstrate the utility of the unified hierarchy+network structure in our HKGs.

Chapter 7. Information Seeking with Error-prone Representations Finally, developing solutions to support users' exploratory search tasks also includes significant challenges in evaluation [129]. We note that a reliable evaluation model needs to assess exploratory search systems based on two complementary aspects: accuracy and effectiveness. Essentially, assessing the accuracy of an exploratory search system is closely coupled with the accuracy of the information it extracts from the retrieved documents to be presented to the searchers. System effectiveness for exploratory search, on the other hand, requires evaluating how well the system aids in the exploratory search tasks it is designed around.

Given that our proposed framework for supporting users with their exploratory search activities combines two main components of *Information Retrieval and Extraction*, which retrieves the relevant information for a query, as well as a *Search UI*, that represents and enables interaction with this information we need to incorporate both the accuracy of the output of our IRE module as well as the efficacy of our Search UI module to truly measure the effectiveness of our proposed solution.

In Chapters 4, 5 and 6 we made a simplifying assumption; that perfect IE systems exist and as a result, the output of our IRE module is as accurate as the source documents that contained the extracted information. The challenge is, however, that algorithms that automatically extract information from a collection of retrieved documents are always error-prone and therefore our previous experiments are limited in their generalizability to real-world scenarios. That is, they do not provide an answer to the following question: *how search outcomes and users behaviors change as a result of leveraging error-prone representations of search results?* To this end, in Chapter 7 we revisit our assumption that perfect IE exists, and through a mixed methods analysis of the effect of precision and recall on the performance of hierarchical knowledge graphs for two different exploratory search tasks, we seek to probe this question.

1.5 Broader Impact

The premise of the internet is to empower the people of all ages, expertise and backgrounds around the globe with universal access to the information, to learn, explore and get tasks done effectively. Web search engines are a primary mechanism by which people seek information and solve problems, However, Search is only a partially solved problem as it's currently optimized for look-up tasks which yields only candidate starting points for learning and cognitive development. In fact, people are still forced to consume and make sense of the information independently from search systems [597].

While there have been many efforts towards designing the next generation of search systems that do not only provide search results, but help users explore, overcome uncertainty, and learn, we are still far from fully realizing the premise of the World Wide Web.

The goal of this research, is to advance the current Web Search paradigm, that is highly effective for factoid look up tasks, towards a knowledge seeking and knowledge exploration platform. This new platform can enable people to achieve more and support the activities that they value the most such that this vision of internet empowering users to find and understand the information they need is realized.

Chapter 2

Background

Our work draws from several areas of literature that we review in this chapter. We start with understanding the types of Web search activities and review some of the established taxonomies of search tasks in Section 2.1. Next, in Section 2.2 we provide an overview of exploratory search research and the related theories from fields of IR, HCI and Psychology that can collectively shape our understanding of what triggers exploratory searches and how this process unfolds. This characterization of exploratory search behavior is essential for designing new search paradigms that can assist information seekers with more complex types of search tasks.

Given a broad characterization of exploratory search behaviors and common strategies, we then shift our focus to reviewing existing research related to the three main requirements of systems that support exploratory search as specified in our thesis statement (Section 1.1). As such, Section 2.3 describes the main approaches that aimed at extending the current query-response paradigm by providing information and not documents; Section 2.4 describes efforts that are focused on designing exploratory search systems that are capable of supporting complex and evolving information needs, particularly through providing richer representations of search results and more advanced interaction paradigms to explore and navigate through these representations.

Finally, in Section 2.5 we review some of the main evaluation methodologies and metrics that can be leveraged for assessing the effectiveness of new search systems supporting exploratory search tasks.

2.1 Taxonomy of Search Tasks

People's day to day search activities can vary greatly in their motivations, objectives and outcomes. Studies of user search behaviour have a long history in Information and Library Science. Specifically with respect to web search, Broder [75] proposed a taxonomy of Web Search in 2002. He was motivated by the idea that the traditional notion of an "information need" might not adequately describe web searching. Broder's taxonomy classifies web searches into navigational, informational and transactional. Similarly, Rose and Levinson [451] analyze user goals to classify web searches into Navigational, Informational and Resource. Drawing upon earlier work by Campbell [86] and Byström [82], web searches can broadly be classified into "Simple" and "Complex" searches. Simple search tasks are similar to "known-item" search tasks and usually involve looking up some discrete, well-structured information object: for example numbers, names and facts [352]. Complex search tasks, on the other hand, involve investigating, learning and synthesizing of information [605].

In contrast to Broder's and Rose and Levinson's taxonomies, Marchionini [352] focuses specifically on a process he terms *exploratory search*. Marchionini broadly separates web search into three categories: Look-up, which includes fact retrieval, navigation and transaction; Learn, which includes knowledge acquisition, comprehension, and comparison; and Investigate, which includes analysis, synthesis and evaluation. The latter two categories, Learn and Investigate, he groups under the umbrella of exploratory search. There are two activities which mediate the process of exploratory search: information foraging theory [412], which describes how searchers collect relevant pieces of information, and sensemaking [140], which describes the process through which people assimilate new knowledge into their existing understanding. In the next section we elaborate on the characteristics of exploratory searches and review the related theories that provide a foundation for understanding this category of search tasks.

2.2 Characterizing Exploratory Search

An initial question we wish to explore in this section is '*what characteristics make a search exploratory?*' Indeed exploration is an important aspect of many search processes. However, not any act of exploring makes a search exploratory. Researchers have identified salient characteristics of such searches based on what motivates these types of search activities and how the process of search is conducted. Exploratory searches are commonly characterized as a class of information seeking activities that are open-ended, complex and multi-faceted. These tasks are often motivated by a complex information problem and a poor understanding of terminology as well as information space structure. They include complex cognitive activities associated with knowledge acquisition and the development of intellectual skills [597].

Marchionini [352] suggested that key components of the exploratory search process are learning and investigation. He also argues that characterizing exploratory search requires describing two main aspects of this type of information seeking: (a) the *problem context* that motivates the search and (b) the *process* by which the search is conducted.

Problem Context. The problem context in exploratory search is often described as ill-structured and open ended where searchers are motivated by a knowledge gap or a desire to learn or solve a problem [556, 612, 306, 597]. Exploratory searches often involve complex situations. The complex nature of these activities commonly leads to a non-linear and dynamic process requiring the information seeker to transition back and forth between induction and deduction [78]. The resolution of vague or complex information problems calls for exploratory search behaviors. In fact, Exploratory search is a specialization of information seeking, which describes the activity of attempting to obtain information through a combination of querying and collection browsing. While searching to learn has been established as an important motivator behind information activities (e.g. as suggested by [50, 352]), White and Roth [597] note that learning in exploratory search is not only about knowledge acquisition, but rather the development of higher level intellectual capabilities that can lead to the application, synthesis and evaluation of this new knowledge.

Search Process. The purpose of exploratory search is typically to create a knowledge product (e.g. a research paper) or make a decision (e.g. choosing a medical treatment) [411]. To this end, much of the search time in exploratory or learning tasks is devoted to examining and comparing results, as well as discovering the key concepts of the domain [597]. Marchionini [352] distinguishes three classes of search activities associated with an exploratory search process: Lookup, Learn and Investigate. As depicted in Figure 2.1, while these activities are separated into three groups there is an interplay between them. Essentially, exploratory searches involve both exploratory browsing and focused searching activities. While learning and investigating activities are commonly associated with exploratory browsing, focused searches can be seen as instances of look-up activities. The majority of current search systems are capable of handling look-up tasks given the significant investment in ranking technologies [611, 597]. However, in order to fully support learning and investigation activities associated with exploratory searches there needs to be more involvement from the user, more advanced interaction possibilities between the user and the search system as well as more advanced UI designs that move beyond simple query specification and a linear results presentation.

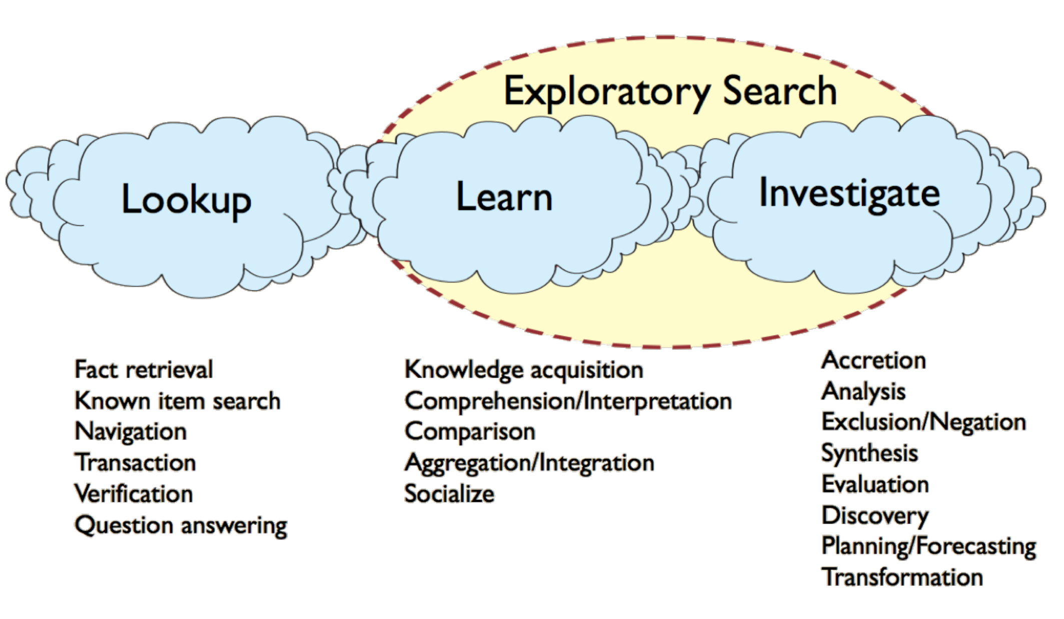


Figure 2.1: Exploratory Search Activities (based on Marchionini's Taxonomy[352]).

Marchionini's model of exploratory search primarily addresses the educational objectives of Bloom's taxonomy [63]. However, it does not examine other types of exploratory behavior where the searcher might traverse the information space with no prior knowledge of the domain and possibly without a defined target. Past research has identified such exploratory behavior as evident in *Wayfinding* tasks [343] that require the navigator to be able to conceptualize the space as a whole as well as *survey knowledge* [540] that requires a scientist to visualize a dataset with no prior understanding of the shape or the organization of the data [597]. In order to provide assistance with these other types of exploratory searches, search systems need to support both exhaustive and directed searches. In the domain of Web search, information seekers that are navigating an unfamiliar document collection can also benefit from such search systems.

Exploratory Browsing versus Focused Searching In the previous two subsections we described two important elements of exploratory search: the problem context and the search process. The problem context is an important motivating factor, but is also highly dynamic in exploratory search scenarios. In order to reduce this dynamism, and as a result the inherent uncertainty in the problem context, the information seekers can engage in two different types of strategies over the course of an exploratory search episode [597]: Exploratory Search Strategies and Focused Searching. Search strategies that are exploratory in nature e.g. berrypicking (see Section 2.2.1.2) or information foraging (see Section 2.2.1.3) can support the gathering and re-representation of information – as is common practice in sensemaking [139]. Alongside these exploratory approaches, a systematic learning mechanism such as hypothesis formulation and testing, as in exploratory data analysis (EDA; [548]) can be used to assist the searcher with better defining the problem context. Essentially, the **exploratory browsing strategy** exposes the user to collection content to help better define their information needs and to promote information discovery and new cognition based on observed content. **Focused searching** on the other hand, may include some degree of navigation, but is generally intended to help the searcher follow a known or expected trail rather than foraging new ground [412, 597]. As such, the searcher can query the document collection, examine search results and documents in close proximity to search results, and extract relevant information to meet their goals. In this regard, exploratory browsing can be considered as a hypothesis generation activity, while focused searching can act as a hypothesis testing step in the process of exploratory search. Effective exploratory search systems thus need to maintain a balance between supporting exploratory browsing activities as well as focused searching.

2.2.1 Understanding Exploratory Search Process and Strategies

In this section, we review relevant theories from related disciplines such as HCI, IR, information science, and psychology that can provide a new perspective for understanding the process of exploratory search and strategies that are common during these types of activities. These theories highlight different aspects of exploratory search behavior including exploration and browsing, locating and collecting relevant pieces of information and making sense of these pieces.

The last subsection highlights the relationship between exploratory search research to each of these related areas and how different aspects of exploratory search behavior can be understood using these relevant theories.

2.2.1.1 Exploratory Behavior and Browsing

Exploratory behavior is defined by the National Library of Medicine as “the tendency to explore or investigate a novel environment”, is driven by curiosity and is evident in most exploratory searches. As we noted in the previous section, exploratory searches are not always motivated by a need to solve a problem; rather curiosity often leads to exploration, investigation and learning. Yet, curiosity is not particular to exploratory searches. In fact, both exploratory and look-up search tasks apply different types of curiosity throughout the search process: *Specific curiosity* is a desire for a particular piece of information, that is best exemplified by an attempt to solve a problem or puzzle. *Diversive curiosity*, on the other hand, is a more general seeking of stimulation or novelty, for example a television viewer flipping between channels [56]. In information seeking, specific curiosity corresponds to well-defined goals and directed searching, while diversive curiosity corresponds to ill-defined goals and exploratory browsing [399].

According to Berlyne [56] there are three stages of exploratory behavior: (1) orienting responses, (2) locomotor exploration and (3) investigatory responses. White and Roth [597] map these stages to three phases of exploratory searches: (1) obtaining overviews of the information space by utilizing techniques such as information visualizations, (2) focusing on a specific object (e.g. a relevant document) and (3) examining the object in more detail. We also note the parallel between these stages of exploratory search and the steps specified by Shneiderman [495] in the Visual Information Seeking Mantra: Overview first, zoom and filter, then details on demand.

Browsing A closely related activity to exploratory behavior is browsing. Browsing is defined as movement in a connected space [312]. Marchionini [351] reviewed the research on browsing and observed three general types of browsing activities based on the object of search (i.e. the information needed) and by the tacticts employed to find this information; Directed browsing occurs when browsing is systematic, focused and directed by a specific target. For example, scanning a list to locate a known item or verifying attributes such as dates on a Web page require directed browsing. Semidirected browsing involves scenarios where browsing is generally purposeful, the target is less defined and the browsing is less systematic. An example is querying a database for a general term and examining retrieved records. Finally, undirected browsing occurs when there is no real goal and very little focus. Channel-surfing or flipping through a magazine are examples of undirected browsing. Similar to Marchionini's categories of browsing, Wilson [612] identifies four categories of information seeking and acquisition after a survey of health information seeking: Passive attention, passive search, active search and ongoing search.

Bates [46] suggests that browsing is a cognitive and behavioral expression of exploratory behavior which has four elements: (1) glimpse a scene; (2) target an element of a scene; (3) examine item(s) of interest; and (4) physically or conceptually acquire or abandon examined item(s). To this end, White and Roth [597] argue that in order to support browsing activities, as an integral part of exploratory searches, exploratory search support systems should offer collection overviews and the ability to traverse through the collection (exploratory browsing) and document examination / retention.

2.2.1.2 Berrypicking

The theory of Berrypicking was introduced by Bates [44] as a way of describing information seeking behavior in online environments. This method depicted information retrieval as a dynamic and evolving process criticizing traditional approaches suggesting a single results set can satisfy a user's query. When introduced in 1989, the berrypicking model was considered revolutionary as it emphasized browsing and navigation as searching modes for which explicit queries do not have to exist [241]. Essentially, the theory of berrypicking exemplifies a pioneering approach to exploratory search [597].

The term "berrypicking" is an analogy to picking berries in a forest; berries are scattered on bushes and not in dense bunches. This analogy points to a resemblance between this model and the information foraging approach to information seeking: similar to information foraging [410] and wayfinding [343] theories, the berrypicking model views the searcher as moving through an information space, gathering fragments of information as they move and seeking cues that guide navigation decisions. In fact, the main distinction between these two models is their focus; while the berrypicking model posits evolving information needs as the main motivator for exploratory search, information foraging focuses on the act of searching itself.

The berrypicking model, as demonstrated in Figure 2.2, is based on the idea that in real-life searches, each new piece of information that is gathered by the searcher gives them new ideas and directions to follow and, as a result, a new understanding of their information need. Similar to picking berries in a forest, the individual picks pieces of information a bit at a time and thinks about the information she has found,

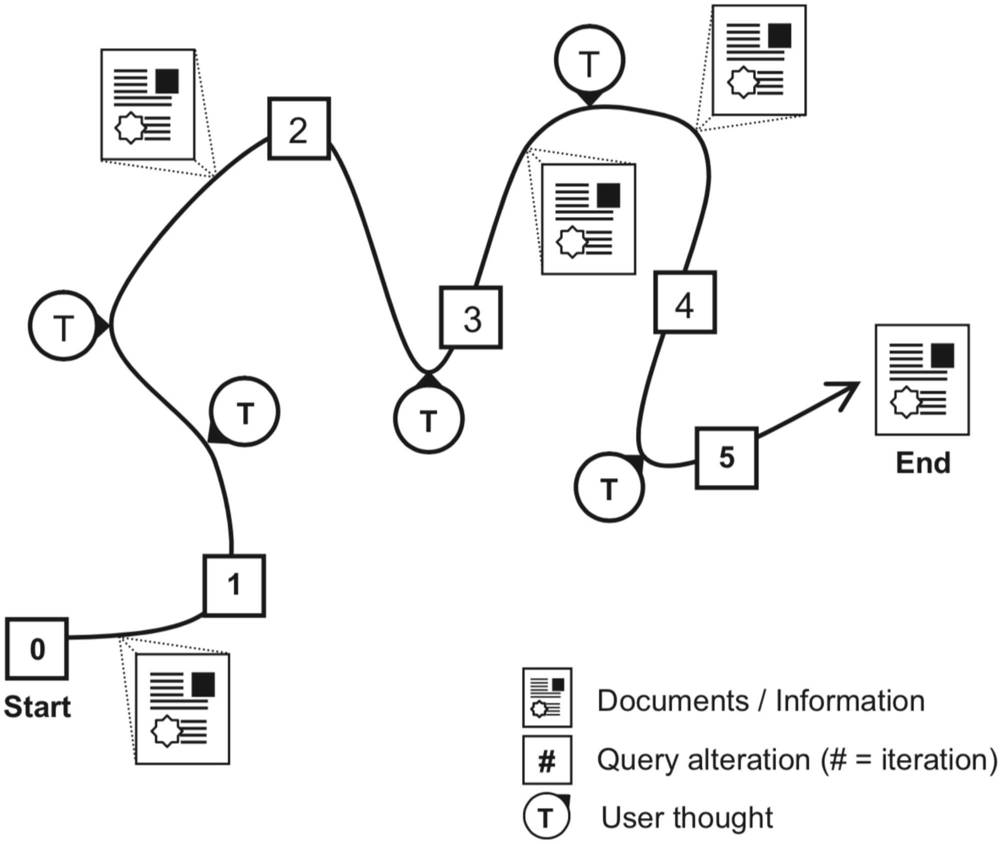


Figure 2.2: An evolving berrypicking search based on Bates' model [44] - figure is adapted from [597]

partially by relating it to what she is trying to achieve through the search process [480].

Berrypicking is a commonly used strategy in exploratory search. During exploratory searches, the evolution of the information need is important, and it is indeed an integral part of gathering and understanding information fragments. As noted, the berrypicking model criticizes the look-up model offered by current search engines and exhibits two main distinctions from look-up searches [597]: (1) the nature of the query (representing the problem context) is evolving, rather than static and unchanging; (2)the nature of the search process follows a berrypicking pattern (see Figure 2.2), instead of leading to a single best retrieved set.

2.2.1.3 Information Foraging

Information foraging theory [412] attempts to explain information seeking behavior in humans by drawing ideas from food foraging mechanisms in a variety of living organisms. Central to the *optimal foraging theory* is the observation that the eating habits of animals revolve around maximizing energy intake over a given amount of time. Similarly, Pirolli and Card [412] proposed information foraging theory based on this assumption that humans' adaptive success depends to a large extent on sophisticated information gathering, sensemaking and problem solving strategies [410].

Essentially, the information foraging theory provides an account of how people, while searching for and browsing information resources, choose to continue their activities in the same region or rather identify a new region in which to look for information. One of the fundamental premises of the information foraging theory is that the identification and selection of information regions (aka patches) is done based on users' assessment of their appropriateness. This assessment is often done by relying on available *'cues'* (e.g. textual snippets or thumbnail images) provided by modern search engines). These cues can be considered as "information residue" [176], which was later refined and labelled as "information scent" by Pirolli [409]. Later, Card et al [89] defined information scent as a user's "imperfect perception of the value, cost, or access path of information sources obtained from proximal cues, such as WWW links." Information scent can thus be understood as the quality that the forager attributes to proximal cues of a variety of information objects [480].

2.2.1.4 Sensemaking

The Oxford dictionary defines Sensemaking as the action or process of making sense of or giving meaning to something, especially new developments and experiences. People are constantly engaged in making sense of the world. Dervin [139] was one of the influential pioneers in focusing on the needs of users of information systems and the way they make sense of the world. She describes the process of sensemaking as a series of continuing gap-defining and gap-bridging activities [139, 140]. This process can be considered as an active two-way cycle of fitting data into a frame (i.e. a schema or a mental model) and fitting the frame around the data.

Inspired by Dervin's view of sensemaking, later research efforts (e.g. [413, 296]) used observational studies and cognitive task analysis to identify the main steps in sensemaking as follows: (1) knowledge gap recognition; (2) generation of an initial structure or model of the knowledge needed to complete the task (i.e. concepts, relationships and hypotheses); (3) search for information; (4) analysis and synthesis of information to create insight and understanding; and (5) creation of a knowledge product or direct action based on the new understanding or insight.

The fundamental intuition behind sense making is also proposed in “assimilation theory of cognitive learning” by Ausubel [33]. To Ausubel, meaningful learning occurs when learners build a new knowledge structure by explicitly constructing new nodes and interrelating them with existing nodes and with each other. Novak [387] demonstrates how Concept Maps make the theoretical principles of Ausubel’s assimilation theory practical. According to Novak [387], concept mapping involves the construction of a conceptual network that represents the learner’s understanding of a certain domain of knowledge.

Essentially, sensemaking is the creation of situational awareness and understanding in situations of high complexity or uncertainty in order to make decisions [597]. This *situational awareness* is commonly achieved through constructing a *conceptual representation* of the problem context and highlights the role of analysis and synthesis as a crucial part of the process [213, 464]. This view of sensemaking, as “the iterative process of formulating a conceptual representation from a large volume of information” [213] is the main inspiration behind the research reported in this dissertation, and is supported by a body of past work. Russell et al. [462] described sensemaking as “the process of encoding retrieved information to answer task-specific questions.” They defined a sensemaking model comprising four main processes: (1) search for representation (structure): the sense-maker creates representations to capture salient patterns of data; (2) instantiate the representations: the sense-maker identifies relevant information and encodes it in the representation; (3) modify the representations: representations are modified during sensemaking when data is ill-fitted or missing in the representation; and (4) consume instantiated representations: the sense-maker consumes the instantiated representation. Qu and Furnas [423] emphasize the bidirectional relationship between search and representation construction and identify two separate sub-processes of ‘search for structures’ and ‘search for data’ during sensemaking. In a more recent work, Russell et al. [463] focus on sensemaking activities in the context of information retrieval tasks that require understanding large document collections and characterize the process of sensemaking as “creating a representation of a large volume of information that allows the analyst to perceive structure, form and content within a given corpus”. They note that when people need to rapidly make sense of a large document collection they usually begin by skimming the documents and organizing the collection into temporary groups (clusters). This sensemaking behavior gives a quick overview of the contents, while creating a fast, easy to use representation for organizing and accessing the accumulated contents.

Like exploratory search, sensemaking can be supported by information search interfaces. Hearst [213] discusses examples of sensemaking interfaces and their constituent elements, which include flexible arrangement and grouping of information, integrating notetaking and sketching, hypothesis formulation and collaborative search.

Evidently, exploratory searchers are constantly engaged in sensemaking activities as they move through the information space. As White and Roth [597] note sensemaking is an individual process of construction, not a process of utilizing existing information. Given this view, exploratory search typically involves a prolonged engagement in which information seekers iteratively look up and learn new concepts and facts. Exploratory search can thus be viewed as a sub-component of sensemaking [597].

2.2.2 Information Seeking

Exploratory search can be more broadly characterized as a class of information seeking. Information seeking is the fundamental process of attempting to obtain information in both human and technological contexts [597]. Marchionini [351] views information seeking as a process driven by human needs for information so that they can interact with the environment. His view of information seeking is inspired by three beliefs about human's existence: Life is active, continuous and accumulative. This perspective implies that we learn by "bumping into the environment"; that information flows, continuously, from the environment, regardless of how we are able to process and store it; and that as the flow of information affects our knowledge structures, these structures are extended, reinforced or altered.

Wilson [612] distinguishes three classes of related research areas, namely information behavior, seeking and searching to further understand user-oriented (cognitive) IR research. His nested model (see Figure 2.3) defines *"information behavior"* as a broad concept representing "the totality of human behavior in relation to sources and channels of information use." This definition includes information seeking as well as other types of behavior (e.g. the passive reception of the information). *"information seeking"* is placed as a subset of information behavior and is defined by Ingwersen and Jarveling [241] as "human information behavior dealing with searching or seeking information by means of information sources and (interactive) information retrieval systems". *"Information searching"* as a sub-field of information seeking in this model

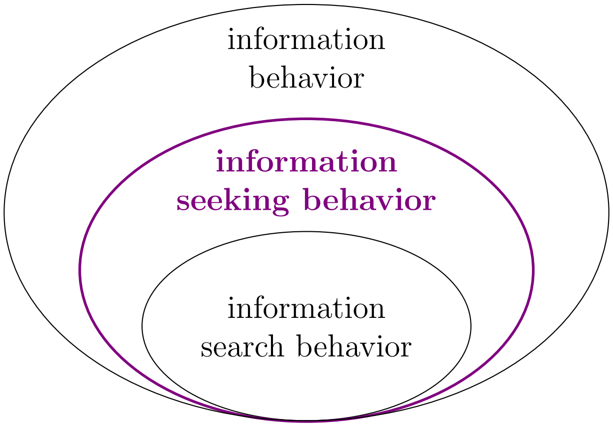


Figure 2.3: Wilson's nested model; figure adapted from [612]

is focused on the interaction between the user and the information system, and is more formally known as Interactive Information Retrieval (IIR) [597].

2.2.2.1 Information Seeking Models

The process of information seeking has been characterized by a number of models including Ellis' behavioral model [157], Dervin's Situation-Gap-Use Model [140], Belkin's Anomalous States of Knowledge (ASK) Model [50], Kuhlthau's Information Search Process (ISP) Model [306], Wilson's Problem Solving Model [612] and Choo et al's integrated model of browsing and searching [110].

Ellis' Behavioral Model. Ellis [157, 158] proposed a general model of information seeking behaviors based on the studies of the information seeking patterns of scientists, researchers and engineers in an industrial firm. One version of this model describes six categories of information seeking activities: starting, chaining, browsing, differentiating, monitoring and extracting. In this model, *browsing* is different from simply viewing some results, and is rather the activity of semi-directed search in areas of potential interest. To elaborate, browsing takes place in many situations in which related information has been grouped together or organized as tables of contents, lists of titles or a set of entities.

Situation-Gap-Use Model. Dervin [139] was one of the influential pioneers in focusing on the needs of users of information systems and the way they make sense of the world. Her Situation-gap-use model posits that users go through three phases in making sense of the world which revolves around facing and solving their information problems. The first phase, called the *situation*, establishes the context for the information need; Next, people find a *gap* between what they understand and what they need in order to make sense of the current situation. These gaps are manifested by questions. The final step involves using answers or hypotheses for these gaps to be able to move to the next situation. Marchionini [351] notes that the situation-gap-use model applies to more general human conditions than information seeking but has been adopted by researchers in information science and communications as a framework for studying the information seeking process.

ASK Model. Related to this concept of a knowledge gap, Belkin et al. [50, 395] constructed a model of information seeking that focuses on information seeker's anomalous states of knowledge (ASK). According to this model, when a search begins, a searcher's state of knowledge is in an "anomalous state", and they have a gap between what they know and want to know. Hence, the searcher must go through a process of clarification to articulate a search request, with the obvious implication that search systems should support iterative and interactive dialouges with users. This model contributes to designing systems for supporting exploratory search in at least two ways: (1) this model was designed to explain generally open-ended information problems and does not directly apply to fact-retrieval tasks; (2) This model serves as a theoretical basis for the design of highly interactive information systems [351].

Marchionini's Information Seeking Process Model. Marchionini [351] proposed another often-cited model of the information seeking process directed towards electronic environments. In this model, the information seeking process is composed of eight sub-processes which develop in parallel: 1) recognize and accept an information problem, 2) define and understand the problem, 3) choose a search system, 4) formulate a query statement, 5) execute search, 6) examine results, 7) extract information and 8) reflect/iterate/stop. It is interesting to note that in this model, Marchionini describes the sub-process of extracting as an *assimilation activity*: "there is an inextricable relationship between judging information to be relevant and extracting it for all or parts of the problem's solution ... To extract information, an information seeker applies skills such as reading, scanning, listening, classifying, copying and storing information. ... As information is extracted it is manipulated and integrated into the information seeker's knowledge of the domain" (pp 57-58).

ISP Model. Based on several longitudinal studies, Kuhlthau [306, 305] developed a multistage model of the Information Search Process (ISP), which "depicts information seeking as a process of construction". Kuhlthau's model is descriptive, documenting "commons patterns in users' experience in the process of information seeking" for complex tasks requiring contraction and learning, with a discrete beginning and ending. This model identifies and emphasizes the importance of the individual stages that learning tasks and problem solving involve. The stages in the ISP model are as follows [241]: (1) initiation: becoming aware of the need for information, when facing a problem; (2) selection: the general topic for seeking information is identified and selected; (3) exploration: seeking and investigating information on the general topic; (4) focused formulation: fixing and structuring of the problem to be solved; (5) collection: gathering pertinent information for the focused topic; and (6) presentation: completing seeking, reporting and using the results of the task. An interesting aspect of the ISP model is that it outlines exploration as one of the primary six tasks that the user executes during search and as noted frequently the notion of exploration is fundamental to exploratory search. Exploration, as used by Kuhlthau, is defined as being an investigational stage of the information-seeking process [597].

Task-based IR Model. In 2001, Vakkari refined the ISP model in the context of information retrieval into a tentative theory of Task-based IR process [558] based on a longitudinal study with twelve students. He refined the ISP model and summarized the six stages into three categories: *pre-focus* (ISP's stage 1, 2 and 3), *focus formulation* (stage 4) and *post-focus* (stage 5 and 6). Essentially, Vakkari emphasizes the crucial role of finding a focus in the search process.

Integrated Model of Searching and Browsing. Finally, Choo et al. [110] developed a model of online information seeking that combines both browsing and searching. The main motivation behind this model is the observation that current Web browsers have already enabled browsing mechanisms that were highlighted as a part of the Ellis's model of information seeking. This model suggests that people who use the Web as an information resource to support their daily work activities engage in a range of complementary modes of information seeking, varying from undirected viewing that does not pursue a specific information need, to formal searching that retrieves focused information to guide action and decision making. In fact, Choo et al. [110] argue that each mode of information seeking on the Web can be distinguished by the nature of information needs, information seeking tactics, and the purpose of information use.

2.2.2.2 Understanding Information Seeking Activities

There is a body of more recent studies that focused on observing users' information seeking behaviour and identifying the challenges searchers face during their search session, common information seeking activities among them and gaining insight into how to support these activities.

Pirolli and Card [413] report on preliminary results from a study using cognitive task analysis to help broadly characterize the processes used in "intelligence analysis". They developed a model that indicates how an analyst comes up with novel information. The overall process is organized into two major loops of activities: (1) a *foraging loop* that involves seeking information, searching and filtering it, and reading and extracting information [412] and (2) a *sensemaking loop*[462] that involves iterative development of a mental model from the "schema" that best fits the evidence. Here, schemas" are defined as a set of patterns developed by experts around the important elements of their tasks. These patterns are built up over time and from extensive experience.

Elsweiler and Ruthven [159] focus on Personal Information Management (PIM) which investigates how people store, manage and re-find information. They designed a diary study to investigate (a) the types of re-finding tasks that were performed when search on email and on the Web; (b) which of these types is performed more often and (c) which types are perceived as "difficult" by the majority of participants. They classified the tasks into lookup, item and multi-item tasks and they did not find any significantly more difficult task among them as perceived by the participants. Overall, their study can offer an increased understanding of PIM behaviour at the task level and an evaluation method to facilitate further investigation.

Diriye et al. [144] employed qualitative and quantitative data gathering methods to investigate the efficacy of different interface features based on the complexity of search tasks. They found the simplest interface provided better support for known-item than exploratory search tasks, while richer search interface features were found to provide better support for exploratory search, but would distract people from the objective of more clearly defined search tasks. The results suggest that "searching is more effective when supported by an interface that is tailored towards the search activities of the task".

Alhenshiri et al. [18] present the results of a study to explore the difficulties users experience during Web information gathering tasks. They argue that while several Information Seeking models are proposed and they have focused on identifying activities performed by users during ”locating of information”, not all these activities are included in these models. They conducted a user study to examine how users manage and organize information during Web information gathering tasks. They identified a set of activities performed by searchers, their frequencies and the reasons behind the most and least frequent activities.

Granka et al. [186] investigated how users interacted with the results page of a Web search engine using eye-tracking. They designed a study to understand what the searcher is doing and reading before actually selecting a document. They investigated how rank influences the amount of attention a link receives as well as how searchers explore a SEarch Results Page (SERP).

Bron et al. [76] study the research cycle of media studies researchers and what activities they perform in order to investigate a research question. They construct a model that identifies sequences of search processes and their influence on the research question. Based on their understanding of these researchers' needs, they propose a subjunctive exploratory search interface to support media studies researchers in refining their research questions in an earlier stage of their research. This interface provides users with support for (1) exploration (i.e., query formulation, query refinement and exploring various aspects of a topic) and (2) discovering patterns in the data (i.e., to compare alternatives and to observe trends in data). They determine that the interaction patterns of the users are different across the subjunctive interface and the baseline. These patterns indicate that users alternate more between formulating queries and inspecting results when they use the subjunctive interface. However, the users of the baseline interface formulate less queries and look through more result pages.

2.2.2.3 Factors Impacting Information Seeking

Information Seeking depends on the interactions among several factors: information seeker, task, search system, domain, setting and search outcomes [353, 351]. There have been a number of observational studies conducted to determine what factors influence information seeking behaviours and outcomes.

Kelly and Cool [280] report the results of a preliminary investigation of the relation between topic familiarity and information search behaviour. The authors argue that while it is commonly acknowledged that topic familiarity is an important factor influencing information seeking, there is limited insight into how topic familiarity can be automatically assessed and how it can be used to improve retrieval. Their work mainly focuses on identifying information search behaviour that may directly be influenced by topic familiarity. They have initially focused on reading time and efficacy as two information seeking behaviours. They have found that as one's familiarity with a topic increases, one's searching efficacy increases while reading time decreases. Although these results are not surprising, the authors suggest that this finding can be used to infer topic familiarity from information search behaviour.

Amadieu et al. [23] examined the interaction effects of prior knowledge and concept map structure (network v.s. hierarchy) on comprehension. They found that for low prior knowledge readers a hierarchical structure improved comprehension performance, helped with reading coherent sequences and lowered their feelings of disorientation. However, the high prior knowledge readers' performance was not affected by this structure. They also found that prior knowledge would support comprehension processes when readers are required to establish semantic relations between text sections in the non-linear document.

Diriye et al. [144] employed qualitative and quantitative data gathering methods to investigate the interplay between the interface features, the user and the search tasks. They found the simplest interface provided better support for known-item than exploratory search tasks, while richer search interface features were found to provide better support for exploratory search, but would distract people from the objective of more clearly defined search tasks. The results suggest that "searching is more effective when supported by an interface that is tailored towards the search activities of the task".

White and Ruthven [598] investigated the concept of "control" and identified the activities for them they would like to have control over. They argue that users need to make decisions for three major search activities: (1) selecting query terms and operators; (2) whether or not formulate a new query and at what point to stop searching; (3) indicating relevance. They presented an experiment in which the participants were asked to interact with three interfaces with varying degrees of user control over how the query is used. The intuition behind the design of this experiment was to investigate how much control users want over the selection of search decisions. They found that users are willing to hand over full responsibility for indicating relevance but want to receive assistance for query formulation and making search decisions. They also found that the users still wish to retain control over search activities they consider important for the effectiveness of their search.

2.2.3 Positioning Exploratory Search

In Section 2.2 we presented past work that contributed to characterizing exploratory searches as a subclass of information seeking activities and differentiated exploratory search from other classes of information seeking. We have also reviewed relevant theories that describe strategies that information seekers leverage to forage for the information and perform sensemaking and synthesis activities. To conclude this section we borrow White and Roth's [597] positioning of exploratory search as a viable sub-discipline of information seeking in order to highlight its relationship with existing disciplines such as IR, information visualization, information foraging and sensemaking. The Venn diagram in Figure 2.4 positions exploratory search in relation to these relevant areas.

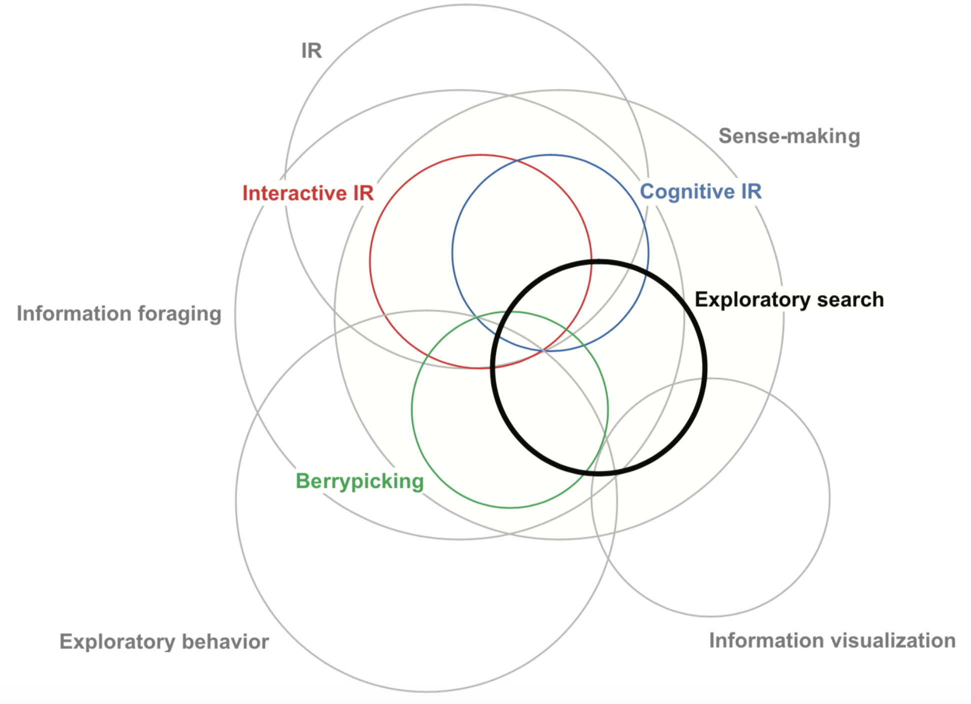


Figure 2.4: Venn diagram positioning exploratory search relative to other related research disciplines. Circle size signifies approximate size of each discipline. Color is used to differentiate interior circles. Figure is adapted from [597].

As depicted in the Venn diagram, *"exploratory search is a type of information seeking and a type of sensemaking focused on the gathering and use of information* *to foster intellectual development."* [597]. Overlap exists with different aspects of information seeking that are essential in exploratory search activities: (1) **information visualization**: is an important asset in generating hypothesis and insight generation as well as in learning about the information space; (2) **exploratory behavior (browsing)** is an important strategy in navigating this information space; (3) **berrypicking and information foraging** describe how searchers find information and adapt to their information environment; (4) **sensemaking** overlaps significantly with information foraging and addresses issues of searcher comprehension and information use; and (5) **IIR and cognitive IR** focus on individuals' complex psychological functions during the retrieval process and is crucial to describe the bahavioral and mental processes involved in finding the information as well as learning and understanding components of exploratory search.

As White and Roth note the exploratory search is a multifaceted concept and is constantly being changed and shaped by all these related areas. System designers can thus draw from the research in all of these related disciplines to better support and understand exploratory search behaviors. In the next two sections we review some of the related efforts to support exploratory search tasks, corresponding to providing information (Section 2.3), richer representations and more advanced interaction models specified in our thesis statement (Section 2.4).

2.3 Going Beyond Document Retrieval

To support the growing complexity of search tasks, researchers in the field of information retrieval developed and explored a range of approaches that extend the traditional document retrieval paradigm. The earlier approaches aimed at extending the document retrieval paradigm by providing information and more focused answers. We review three main classes of such extensions in this section.

To elaborate, the recognition that there is more to search than basic Information Retrieval has led to many extensions and alternatives to the keyword search paradigm. These extensions aimed at supporting users by providing "information" and not documents and also involving users more actively in the search process. Question Answering (QA), Summarization and Information Extraction (IE) all generate a focused response to a user's information need in the form of sentences, text snippets or entity-relationship triples. We review these approaches next.

2.3.1 Question Answering Systems

People widely use IR systems on a daily basis to find documents that answer their questions. Compared to these IR systems, QA systems aim to speed the rate at which users find answers by retrieving answers rather than documents. QA systems respond with a short, focused answer to a question formulated in a natural language.

QA systems are mainly composed of three main components: question classification, information retrieval, and answer extraction. Therefore, each of these three components attracted the attention of QA researchers [622]. For a survey of different QA systems we refer the reader to [19] and [298] which is focused on the IR perspective.

Different approaches to QA range from statistical and classification based methods (e.g. [601]) to linguistics based and deep text processing algorithms (e.g., [322, 591]). Many question answering systems translate questions into triples which are matched against the RDF data (i.e., the data format for linked data and entity-relationship triples) to retrieve an answer, typically relying on some similarity metric. Unger et al. [554] presented a novel approach to question answering over Linked Data that relies on a deep linguistic analysis yielding a SPARQL template with slots that need to be filled with URIs. In order to fill those slots, possible entities were identified using string similarity as well as natural language patterns extracted from structured data and text documents. Therefore, these templates capture the semantic structure of the natural language input.

A major problem with the present QA system is that the answer to a question is also limited to pre-defined categories. They thus suffer from low recall. Furthermore, they mostly support a limited range of questions requiring a factoid answer. For example: “who is the president of US?”.

In 2006, the “complex interactive question answering” track was developed by TREC that aimed at providing support for more complex questions. A question is considered complex if it contains a relationship between two or more entities. For example “what effects does [Aspirin] have on [heart disease]?”. This track also focused on “interactive” QA systems that involve users in the process of finding answers. There has been different attempts towards developing evaluation methods and metrics for Interactive QA systems.

Kelly et al. [283] argue that the Cranfield Model shouldn’t be followed to study interactive systems with real users. Instead, metrics that are sensitive to individual users, tasks and contexts need to be developed. To this end, they conducted a two-week evaluation on three QA systems and a Google baseline system. They generated a set of hypotheses to describe a good interactive QA system (e.g., "a good interactive QA system should support information gathering with lower cognitive workload"). They designed a study to collect data in order to support their hypotheses. They identified the most effective methods of collecting data for supporting different hypotheses. Overall, they have sketched a method for evaluating interactive analytic question answering systems, identified key design decisions and described the effectiveness of data gathering methods.

Smucker et al. [507] aim at understanding how IR systems compare to QA systems. To this end, they designed a study to measure the performance of humans using an interactive IR (IIR) system to answer questions. They ran four experiments that differed in the choice of participants (i.e., NIST assessors, expert and non-expert) and the time constraints for task completion (i.e., 5 minutes, 10 minutes, unlimited for each question). They found that IIR systems are competitive with automatic QA systems for users with complex information need. They propose a better performance can be achieved by combining the flexibility and precision of IR systems with the ease-of-use and recall advantages of QA systems.

Complex Answer Retrieval. More recently, Complex Answer Retrieval (CAR) [142], a new track at TREC, has aimed at answering more complex information needs through retrieving longer answer passages. The ultimate goal of this track is to enable systems to collect relevant information from an entire corpus and create synthetically structured documents by collating these results automatically. In this regard, CAR can be considered a solution at the intersection of QA systems and automatic summarization algorithms described in Section 2.3.2. An example scenario (adapted from CAR homepage [1]) involves a user interested in *learning about water pollution through fertilizers, ocean acidification, and aquatic debris and the effects it has.* A desired answer for this question can be a structured body of text that covers the topic with its different facets and elaborates on pertinent connections between relevant concepts or entities. This scenario can be envisioned as an automatic essay generation task given a topic and its related facets. CAR is still in its early stages of accomplishes this goal of synthesizing complex answers to support a range of learning tasks. While the first and second years of this track were dedicated to producing passage and entity rankings for the query and facets given in the outline, the third year is now focused on the task of arranging paragraphs into a topically coherent article. Assuming the future efforts can realize the ultimate goal of CAR, these solutions can provide a starting point for exploratory searches that are motivated

[1] http://trec-car.cs.unh.edu/

by learning tasks around general topics where the searcher has a relatively good understanding of what aspects their topical need involves and what is required to be covered in the generated output.

2.3.2 Automatic Summarization Systems

Automatic summarization is a process that creates a shortened version of one or more documents. These summaries contain the most important parts presented in a concise and coherent way. When compared with QA, these summaries provide more context and coherent text and thus offer better support for search tasks that require learning and understanding in order to generate answers to more broad questions.

A comprehensive overview of research in summarization is provided in [346] and a more recent one can be found in [381]. Text Summarization methods can be classified into extractive and abstractive summarization. An extractive summarization method consists of selecting important sentences or snippets from the original document and concatenating them into a shorter form. Statistical and linguistic features of sentences are used for ranking the sentences based on "importance". An Abstractive summarization on the other hand, attempts to develop an understanding of the main concepts in a document and then express those concepts in a coherent natural language content. These approaches use linguistic methods to examine and interpret the text and then to find the new concepts and expressions to best describe it by generating a new shorter text that conveys the most important information from the original text document [193]. For a survey of extractive and abstractive summarizarion approaches one can refer to [193] and [284] respectively.

Given a collection of documents, most of existing multi-document summarization methods automatically generate a static summary for all the users using unsupervised learning techniques such as sentence ranking and clustering. However, these techniques suffer from two major limitations: (1) the generated summary is the same for all users and does not leverage users feedback and interaction; (2) the generated summary is presented as a text snippet that lacks a structure which can reveal how extracted sentences are connected to one another. There has been a few attempts that addressed these two limitations. The next subsection reviews these attempts.

2.3.2.1 Interactive / Dynamic Summarization

Zhang et al. [644] develop an interactive summarization system called iDVS which improves the summarization performance using users' feedback and assists users in document understanding using visualization techniques. The system first generates a 2D view graph of current sentence set of the given documents that initially contain all the sentences in all the documents. In this graph, the nodes are the sentences and the edges indicate the cosine similarity between the corresponding sentence pair is above 0. The similar nodes (i.e. sentences) are positioned close to each other using a maximizing graph modularity based clustering algorithm. The system then picks the most important sentences and present them to the user.

The user provides his feedback to the system in two ways: (1) among comparable sentences, he selects the one which he most prefers; (2) the user partially reorders the sentences selected by the system based on his understanding of their context in the corresponding documents. Next, a semi-supervised ranking algorithm uses the users' preferences and provided ordering and ranks all sentences in all documents. Top n sentences from each document are then selected as candidates and recommended to the user. Finally, the user selects the candidates sentences which he is satisfied with. These sentences will then be included in the summary. A new iteration starts until the required length of the summary is reached.

They evaluate their system by comparing it with other most widely used unsupervised and supervised summarization systems. They observed their system outperformed all the unsupervised systems, while maintained a comparable performance with the supervised summarization approach. They also conducted a user study and asked fifteen participants to assign a score of 1 to 5 according to their satisfaction of the use of iDVS as compared with four other summarization systems. They reported an average score of 4.07 for iDVS, followed by 3.47 for the supervised system and 2.80 for the leading unsupervised system.

Jones et al. [254] develop an interactive summarization system called IDS that provides dynamic control over some summarization features (e.g., length and topic focus) and allows users to move flexibly between summaries and their source document. The authors consider any system that allows the users to specify a range of summariser options (e.g. summary length) interactively, an "interactive" summarizer. They also characterise such a system by the ability to bias the summary towards a particular topic or a set of topics.

The IDS system adopts the sentence extraction approach to the construction of summarise and has two components: (1) a keyphrase extraction module that uses machine learning to extract more salient phrases that incorporates TF-IDF values of different candidate phrases. Next, each sentence from a document is scored based on the keyphrases it contains; (2) a user interface that supports the transition between summaries and full text. This "summary in context" functionality uses text shading to reveal the summary sentences yet retain the context in which they occur in the source documents. It thus enables the users to tailor the topics of the documents that are emphasised in the summary.

In order to evaluate IDS they recruited 18 participants to read through two randomly assigned articles and highlight the key sentences in these articles. These assignments were done such that each article would be reviewed by six different participants. They then constructed an “ideal” summary by taking the majority opinion of the participants on which sentences are to be included. They measured the performance in terms of precision and recall of “relevant” sentences. They considered a sentence “relevant” if it was selected by at least two participants. For the baseline they used a simple ranking of sentences by the order they appear in the text. They found that the IDS summaries outperform baseline F-scores (by a maximum of 7%).

NewsInEssence [424] is a digital news system that produces domain-independent multi-document summaries of news articles related to the user’s current news story of interest. Therefore, it builds a personalized view of news sources. It also provides an interface with which the user can visualize the cluster of retrieved news articles related to his seed article. Each article is represented as a data point in a two-dimensional space which indicates the time and the news resource associated with this article.

2.3.3 Information Extraction

Information Extraction has focused on the extraction of entities and relations between them from natural language texts. IE techniques are exploited as the first step for QA and summarizarion systems. These systems leverage IE to improve their effectiveness by finding key entities and the relationships between them to generate candidate answers. In fact, it is envisioned that in the future, the main source of structured data to build knowledge bases will be automatically extracted from natural language sources [145]. By applying an IE system to a collection of documents a set of triples in the form of (entity, relation, entity) will be generated. These triples, when combined, can generate a Knowledge Graph. These graphs organize the entities present in text in a 2 dimensional networks by connecting each pair by the underlying semantic relationship between them.

While there are many external knowledge bases (e.g., DBPedia [2]) are available that provide a collection of these triples, they cannot necessarily represent the same

[2] http://dbpedia.org

information space that the user is interested in. That is, in order to provide better support for users' complex search activities, it is beneficial to derive these knowledge graphs from the information retrieved for the user's query. These graphs can provide an alternative representation of the user's domain of interest and they can support interaction with the corresponding documents retrieved by the search engine. To this end, we are interested in methods that extract entities and relations between them from text.

Many studies have been proposed to extract these triples from text. While some approaches extract entities or relations independently [648, 11], others aim at joint extraction of entities and relations using a dual approach [109, 269, 427].

One example of methods that automatically derive lexical knowledge graphs from text is MindNet [530]. This approach introduces a new "model" to extract semantic relations fully automatically from text using the Encarta encyclopedia and lexical-semantic relations discovered by MindNet. MindNet is a lexical knowledge base that can be constructed fully automatically from a given text corpus without any human intervention. Through MindNet, words are connected to other words within a sentence, across sentences and even across documents. For example, if the word "car" occurs in a particular sentence then it will be connected to words in the same sentence and also to words in other sentences present anywhere in the corpus wherever the word "car" is present. In this way one is able to retrieve how a particular word is related to other words in sentences across documents in a corpus. Unlike MindNet, our Information Extraction component is not limited to any knowledge resource (e.g., Encyclopedia) to extract the relations. Also, the relations we extract imply a "semantic" connections between two entities and it is based on deep linguistic features as opposed to the occurrence of those entities in the same sentence only.

More recently, IE researchers concentrate especially on Open Information Extraction (OIE), where information is automatically extracted from large textual resources, e.g. web news articles, which are not restricted to any particular domain or terminology.

One of the first successful systems for the fast and scalable fact extraction from the Web is the domain-independent system, KnowItAll [162]. KnowItAll starts with the extraction of entities of pre-defined entity types (e.g. CITY, MOVIE) and then discovers instances of relations between extracted entities using handwritten patterns. Another system called TextRunner [628] applies a technique of extracting all meaningful instances of relations from the Web.

The system ReVerb [163], in turn, overcomes some limitations of the mentioned systems due to a novel model of the verb-based relation extraction. However, Re-

Verb suffers from two key drawbacks. Firstly, it uses a limited subset of sentence constructions for expressing relationships and they all are mediated by verbs only. Secondly, it only analyzes a sentence locally so it often extracts relations that are not asserted as factual in the sentence (e.g. extracting (Romney, being the president of, US) from "if Romney wins the votes of 5 states, he will be the president of US."). A newly developed OpenIE system called OLLIE [483] overcomes the limitations of previous Open IE systems by (1) extracting relations mediated by nouns, adjectives, and more which leads to a higher yield and (2) a context-analysis step increases precision by including contextual information from the sentence in the extractions.

2.3.3.1 Ranking Extractions

Extraction engines, similar to search engines, intermix relevant and irrelevant information. This problem is exacerbated in IE systems because they use heuristic methods to extract phrases that potentially contain entities and relationships [334]. Therefore, there is a need to rank these extractions in order to (1) filter out the not "useful" assertions as well as (2) providing an ordered list of suggestions for the users as candidates for exploration. There are different criteria to rank these extractions. These criteria can be classified into two groups: Subjective and Objective. We can also classify approaches in these groups into Graph based and Semantic based Measures. An overall classification of ranking approaches that we review is as follows.

1. Objective
   (a) Graph-based
      - Betweenness (e.g., [650])
      - Centrality (e.g., [637, 68])
      - PageRank (e.g., [650])

   (b) Semantic-based
      - Co-occurrence and Frequency (e.g., [417])
      - Using Ontologies (e.g., [531])
      - Semantic Similarity
2. Subjective
   (a) Informativeness (e.g., [267])

(b) Interestingness (e.g., [334, 426])

(c) Unexpectedness (e.g., [336, 184])

(d) Actionability (e.g., [212])

There are different combinations of the mentioned graph-based and semantic-based ranking methods used in different OIE systems. NAGA [268] is a graph-based search engine, which ranks subgraphs, based on “confidence”, “informativeness” and “compactness” based on a user’s query.

Kasneci et al. [267] first formulate a measure for representing the informativeness of relations between entities in an ER graph and then employs this measure to determine the most informative subgraph for a given query. The authors argue that in order to compute the informativeness of nodes in ER graphs, the link structure has to be taken into account. However, they also argue that the link structure on ER graphs represents a fraction of the real world only and thus it is not sufficient for an effective informativeness ranking method. They propose to assign weights to the edges based on co-occurence statistics for entities and relationships. These weights will then guide a random walk process on the adjacency matrix of the ER graph (similar to PageRank). To this end, they use a similar measure as “distinguishing assertions” employed by [334]. In order to evaluate their ranking algorithm (MING), they simply compared the answers provided by MING to the ones provided by a different ranking system (CEPS [R]). They focused on the user perceived quality of MING’s answers. Therefore, they conducted a user study and collected 210 assessments in total, out of which MING’s results were marked as informative in 185 of the cases.

Zouaq et al. [650] experiment with different graph-based measures for ranking nodes and entities. They use nodes’ degree, betweenness centrality for nodes, betweenness centrality for edges, Page Rank and etc and compare the results against a different set of ranking approaches based on Pointwise Mutual Information (PMI) [114] and frequency of co-occurence to prove to the superiority of graph-based measures.

Lin et al. [334] develop three distinct models of what assertions are likely to be interesting in response to a query. They apply these three models to filter out less useful statements extracted by TextRunner [628] extraction engine. The authors argue that the problem of ranking based on interestingness is especially challenging because “interestingness” can be subjective, personal and context specific. They define interesting assertions to be those that a person may find useful or engaging. They identify three qualities of interesting assertions as follows:
(a) they tend to provide more *specific* information. For example “Steve Jobs was the

CEO of Apple" is more interesting than "Steve Jobs was the CEO of a company". They formulate *specific* assertions as the ones that either relate multiple proper nouns or contains a year;
(b) they might be providing *distinguishing* information about an object. For example "Obama is the president of US" is more interesting than "Obama is a man", because it sets him apart. They used a technique similar to TF-IDF weighting to formulate the notion of *distinguishing*.
(c) "basic facts" can also be interesting for a person learning about an object. They trained a classifier based on the facts included in Wikipedia infoboxes in order to distinguish these facts.

They evaluated these three models by conducting a user study to collect human ratings of assertion interestingness. They gathered a total of two or three ratings for every assertion. They reported an inter-annotator agreement of 71%. They also used these human labels to evaluate a classifier than combines all these three models and concluded that the hybrid model outperformed each of these three models when applied in isolation.

Ram [426] presents a theory of interestingness that serves as the basis for two story understanding programs. This theory is based on the analysis of the knowledge goals that underly the understanding process.

"Unexpectedness" is another criterion that can be used for ranking extractions and suggesting a next candidate for exploration.

As defined in [336] "A piece of information is unexpected if it is relevant but unknown to the user, or it contradicts the user's existing beliefs or expectations." Liu et al. argue that retrieving the information that is explicitely specified in a user's query is not sufficient to fully satisfy user's information need. Unexpected information can also be interesting and useful for the user. They have developed a system that for given competitive Web sites, merges all the pages of the sites, and then clusters the pages hierarchically according to the feature vectors of the pages. The system presents the clustering result in a tree form, in which the Web site that each page belongs to is indicated by the node color. Their system is useful for browsing through the entire contents of competitive Web sites and grasping the difference between the contents.

Another approach towards identifying unexpected events is the realization of what events are *expected* based on how frequently they occur in people's everyday life. Gordon and Schubert [184] aimed at releasing a collection of the resulting event frequencies, which are evaluated for accuracy. They utilize a set of patterns that contain words such as “always”, “usually”, “hourly”, “daily” and etc to extract mentions of different events from large text corpora and estimate their frequency. Their results can be used as the first step towards providing commonsense knowledge datasets needed for many AI applications.

How topic-centric, or informative, a word is can be valuable information for ranking entities. One extension of IDF measure called Residual IDF (RIDF) introduced in [349] has proven effective for automatic summarization [398] and named entity recognition [436]. In rIDF, the usual IDF component is substituted by the difference between the IDF of a term and its expected IDF according to the poisson model.

In a broader sense, Informativeness can be defined in an objective fashion. While evaluating an extraction as an informative assertion can be different for different users, we are able to apply objective semantic-based measures to rank the relations based on the informativeness of the corresponding entity pair. Such approaches usually take advantage of co-occurrences and statistical frequencies to rank entities / entity pairs.

Cumby et al. [130] focuses on task specific entity retrieval for Enterprise domain. They model the target topic and each extracted entity as a vector and rank entities based on the cosine distance between the entity’s vector and the target topic’s vector. Jin et al. [251] created a graph by aggregating multiple relations between the same entity pair. In order to combine multiple relations they proposed three different approaches (1) Choosing the most predictive type of relation (using different measures for defining “importance”, including degree centrality, betweenness centrality and closeness centrality) (2) Learning ranking using a probabilistic model (performing a random walk on the graph) (3) Integrating multiple indices from the network (using SVM regression).

2.4 Designing Exploratory Search Systems

Alongside efforts that provided more focused textual representations relevant to a user’s query, researchers from a variety of disciplines including information and library sciences, human computer interaction, psychology and cognitive sciences focused on the unique nature of complex tasks where the target is unknown and information need is evolving. Consequently, a number of desired features of systems for supporting complex and exploratory search tasks were identified and studied. We review these features in this section and elaborate on the past work that supports a subset of these features that are more closely related to the focus of this dissertation.

The design of Exploratory Search Systems (ESSs) presents unique demands, unlike searchers where the target is well known or where a single document or fact will suffice. The primary observation is that supporting users during their interaction with the information space requires a more advanced design than the classic ranked list of documents provided by current search engines. White and Roth [597] presented a set of features that must be present in systems that support exploratory search activities. Among them we can refer to (1) supporting querying and rapid query refinement (e.g., [366, 461]); (2) provide groupings and organization of search results; (3) offer visualizations to support insight and decision making (e.g., [13]); and (4) support learning and understanding.

A review of available ESSs indicate that only a subset of these features are supported by such systems. The research presented in this dissertation is our attempt to addressing the need for meaningful organizations of search results that are amenable to learning and comprehension of these results. To this end, as described in Chapter 1, we leverage information extraction and visualization techniques to construct and evaluate the efficacy of alternative representations of search results. In the rest of this section we review the related efforts in supporting exploratory search tasks through organizing and structuring the search space (Section 2.4.1), offering visualizations of information to support insights and decision making (Section 2.4.2) and providing spatial representations of information to support browsing as well as learning and sensemaking activities (Section 2.4.3).

2.4.1 Organizing Search Results

Information seekers often express a desire for interfaces that organize search results into meaningful groups, in order to help make sense of the results and guide decision making [215]. A taxonomy of techniques for organizing search results was proposed by Wilson et al. [611]. They identify two main classes of approaches: (1) Coupling results with additional metadata and classifications such that searchers can interact and control the presentation of results. (e.g. faceted browsing or categories), or (2) providing alternative or complementary representations of search results (e.g, a network representation). Essentially, the first class of approaches involve efforts that add classifications to structure the search results, whereas the second group attempts to directly organize search results into alternative representations including spatial representations of search results.

Following on the first class, Wilson et al. also present four common approaches to structured classification [611]: hierarchical classifications, faceted classifications, automated clustering, and social classifications. In this class, we look at faceted categorization and clustering as two popular approaches for generating useful document groupings. In Sections 2.4.2 and 2.4.3 we focus on information visualization techniques as well as existing spatial representations of information as a way of directly organizing search results and generating complementary representations of these results.

2.4.1.1 Adding Classification

Looking first at structured classification, early forays into the domain of structuring search results contrasted categories with automatic clustering to support search. Hearst [214] showed that categories, because they were more interpretable for the user, captured important information about the document but became unwieldy when the document corpus was too large. Clusters, by comparison, were highly variable with respect to quality and were often less meaningful for the user.

Clustering. Clustering refers to the grouping of items according to some measure of similarity. In document clustering, similarity is commonly computed using associations and commonalities among features such as words and phrases [131]. The clustering procedure is fully automated, can be easily applied to any document collection, and can reveal interesting and unexpected trends in a group of documents [215]. Clustering of search results have been used as a way of making search more interactive (e.g. [216, 638, 415]). For example, Pirolli et al. [415] designed an interface called Scatter/Gather to support search results exploration through text clustering. In order to evaluate this interface they attempted to measure learning and understanding in terms of topic structure and query formulation capabilities at various points during subject interaction with the system. In comparison to a control group that performed the same task using the standard interface of a search engine, users of the Scatter/Gather system showed larger gains in understanding the underlying topic structure and in formulating effective queries.

Despite some benefits of clustering including domain independence, scalability, and the potential to capture meaningful themes within a set of documents, the resultant clusters can be highly variable [214]. As well, generating meaningful groups and effective labels is a recognized problem [446, 215].

Faceted Categorization. Given the lack of intuitiveness associated with clustering [215] and a desire for understandable hierarchies in which categories are presented at uniform levels of granularity [422, 448], alongside specified hierarchies such as tables-of-contents, researchers have explored faceted categories, i.e. categories that are semantically related to the search task of the user, to organize search results. These include systems that define faceted categories [629], research that studies the use of facets to support browsing [87], and research that identifies strengths and weaknesses of faceted browsers [610].

Capra et al. [87] present the results of a study that investigates the relationships between search tasks, information architecture and interaction style. They tested three different user interfaces (standard website, text-based faceted interface and dynamic query faceted interface) by performing three kinds of search tasks (simple lookup, complex lookup and exploratory). This accounts for two different architectures (content-driven semi-hierarchical and faceted structure) and three different interaction styles (hypertext selection, faceted navigation and dynamic query). They found that (1) handcrafted faceted interface is effective in supporting all three search tasks even when the overall design is complex and information intensive; (2) interfaces should support familiar interaction styles such as keyword search, but that users gain benefits from support for facets and topic organization implemented in a flexible fashion; (3) automated metadata and facet extraction presented in generic screen display forms is a feasible alternative to handcrafted structures because they did not penalize effectiveness or efficiency for the participants in their study.

Villa et al. [565] investigated the effectiveness of the mechanisms which enable the users to categorize their search environment during search. They designed an "aspectual" interface which allows the users to create search *aspects*. These aspects can be representatives of independent subtasks of some larger tasks. They conducted a between-subjects user study to find answers to three research questions:
(1) does the aspectual interface allow a user to better explore the task?
(2) does the aspectual interface aid the user in better understanding of the search task?
(3) what features of the aspectual interface are used by the users carrying out the search tasks?

When evaluating the performance of their interface (i.e., the first research question), they found that for the tasks where multiple solutions exist and aspects were implicit in the description, a significant difference was observed between the aspectual interface and the baseline. However, where the aspects were mostly specified in the task description or there was a single solution, no significant difference was observed between these two interfaces. They also found the user perception of the difficulty of the task dropped significantly for the aspectual interface with the search task was complex, had multiple solutions and the aspects were implicit in the description.

In terms of strengths and weaknesses, faceted browsing has proven beneficial for users already clear about their search task [610]; additional information on interactions between facets (e.g. inter-facet relationships) is helpful when users are unfamiliar with a domain and need 'sensemaking'. In other words, exploratory tasks (e.g.learning or investigating [352]) are precisely those tasks where interactions between facets are needed.

2.4.2 Information Visualization

A different class of desired features to be supported by exploratory search systems is to offer visualizations in order to support insight and decision making. White and Roth [597] note that exploratory search interfaces must present customizable visual representations of the collection being explored to support hypothesis generation and trend spotting. In fact, in the domain of digital information seeking, while text as a representation is highly effective for conveying abstract information, reading and even scanning text is a cognitively taxing activity, and must be done in a linear fashion. By contrast, visual information and images can be scanned quickly and the visual system perceives information in parallel. Information and Knowledge visualization techniques are powerful tools for generating these visual representation of information from raw data. Information visualization focuses on the visual representation of large collections to help people understand and analyze data.

Information visualization is an important tool to support exploratory searches. In this section we first motivate the application of visualization techniques to support exploration, comprehension and sensemaking activities. Next, we briefly describe a reference model that formally defines the information visualization process. Finally, we focus on the visualization of large graphs, as it is closely related to the research described in this dissertation, and expand on interaction models that aim at mitigating some of the challenges of representing large graph datasets.

2.4.2.1 Why Visualization?

Visual representations, in general, are structures for expressing knowledge. Information visualization uses graphical techniques to visually represent large-scale collections of non-numerical information and help searchers attain new insights in support of decision making or other related complex mental activities [597]. According to

Card, Mackinlay, and Shneiderman [88], information visualizations can be characterized as "*computer-supported, interactive, visual representations of abstract, non-physically based data to amplify cognition*"(p. 6).

Information visualization amplifies cognitive capabilities in six basic ways [88]: (1) by increasing cognitive resources, such as by using a visual resource to expand human working memory; (2) by reducing search, such as by representing a large amount of data in a small space; (3) by enhancing the recognition of patterns, such as when information is organized in space by its time relationships; (4) by supporting the easy perceptual inference of relationships that are otherwise more difficult to induce; (5) by perceptual monitoring of a large number of potential events; and (6) by providing a manipulable medium that, unlike static diagrams, enables the exploration of a space of parameter values.

The efficacy of visualizations for enhancing our cognitive processing power has been noted by many researchers in the past. According to Ware [582], the "*power of a visualization comes from the fact that it is possible to have a far more complex concept structure represented externally in a visual display than can be held in visual and verbal working memories*". In this regard, visualizations are cognitive tools aiming at supporting the cognitive system of the user. In other words, visualizations can make use of the automatically human process of pattern finding.

Others have highlighted the role of externalizing relationships between concepts in supporting cognition, i.e., the acquisition or use of knowledge. Essentially, visualizations can enhance our processing ability by visualizing abstract relationships between visualized elements and may serve as a basis for externalized cognition [481, 125]. External representations may also help in "computational offloading" [482]. That is, compared with a textual representation of the same underlying information, spatial visualizations may allow users to avoid having to explicitly compute information because users can extract information 'at a glance'. "*Such representations work best when the spatial constraints obeyed by representations map into important constraints in the represented domain in such a way that they restrict (or enforce) the kinds of interpretations that can be made*" [482].

Overall, the idea behind all visualization methods is that orientation, visual search, and cognitive processing of complex subject matter may be enhanced if structures behind ideas, knowledge, and information, as well as their relevance for coping with a particular task, are made explicit [538]. In the next subsection we look at the process of getting from raw data to visual representations that can enhance our ability to process information more effectively.

2.4.2.2 Process of Information Visualization

In striving for a better understanding of information visualization, a variety of classification schemes have been proposed over the past years. Depending on provenance and intention, they shed light on the information visualization process, its application, or its utility. Information visualization techniques, applications, systems, and frameworks can be classified according to the data types they can display, user tasks they support, characteristics of visual representations they deploy as well as cognitive aspects of their visual appearance [538].

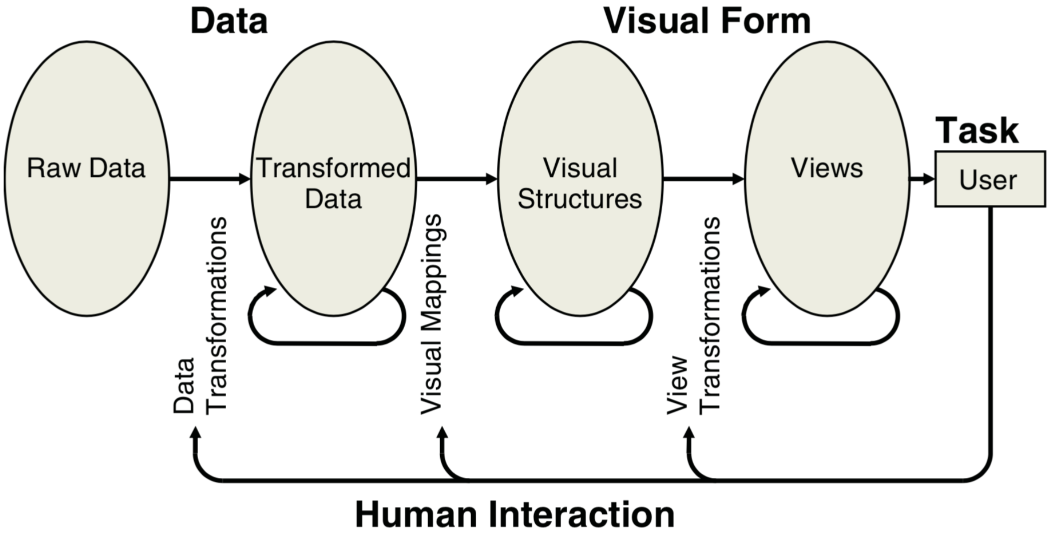


Figure 2.5: Reference Model for Visualization. Figure adapted from [538]

Reference model for visualization. Card, Mackinlay and Shneiderman [88] introduced a reference model for information visualization (Figure 2.5), which provides a high-level view on the (information) visualization process. This model assumes a repository of raw data, which exist in a particular format (e.g. based on an encoding scheme) and can be structured or unstructured. To get to a visualization of this data, data has to first undergo a set of transformations including filtering of raw data, computation of derived data as well as data normalization. Next, visual transformations map the transformed data onto a corresponding visual structure, i.e., a representation. From this visual structure, a set of views can now be generated, which allow users to navigate through the representation. The cyclic arrows in the diagram refer to the fact that the processes involved in the distinct steps are of an iterative nature and can occur repeatedly before the next step follows.

A key takeaway from this reference model is that different types of data require different visual forms which in turn impact the choice of interaction models that can best support manipulating these visual representations. Shneiderman [495] suggested a taxonomy for information visualization designs built on data type and task, the type by task taxonomy (TTT). He distinguished seven data types: 1-dimensional (e.g. textual documents), 2-dimensional (e.g. geographical maps), 3-dimensional (e.g. real world objects), Temporal (e.g. time series), Tree (e.g. hierarchical structures such as table-of-contents) and Network (e.g. entity-relationship maps).

Given our interest in representations of entity-relationship data we focus specifically on network based representations and elaborate on generating appropriate views of these representations and interaction models that are tailored to them in the next subsection.

2.4.2.3 Visualizing Large Graphs

Graph visualization research concentrates on the development of effective graph layouts and visual mappings between concepts and relationships to their corresponding visual elements on the display. The visualization of large graphs is indeed challenging, in particular, in cases where the whole graph is too complex or large to be visualized in one static view. To mitigate this challenge, large graphs visualizations are often accompanied by providing **multiple views of the underlying data** as well as **effective interaction techniques** to transition between these views. Essentially, the interface provides a visualization of the underlying data such that the users are able to manipulate the view in order to highlight patterns, investigate hypotheses, and drill down for more details. Further, users must be able to select items or data regions to highlight, filter, or operate on them. Large information spaces may require users to scroll, pan, zoom, and otherwise navigate the view to examine both *high-level patterns* and *fine-grained details*. We elaborate on commonly generated views of a graph representation as well as effective interaction techniques that can be coupled with these selected views of underlying data next.

Global and Local Views. There has been a lot of debate over the impacts of global and local views that a graph can provide for an information space. Most network visualizations tend to provide a global perspective on a graph by attempting to represent an overview of the information space so no information is missing and the data can speak for itself. However, there has been some work favouring local views. In the context of online communities it has been shown that "*starting with what you know*" can serve a more useful approach [217] than the established principle of "*overview first*" [495]. It has also been shown that when dealing with particularly large networks, "search relevance" can be used to initially establish a partial context and then expand the visualization from there [562].

Coupled with these two views of large graphs, theories often suggest two predominant navigation paradigms: top-down or overview-first and bottom-up or expand-from-known. We elaborate on these two navigation paradigms, as means of exploring graph representations, next.

Top-Down or Overview-first Navigation Paradigm. Top-down approaches are coupled with global views of the data and are best characterized by Shneiderman's mantra "*overview, zoom & filter, details-on-demand*" paradigm in visual information seeking [495]. These approaches have conventionally received much attention and have worked well for numerous kinds of data in many domains (e.g. [108, 35, 277]). However, in the era of big data, top-down approaches that focus on providing overviews of global information landscapes face significant challenges when applied to graphs with millions or billions of nodes and edges [277, 278]. The seminal work on graph clustering by Leskovec & Faloutsos [324] suggests there are simply no perfect overviews (i.e., no single best way to partition graphs into smaller communities), a view echoed by sensemaking literature in that people may have very different mental representations of information depending on their individual goals and prior experiences [275]. Graph sensemaking is a complex and abstract task, highly dependent on both domain and data. For this reason, it is highly unlikely that a single visualization will be sufficient for all sensemaking tasks.

Bottom-Up or Expand-from-Known Paradigm. Bottom-up approaches are coupled with local views of the graph and generally support the idea of starting with a small subgraph and expanding nodes to show their neighborhoods. These approaches are particularly helpful in scenarios where the user is not aiming to learn about global patterns in the data, and rather is interested in learning something about a particular data-point, or an entity, in the graph and how this entity relates to the rest of the graph. A different advantage of bottom-up approaches is that they don't impose a structure to the information space and help users construct their own landscapes of information. This way of exploring the information space may be preferred in certain scenarios because people may have very different mental representations of information depending on their individual goals and prior experiences. Bottom-up exploration in hierarchical graphs was first investigated in [175] and later expanded on by [562] to incorporate the idea of "degree of interest" to help users identify which nodes to explore. Other systems like Apolo [102] do not impose an hierarchy on the data, allowing users to freely define their own clusters, which Apolo incorporates into its machine learning algorithm to infer which nodes the users may want to explore next. Among other approaches in this space we can refer to research that has explored local exploration of graphs, including Treeplus [320] and Vizster [217].Treeplus [320] supports the exploration of the local structure of a graph based on a guiding metaphor of "*Plant a seed and watch it grow*". This interaction model allows users to start with a node and expand the graph as needed, which complements the classic overview techniques that impose a global view of the data as a starting point. Similarly, Vizster [217], focusing primarily on the domain of social networks, is also promoting a "*start with what you know, then grow*" approach to graph exploration as an overview of the full network is not helpful in this personal context.

2.4.3 Spatial Representations of Information

In the previous section we described related efforts for visualizing large graphs as a way of generating spatial representations of entity-relationship data. In this section, we specifically focus on the use of these spatial representations for supporting exploration, investigation and sensemaking activities, essentially to enable effective exploratory search.

Since their development, concept mapping [391], knowledge graphs [246] and linked data [62] have been widely used in education and capturing knowledge, and as a method for knowledge examination, sharing and browsing. These graph-based datasets allow a natural visualization and browsing of information and simplify the implementation of learning and investigating strategies for knowledge acquisition and discovery. Concept Maps, Linked Data and Knowledge Graphs all express concepts/entities and relationships in a network; they use natural language for node and link labels, and the concept-link-concept triples of these graphs form simple natural language sentences. Amadieu and Salmerón [22] provide a comprehensive survey to examine the effects of concept maps on navigation, comprehension, and learning from hypertexts. Despite the variability of concept maps used in hypertexts, some findings converge: *"Concept maps reduce the cognitive requirements for processing hypertexts. They support outcomes as well as guiding learner navigation. They* *convey a macrostructure of the semantic relationships between content that supports more coherent navigation and promotes the construction of a mental representation of the information structure of hypertexts"* [22].

While these graphs appear to be effective as a cognitive strategy to stimulate learners to make cognitive progress in organizing and understanding new information [390], there is limited insight into how they can be used by the searchers to support learning and sense making during an exploratory search task. Here we review two groups of approaches that leveraged these spatial representations to support searching and browsing (Section 2.4.3.1) as well as sensemaking and learning activities (Section 2.4.3.2).

2.4.3.1 Structuring Information for Search and Browsing

The models most similar to our work are those which make use of entities and the relations between them to support search. Dimitrova et al. [143] designed a semantic data browser based on external Linked Data resources to support exploratory tasks. Their study is qualitative and exploratory in nature and examines (1) obstacles and challenges related to user exploratory search in Linked Open Data (LOD) and (2) the serendipitous learning effect and the role semantics play in that. They designed two search tasks for which the participants were expected to find the main characteristics of a particular musical instrument, its similarities and distinctions to other instruments, and usage and features of a different musical instrument. The participants reported their findings in a structured form which were then evaluated by two musical instrument experts.

Yan [623] argues that (a) entity data could be used by Web users for navigation purposes (e.g., browsing docs retrieved from a search engine); (b) entity data could also be used for better understanding of the data itself. They proposed to produce a faceted interface for exploring the documents retrieved for a keyword query automatically and dynamically by exploiting "collaborative vocabularies" in Wikipedia. They also propose to summarize Entity-Relationship (ER) graphs into multiple relational tabels. However, the ER graphs they work with are limited to graphs with Named Entities as nodes and simple predicates as edges. They do not provide any empirical study or evaluation for this design.

Yogev et al. [632] describe an extended faceted search solution that allows to index, search and browse rich ER data. The output of the search system is a ranked list of entities that are distributed over different facets. These facets can be used by the user to focus the search on a specific entity type or to explore another direction by navigating to another related entity in the ER graph.

With the introduction of the so-called “Knowledge Graph”, Google has made a significant paradigm shift towards “things, not strings” [502]. Entities covered by their graph include landmarks, celebrities, buildings and more. The “Knowledge Graph” enhances Google’s search in three main ways: query disambiguation, providing a summary of related facts to the user’s query, and exploratory search suggestions (based on what other users explored next).

2.4.3.2 Structuring Information for Sensemaking

People spend a significant amount of effort capturing and organizing relevant information during exploratory activities. As observed by Bates [44] this information is collected in pieces and they need to be stored and processed. Structuring these collected pieces of information is shown to assist the problem solving and reasoning activities [44, 64, 413].

Similarly, the process of analysis is one of sensemaking [413] in which analysts constantly forage for relevant information, integrate that information into schemas or hypotheses that explain what they have found, and use these schemas to guide decisions. A variety of tools have been developed to improve the sensemaking and analysis process, many of which have focused on the area of crime analysis [185, 517, 526]. Solving a crime requires the analyst to “connect the dots” by identifying links between facts across documents, time and space [185]. Jigsaw [517] is a visual analytics system that represents connections and relationships between entities in document collections. It was developed to help analysts search, review and understand the crime reports better.

Goyal et al. [185] compared the utility of a visualization of relationships among entities and documents and a notepad for collecting and organizing annotations. The Visualization mode shows all the documents in the dataset that contain entities in common with the active document as edges between the document nodes. They calculated TF-IDF for the unique entities in common between two documents to assign weight to these edges reflected by their thickness. The Notepad mode is a text editor where evidence can be collected by highlighting important text found while reading a document. They found that the visualization of relations between documents significantly improved participants’ ability to solve the crime whereas the notepad did not. Piolat et al. [408] also showed that a matrix structure for recording information is more beneficial than an outline structure, which is in turn more beneficial than a linear structure.

Graphic organizers that resemble a concept map have been proved to aid memorization and memory offloading through facilitating organization of the mental representation of texts [154, 64, 498]. Shrinivasan and Wijk [498] present a new information visualization framework that supports the analytical reasoning process through a data view, a knowledge view and a navigation view. In this design, the user records the findings in the knowledge view using a mind map. Four analysts tested this system and found recording the findings, linking them to the visualization and organizing them very important for their analysis process and it improved their quality of results.

While the strategies of note-taking and the structures used for memory offloading is crucial for developing effective techniques for sensemaking and analysis, the stage in which the searchers need to employ these strategies is as essential. Kittur et al. [294] introduce a novel interface for capturing online information in a structured but lightway way to empirically characterize the costs and benefits of structuring information. They found that it is more effective and less costly to start structuring the collected information *after* information foraging is done. The data they gathered by the user study they conducted indicated that the foraging process is very dynamic and people's mental models change significantly over time. Therefore, there are significant costs to eliciting structure early in the foraging process.

On the other hand, another product of sensemaking and analysis process is generating coherent schemas. Hummel and Holyoak [238] suggest that the schema induction involves the alignment of many examples in order to find the commonalities and overlaps between them. Hence, Kittur et al. [294] propose the need for a two-stage process in which information is first saved and then later structures. Zhange et al. [640] also argue that a user doesn't have a clear idea of structure initially. Therefore, they designed an environment to support literature search and analysis such that the user can organize papers into informal clusters by simply moving them in different areas in the working space. Later, he gets a better understanding of the structure in the collection, and creates hierarchical sections to relate the items he collected. Indeed, encouraging a second stage of structuring information can promote induction and the formation of a better structured information space which could be useful for both the current searcher and others interested in the same information [294].

Given the benefits of structuring information in supporting analysis and sensemaking, a main research question we wished to explore in this research is that whether representations of information that are generated automatically and correspond to a searcher's information need can enable more effective extraction and assimilation of information? We address this question through a series of designing and evaluating alternative representations of search results in Chapters 4, 5 and 6. The research described in these chapters is inspired by the premise that exploratory search tasks require sensemaking and sensemaking involves constructing and interacting with representations of knowledge.

> *Most cognitive scientists believe, learning best begins with a big picture, a schema, a holistic cognitive structure, which should be included in the lesson material-often in the text. If a big picture resides in the text, the designers' task becomes one of emphasizing it. If this big picture does not exist, the designers' task is to develop a big picture and emphasize it. (West, Farmer and Wolff, 1991, p. 5) [592].*

2.5 Evaluating Exploratory Search

A final aspect of developing solutions for supporting exploratory search tasks involves significant challenges in evaluation. Over the last decade, researchers have focused on the development of systems and interfaces to support exploratory search activities (see Section 2.4). Yet designing new evaluation metrics and methodologies that are tailored to these complex, evolving and highly interactive search scenarios is much less explored [595, 558]. As White [595] posits. it is necessary to shift the focus of research in exploratory search towards understanding the behaviors and preferences of users engaged in exploratory searching, the tasks supported by exploratory search systems and can elicit exploratory behavior, and on measures of exploration success.

When evaluating exploratory search systems, it is impossible to completely separate human behavior from system effects because the tools are designed such that they are closely related to human acts and their intentions [597]. Evaluating the success of search systems in supporting a range of information seeking tasks involves leveraging two related aspects of an evaluation paradigm: Evaluation Measures and Evaluation Methodologies. The performance of search systems is often measured using a wide range of **evaluation measures or metrics**. Cognitive load, engagement, precision and recall are all examples of common IR metrics. These metrics facilitate benchmarking a system's performance as well as quantify the impact of any changes. **Evaluation methodologies**, on the other hand, are tightly connected to how user interaction behavior is presented and is also dependent on the metrics adopted for measuring success. Essentially, evaluation methodologies connect models of interaction and success metrics by specifying the rules, methods employed in evaluation, as well as, rationale and philosophy behind the evaluation [595].

In the rest of this section, we review a subset of metrics (Section 2.5.1) and methodologies (Section 2.5.2) that are commonly leveraged for evaluating search systems and reflect on the ones that are deemed more suitable for assessing exploratory search systems. We end this section by expanding on two areas of past research that contributed to opening new venues for evaluating exploratory search systems (Section 2.5.3.

2.5.1 Evaluation Measures

Evaluation metrics facilitate tracking the incremental improvement of search systems and comparisons between experimental systems by providing a way to assess system performance. White [595] categorize evaluation metrics into two groups: (1) process-oriented metrics and (2) outcome-oriented metrics. We review these two classes of metrics next.

2.5.1.1 Process-oriented Metrics

These measures are calculated based on the value of the process that searchers engage in to meet their information goals. These metrics can capture behavioral traces of searchers during their interaction with the system and in certain scenarios (e.g. in controlled lab studies) additional information about the process (e.g. searcher's rationale for a certain action) can also be collected. This data can be captured as the search process is unfolding (e.g. using think-aloud protocols) or after it is completed (e.g. using stimulated recall). White [595] surveys a number of metrics that were used in the evaluation of search systems. The metrics range from learning and cognitive load to efficiency and also include subjective metrics such as engagement, enjoyment and frustration.

While process-oriented metrics are not commonly used for the evaluation of IR systems in the past, they are essential for investigating the effectiveness of the next generation of the search systems that are designed for supporting information seekers engaged in more complex and exploratory search tasks. In 2006 a workshop entitled "Evaluating Exploratory Search Systems" organized by Ryen White, Gary Marchionini and Cheorghe Muresan was one of the first joint efforts to discuss exploratory search evaluation and it led to identifying metrics that are affiliated with the process of exploratory search tasks. Among them we can refer to *engagement and enjoyment*, described as the degree to which the users are engaged and are experiencing positive emotions throughout the search process, and *task time*, that is computed as the time spent to reach a state of task completeness as an effective way to asses the efficiency of exploration activities.

2.5.1.2 Outcome-oriented Metrics

A different group of metrics reflect search outcomes and are often computed after the search process is completed. Traditional IR metrics are focused on the output of the search systems and assess the retrieval performance given a user's query. The two main classes of these IR metrics assess the quality of retrieved search results based on their *relevance* to the query as well as their *novelty and diversity*[115].

In the classic IR paradigm, measuring the quality of the output of retrieval systems was synonymous with assessing the outcome of the search task. A broad range of metrics have been proposed for evaluating the relevance of results returned by search systems. We can refer to precision and recall, mean reciprocal rank (MRR) and discounted cumulative gain (DCG) as three popular relevance metrics. The main premise behind all these metrics is a user model describing how searchers examine the search results as well as the impact of the search results quality on the outcome of the search (e.g. the amount of user effort and information gained per rank positions). For example, measures such as precision and recall consider all relevant items up to a rank position (see [35]), whereas, MRR [571] is the multiplicative inverse of the rank of the first relevant result (or correct answer), averaged across all queries. DCG, on the other hand, is a measure of search engine effectiveness that uses a graded relevance scale of documents in the search results set, unlike precision, recall or MRR metrics that assume binary relevance labels.

More recent efforts proposed a new class of relevance based metrics that leverage more realistic models of users interaction with the search results. For example, Smucker and Clarke [509] introduced a measure known as time-biased gain which criticizes the earlier models that assume searchers carefully examine each of the search results at a constant speed. Essentially, the time-biased gain metric considers temporal effects during the search process and builds on a range of previous research that utilized time to represent the cost of interaction (e.g. [43]).

Relevance based metrics contributed to designing effective ranking algorithms and powerful retrieval systems that are capable of retrieving a set of documents that are ordered based on their perceived relevance to a user's query. These metrics, however, exclude the user from the evaluation process and are only concerned with the system based utility. While evaluation metrics are meant to reflect the searcher preferences, a number of studies by Hersh et al. [222], Turpin and Hersh [550] and others (e.g. [469, 510]) revealed that the ranking of search systems based on offline evaluation metrics does not necessarily match the online evaluation results. These studies have shown that the Cranfield-style metrics can have low predictive power about the performance of search systems [595] mainly because they do not attempt to model a searcher or the interface with which they are engaged at the query time [510] and the complexity of searchers' search strategies [504].

In the end, as the range of tasks for which search systems are used continues to grow, the search process will become more involved and thus play an important role in the evaluation of search systems. Essentially, when designing search system evaluation, designers should consider utilizing both process-oriented and outcome-oriented measures [595], where outcome-oriented metrics should leverage search outcomes beyond the output of retrieval systems. We elaborate on the relation between the output of search systems and the outcome of search tasks in Section 2.5.3.2.

2.5.2 Evaluation Methodologies

A number of methodologies have been used to evaluate search systems that range from offline test collections and simulated search models to living labs, ethnography and large-scale log analysis. White [595] classifies the available methodologies across three axes: **stage**, that is the point in the design process that the method is used; **scale**, that is the scale at which the method is employed and is one of small, medium or large; and **participants**, that are the people involved in the evaluation of the search systems. Dumais et al. [153] focus on the analysis of behavioral logs of searchers interactions and categorize experimental methodologies into **(1) observational**, where people may be observed searching naturally; and **(2) experimental**, where the search experience may be manipulated using an experimental design.

Various evaluation methodologies provide different perspectives on search system performance and the goal of this section is not to provide a comprehensive survey of all available methodologies. We can refer the reader to different surveys of IR systems evaluation methodologies that focus on interactive IR [281], using test collections [469, 199] or are based on behavioral data [153].

In order to effectively evaluate a search system, researchers may choose to express their experimental objectives in the form of hypotheses. In fact, a very first step to evaluating the ESSs is to formulate clear research questions that can guide the experiment and inform the choice of suitable metrics and methodologies for evaluation. Once a descriptive list of research questions or hypotheses is specified, the researchers can choose a set of metrics to assess the performance of a search system as well as the experimental methodologies that can lead to collecting reliable types of data and providing supports for these hypotheses. There are many different scenarios where the research questions are complex or multifaceted which requires a heterogeneous collection of data sources. Mixed methods approaches [126] are great candidates for providing different perspectives on different variables of interests and enabling a holistic characterization of the studied users behaviors. We elaborate more on these approaches at the end of this section.

Similarities between understanding the utility of search systems for humans who are engaged in a variety of search tasks and other fields that involve understanding humans or machines or both has led to leveraging many existing experimental methodologies from areas such as IR, HCI and Psychology. Essentially, in classic IR, experiment and evaluation have been used interchangeably, but as Kelly [281] argues these two types of studies need to be separated when discussing Interactive IR scenarios. That is, one can conduct an evaluation without conducting an experiment and vice versa. Evaluations are conducted to assess the utility of a system, interface or interaction technique, whereas, experiments have historically been the main method for interactive system evaluation and for understanding human behaviors.

As a result, most of evaluation methodologies that are suitable for interactive search scenarios are very similar to those conducted in social science disciplines such as psychology and education. Focusing on interactive search systems, White [595] surveys these methodologies and expands on the strengths and weaknesses of each evaluation method based on the prior work done by Grimes et al. [192]. These methods ordered from small scale to large scale include interviews and focus groups, instrumental panels, lab studies, crowdsourced studies, surveys, retrospective log analysis, online evaluation and offline evaluation. In the remainder of this section we elaborate on some of these methodologies that are more closely related to our work and are utilized for a number of experiments reported in this dissertation.

Interviews. In Interactive IR (IIR) scenarios, interviews are a common component of many study protocols. Interviews involve one-to-one dialogs between the experimenter and the experimental participant. Interviews can be structured, unstructured or semi-structured and the questions can range from open-ended and abstract to more focused enquiries with a possibility of open-ended follow up questions. Interviews are commonly conducted after an experiment is completed, giving the participant an opportunity to reflect on their experience using the experimental system. While interview questions can be delivered via print or electronic questionnaire, the in-person interviews allows one to get more individualized responses and offers some flexibility with respect to probing and follow-up. Kelly et al. [282] compared participants' responses to a set of open-ended questions across three modes: interview, pen-and-paper,and electronic and found that while participants' responses were longer in the interview mode than in the other two modes, the number of unique informative statements they made in each mode were about equal. This observation indicates that there are certain scenarios where an in person interview can be a preferred method (and that for some scenarios a follow-up questionnaire or survey will suffice). To elaborate, when an experimenter is interested in asking more complex, abstract questions then it is likely that the interview mode would be more appropriate. Another place where interview techniques can be used in Interactive IR evaluations is during stimulated recall where participants can verbalize their decision-making processes and thoughts while watching a video recording of a search they recently completed [281]. During this process, the researcher might interrupt with specific pre-planned questions. These questions might be used to investigate something specific, or to probe remarks or actions made by participants. While in-person interviews often result in deep understanding of users behavior and rich and detailed datasets, the cost associated with the process of conducting the interviews and transcribing the data usually leads to including a small number of experimental participants. Hence, triangulating interview data with other sources of behavior data (e.g. quantitative logs of searchers interaction with the system) can improve the generalizability of the findings.

Lab Studies. Controlled lab studies are the primary means of evaluating interactive search systems and they have been used extensively for this purpose (e.g. [52, 283, 598]). In a typical lab study, participants use one or more experimental search systems to find information described in a small number of prespecified topics. These topics can be derived from common topic sets (e.g. from different TREC tracks), or can be designed by the researchers following existing guidelines for developing simulated work tasks (e.g. [66, 307]). Given these search tasks or topics, the participants interactions with the system(s) are recorded for a later analysis, and they may provide feedback during the search process (e.g. via thinkaloud protocols) or at the end of the study (e.g. via surveys or post-task interviews). A variety of metrics are computed in order to characterize the usability of the systems and the performance of the participants. Typical performance metrics include number of relevant documents found, time spent on task, or quality of the answers provided.

Qualitative measures can also be applied to get a more in-depth insights regarding the perception of the process, the usability of the systems or the complexity of the search tasks. Kelly [281] distinguishes between lab studies and controlled experiments that are conducted in a laboratory. According to her the main factor that makes a study an experiment is the existence of experimental conditions or manipulations in the process of experimental design. Other types of lab study, e.g. usability tests, are thus not an instance of an experiment. Throughout this dissertation we alternate between the use of terms *lab study* or *experiments* to refer to a controlled user study where experimental conditions exist. For a detailed description of common steps in conducting a lab study and the application of user studies for evaluating interactive search systems we refer the reader to [595, 281].

Crowdsourced Studies. The recent emergence of crowdsourcing platforms such as Mechanical Turk and CrowdFlower has enabled low cost, carefully controlled studies of human behavior [292, 400]. These platforms provide access to crowdworkers who are paid a small amount of money to complete some simple human intelligence tasks such as assigning a category label to a query or performing relevance judgement of a set of documents given a search query. In the domain of IR, crowdworkers can also be assigned as surrogate assessors to perform annotation or evaluation tasks. Essentially, crowdsourcing has grown into a viable alternative to expert ground truth collection, as crowdsourcing tends to be both cheaper and more readily available than domain experts. More related to the research described in this dissertation, crowdsourcing platforms can be a promising approach to creating reference datasets for intrinsically evaluating Open ended entity and relation extraction as well as manual assessment of the output of information extraction systems. We describe our developed evaluation framework that applies crowdsourcing workflows to the task of assessing extraction errors in Appendix A.

Mixed Methods Analysis. One of the strengths of the experimental framework that is designed for all of the experiments described in this dissertation is the application of mixed methods approaches as our primary evaluation methodology. Essentially, we leverage mixed methods research to understand how well the exploratory search systems we design satisfy this ultimate goal of enabling searchers to obtain knowledge more effectively. **Mixed methods research** is a methodology for conducting research that involves collecting, analysing and integrating quantitative (e.g., retrospective log analysis, surveys) and qualitative (e.g., think-alouds, interviews) data. This approach to research is used when this integration provides a better understanding of the research problem than either of each alone. To elaborate, **Quantitative data** includes close-ended information such as that found to measure attitudes (e.g., rating scales), behaviours (e.g., number of clicks, number of documents visited), and performance instruments. The analysis of this type of data consists of statistically analysing scores collected on instruments (e.g., questionnaires) or checklists to answer research questions or to test hypotheses. **Qualitative data**, on the other hand, consists of open-ended information that the researcher usually gathers through interviews, focus groups and observations. The analysis of the qualitative data (words, text or behaviours) typically follows the path of aggregating it into categories of information and presenting the diversity of ideas gathered during data collection. In these approaches, research starts with data collection and is motivated by questions that are broad and non-leading, e.g., How do searchers perceive the representation of a knowledge graph as a means of supporting exploratory search tasks? Next, the researcher establishes meaning from views of participants by looking for common patterns, linking these patterns through axial coding and eventually building a theory from ground up. By mixing both quantitative and qualitative research and data, the researcher gains in breadth and depth of understanding and corroboration, while offsetting the weaknesses inherent to using each approach by itself.

2.5.3 Exploratory Search Systems Evaluation

Given the overview of standard evaluation measures and methods to assess the performance and utility of (interactive) search systems in previous subsections, in this section we primarily focus on requirements of evaluation paradigms for assessing exploratory search systems. The role of evaluation in exploratory search is primarily to assess the success of the information seeking process at reaching the information objectives for the current session, if those exist, and achieving higher order learning objectives for the searcher, including the application of the newly gained knowledge to solve problems or to synthesize knowledge in order to design a new knowledge product (e.g. writing a research paper) [597]. As a class of information seeking which involves highly interactive, complex and dynamic search scenarios there is a pressing need for evaluation metrics and methodologies that move beyond minimal human-machine interaction and focus on assessing the outcomes and the process of the information seeking activity as opposed to the output of retrieval systems.

We can refer to at least two groups of approaches that contributed to the evaluation of new generations of search systems. The first body of work involves efforts that criticized the traditional system-oriented approaches and the Cranfield methodology to the evaluation of search systems and motivated techniques that involve users and their interactions with the search systems as an integral part of the information seeking process. The second group of approaches motivated a shift from the assessment of search systems outputs to the evaluation of search process outcomes. We elaborate on some related efforts in these two groups next.

2.5.3.1 From System-Focused Methods to User-oriented Evaluation Approaches

The first body of past research that contributed to the evaluation of exploratory search systems motivated a shift from system-focused methods to user-centered approaches. System-oriented approaches to evaluating the effectiveness of search systems such as the Cranfield method [118], the Text Retrieval Conference (TREC), and other initiatives such as the Cross-Language Evaluation Forum (CLEF) have been central in driving innovation in information retrieval. At the core of these approaches is to sample a set of queries that are deemed representative of the searcher needs, use them as the input to a group of experimental IR systems, pool and judge the ranked results from these systems, and evaluate the quality of these outputs based on their ability to retrieve and rank a set of documents that are judged as relevant by a set of assessors. Similar to the shortcomings of system-oriented metrics, past work has highlighted some of the limitations of the offline evaluation methodologies (e.g. [550, 51]). The main criticism against this group of approaches is that they exclude users from the evaluation of search systems and, as a result, it is not clear whether the improvements noted in offline evaluation settings will actually translate to a better search outcome for the searchers. To elaborate, system-oriented approaches mainly ignore how people formulate their information needs, examine search results and essentially explore the information space to make sense of the retrieved results.

A group of evaluation methodologies, including the approaches described in Section 2.5.2, offer mechanisms to involve the human searchers in the experimental design and incorporate their perception of the search process as well as the perceived utility of search results for their task at hand in the assessment of the experimental search system.

2.5.3.2 From Search Systems Output to Search Task Outcome

A second group of past work that transformed the ways to evaluate exploratory search systems has distinguished between the assessment of search systems output and evaluating the search tasks outcomes. To elaborate, *outputs* are the products delivered by a system, whereas *outcomes* are the benefits the system produces for its users [454]. In information retrieval, the major aim of search systems is to retrieve information that is useful for performing larger tasks (e.g. writing an essay on a given topic). It is often assumed that by assimilating the information obtained from a list of retrieved results set, i.e., from the output of the search system, users are able to proceed in their task and that the search is successful if the items retrieved (i.e. the *output*) contribute to the task and the searcher achieves the desired *outcome* [51, 241, 281, 559].

The assumption that good system output is associated with good task outcome has led to designing numerous information retrieval algorithms that is assessed by their quality of the output in terms of relevance to a user's query through established metrics such as precision and recall. In recent years, however, there have been calls to extend the evaluation framework from the output to the outcome of searching [51, 241, 281, 447, 559]. Robertson ([447], p. 453) posits "From the point of view of a user engaged in a larger task, the retrieval of items of information must at best be a sub-goal. Our understanding of the validity of this as a subgoal, and how it relates to the achievement of wider goals, is limited and deserves more analysis.".

There has been a long history of research (e.g. [510, 469, 222, 223, 549, 504, 559]) that examined the correlation between system-oriented effectiveness metrics (based on the output of retrieval systems) and users performance (as reflected in their search outcome). While the reasons behind these studies' conflicting results are not clear, in the domain of complex search tasks we can refer to at least two factors contributing to the growing evidence that good retrieval performance does not necessarily lead to successful search outcomes (and that weak retrieval systems will not always lead to poor search outcomes); First, in the IR community there is a growing realization that users' search activities are motivated by a *work task* which provides a problem context within which the searcher operates [241]. This view of the search process has motivated new evaluation methods that go beyond assessing the performance of the search system based on independent queries and consider a broader scope of interactions between the searcher and the search system. Alongside a better understanding of work tasks as a catalyst behind search activities, the advances in the design of new exploratory search interfaces has led to a non trivial relationship between the performance of retrieval systems and the success of the search outcomes.

We conclude our review of existing research that contributed to designing new evaluation frameworks for assessing exploratory search systems by elaborating on these two factors and motivate an in-situ evaluation model that we leveraged for evaluating the efficacy of tools we developed throughout this research, which evaluates the output of IE systems in situ using a balanced mix of quantitative and qualitative methods. This approach is detailed in Chapter 7.

2.5.3.3 Understanding a Holistic Evaluation Approach

Synthesizing past research we see ambiguity in the correlation between system-oriented effectiveness metrics, i.e., *the assessment of outputs*, and users performance, i.e., *the assessment of outcomes*. In this section we elaborate on two factors that could explain the reasons behind observing non-trivial relationships between retrieval performance and the success of search outcomes. We conclude with motivating the need for leveraging evaluation approaches that incorporate both system-based assessments of outputs as well as the efficacy of search system in leading to successful outcomes.

From Independent Queries to Work Tasks. The first factor that highlights a non-trivial correlation between system focused metrics and the success of search tasks is the limitations of the query-based approach to evaluation of search systems as evident in the Cranfield Model. Traditionally, the unit of retrieval evaluation is an individual query. More recently, there has been a shift from considering queries independently and satisfying task-relevant information needs one query at a time to supporting the completion of information seeking tasks end-to-end [508]. There are two venues of related research that investigated ways to extend the evaluation of ESSs by considering the searcher's broader problem context and not individual queries. In the domain of observational methodologies (e.g. log based analysis of search interactions) there have been some work to develop metrics that can handle system performance during a search session [237, 205, 206, 249, 508].

In the context of experimental methodologies and more related to our work, simulated search tasks have been leveraged as a primary component of assessing the efficacy of search systems. To elaborate, as we noted in Section 2.5.2, experiments have dominantly been the main method for evaluating interactive search. Similar to classic IR experiments where test collections were used to evaluate the output of retrieval systems based on a given query, most existing experimental settings for interactive and complex search scenarios are based on the assigned task paradigm

[597]. Research on task development (e.g. [66, 307] allows for the creation of simulated work task situations that are well-suited for exploratory search scenarios, as they are comparable between the experiment participants, while they allow for personal assessment of relevance.

One of the main challenges in evaluating exploratory search systems is defining realistic tasks such that (1) they elicit exploratory search behaviour and (2) participants can understand and relate to them. Wildemuth and Freund [605] provided a set of recommendations for designing search tasks that lead to exploratory behaviour in the users. Their main rationale behind designing a formal method for assigning appropriate search tasks is that different studies use different definitions of exploratory search and they all design their own search tasks to evaluate their systems. Therefore, one cannot compare the results of one exploratory search system to another. The authors discuss the attributes of exploratory search tasks and provide a compilation of empirical studies and the tasks which were designed for them. They recommend that exploratory search experiments incorporate simulated work task situations as stimuli for eliciting exploratory behaviour. In addition, they suggest that when a consensus is reached regarding the definition and attributes of exploratory search, it is possible to design the search tasks that are re-usable across studies and will result in a better understanding of exploratory search.

Kules and Capra [307] argue that creating a realistic, representative search task is challenging. They suggested a principled way of designing tasks for evaluating Exploratory Search systems. Their tasks were designed as “writing a paper for a class” and defined such that (1) there was ambiguity to the answer; (2) students needed multiple interactions to complete the task; (3) students were not very familiar with the topic before the experiment. They used these criteria to extract task candidates from log data. Then they refined these candidates and compiled a list of qualified search tasks. Finally, they conducted a user study to investigate if the selected exploratory tasks were indeed different from known-item tasks. The results confirmed that these tasks did elicit exploratory search behaviour for the participants.

Development of New Exploratory Search Interfaces. A different contributor to observing a nontrivial relationship between the retrieval outputs and the outcome of the search process is the development of new Exploratory Search Support (ESS) UIs. These new ESS UIs leverage structured organization of search results and enable more advanced interaction mechanisms than the traditional ranked list of documents. In these exploratory search systems the retrieval of the relevant information is indeed only the first step towards enabling effective information seeking activities.

There are a number of studies that highlight the role of complex interactive search strategies, as a central element of exploratory searches, in the success of the final search process outcome. For example, users are able to adapt their searching to system performance and differences in retrieval effectiveness can be compensated by human effort [504, 550]. Section 7.2.1 reviews similar studies.

Further, richer representations of search results and more advanced search UIs have also impacted the ways searchers examine the search results and perform sense-making activities. Given the new capabilities of these search interfaces and more advanced information seeking strategies, information retrieval accuracy can serve as only one element in assessing the effectiveness of the designed ESS for supporting information seekers. Essentially, the efficacy of search UI and the ways that information is structured and presented to the searcher and the ways that user can interact with this information can indeed impact the success of the searcher.

Focusing on the scope of research presented in this dissertation, we are inspired by the existing research, described in Section 2.5.3, acknowledging the limitations of the traditional IR metrics such as precision and recall for evaluating exploratory search systems. Given that any assessment of the effect of information extraction errors on exploratory search interfaces is absent from past research, we contribute to this space by conducting a mixed-methods analysis of how varying levels of precision and recall in the output of information extraction systems impact user behaviors and outcomes in exploratory search. We elaborate on a proposed evaluation framework that incorporates both the accuracy of extraction systems as well as the efficacy of information representation to assess the effectiveness of an exploratory search support system in Chapter 7.

Chapter 3

Extending the Document Retrieval Paradigm

> *Any piece of knowledge I acquire today has a value at this moment exactly proportional to my skill to deal with it.*
>
> – Mark Van Doren, Liberal Education

This chapter describes our efforts in developing the first component of our proposed Exploratory Search Support (ESS) Framework, as motivated in Chapter 1. Our goal is to design an enhanced IR module that goes beyond document retrieval by extracting relevant pieces of information from textual corpora. To this end, we extend the current Document Retrieval framework by developing an Open Information Extraction (IE) tool that can extract semantic information from textual content of a set of retrieved documents. Our Information Retrieval and Extraction tool can be used to automatically generate entity-relationship triples in order to populate knowledge bases of linked data, while they can also be visualized as knowledge graph representations pertaining to the same set of retrieved documents.

3.1 Motivation

Given the massive increase in information availability, it gets more and more difficult to make sense of the available information. The Web has provided the opportunity to browse and navigate through an extensive information space by utilizing modern search engines. This in turn has led to increasing expectations to use the Web as a source for learning and exploratory discovery. Further, as noted in the previous chapter, the current document retrieval paradigm offered by major search engines is generally sufficient when the searcher's information need is straightforward and well-defined. However, when the information is sought for learning, investigation or other complex mental activities, retrieval is necessary but not sufficient [31, 597].

Overall, there is a recognized need for search systems to provide an effective user experience that enables the users to explore the information space and analyse the fragments of information they find relevant to their information need. One of the main limitations of the current document retrieval paradigm, is that it provides a ranked list of documents as a response to the searcher's query with no further support for locating and synthesizing relevant information. Therefore, the searcher is left to find and make sense of useful information in a massive information space that lacks any overview or conceptual organization. This is particularly challenging when the information need is complex and requires investigation and analysis (i.e., exploratory tasks).

In these exploratory scenarios, searchers are known to leverage two major strategies to mediate their exploration [413, 412, 462]: information foraging and sensemaking. While theories of information foraging [412] and berrypicking [44] describe the process of collecting relevant pieces of information as a step by step journey using multiple sources along the way, this is only the first step; users must also make sense of what they encountered [462, 352]. Sensemaking is described as the process of assimilating new knowledge into one's existing knowledge of a domain being explored [33]. To this end, structuring relevant content as inter-connected networks of concepts and semantic relationships describing their connections (e.g. Concept Maps [391]) can be an effective mechanism in assisting people with this assimilation process [387, 90].

Evidently, the primary source of information in the current IR paradigm is documents. These documents offer two types of relationships that can assist with comprehension and sensemaking: *Discourse Relations* and *Implicit Semantic Relations*. While reading the textual content of a document, the reader can see how sentences are connected. For example, "John is a great cook. He was a cook in the army." indicates an elaboration relation between these sentences. Elaboration, explanation and contrast are some instances of *Discourse Relations* that connect sentences in a coherent text. These textual clues provide a global view of the document content and its rhetorical structure [347]. These rhetorical relations provide a systematic way for an analyst to analyse the text, mostly in a linear fashion, to understand the content.

On the other hand, *Semantic Relations*, i.e. relationships that connect each term or concept with other terms and concepts in a document or a domain of interest, are not immediately apparent in the text. Therefore, while coherence and discourse relations help with comprehension, the lack of explicit semantic structure in textual resources forces the searchers to read through the content in a linear fashion. This makes the process of locating relevant information inefficient and exhausting.

In order to provide the searchers with more support in performing exploratory activities, we need to automate the process of locating and extracting relevant information pieces as well as the process of externalizing the semantic relations that connect terms and concepts that are discussed in retrieved documents.

A dominant technique towards automatic retrieval of semantic information between terms and concepts is Information Extraction (IE). IE aims at (semi)automatic collection of triples from textual corpora of a given domain (for example, the tuple $< Napoleon, invaded, Russia >$ is extracted from “Napoleon invaded Russia.”). These triples indicate the semantic relationship (i.e., “invaded”) between two entities (i.e., “Napoleon” and “Russia”). The outcome can be represented as a Knowledge Graph [246] (similar to a Concept Map [391] except the edges are not restricted to hierarchical relationships), that is a network of some domain knowledge represented by labelled nodes and labelled links between them. When these maps are available, they can provide a structured overview for understanding new documents, and the new documents can provide coherent context to the knowledge models [560].

In fact, for many years, these knowledge maps have been widely used in education, capturing knowledge [22] as well as browsing and facilitating information finding [90]. These knowledge maps represent concepts and relationships between them and promote the construction of a mental representation of the underlying content [22].

In this chapter we begin to develop the premise that different types of information representations provide different types of support for exploratory search. We argue that, while extracting facts and relevant information provide a structured view of the underlying content which can in turn facilitate exploration and browsing, they are not meant to entirely replace the textual representation of retrieved documents. Instead, coupling these extracted entity-relationship triples with their corresponding sentences and documents can be leveraged to extend the current document retrieval paradigm by enabling an interplay between alternative representations of search results.

In the rest of this chapter, we first provide some background on Information Extraction techniques and motivate them as a powerful methodology towards generating structured representations of the search space. Next, we describe our design requirements for developing an extension to the document retrieval paradigm. Finally, we detail the architecture of our proposed Information Discovery Framework and the implementation of related components of this framework.

3.2 Information Extraction

In Natural Language Processing (NLP), Information Extraction (IE) is the task of generating a structured, machine-readable representation of the information in text [260], usually in the form of triples or n-ary propositions. A proposition can be considered as a natural language representation of a potential fact (e.g., "Kenya is the leading exporter of coffee in the world.") in a format that is suitable for a variety of downstream systems to process (e.g. <Kenya, is the world's leading exporter of, coffee>). [1] More recently, IE researchers concentrate especially on Open Information Extraction (OIE), where information is automatically extracted from large textual resources, e.g. web news articles, which are not restricted to any particular domain or terminology and can scale to large, heterogeneous corpora such as the Web.

By applying an Open IE system to a collection of documents a set of triples in the form of (entity, relation, entity) will be generated. We note that, while there are many external knowledge bases (e.g., DBPedia[2]) available that provide a collection of these triples, they cannot necessarily represent the same information space that the user is interested in. That is, in order to provide better support for users' complex search activities, it is beneficial to derive these triples from the text of the documents retrieved for the user's query. The generated entity-relationship triples can be structured as semantic networks (also known as knowledge graphs) where entities are the nodes and relationships correspond to the edges in this network. These knowledge graphs can serve as an alternative representation of the user's domain of interest and they can support interaction with the corresponding documents retrieved by the search engine. To this end, we are interested in methods that extract entities and relations between them from text.

[1] The validity of a 'potential fact', extracted by an IE method, is only as good as the original sentence as the source of this extraction.

[2] http://dbpedia.org

3.3 Main Design Considerations

The first module in our proposed Exploratory Search Support (ESS) framework is the Information Retrieval and Extraction (IRE) module that leverages a Search and Document Retrieval (SDR) component as well as an Information Extraction component. In this section we review our main design considerations to develop this module such that we can build a search tool for effective *Information Discovery*.

Search and Document Retrieval.
Major search engines such as Google and Bing are powerful search systems that are effective in matching users' queries to a subset of documents that are ranked in the order of predicted relevance to users' information need. Our ESS framework leverages these engines through the available APIs in order to retrieve a ranked set of documents based on a given search query formulated by a user.

Information Extraction.
Our main contribution in designing an extension to the current SDR framework is through developing an Information Extraction component that supports searchers with their complex information seeking activities.

To this end, the IE component needs to satisfy different constraints:

1. [*Input*] The IE tool needs to be tailored to the search query that is submitted by the user of our information discovery framework;

2. [*Output*] The output of the IE tool is expected to support locating fragments of information by externalizing semantic relationships between different entities and concepts that are discussed in the textual content of retrieved documents.

To satisfy the first constraint, the IE component is designed to extend the SDR component by directly applying the IE algorithms to the output of SDR in order to generate triples from the text of documents that are retrieved for a user's query. Therefore, we are not interested in incorporating pre-existing knowledge bases such as DBPedia as they do not necessarily correspond to the space the searcher is interested in.

In order to satisfy the second constraint, we conducted a review of the state-of-the-art OpenIE systems that were available at the time of developing our ESS framework in order to evaluate their efficacy in supporting an information discovery platform.

Among the available tools (see Section 2.3.3 for a survey), Ollie [163] appeared to be the most general purpose information extraction tool that overcame the limitations of its predecessors (e.g. learning only verb-based relational phrases in ReVerb). However, Ollie came with its own limitations. Given that this tool was not specifically designed to support searchers with information seeking tasks, we found Ollie to be more focused on relation extraction while the arguments that are connected by these relations could be any noun phrase extracted from sentences. To clarify, as can be seen in examples below, the triples extracted by Ollie can be used to express a *relationship* between two *arguments* where these arguments could be a noun phrase with very loose boundaries (which can include multiple or no entities at all).

`Sentence: In addition, their kidneys have small glomeruli or lack glomeruli entirely.`
`Extraction: (their kidneys; have; small glomeruli or lack glomeruli)`

`Sentence: On July 16 2008, Hezbollah transferred the coffins of captured Israeli soldiers , Ehud Goldwasser and Eldad Regev, in exchange for Samir Kuntar and four other Hezbollah members captured by Israel during the 2006 Lebanon War.`
`Extraction: (the coffins of captured Israeli soldiers, Ehud Goldwasser and Eldad Regev; be transferred in; exchange for Samir Kuntar and four other Hezbollah members)`

3.4 Our Information Discovery Tool

Through a 9-month research collaboration with an industry partner (InsightNG company[3]) we developed an Information Discovery tool, as an extension to the current search and document retrieval framework offered by major search engines. This tool was designed to help users with finding relevant information regarding the goals they intend to achieve. We first briefly describe the task of entity-relationship extraction from a set of reference sentences and clarify the main terminology used for rest of this chapter. Next, we describe the architecture and the main components of this framework.

[3] https://www.insightng.com/

3.4.1 Terminology and Task Description

Before we describe the details of how this information discovery tool is developed and how the underlying components interact with one another, we need to clearly specify the task we are trying to achieve and define the terminology we are using to describe our framework.

Task Description. Given a search query, extract a set of entity-relationship triples that convey the same information as their corresponding set of retrieved documents for that query. To simplify this task, we elaborate on the extraction of related entities and the relation label describing their relationship given a set of reference sentences. These reference sentences are a subset of all sentences in the documents retrieved for a user's query.

Entity. There are many definitions for an *entity*. For example, "a thing with distinct and independent existence" or "any singular, identifiable and separate object. An entity refers to individuals, organizations, systems, etc." In this dissertation, we simply consider any noun phrase that could be selected as the title of a Wikipedia article a valid entity. For example, president, coffee, Canada, freedom, and assembly of experts are all valid entities.

Reference Sentence. While each document that is retrieved in response to a user's query is a collection of sentences, in our extraction task we are particularly interested in candidate sentences that contain at least two entities and hence can potentially describe a semantic relationship between these two entities.

We define such a Reference Sentence S as a sentence that contains two entities E1 and E2 (and possibly some other entities). For example, given the entities "Kingston" and "United Canadas", the following sentence provides some context on how these entities might be related:

> *"The beautiful **Kingston** was chosen as the first capital of the **United Canadas** and served in that role from 1841 to 1844."*

Relation Label. A Relation Label is a simplified sentence that is expected to describe the relationship between E1 and E2 based on what can be inferred from the Reference Sentence S. For example, *"Kingston was chosen as the first capital of the*

United Canadas" can be extracted as a Relation Label describing the relationship between Kingston and United Canadas based on the reference sentence S above.

Related Entities. We consider an entity pair, E1 and E2, related if they appear in the same reference S and there is a *direct relationship* between them.

Direct Relationship. Two entities E1 and E2 are considered to be directly related if the following requirements are met:

- It is possible to write an independent clause such that:
    1. the clause contains E1 and E2;
    2. E1 and E2 appear in the same order as in S;
    3. no external knowledge is used to write this independent clause;

Independent Clause. Like a phrase, a clause is a group of related words; but unlike a phrase, a clause has a subject and verb. An independent clause, along with having a subject and verb, expresses a complete thought and can stand alone as a coherent sentence. In contrast, a subordinate or dependent clause does not express a complete thought and therefore is not a sentence. A subordinate clause standing alone is a common error known as a sentence fragment. For example, grammatically complete statements such as "*He saw her*", "*The Washingtons hurried home*", or "*Free speech has a price*" are sentences and can stand alone. When such statements are part of longer sentences, they are referred to as independent (or main) clauses.

Given the task described above we can now detail our algorithm for extracting related entities and their relation labels and how it is incorporated in the overall ESS framework.

3.4.2 Architecture and Design

Our information discovery system is implemented in four phases. During the first phase we create the input corpus by collecting retrieved documents based on a given query. Next, we extract entities from text using state-of-the-art entity taggers. We then select the sentences that contain at least two entities in them and parse them using Stanford Dependency Parser. For each sentence, we extract meaningful relations between the entities by finding the shortest path in the corresponding parse tree. We constructed a set of patterns based on dependency triples that lead to semantically meaningful relations. In the final phase we generate labels for the extracted relations and rank them based on relevance to the query and the informativeness of the extraction.

In a nutshell, this system is designed to address the following tasks:

1. Extracting “important” entities from a set of top ranked documents retrieved for a query (i.e. **top ranked entities**);
2. Identifying meaningful relations between each entity pair (i.e. **direct relationships**);
3. Generating concise and readable labels for each relation which briefly explain the underlying connection between the corresponding entities (i.e. **relation label**);
4. Ranking extracted entities and relations based on different measures including TF-IDF and NPMI.

Figure 6.1 illustrates the architecture of this system. The following subsections describe these phases and their building blocks in more detail.

3.4.2.1 Phase I: Creating the Source Corpus

The first phase of this system is designed such that it can process different source documents. The input corpus can be provided by the user as a collection of documents or the system can collect relevant documents from the web using the query provided as an input.

The input query can be formulated by the user or, alternatively, the system can automatically generate a query based on user’s previous interactions with a set of related concepts. To elaborate, in the first scenario, the user simply formulates the query as a list of keywords or a natural language sentence while in the second case the query can be generated by the system using the entities that the user has already added to a list of concepts they are interested in exploring.

In both scenarios, Bing is used to retrieve the top 50 documents related to the user’s query. The source documents can be either in HTML format or plain text. The HTML format can help with creating a higher quality corpus as our system can pre-process the text more effectively.

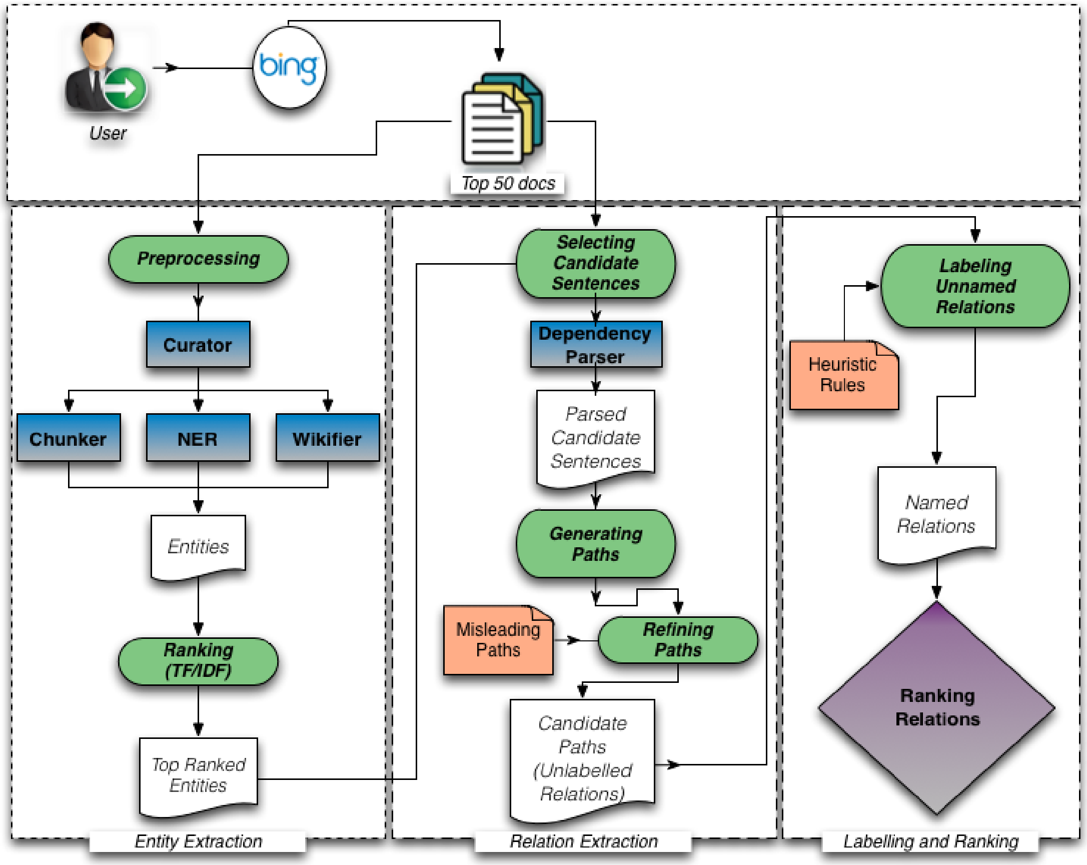


Figure 3.1: The System Architecture

Retrieving Documents from the Web. Microsoft provides a method for retrieving documents using the Bing search engine via the Bing API. The documentation and account registration can be found using the link below [4]. Once an account has been created one can login to determine the query URL and API Key. Bing uses OAuth to validate the API key; therefore, in order to use the API, the key must be encoded in the header of the http request (see Bing documentation for code snippets). Since the Bing API does not return the document itself the contents must be retrieved using the given URLs. In order to identify each URL the API does provide a unique identifier for each result.

[4] https://datamarket.azure.com/dataset/5BA839F1-12CE-4CCE-BF57-A49D98D29A44

Preprocessing. Most Named Entity Taggers and Parsers are designed to work on non-web documents, so it is important to remove html tags as well as special and non ASCII characters. These can be removed by a set of simple regular expressions (Table 3.1).

Regular Expression Pattern	HTML Tag Match
$< script.*? > .*? < /script >$	Client side scripting
$< style.*? > .*? < /style >$ $< link.*?/? >$	HTML Style sheets
$<?/div.*? >$	Div Tags
$< head > .*? < /head >$	HTML Header
$<! - - - .*? - - >$	HTML Comments
$\#\&[0-9]+$ $\&nbsp$	Special Characters
¡/?a.*?¿	Hyperlinks

Table 3.1: Sample expression patterns used for cleaning HTML documents

Bing API Script. We have developed a script to access the Bing API from the command line. The input arguments are the output folder and the query terms. Similar to a search engine, quoted terms are treated as phrases to be searched. The script produces both the raw and cleaned documents in the output folder with each document identified by its Bing ID. A mapping file from document URL to Bing ID is also included so that the original document address can be linked to the document itself. The script also contains a number of tunable parameters which can only be accessed from within the code. Table 3.2 summarizes each of these parameters.

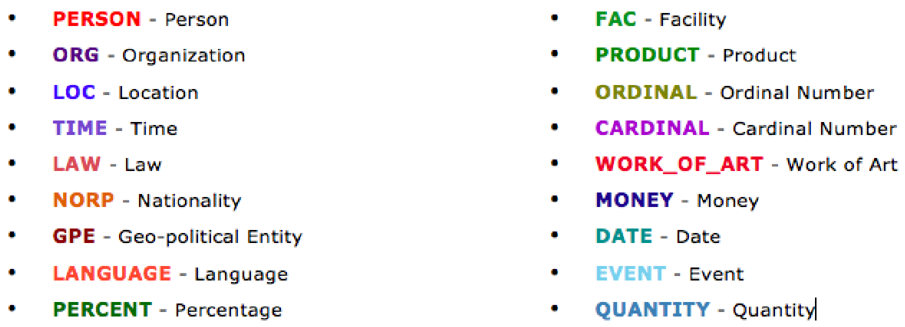
3.4.2.2 Phase 2: Entity Extraction

In this phase entities are identified and tagged in input texts. Informally, any noun phrase that could be considered as the title of a Wikipedia article is a valid entity. We extract a wide range of entities which include both Named Entities (people, locations, organizations, ...) and more general entities (e.g., coffee, investment, project, ...). In order to identify and tag entities we use a Chunker, a Named Entity Tagger (NER) and Wikifier which are all developed by University of Illinois at Urbana-Champaign and they are available for download at http://cogcomp.cs.illinois.edu/page/software. These tools are briefly described in the following subsections.

Parameter	Description
DOCUMENT_NUMBER	The number of documents to retrieve from the API
AUTH_KEY	Bing API Key
DELAY_RANGE	Delay between subsequent calls to the Bing API.
THRESHOLD	For phrase queries the minimum number of documents needed before trying each term as a separate query term e.g. If the phrase "pizza hut" was searched and below this number of documents were returned "pizza" and "hut" would be searched.
BATCH_SIZE	The number of documents to be fetched by a single thread
THREAD_POOL_SIZE	Size of the thread pool used for fetching and cleaning documents from the web

Table 3.2: Tunable Parameters for the Bing API

Named Entity Tagger. The utilized Named Entity Tagger is a state of the art NER tagger that tags plain text with named entities. The newest version tags entities with either the "classic" 4-label type set (people / organizations / locations / miscellaneous), while the most recent (Extended NER) can also tag entities with a larger 18-label type set (based on the OntoNotes corpus). It uses gazetteers extracted from Wikipedia, word class models derived from unlabeled text, and expressive non-local features. The best performance is 90.8 F1 on the CoNLL03 shared task data. The tagger is robust and has been evaluated on a variety of datasets [430]. In our implementation we used the Extended NER while we only considered a subset of tags including PERSON, LOC, ORG, NORP, LANGUAGE, PRODUCT and WORK_OF_ART. Here is a full list of tags provided by the Extended NER:



- PERSON - Person
- ORG - Organization
- LOC - Location
- TIME - Time
- LAW - Law
- NORP - Nationality
- GPE - Geo-political Entity
- LANGUAGE - Language
- PERCENT - Percentage
- FAC - Facility
- PRODUCT - Product
- ORDINAL - Ordinal Number
- CARDINAL - Cardinal Number
- WORK_OF_ART - Work of Art
- MONEY - Money
- DATE - Date
- EVENT - Event
- QUANTITY - Quantity

Wikifier. The Wikifier identifies important entities and concepts in text, disambiguates them and links them to Wikipedia. Wikification is an important step in helping to facilitate Information Access, in knowledge acquisition from text and in helping to inject background knowledge into NLP applications. The main decisions the Wikifier must make are: (1) What expressions to link to Wikipedia. (2) Disambiguating the ambiguous expressions and entities [431]. Wikifier is implemented in Java and the code can be modified in order to consider more general entities such as “school” or “park” which are not referring to a particular instance according to the surrounding context. As an example consider the following sentence: “Kenya is a capitalist country with an economic policy that emphasizes the role of the free market.”; Applying Wikifier we generate the following output:

```
Term from text: ‘economic policy’
Label: http://en.wikipedia.org/wiki/Economics
Properties:
RankerScore, 1.5369297060473759;
IsLinked, true;
SurfaceFormWikiCatAttribs,  policy;
TitleWikiCatAttribs,  science;
LinkerScore, 0.35744175439733744;
----------------------
Term from text: ‘free market’
Label: http://en.wikipedia.org/wiki/Free_market
Properties:
RankerScore, 2.2444808231094133;
IsLinked, true;
SurfaceFormWikiCatAttribs,  free market;
TitleWikiCatAttribs,  market;
LinkerScore, 1.7701680178468986;
----------------------
Term from text: ‘Kenya’
Label: http://en.wikipedia.org/wiki/Kenya
Properties:
RankerScore, 0.7606882804807147;
IsLinked, true;
SurfaceFormWikiCatAttribs, ;
TitleWikiCatAttribs,  country state nation member  territory;
LinkerScore, 0.4709923510089708;
```

```
----------------------
Term from text: ‘capitalist’
Label: http://en.wikipedia.org/wiki/Capitalism
Properties:
RankerScore, 2.3268430680273595;
IsLinked, true;
SurfaceFormWikiCatAttribs,  capitalist;
TitleWikiCatAttribs,  economy system ideology;
LinkerScore, 1.3837449732297202;
----------------------
Term from text: ‘country’
Label: UNMAPPED
Properties:
RankerScore, -999.0;
IsLinked, false;
SurfaceFormWikiCatAttribs,  country;
TitleWikiCatAttribs, ;
LinkerScore, -999.0;
----------------------
Term from text: ‘policy’
Label: UNMAPPED
Properties:
RankerScore, -999.0;
IsLinked, false;
SurfaceFormWikiCatAttribs,  policy;
TitleWikiCatAttribs, ;
LinkerScore, -999.0;
----------------------
Term from text: ‘role’
Label: UNMAPPED
Properties:
RankerScore, -999.0;
IsLinked, false;
SurfaceFormWikiCatAttribs,  role;
TitleWikiCatAttribs, ;
LinkerScore, -999.0;
----------------------
Term from text: ‘free’
```

```
Label: UNMAPPED
Properties:
RankerScore, -999.0;
IsLinked, false;
SurfaceFormWikiCatAttribs,  free;
TitleWikiCatAttribs, ;
LinkerScore, -999.0;
----------------------
Term from text: ‘market’
Label: UNMAPPED
Properties:
RankerScore, -999.0;
IsLinked, false;
SurfaceFormWikiCatAttribs,  market;
TitleWikiCatAttribs, ;
LinkerScore, -999.0;
```

As you see Wikifier did not map terms such as “country” or “policy”, even though there are corresponding webpages in Wikipedia. However, one might modify the code to include such general terms in the list of entities.

Curator. The University of Illinois has also developed a tool called Curator. The Curator is a system that acts as a central server in providing annotations for text. It is responsible for requesting annotations from multiple natural language processing servers, caching and storing previous annotations and refreshing stale annotations. The Curator provides a centralized resource which requests annotations for natural language text [116]. The Curator architecture defines multiple data types and service interfaces for creating new annotation servers and communicating with the Curator. The interfaces are defined using Apache Thrift which provides a software stack and code generation for cross-language deployment. This allows annotation servers and Curator clients to be implemented in multiple languages. Currently Thrift supports C++, Java, Python, PHP, Ruby, Erlang, Perl, Haskell, C#, Cocoa, Smalltalk, and OCaml. The Curator package comes bundled with annotators capable of performing the following annotations:

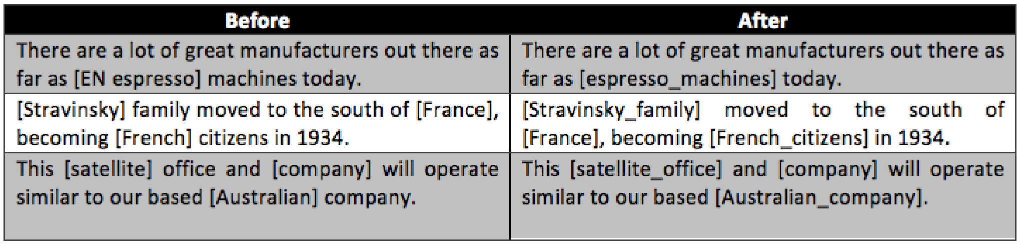
- Tokenization and Sentence Splitting (via Illinois NLP tools)
- Part-of-speech tags (via Illinois POS Tagger)

- Chunk (shallow parse) analysis (via Illinois Chunker)
- Named Entities (via Illinois Named Entity Recognizer)
- Coreference (via Illinois Coreference package)
- Parse trees (via Stanford Parser and Charniak Parser)
- Dependency trees (via Stanford Parser)
- Semantic Role Labels for verbs and nouns (via Illinois SRL)
- Reference Entities (via Illinois Wikifier)

Some of these components require significant amount of memory; one motivation for creating the Curator was the need to distribute such components across multiple machines. However, a server with 32G of RAM should be able to run all components together. We use Curator in order to integrate the output of the NP Chunker, NER and Wikifier for every input text. One can also configure Curator such that it integrates annotators (e.g., Stanford Parser) which are not developed by The University of Illinois. The code is open-source and is implemented in Java, making it easy to modify and add or remove annotators.

Improving Extracted Entities Not all entity mentions that are tagged by the NER and Wikifier are accurate. We identified two main boundary detection errors in the output of these entity taggers. In the first case the head of a noun phrase is tagged as an entity that could be expanded to include a more specific entity mention. For example in the following sentence "There are five main components in any information system.", "system" is not the most specific entity mention and an accurate tagger needs to include "information" in the extracted entity mention. In the second case, a modifier (adjectives, adverbs and nouns modifying another noun) is tagged as entity while the head noun that is modified by this modifier is left out. Correcting these errors is even more crucial than the first case since entities by definition need to be a noun phrase and other parts of speech such as adjectives, adverbs or verbs are not acceptable.

Such incomplete extractions should either be extended to a meaningful entity or be removed from the list of entities. We developed a subroutine which improves the quality of these extractions. This subroutine fixes entities in two steps. First, we parse sentences and obtain POS tags, chunks and dependencies. Next, we identify the tagged entities which are a modifier (participating in amod or nn dependency relation); and merge these modifiers with the rest of the noun phrase containing them until the head of the phrase is added to this entity. The extended phrase replaces the incomplete entity. Figure 3.2 provides some sample sentences before and after applying this subroutine. In the case that a noun is modifying a noun (nn dependency relation), even when the governor noun is tagged by the entity we extend it by adding in the modifying nouns. The reason for this is that governor nouns participating in an nn dependency can be too general and do not necessarily refer to a specific entity. For example "system" is a very general noun and shouldn't be considered as an entity. While "Information system" is a specific type of system and therefore should be an entity. The merging process is done in two stages. During stage 1, we are merging "nn", "amod" and "poss" relations while in stage 2, prepositions and conjuncts (prep_ and conj_ relations) are merged to form a complete noun phrase. As for stage 2, some statistical measures (including mutual information and ttest) are used to select better candidates for merging.



Before	After
There are a lot of great manufacturers out there as far as [EN espresso] machines today.	There are a lot of great manufacturers out there as far as [espresso_machines] today.
[Stravinsky] family moved to the south of [France], becoming [French] citizens in 1934.	[Stravinsky_family] moved to the south of [France], becoming [French_citizens] in 1934.
This [satellite] office and [company] will operate similar to our based [Australian] company.	This [satellite_office] and [company] will operate similar to our based [Australian_company].

Figure 3.2: Examples for Improving Entity Extraction

Please note that entities are surrounded by "[ ]" and parts of multiple word entities are joined by "_".

Ranking Entities. As described above, Curator integrates the entities extracted by the NER with that of the Wikifier. Not all these entities are meaningful and informative. TF-IDF is used to rank the entities. This metric is a statistical measure used to evaluate how important a term is to a document in a collection or corpus. We calculate the term frequency (TF) as the number of occurrences of each tagged entity in the input corpus. We also calculate the document frequency (DF) for each term and query ClueWeb to find inverse document frequency (IDF) for those terms. Then we rank the extracted entities based on the TF-IDF measure calculated as:

$$tfidf(t, d, D) = tf(t, d).idf(t, D) \tag{3.1}$$

$$idf(t, D) = log[\frac{|D|}{|d\epsilon D : t\epsilon d|}] \tag{3.2}$$

Where, t is the term, d is a document and D is all the documents in the reference corpus, in our case ClueWeb, (which has 13712600 documents). Please note that the term frequency (tf) can be calculated using the documents from the corpus we created based on a query, while idf is calculated from ClueWeb. This ranking measure tends to assign higher scores to entities which are frequent in one document but not in all documents. In order to obtain these frequencies we used the Wumpus search engine which is an Information Retrieval system developed at the University of Waterloo. Wumpus is freely available under the terms of the GNU General Public License (GPL). For more information about Wumpus you can refer to http://www.wumpus-search.org/.

3.4.2.3 Phase III: Relation Extraction

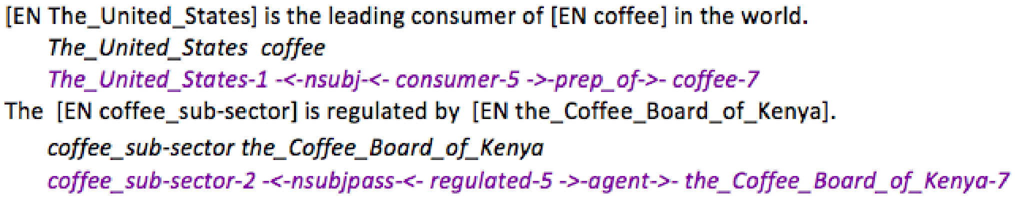
We start this phase by selecting the sentences which contain at least 2 top ranked entities extracted from Phase I. Once we have a collection of sentence candidates, we aim at finding meaningful relations expressed in those sentences. To this end, we generate all possible entity pairs for each sentence. Next, we parse the sentence containing this entity pair and find the shortest path from the first entity to the second entity in the generated parse tree. This path indicates the underlying relation between these two entities. We apply a set of heuristics to remove paths which do not correspond to a meaningful relation. The result will be a list of paths between entity pairs. Each entity pair, along with its underlying path, indicates an unlabeled relation which will be an edge in our knowledge graph. Finally, we apply two different ranking methods to rank the extracted relations. We also generate readable labels for a subset of extracted relations which can reveal the existing relation between an entity pair. Following subsections provide more details for implemented methods.

Parsing Sentences. We applied Stanford Dependency Parser [136] to parse the candidate sentences. A natural language parser is a program that works out the grammatical structure of sentences, for instance, which groups of words go together (as “phrases”) and which words are the subject or object of a verb. Probabilistic parsers use the knowledge of language gained from manually labelled sentences to try to produce the most likely analysis of new sentences. The package we used is a Java implementation of probabilistic natural language parsers, both highly optimized Probabilistic Context Free Grammar (PCFG) and lexicalized dependency parsers, and a lexicalized PCFG parser. The latest version can be downloaded from http://nlp.stanford.edu/software/lex-parser.shtml . The parser provides Stanford Dependencies output as well as phrase structure trees. Typed dependencies are otherwise known grammatical relations. The Stanford dependencies provide a representation of grammatical relations between words in a sentence. They have been designed to be easily understood and effectively used by people who want to extract textual relations. Stanford dependencies (SD) are triplets: name of the relation, governor and dependent. As an example the standard dependencies for the sentence "*Many local entrepreneurs made tons of money bringing the Starbucks coffeehouse concept to their hometowns and then expanding from there.*" are given below:

```
amod(entrepreneurs-3, Many-1)
amod(entrepreneurs-3, local-2)
nsubj(made-4, entrepreneurs-3)
root(ROOT-0, made-4)
dobj(made-4, tons-5)
prep_of(tons-5, money-7)
dep(made-4, bringing-8)
det(concept-12, the-9)
nn(concept-12, Starbucks-10)
nn(concept-12, coffeehouse-11)
dobj(bringing-8, concept-12)
poss(hometowns-15, their-14)
prep_to(bringing-8, hometowns-15)
advmod(expanding-18, then-17)
dep(made-4, expanding-18)
conj_and(bringing-8, expanding-18)
prep(expanding-18, from-19)
pobj(from-19, there-20)
```

Generating Paths

In order to find the connection between two candidate entities we implemented a getPath subroutine which finds the shortest path from the first entity to the second entity using the corresponding parse tree generated by the dependency parser. This algorithm first generates all possible paths between two entities and then selects the ones that can potentially lead to meaningful relations. We use a set of heuristic rules to remove meaningless path. These rules can be customized to select the valid paths based on different criteria. If there are no particular preferences for the dependencies that can appear on the path, the method will generate all paths and then select the shortest one. Other possible scenarios could be selecting the paths that contain a verb, selecting the paths of the length longer than 2 or selecting the paths for which not all intermediate nodes are stop words. We developed a different set of rules which can select the potentially more meaningful relations and we use it as the first step for generating labels for the relations. These rules are presented in Section 3.4.2.4. The following examples illustrate the paths generated for two sample sentences:



[EN The_United_States] is the leading consumer of [EN coffee] in the world.

The_United_States coffee

The_United_States-1 -<-nsubj-<- consumer-5 ->-prep_of->- coffee-7

The [EN coffee_sub-sector] is regulated by [EN the_Coffee_Board_of_Kenya].

coffee_sub-sector the_Coffee_Board_of_Kenya

coffee_sub-sector-2 -<-nsubjpass-<- regulated-5 ->-agent->- the_Coffee_Board_of_Kenya-7

3.4.2.4 Phase IV: Labeling and Ranking Relations

As described in the overall algorithm for identifying relations between entities, we select sentences that contain at least two highly ranked entities. We then find the shortest path between entities for all possible entities in a candidate sentence. The fact that two entities occur in the same sentence does not always mean there should be a meaningful relation between them. In fact, a sentence with multiple entities can lead to many meaningless relations that do not communicate any information to the user. Therefore, we need to rank the extracted relations to make sure the informative relations will be shown to the user before the ambiguous or meaningless ones. Moreover, representing relations as unlabeled connection are not very useful for the user. While these links indicate a connection between two entities, it is not clear how these entities are related. Therefore, we aim at generating a readable label for the relations we extract. However, not all relations can be labeled effectively due to the complexity of some sentences and the errors made by the utilized annotators. The next two subsections provide more details about the methods we developed for ranking and labeling relations.

Ranking Relations

There are different criteria to rank extracted entity-relationship tuples. These criteria can be classified into two groups: Subjective and Objective. We can also classify approaches in these groups into Graph based and Semantic based Measures. We reviewed these approaches in Section 2.3.3.1. An overall classification of these ranking approaches is provided here as a reference.

1. Objective
   (a) Graph-based
      - Betweenness (e.g., [650])
      - Centrality (e.g., [637, 68])
      - PageRank (e.g., [650])
   (b) Semantic-based
      - Co-occurrence and Frequency (e.g., [417])
      - Using Ontologies (e.g., [531])
      - Semantic Similarity
2. Subjective
   (a) Informativeness (e.g., [267])
   (b) Interestingness (e.g., [334, 426])
   (c) Unexpectedness (e.g., [336, 184])
   (d) Actionability (e.g., [212])

In a broader sense, Informativeness can be defined in an objective fashion. While evaluating an extraction as an informative assertion can be different for different users, we are able to apply objective semantic-based measures to rank the relations based on the informativeness of the corresponding entity pair. Such approaches usually take advantage of co-occurrences and statistical frequencies to rank entities / entity pairs. In this system we implemented two ranking methods: Average TF-IDF and Normalized Pointwise Mutual Information (NPMI) for the entity pairs. These methods are described in the following subsections.

Selecting Meaningful Relations. Prior to ranking the relations, we select a subset of extractions which can potentially lead to more meaningful relations. We objectively define a relationship as meaningful, if it is a direct relationship between two entities that appear in the same reference sentence.

We developed a set of heuristics which are mostly observed in paths that lead to more comprehensive and meaningful relations and labels. First, we remove the paths containing "dep", "rcmod" and "ccomp". While "dep" indicates the parser did not identify all dependencies in the sentence, "rcmod" and "ccomp" usually correspond to indirect relations. For example, the sentence "John believes that Kenya exports coffee to US." Will obtain the following dependencies:

```
nsubj(believes-2, John-1)
root(ROOT-0, believes-2)
mark(exports-5, that-3)
nsubj(exports-5, Kenya-4)
ccomp(believes-2, exports-5)
dobj(exports-5, coffee-6)
prep_to(exports-5, US-8)
```

Therefore, “ccomp” will appear on the path between “John” and “coffee”, while there is no direct relation between these two entities and the extraction has to be removed. Next, we select paths that contain a verb and the entities are connected to this verb via a particular set of dependencies including nsubj, xsubj, nsubjpass, dobj, iobj, agent, partmod and etc.

Ranking by Average TF-IDF. While ranking individual entities by TF-IDF (as described in Phase II) will result in selecting sentences that contain high rank entities only, it does not guarantee the extracted relation is meaningful and informative. We define informativeness of a relation as a sum of TF-IDF scores of all entities connected by that relation (i.e., the entity pair) divided by the number of those entities (i.e., 2). Therefore, for every selected entity pair that has a corresponding path, we calculate the average TF-IDF as follows:

$$Informativeness_R = \frac{(TFIDF(entity_1) + TFIDF(entity_2))}{2} \tag{3.3}$$

Ranking by NPMI. Mutual information (MI) is a measure of the information overlap between two random variables. Pointwise mutual information (PMI) is a measure of how much the actual probability of a particular co-occurrence of events p(x; y) differs from what we would expect it to be on the basis of the probabilities of the individual events and the assumption of independence p(x)p(y). Mutual information can be used to perform collocation extraction by considering the MI of the indicator variables of the two parts of the potential collocation [113]. To give MI and PMI a fixed upper bound, Gerlof Bouma [69] normalized the measures to have a maximum value of 1 in the case of perfect (positive) association. This measure is called Normalized Pointwise Mutual Information (NPMI). NPMI can be calculated for two terms and its value is within the range of -1 and 1. “Some orientation values of NPMI are as follows: When two words only occur together, NPMI = 1; when they are distributed as expected under independence, NPMI = 0; finally, when two words occur separately but not together, NPMI is defined to be -1". We applied this measure to our relation ranking problem. We calculate NPMI for every entity pair (e1,e2) using the following formula:

$$NPMI_(e_1, e_2) = \frac{log(\frac{P(e_1,e_2)}{P(e_1)\times P(e_2)})}{-log(P(e_1, e_2))} \tag{3.4}$$

Where:

$$P(e_1, e_2) = \frac{TF(e_1 \; and \; e_2 \; occuring \; together)}{window \; size \; \times \; total \; number \; of \; docs} \tag{3.5}$$

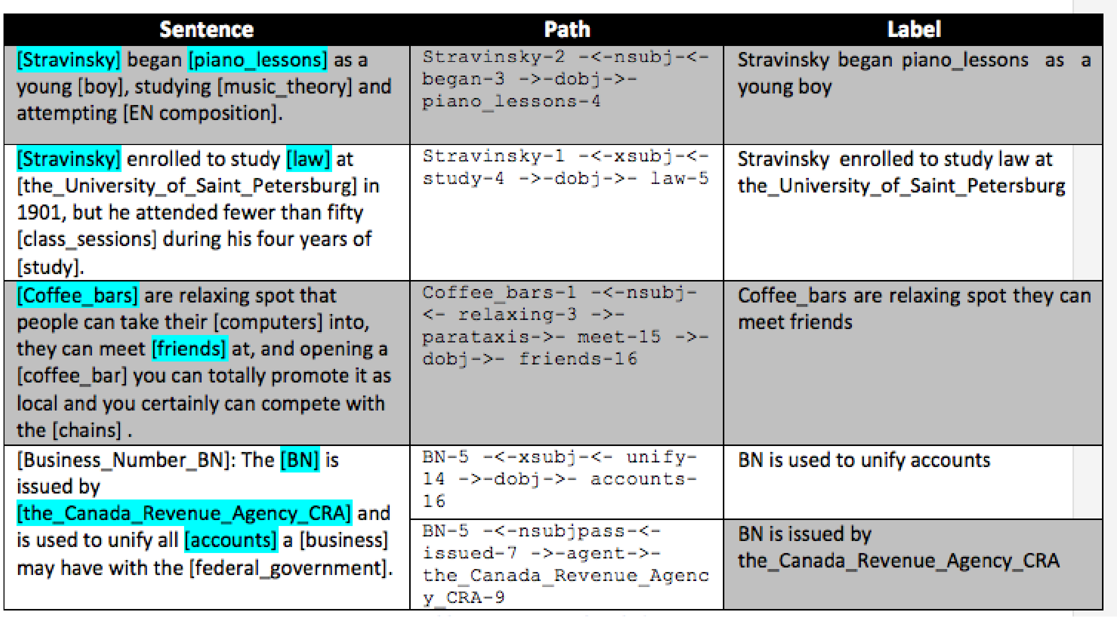
$$P(e_i)_{i\epsilon\{1,2\}} = \frac{TF(e_i)}{total \; number \; of \; docs} \tag{3.6}$$

The window size indicates the maximum distance between the related entities in the corresponding sentence. In our experiments we set the window size to 10. Reviewing final results suggests the superiority of TF-IDF for ranking relations which are related to the topic / goal (e.g., "investment for coffee in Kenya"). On the other hand, NPMI tends to assign higher scores to the entity pairs which have a strong association with each other (e.g., "wine" and "beer") regardless of their relevance to the topic (e.g., "investment in coffee"). A hybrid approach can be designed to combine these two measures into one. We suggest a linear interpolation of these two ranking measures as:

$$Ranking\_Score(r) = \alpha \times TFIDF\_score(r) \; + \; (1 - \alpha) \times NPMI\_score(r) \tag{3.7}$$

The weight α can be learned by conducting different experiments and evaluating the final results. Also, TFIDF scores have to be normalized such that the values range between 0 and 1 which is compatible with the values calculated by the NPMI measure.

Labeling Relations. Labeled relations can reveal the underlying connection between an entity pair for the users. The default label for every relation is the sentence expressing that relation. Since sentences are sometimes very long and contain many terms that do not contribute to the core connection between the entities, we aim at generating a label which is a simplified version of the sentence. This label is expected to be shorter than the original sentence while it is still readable and meaningful for the user.



Sentence	Path	Label
[Stravinsky] began [piano_lessons] as a young [boy], studying [music_theory] and attempting [EN composition].	Stravinsky-2 -<-nsubj-<- began-3 ->-dobj->- piano_lessons-4	Stravinsky began piano_lessons as a young boy
[Stravinsky] enrolled to study [law] at [the_University_of_Saint_Petersburg] in 1901, but he attended fewer than fifty [class_sessions] during his four years of [study].	Stravinsky-1 -<-xsubj-<- study-4 ->-dobj->- law-5	Stravinsky enrolled to study law at the_University_of_Saint_Petersburg
[Coffee_bars] are relaxing spot that people can take their [computers] into, they can meet [friends] at, and opening a [coffee_bar] you can totally promote it as local and you certainly can compete with the [chains] .	Coffee_bars-1 -<-nsubj-<- relaxing-3 ->-parataxis->- meet-15 ->-dobj->- friends-16	Coffee_bars are relaxing spot they can meet friends
[Business_Number_BN]: The [BN] is issued by [the_Canada_Revenue_Agency_CRA] and is used to unify all [accounts] a [business] may have with the [federal_government].	BN-5 -<-xsubj-<- unify-14 ->-dobj->- accounts-16	BN is used to unify accounts
	BN-5 -<-nsubjpass-<- issued-7 ->-agent->- the_Canada_Revenue_Agency_CRA-9	BN is issued by the_Canada_Revenue_Agency_CRA

Figure 3.3: Some sample labels.

While sentences can be an upper bound for generated labels, shortest paths between entities can be considered as a lower bound. Although a label must include all the terms which are located on the path, constructing a label by combining these terms only will not always lead to a complete and meaningful label. Our label generation algorithm takes the entity pair, shortest path between them and the corresponding sentence as an input and progresses as follows: First it parses the sentence using Stanford Parser and obtains dependency triples relating all words and phrases in this sentence. Next it processes the words appearing on the path:

- For every noun
    - the algorithm first adds the noun to the label;

  - next, all the prepositions that have this noun as a dependent (second argument in the dependency triple) are also added to the label. For example in “Kenya exports coffee to US.”, “US” is a noun which has a prep dependency as prep_to (coffee, US). Therefore, if “US” is on the path, we will add the preposition “to” to the label.
  - similarly, we add all the conjuncts related to this noun.
  - if there is a copula verb (cop) associated with this noun, we add it to the label a well.
  - finally, we extend the nouns to include modifiers (amod / nn) or prepositions (e.g., prep_of).

- For every verb
  - the algorithm first adds the verb to the label;
  - if the verb is negated, the proper auxiliary verb is also added to the label to negate the verb;
  - for active verbs, direct object and indirect object (if applicable) will be added to the label;
  - for passive verbs, the agent will be added to the label;
  - for the verbs with open clausal complement, both xcomp and xsubj are added to the label;
  - if the verb has a subject (nsubj) which is missing from the path, has to be added to the label;
  - all phrasal particles (prt) and auxiliary verbs (aux) related to the verb have to be added;
  - we also expand on rcmod, infmod, partmod and prepc_of, since these are relative clauses and convey important information.

  - For every adjective
  - If it’s not already added to the label, we add it;
  - If there is a copula verb associated with it, it is also added.

Please note that all these words are added to the label such that the order of them in the original sentence is preserved. Therefore, the generated label is readable and can be considered as a simplified version of the sentence. Figure 3.3 indicates some sample labels along with the original sentence and the shortest path used to generate that label.

Please note that the entity pair corresponding to the relation is highlighted in blue. As you see the labels are mostly shorter than the sentences while they do not necessarily start from the first entity or end with the second entity. Going beyond the entity pair will lead to more meaningful labels. Yet, they are shorter and simplified, thereby easier to read and understand the underlying meaning of the relation.

3.5 Evaluation Considerations

As with many NLP tools, the main evaluation question to address is the extent to which they produce the results for which they were designed. It must be noted that the design or application of an NLP system is commonly connected with a broader task; For example applying a parser to a set of documents to identify answer candidates for a user's query. Given that the underlying methods involve a component that analyzes sentences to produce $< subject - relation - object >$ dependency tuples, they have the potential to allow more concept level matches. For example the question "*Who patented the light bulb?*" can match "*Thomas Edison's patent of the electric light bulb*" via the tuple $< ThomasEdison, patented, bulbs >$.

Once a system is in place to generate these triples based on a set of input sentences, a major question is how might the quality of dependency tuple analysis be evaluated? There are many ways that this evaluation can be done. The simplest and the most intuitive approach could be a live demonstration of a '*semantic*' search engine that uses the parsing algorithm in order to match the users' questions with the facts that appear in documents. While the target audience can directly observe the value of such a system, it is difficult to assess the full range of the system capabilities and can potentially hide known flaws.

A more systematic approach would be to perform a standard *intrinsic* evaluation of the dependency extraction component. To this end, a test set can be created which contains a sample of test sentences, along with ground truth labels, i.e. the tuples that the system is expected to extract from those sentences. Standard metrics such as precision, recall and F-measure can be employed to measure the performance of the extraction algorithm; either against a different version of the same algorithm (i.e. a *formative evaluation*) or against competing techniques (i.e. a *summative evaluation*).

While an intrinsic evaluation helps to analyze the quality of the tuples automatically, there are at least three challenges regarding this type of evaluation. First, for many NLP tasks, including Open IE, there are no pre-existing test sets that can be leveraged as ground truth labels. In fact, most Open IE systems are designed with different goals in mind and the tuples they extract are not directly comparable. Second, creating a test dataset that includes ground truth labels for a set of sentences is very challenging. There are no standard guidelines for what constitutes a valid tuple. Essentially, these ground truth labels need to be manually created for any new domain or task and they cannot be transferred to new use cases. Finally, even if a test dataset is available or created, while it can enable an automatic assessment of the quality of the output of the extractor algorithm according to the ground truth labels, and quantify the impact of incremental changes to the algorithm, how do we know that improvements in the accuracy and coverage of the extraction algorithm output actually makes a difference in the overall quality of the system? That is, whether a higher extraction accuracy will lead to a better support for the search task and a meaningful experience for the user?

A more robust approach to the evaluation of tuple extraction algorithm is thus to perform an *extrinsic* evaluation, which measures the quality of the extractor by looking at the impact on the effectiveness of the search task. For certain applications or tasks, e.g. simple QA, the extrinsic evaluation can be automated. For such scenarios a new test set can be created by using the user questions as an input and the answers that should be produced by the system will be the ground truth labels. In the context of complex and exploratory search tasks, however, since the outcome of the search task involves synthesis of multiple information fragments and high levels of analysis and sensemaking, we cannot fully automate the evaluation process. Essentially, any evaluation methodology needs to involve the user in the process. A reliable way to address this evaluation challenge is to conduct a lab study involving real users employing the search system to perform an assigned task, e.g. finding answers to a set of test questions. This approach, in effect, is a variation of extrinsic evaluation and we can refer to it as *user-centered extrinsic evaluation.*

3.5.1 Evaluating Our IRE Module

In this section we have elaborated on a few possible directions to take in order to evaluate the effectiveness of our designed information retrieval and extraction tool. In particular, we contrasted intrinsic and extrinsic evaluations; While intrinsic criteria relate to a system's objective, extrinsic criteria relate to its functions and its main purpose [179]. To put this another way, Resnik [437] notes that the extrinsic evaluation treats the analyzer (e.g. the information extraction tool) as an *enabling technology*, whose value is not intrinsic but rather resides in its contribution to a larger application. In the rest of this section we expand on ways of conducting intrinsic and extrinsic evaluations of our developed IRE module and describe our next steps.

3.5.1.1 Intrinsic versus Extrinsic Evaluation

In the context of designing a search framework to support exploratory and complex information seeking activities, an intrinsic evaluation of our information extraction tool would assess the accuracy of the results returned by this tool as a stand-alone system, whereas the extrinsic evaluation would focus on the impact of the extracted information (i.e. entity-relationship tuples) within the context of an exploratory search support framework.

In an intrinsic evaluation we can ask questions such as the following:

1. based on the given reference sentence, are the connected entities directly related?
2. based on the given reference sentence, are the tagged entity mentions as specific as possible?
3. based on the given reference sentence, is the generated relation label readable?
4. does the relation label convey the same information as the reference sentence?

We explore this type of evaluation for our system in Appendix A.

On the other hand, the extrinsic evaluation scenario can focus on questions such as:

1. are users who take advantage of the search UIs, which visualize and enable interaction with these extracted entity-relation tuples, more successful in completing their search tasks?
2. will a coupling of entity-relationship based representation of search results and the textual content of document result in less time spent on reading the returned documents?

Both of these types of evaluation are valuable and support different purposes. In fact, it is important to separate out the impact of entity-relationship based representations on the outcome of search tasks from the effect of possible errors and omissions in the output of our information extraction tool on these outcomes.

Further, as we argued in Section 2.5.3.2, there is a non-trivial relationship between the accuracy of the output of information retrieval systems and the outcome of exploratory search tasks. Similarly in NLP systems, since different components (e.g. a parser, an entity tagger, etc) often interact in complex ways, we cannot simply assume that there is a linear correlation between the accuracy of individual components and the quality of the final output [438]. For example, there are cases that the effects of errors of different components are multiplicative, that is errors propagate down a processing pipeline and the final output may be quite poor despite high effectiveness for each of the individual components. For other systems, however, the overall effectiveness can be much higher than one would expect given individual component-level effectiveness. As Resnik [438] notes these systems represent cases where some components are able to compensate for the poor quality of other components. One example of this is in cross-language information retrieval (CLIR) where the user issues a query in one language to retrieve documents in another language. If we consider CLIR systems as having a translation component and a search component, given that the accuracy of automated translation systems is quite poor (relative to humans performance), one might expect the performance of a CLIR system to be lower than a monolingual IR system. Yet it is shown that both cross-lingual and monolingual IR systems have comparable performances. Resnik attributes this to the inherent redundancy in documents and queries as a means of compensating for the poor translation quality.

3.5.1.2 Next Steps

Given the complementary benefits of intrinsic and extrinsic evaluation methods and the observation that a poor component-based evaluation results does not necessarily correspond to a poor end-to-end performance, we consider three main evaluation criteria to reliably investigate the efficacy of knowledge-graph based representations of search results on the exploratory search outcomes.

1. [**Evaluation Requirement 1**] Efficacy of knowledge graph based representations of search results in supporting exploratory search tasks should be evaluated independently from the effect of errors in the output of current IE systems.

That is, we consider an ideal version of our IE algorithm's output, where there are no extraction errors and no extraction is missed so that we can isolate the effect of extraction errors from the efficacy of perfect knownledge graph representations on search outcomes.

2. [**Evaluation Requirement 2**] Since the main goal of designing the information retrieval and extraction tool is to support exploratory search tasks, an extrinsic evaluation method is preferred as it can assess the impact of extracted entity-relationship tuples on the outcome of exploratory search tasks.

3. [**Evaluation Requirement 3**] Measuring the impact of the information retrieval and extraction tool under real-world circumstances (i.e. erroneous knowledge graph representations) is also useful as it can provide an ecological validity for our findings. An intrinsic approach that directly evaluates the output of our designed IE tool can provide a mechanism to artificially control and inject errors to the output of an IE algorithm and observe its impact on the final search outcome.

In order to satisfy these three requirements we outline an evaluation framework as follows;

- To satisfy the first requirement, two experts manually revise the output of the IE algorithm such that the extraction errors including incorrect entity boundaries and inaccurate relation labels are fixed. Further, any sentence that contained a valid triple but did not lead to an extraction by the system is manually processed by the experts following the same steps as our designed algorithm.

- To satisfy the second requirement, we use the gold dataset that is resulted from the manual refinement of the output of our IE tool to populate our knowledge graph representations which are then incorporated by the second component of our search framework, the Search UI. Next we evaluate the efficacy of these representations as a part of our search UI using representative search tasks and controlled lab studies. Essentially, representations that are populated with gold entity-relationship data provide an opportunity to focus on improving the interactive and visual aspects of knowledge graphs while the underlying information is as accurate as the textual content of retrieved documents. Chapters 4-6 elaborate on these efforts.

- Finally, using the final version of our knowledge graph interface we look at the effect of error-prone representations in supporting information seeking. To this end, we create multiple versions of our entity-relationship datasets with varying levels of errors and omissions, and evaluate the impact of these erroneous representations on the outcome of search tasks. Intrinsic evaluations are used to provide a lower bound for the accuracy and coverage of automatically generated knowledge graphs, while the gold dataset can serve as an upper bound. Chapter 7 reports the findings of our investigation of impact of errors in information seeking.

- As a side project, we also explored the idea of evaluating the output of IE systems using non-expert annotators. We designed guidelines for assessing the quality of different system extractions and recruited crowdworkers to perform this annotation task. We present this work in Appendix A.

3.6 Chapter Summary

In this chapter we motivated the need for approaches that extend the document retrieval paradigm by automatically extracting semantic information from the textual content of documents retrieved for a query. This extracted information can then be leveraged to provide a conceptual overview of the documents and assist searchers with locating relevant fragments of information and how they relate to other concepts discussed in the documents.

To this end, we developed an information extraction tool as an extension to the search and document retrieval paradigm and elaborated on the underlying algorithm. In a nutshell, the work described in this chapter has addressed the first requirement of designing solutions for supporting exploratory search activities specified in our thesis statement, essentially to provide information and not documents in response to users' complex information needs. In the next chapter, we utilize our developed IRE module to populate an entity-relationship dataset given a set of documents retrieved for some sample topics. This dataset is then used to generate knowledge graph representations corresponding to these topics.

Chapter 4

Exploring a Knowledge Graph Extension

> *The power of the unaided mind is highly overrated. Without external aids, memory, thought, and reasoning are all constrained. But human intelligence is highly flexible and adaptable, superb at inventing procedures and objects that overcome its own limits. The real powers come from devising external aids: it is things that make us smart.*
>
> – Norman, 1993, p. 43

Chapter 3 motivated the design of a new search paradigm which combines the textual representation of retrieved documents with their corresponding knowledge graphs in order to better support sensemaking activities. Given our developed information Discovery tool, Open IE algorithms can be employed to automatically extract entity-relationship tuples from the text of retrieved documents. In this chapter, we focus on representing these tuples as a Knowledge Graph extension to documents content. Essentially, we begin to examine the efficacy of such a search framework that *extracts* and *represents* semantic information from a set of documents in supporting information seeking activities.

The main idea behind this framework is based on combining knowledge graphs with document retrieval in order to provide a conceptual overview for the information space. Knowledge Graphs have been widely used to promote meaningful learning as well as browsing knowledge and navigation. However, there is limited insight into how these graphs can be utilized by searchers to aid with locating relevant information and making sense of them. Our initial study, presented in this chapter, challenges the models that focus on either traditional document retrieval or the use of linked data for finding relevant information. The findings demonstrate that knowledge graphs and the coherent content of textual documents are both crucial for supporting users during their exploratory activities.

4.1 Motivation

There is a growing realization in the IR community that the current paradigm of retrieving a ranked list of documents is inadequate in solving complex information needs [21]. Examples of complex and exploratory search tasks include: learning about a new domain (e.g., "astronomy 101") or finding hidden connections between two events or concepts (e.g., "impacts of WWI on economy"). It can be argued that current search engines are generally sufficient when the need is well-defined in the searcher's mind. However, when information is sought to address broad curiosities, for learning and other complex mental activities, retrieval is necessary but not sufficient [597].

In order to bridge the gap between what search engines currently offer with the support needed for more complex search activities, different extensions have been proposed (Section 2.3). These solutions focus on retrieving information as opposed to documents to address the user's information need. You may recall from Chapter 2 that a dominant technique towards automatic retrieval of information is Information Extraction. The outcome of these algorithms can be represented as a Knowledge Graph, that is a network of some domain knowledge represented by labelled nodes and labelled links between them. Knowledge Graphs (also referred to as Concept Maps or repositories of Linked Data) have been widely used to promote meaningful learning as well as browsing knowledge and navigation. As observed by Carnot et al. [90] the structure of Concept Maps that are carefully constructed may assist learners in finding information more quickly. When these maps are available, they can provide a useful structure for understanding new documents, and the new documents can provide useful context to the knowledge models. [560]

The problem of automatically generating knowledge graphs and databases of linked data from the web has been well studied. However, there is limited insight into how these graphs can be utilized by searchers to aid with locating relevant information and making sense of them. Indeed, better integration of structured and unstructured information to seamlessly meet a user's information needs is a promising, but underdeveloped area of exploration [21].

There have been some efforts (e.g., [143]) to utilize Linked Data to enable user-oriented exploratory search systems. However, we believe these graphs, when applied in isolation, are not sufficient for an effective information finding and sense making, particularly for more complex search tasks. That is, a hybrid approach that combines the coherent content of text with the organized structure of graphs should be taken to better support complex tasks. Therefore, we aim at exploring a new search framework (Section 4.2) and observe how the provided Knowledge Graphs and their mappings to corresponding documents will be utilized by different searchers to complete both simple and exploratory search tasks. In this chapter we focus on the interplay between each document and its corresponding graph to gain insight into how this coupling can support finding and analyzing information. Investigating how people make sense of information by utilizing this new framework can help us design an interaction model that facilitates comprehension, analysis and insight.

The main goal of this experiment is to develop a better understanding of how users search for relevant information using a new design based on Knowledge Graphs that are derived from text. There are two areas of past research that are motivating the work presented in this chapter: (1) understanding information seeking behaviors given different search user interfaces; and (2) approaches that specifically leverage entities and relationships to support search and browsing. Among the first group, there is a body of work that focuses on observing users' behaviour and identifying the challenges searchers face during their search session, common information seeking activities among them and gaining insight into how to support these activities (See Section 2.2.2.2). These findings can help guide the design of search interfaces. Similarly, in this preliminary study we are interested in understanding how a new search UI that leverages knowledge graph representations of search results along with their corresponding documents performs across simple and complex search tasks. Further, our proposed search framework is inline with the second group of approaches that demonstrated the efficacy of entity-relationship data for supporting exploratory search and browsing (See Section 2.4.3).

The main distinctions between our work and these related work are as follows: (1) Approaches based on faceted search and linked data are mainly limited to named entities and basic relations (simple predicates or hierarchical) between them. However, we extract a broader set of entities and concepts and we identify semantic relations based on dependencies between them. These relations are not limited to a predefined set of predicates and provide context for understanding the connections between entities. (2) We generate graphs automatically using the documents collection retrieved for the user's query. Our knowledge graphs thus are derived from the same information space that the searcher is interested to explore. This is beneficial because first, the graphs contain the information related to the user's information need and second, it provides an interplay between the text and the graph which can support "comprehension" through discourse relations [348] which is not preserved in linked data.

The rest of this chapter is organized as follows. In Section 4.2 we introduce our designed framework that provides two alternative representations of search results side by side. To evaluate this new interface, in Section 4.3, we conducted a user study which is exploratory and observational in nature and provides the opportunity to document and analyze interesting interaction patterns. We also identified frequent interaction patterns performed during an information seeking session (Section 4.4). Further investigation of the similarities and differences observed between simple and complex search tasks can be utilized to understand the reasons behind the lack of support from the current search engines for complex search tasks. Finally, we examined the obstacles and challenges faced by the participants during their exploration and propose future directions that can lead to better understanding of the requirements of a new search UI that supports information seeking activities (Section 4.6).

4.2 Enabling a New Search Paradigm

We propose a new search framework that takes advantage of knowledge graphs to mitigate the problem of information overload by providing a semantic organization of the information space. We also argue that knowledge graphs cannot enable an effective framework for supporting complex search tasks if applied in isolation. In the following subsections we provide an outline of our general framework and we describe how this new framework can be employed to support searchers during information seeking activities.

4.2.1 The Proposed Framework

Given the current document retrieval paradigm, searchers need to make sense of the long lists of ranked results provided by search engines. In fact, the lack of effective overviews challenges users who seek to understand these results. We envision the following qualities that Knowledge Graphs can offer to minimize this challenge:

1. They provide a fine grained representation of articles and enable searchers to retrieve relevant pieces of information (rather than documents) for their query;
2. They visualize how different entities and concepts are connected in a domain;
3. They provide an overview (i.e., the big picture) of the information space related to the user's topic of interest;
4. They demonstrate the salient entities related to a topic.

Although Knowledge Graphs could be powerful tools to support navigation and learning for exploratory search, they cannot replace the document search and retrieval for searchers. Each document represents facts (described in sentences) in a particular order, which is coherent and meaningful. This ordering helps with identifying the connections between different facts, which are not preserved in the graph representation.

When we extract information from text and restructure it as a knowledge graph to visualize semantic relations between concepts, we lose discourse relations (i.e., information on how two segments of discourse are logically connected to one another) which are crucial for comprehension and inference from a text.

Hence, there should be an interconnection between the documents and their corresponding graphs in order to overcome the shortcomings of each representation of search results in isolation. We hypothesise that a hybrid approach, which combines the structure of graphs and the coherent context of text, should be used to better aid information seeking activities. We believe such a framework can engage users more fully in the search process. As the searcher explores, each graph provides a graphical summary for each document. They could be considered as advanced tables of content that point to the more interesting parts of a possibly long article and help with getting the big picture at a glance.

In order to design a framework which supports a seamless interaction between documents and their corresponding graphs, we identified different types of edges and connections:

Connecting each document to its corresponding graph: In each document, only the sentences containing an extracted triple (entity1, relation, entity2) are linked to their corresponding part of the graph and vice versa. Therefore, when the user skims a document these sentences are highlighted and linked to the graph. So the user can switch to browsing the graph to explore a particular entity (most commonly, a named entity), the related entities and how it connects to the other parts of the article.

Connecting the graphs: Documents fetched for the user's query may discuss different aspects of the same topic or provide different perspectives. Therefore, the corresponding graphs are not independent of one another. The same domain terms and entities appear in these graphs and they have to either be represented as a single node in the aggregated graph, or mapped through a set of inter-graph links to preserve these connections. These links indicate how different parts of different documents are related to each other. These links can also be helpful when a user is analyzing the documents in a sequence. The user can start with the first document and take advantage of the corresponding graph as a structured summary, which guides her through understanding this document. Then she moves on to a new document and extends the current graph to represent both documents. This way, the user can keep track of the facts "she already knows" and the ones which are covered in the new document. She can also identify the common facts covered by both documents. She can build on this graph by incrementally adding a document to his collection. The research described in this dissertation leverages the first type of mappings, i.e., the links between different elements of the knowledge graphs and their corresponding segments in the documents and reserves the second type of mappings, that connect knowledge graphs derived from different documents for future work. **A Sample Search Scenario.**

Consider the following scenario: while reading an article about "Napoleon's invasion of Russia", the user comes across the entity "Treaty of Tilsit". By traversing to this node in the graph, she analyzes a set of related facts, which includes:

- Which countries first signed this treaty? Which countries followed later?
- When was this initially signed and why?
- What were the terms of this agreement?

The user can then go back to the article and resume reading (while she now has a better understanding of this topic) or navigate through the graph and explore a different part of the article based on where the graph takes her. The nodes and facts in the graphs are also linked to their corresponding text in the documents. Therefore, the readers can clarify the interesting facts they observe in the graph.

They can also navigate to the sections of the texts that are more appealing to them without needing to read through an article in a linear fashion.

Also, consider a scenario in which a searcher is trying to find out the "impacts of WWI on economy". The following two facts are extracted from an article: "many of America's men were serving overseas in the war" and "companies allowed women to work in previously male only jobs". While these two facts are located in close proximity in this article, they are positioned at two disconnected parts of the graph. By traversing the graph only, the searcher cannot discover any connection between these two seemingly unrelated facts. Hence, there should be an interplay between the graph and the document in an effective search paradigm.

4.2.2 Evaluating the New Framework

In the previous section we hypothesized that coupling a set of retrieved documents with their corresponding knowledge graphs, which represents the salient entities and underlying relations in a domain of interest, can provide a more effective search experience for the user, especially when investigating more complex search tasks. In this experiment we focus on the interplay between each document and its corresponding graph to gain insight into how this coupling can support information seeking. Therefore, we do not investigate the effects of connecting different documents through knowledge graphs and the incremental extension of a graph in this work. We formulated a list of research questions to investigate our hypothesis:

1. How is this framework used for finding relevant information?

(a) Which features of the graphs are used more frequently by the participants?

(b) Is there a difference in this usage across two different types of search tasks?

(c) What is the most common starting point for the searchers? The graph side or the document side?

(d) Is the starting point affected by the complexity of the search task?

(e) What are the common activities across the searchers who start their exploration from the same side?

2. How does this framework provide support for locating relevant information?

(a) What are the most common interaction patterns that correspond to finding relevant information?

(b) Are these patterns affected by the complexity of the search tasks?

(c) To what extent do the graphs contribute to locating relevant information?

(d) Do nodes and edges in the graphs provide different types of support for finding relevant information?

(e) Does the complexity of the search task affect the effectiveness of the graphs (and in turn nodes and edges) in locating relevant information?

In order to find answers to these questions we designed a user study in three steps: (1) Extracting Knowledge Graphs from Text, (2) Mapping Graphs and Documents and (3) Employing search tasks with different levels of complexity. The following subsections discuss these steps.

4.2.2.1 Generating Knowledge Graphs and Mappings

We designed an Open Information Extraction system that processes a text collection and generates (entity-relation-entity) triples [474]. This tool, as a part of an Information Discovery framework, was described in Chapter 3. Here we provide a high level overview of the underlying algorithm.

This module is implemented in four phases. During the first phase we create the input corpus by collecting retrieved documents based on a given query. Next, we extract entities from text using state-of-the-art entity taggers. We then select the sentences that contain at least two entities in them and parse them using Stanford Dependency Parser. For each sentence, we extract meaningful relations between the entities by finding the shortest path in the corresponding parse tree. We derived a set of heuristics from the parser's dependency patterns that lead to semantically meaningful relations. In the final phase we generate labels for the extracted relations and rank them based on relevance to the query and the informativeness of the extraction. Since we are investigating the effectiveness of employing knowledge graphs to provide support for exploratory search tasks, we assure the generated graphs are accurate. Therefore, some minor errors that were caused by ER extraction and label generation modules were revised by an expert.

For each document in our result list, we created a corresponding knowledge graph and mapped all entities and relations to their corresponding parts of text. All nodes and edges in the graphs as well as their mentions in text are clickable. These mappings provide an interplay between text and graphs and are made possible because our graphs are derived from the same set of documents. This is one key advantage of this framework as compared with systems that employ external resources such as DBPedia to aid information seeking.

4.2.2.2 Simple and Complex Search Tasks

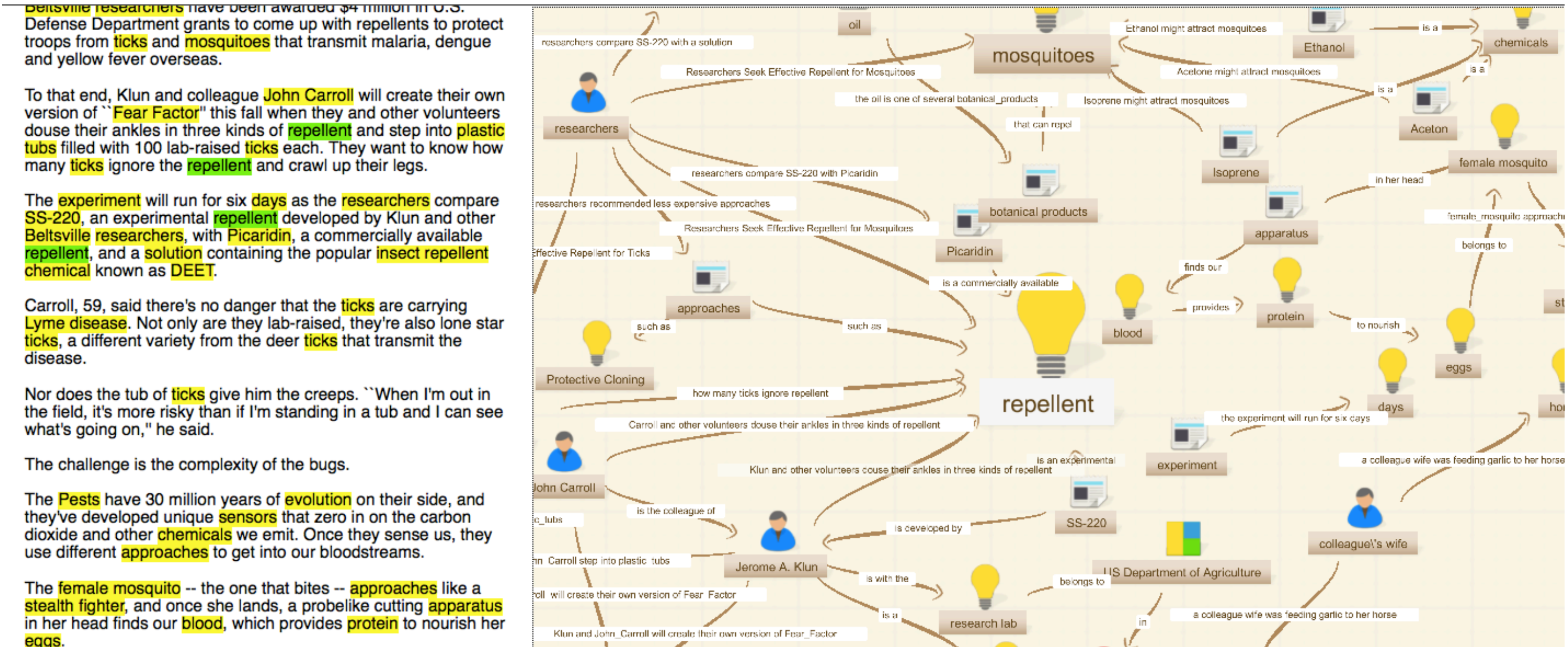
People's day-to-day search activities can vary greatly in their motivations, objectives, and outcomes. These search activities can be broadly classified into two groups: "Simple" and "Complex". Simple search tasks are similar to "known-item" search tasks and usually involve looking up some discrete, well-structured information object: for example numbers, names and facts [352]. Complex search tasks, on the other hand, are seen to be more exploratory and involve investigating, learning and synthesis of information [605].

What really differentiates simple and exploratory search tasks is the clarity of the information need, the familiarity the searcher has with the task domain, and the analysis and understanding involved [597]. These factors invariably affect how searchers interact with information, and how they search and browse. In this experiment we investigate (1) how the complexity level of a given search task affect the searchers' information seeking behaviour and (2) how well the designed framework supports these two types of search tasks.

4.3 Designing the Experiments

We designed a within-subject study in which each participant needed to complete two search tasks using the same interface. We conducted our experiments within the framework provided by the TREC 2007 Question Answering track. We followed the guidelines from QA and CiQA tracks [133] to design a "simple" task and a more complex and open ended search scenario ("complex" task).

For the *complex task*, the searchers were required to find as many relevant sentences as possible within a 10 minute time limit. They were given the following task description: "What is the position of [California] with respect to [stem cell research]? – "The analyst wishes to know if Californians generally support stem cell research and what actions they are taking to accomplish the research." For the *simple task*, the searchers were given the topic "Lyme disease" and a short list of questions (e.g., "what organism causes Lyme disease?") and they were required to find answers to these questions by providing the corresponding sentence and document number from the given list based on a fixed query.



Beltsville researchers have been awarded $4 million in U.S. Defense Department grants to come up with repellents to protect troops from ticks and mosquitoes that transmit malaria, dengue and yellow fever overseas.

To that end, Klun and colleague John Carroll will create their own version of ``Fear Factor'' this fall when they and other volunteers douse their ankles in three kinds of repellent and step into plastic tubs filled with 100 lab-raised ticks each. They want to know how many ticks ignore the repellent and crawl up their legs.

The experiment will run for six days as the researchers compare SS-220, an experimental repellent developed by Klun and other Beltsville researchers, with Picaridin, a commercially available repellent, and a solution containing the popular insect repellent chemical known as DEET.

Carroll, 59, said there's no danger that the ticks are carrying Lyme disease. Not only are they lab-raised, they're also lone star ticks, a different variety from the deer ticks that transmit the disease.

Nor does the tub of ticks give him the creeps. ``When I'm out in the field, it's more risky than if I'm standing in a tub and I can see what's going on,'' he said.

The challenge is the complexity of the bugs.

The Pests have 30 million years of evolution on their side, and they've developed unique sensors that zero in on the carbon dioxide and other chemicals we emit. Once they sense us, they use different approaches to get into our bloodstreams.

The female mosquito -- the one that bites -- approaches like a stealth fighter, and once she lands, a probelike cutting apparatus in her head finds our blood, which provides protein to nourish her eggs.

oil
researchers compare SS-220 with a solution
mosquitoes
Ethanol might attract mosquitoes
Ethanol
is a
chemicals
Researchers Seek Effective Repellent for Mosquitoes
Acetone might attract mosquitoes
is a
is a
researchers
the oil is one of several botanical_products
Isoprene might attract mosquitoes
Aceton
that can repel
female mosquito
researchers compare SS-220 with Picaridin
Isoprene
in her head
researchers recommended less expensive approaches
botanical products
apparatus
female_mosquito approach
Researchers Seek Effective Repellent for Mosquitoes
belongs to
Effective Repellent for Ticks
Picaridin
finds our
is a commercially available
approaches
provides
protein
st
such as
such as
blood
to nourish
eggs
Protective Cloning
how many ticks ignore repellent
repellent
days
ho
Carroll and other volunteers douse their ankles in three kinds of repellent
the experiment will run for six days
is an experimental
experiment
a colleague wife was feeding garlic to her horse
Klun and other volunteers douse their ankles in three kinds of repellent
John Carroll
is the colleague of
c_tubs
is developed by
SS-220
colleague\'s wife
Jerome A. Klun
nn Carroll step into plastic tubs
US Department of Agriculture
is with the
belongs to
roll will create their own version of Fear Factor
is a
research lab
in
a colleague wife was feeding garlic to her horse
Klun and John_Carroll will create their own version of Fear_Factor

Figure 4.1: A sample coupling of a document and its corresponding knowledge graph in the designed interface

4.3.1 Construction of Results Lists

We created a list of 10 documents for each search task. In order to investigate the contribution of the graphs in identifying *relevant* documents we used different procedures to construct *artificial result lists*. For the complex task we were interested to see how many *nuggets* (i.e., a piece of text containing relevant information) would be found by the searchers. Therefore, we retrieved all documents related to the given topic and ordered them based on the number of nuggets they contained (as identified by NIST assessors). We then created our list by selecting the top 10 documents and shuffling them so they are not placed in that order.

For the simple task we constructed a list by adding a mix of relevant documents that contained nuggets (answers to the factoid / list questions) and the ones that did not. We refer to them as "good" and "bad" documents respectively. We distributed the "good" documents evenly throughout the list. This will assure the participants will have to explore the entire list to be able to find all the answers. Hence, we are able to investigate if the graphs can be used to identify "good" documents faster, while they can also be used for locating relevant sentences inside each document.

4.3.2 Experiment Setup

We recruited 20 (6 female) participants from a very diverse pool for this study, all of whom use the Internet on a regular basis to search for information. They ranged from students to senior engineers and faculty members. Their study area covered many major areas of Computer Science, Architecture, Chemistry, Biology and Physics.

The participants were given 10 minutes to complete each task. They also needed to complete two post-task questionnaires that would assess their familiarity with the topic and the experience they just had. We used questionnaires provided by TREC-9 Interactive Searching track[1] and modified them to fit into our experiment design. Two of the participants did not finish the experiment so we excluded their data from our analysis. In order to control for order effects, we rotated the order in which the tasks were assigned to the participants. That is, participants were randomly divided into two groups of equal size and one group started with the "simple" task while the other group started with the "complex" task. By the end of both tasks, the participants were asked to share their feedback about the effectiveness of the search interface and knowledge graphs for search by responding to a more detailed questionnaire. Participants were also monitored during the search sessions and the computer screen was captured for a later review.

4.3.3 Search Interface

We utilized a search interface implemented by the InsightNG company based on our design for conducting the user study. This interface has two panes (Figure 4.1). The left pane resembles a modern search engine's result page which provides a list of 10 documents along with their titles and text snippets. Clicking on each of these items would load the full document on the left and the corresponding graph on the right side. On the document side, all terms with corresponding nodes in the graph were highlighted in yellow while the parts of the text corresponding to an edge (i.e., relationship) were highlighted in grey. The user could explore the article by either browsing the graph or reading through the documents. We implemented mappings between corresponding elements of the text and the graph such that clicking on one would highlight the other. For example, clicking on a node in the graph will highlight all corresponding terms in the text in green. Besides, clicking on an "entity" or "relation" from the text would center and zoom on that element in the graph.

[1] www-nlpir.nist.gov/projects/t9i/qforms.html

4.4 Results and Observations

While retrieving relevant information is the goal of all information seeking systems, we cannot successfully design an effective approach unless we learn about searcher's information seeking patterns. By assessing the designed interface we were mainly interested to learn how people go about finding information. For example, will they navigate through the graphs to find an answer to their question? Or will they make use of the graphs to locate where to read in the text. By looking through the data collected by questionnaires and observing the participants using the designed search interface, we gained interesting insight into the common user behaviour during information foraging activities.

GCN	Clicking on a node in the graph
GCE	Clicking on an edge in the graph
DCN	Clicking on an entity mention in the document
DCE	Clicking on a relationship mention in the document
SfG	Starting from the Graph side
SfT	Starting from the Document side
B2G	Switching to the Graph side
B2T	Switching to the Document side
RM	reading the mention of an entity/relation in text
RF	exploring an entity in the graph by looking at related nodes and connection of this entity
DC	dragging the canvas around
DR	reading the document text
GR	reading the node/edges labels in the graph
CP	Copy-Paste an answer to the answer sheet
SD — SU	Scrolling down / up the text
B2R	going back to the SERP

Table 4.1: Actions and their notations

We defined a set of actions by observing the activities performed by different participants over the course of their interaction with the system. Table 2 lists the more prominent actions. For each participant 2 sequences of actions (one per search task) were generated by using the logs of screen videos and observing the users' interaction with the system during the experiment. For each search task we calculated the frequency of all subsequences of length 1 to 5. We call each of these subsequences an "interaction pattern" (or "pattern" in short). We filtered out the patterns with a frequency below 5. Since none of the patterns of length 5 passed this threshold we did not consider the patterns of length more than 4 in our analysis. The following subsections discuss frequent patterns observed during each search task and how participants exhibited different behaviour during four main activities: (1) switching between the graph and the text mode; (2) taking advantage of nodes and edges to locate relevant information and (3) getting started with the exploration; (4) in the end, we investigate the common patterns that led to locating an answer and compare them across two tasks. We examined the statistical significance of our observations using a paired t-test. Also, for all the tables Simple and Complex tasks are denoted by **SP** and **CX** respectively.

4.4.1 General Characteristics of Interaction Patterns in Simple and Complex Tasks

In this work, we are interested to identify the "interaction patterns" that are common in both search tasks and the ones that are more pertinent to one of the tasks. We hypothesise that identifying these similarities and differences can help us understand the characteristics of simple and complex search tasks and how this will affect searchers interaction behaviour.

Table 4.2 lists the top five frequent patterns of length 1 to 4. As can be seen in this table the pattern B2T→RM (i.e., switch from the graph to the text and read the mentions of the recently clicked node / edge) was the most frequent pattern of length 2 for both tasks. This pattern corresponds to making use of the graphs to highlight the areas in the text that user is interested to read. Also, we can see that patterns starting with a GCN (i.e., clicking on a node in the graph) are the next top two frequent patterns of length 2 for the simple task. These patterns correspond to clicking on a node in the graph and then either going back to text (GCN→B2T) or exploring the related entities and the connection to the current entity (GCN→RF).

Among the longer patterns we observe clicking on a node, switching back to the text and reading the mentions is the most frequent pattern of length 3 for the simple task. This pattern repeats followed by a CP (i.e., locating an answer) as the top frequent pattern of length 4 for this task. Interestingly, a similar pattern occurs for the Complex task with one distinction: while clicking on a node is the most likely pattern to end in locating an answer for the simple task (i.e., GCN→B2T→RM→CP), clicking on an edge is more effective for the Complex task (i.e., GCE→B2T→RM→CP). We will discuss these patterns in Section 4.4.5.

	Length of the Pattern							
	1		2		3		4	
	pattern	freq	pattern	freq	pattern	freq	pattern	freq
SP	GCN	190	$B2T \rightarrow RM$	93	$GCN \rightarrow B2T \rightarrow RM$	61	$GCN \rightarrow B2T \rightarrow RM \rightarrow CP$	20
	B2T	151	$GCN \rightarrow B2T$	70	$B2T \rightarrow RM \rightarrow CP$	34	$B2T \rightarrow RM \rightarrow CP \rightarrow B2G$	17
	B2G	123	$GCN \rightarrow RF$	52	$SFG \rightarrow GCN \rightarrow RF$	26	$DC \rightarrow GCN \rightarrow B2T \rightarrow RM$	14
	DC	117	$SFG \rightarrow GCN$	49	$RM \rightarrow CP \rightarrow B2G$	18	$GCN \rightarrow B2T \rightarrow RM \rightarrow B2G$	13
	RM	115	$RM \rightarrow CP$	42	$GCE \rightarrow B2T \rightarrow RM$	16	$GCN \rightarrow B2T \rightarrow RM \rightarrow B2R$	10
CX	B2T	135	$B2T \rightarrow RM$	76	$B2T \rightarrow RM \rightarrow CP$	42	$GCE \rightarrow B2T \rightarrow RM \rightarrow CP$	24
	CP	129	$RM \rightarrow CP$	65	$GCN \rightarrow B2T \rightarrow RM$	39	$B2T \rightarrow RM \rightarrow CP \rightarrow B2G$	17
	GCN	122	$GCN \rightarrow B2T$	48	$GCE \rightarrow B2T \rightarrow RM$	32	$GCN \rightarrow B2T \rightarrow RM \rightarrow CP$	16
	RM	110	$CP \rightarrow B2R$	38	$RM \rightarrow CP \rightarrow B2G$	18	$GCN \rightarrow B2T \rightarrow RM \rightarrow B2G$	10
	B2G	98	$GCE \rightarrow B2T$	37	$B2T \rightarrow RM \rightarrow B2G$	17	$SFG \rightarrow GCN \rightarrow B2T \rightarrow RM$	9

Table 4.2: Top 5 frequent "interaction patterns" of length 1 to 4

We also looked at the distribution of main activities between the two tasks (Table 4.3). For each pattern we report its conditional probability followed by its frequency (aggregated over all 18 participants). We calculated the conditional probability of (A→ B) by applying Equation 4.1. For example, the action B2G was observed 123 times in total during the simple task and it was followed by the action DC in 27% of the cases (33 out of 123). Please note that the same equation is used for calculating the percentages in Tables 4.4 to 4.6. Also, in all these tables the frequency of patterns are reported in parentheses and the shaded rows indicate the frequency of the preceding action (i.e, A in A→ B).

$$P(B|A) = \frac{freq(A \rightarrow\ B)}{freq(A)} \tag{4.1}$$

These results indicate:
(1) *Switching between Graphs and Documents:* overall, switching back and forth between two sides was done similarly in both tasks ($\rho > 0.1$);
(2) *Click patterns: Nodes v.s. Edges:* In both tasks participants tended to click on nodes more than the edges. This trend was strongly significant for the simple task ($\rho < 0.001$)
(3) *Click patterns: Simple v.s. Complex:* on the other hand, the edges were used more frequently during the Complex task than the Simple task ($\rho < 0.1$);
(4) *Starting the exploration: Graphs v.s Text as a starting point:* finally, while participants started their search from the graph more than the text, it was a strongly significant trend for the simple task ($\rho < 0.001$). Also, starting from text was done significantly more for the Complex task than the simple task ($\rho < 0.05$).

	Switch Sides		**Clicks**				**Starts**		
	B2G	B2T	GCN	GCE	DCN	DCE	SfT	SfG	Total
SP	8% (123)	10% (151)	12% (190)	2% (31)	3% (43)	0.3% (5)	2% (30)	5% (78)	1531
CX	7% (98)	10% (135)	9% (122)	3% (47)	3% (46)	1% (11)	3% (43)	4% (57)	1347

Table 4.3: Distribution of Actions between Tasks

4.4.2 Switching between Graphs and Documents

As we discussed in Section 4.2.1 Graphs and Documents both provide different types of support for users who are searching for information. We were interested to identify the main activities that led the participants to switch from the text to the graph or vice versa. To this end, we analyzed the frequent patterns starting with a "B2G" (switching from text to graph) or a "B2T" (switching from graph to text).

	action	%	action	%
SP	$B2G \rightarrow DC$	27% (33)	$B2T \rightarrow RM$	62% (93)
	$B2G \rightarrow GCN$	27% (33)	$B2T \rightarrow DR$	8% (12)
	$B2G \rightarrow RF$	15% (19)	$B2T \rightarrow DCN$	6% (9)
	B2G	(123)	B2T	(151)
CX	$B2G \rightarrow DC$	29% (28)	$B2T \rightarrow RM$	56% (76)
	$B2G \rightarrow GCN$	23% (23)	$B2T \rightarrow DR$	10% (13)
	$B2G \rightarrow B2T$	13% (13)	$B2T \rightarrow DCN$	8% (11)
	$B2G \rightarrow RF$	8% (8)	$B2T \rightarrow CP$	6% (8)
	$B2G \rightarrow GCE$	7% (7)	$B2T \rightarrow SD$	7% (9)
	B2G	(98)	B2T	(135)

Table 4.4: Reasons for Switching to Graph / Text

As indicated in Table 4.4 (1) for both tasks clicking on a node (GCN) and dragging the canvas (DC) were the most frequent actions taken by the participants right after switching to the Graph side; (2) while learning about an entity (RF) was the third frequent action for the simple task, going back to the document again (B2T) came third for the Complex task. This distinction is significant ($\rho < 0.01$). That could be an indicator of the fact that after switching to the graph the participant was not sure where to start from or where to go next. Therefore, they decided to go back to the text again; (3) one interesting observation is that the top three main activities after going back to the text were similar for both tasks, with similar likelihood. Furthermore, RM (i.e., "reading a mention") was by far the most dominant activity once the participants switched to the document side ($\rho < 0.001$). This corresponds to making use of nodes or edges in the graph to find out where to read in the text.

4.4.3 Click Patterns: Node v.s Edges

This user study revealed that different searchers take advantage of the provided graph in a variety of ways: while some participants found the edges more effective to locate relevant pieces of information, others made use of the mappings between nodes and text to find important terms more quickly in text. We used the click patterns to investigate if nodes and edges are used differently across the two tasks. Overall, we observed similar patterns for clicking on nodes and edges for both tasks.

	action	%	action	%
SP	$GCE \rightarrow B2T$	74% (23)	$GCN \rightarrow B2T$	37% (70)
	$GCE \rightarrow B2T \rightarrow RM$	52% (16)	$GCN \rightarrow RF$	27% (52)
			$GCN \rightarrow GCN$	8% (15)
			$GCN \rightarrow DC$	12% (22)
	GCE	(31)	GCN	(190)
CX	$GCE \rightarrow B2T$	79% (37)	$GCN \rightarrow B2T$	39% (48)
	$GCE \rightarrow B2T \rightarrow RM$	68% (32)	$GCN \rightarrow RF$	25% (31)
			$GCN \rightarrow GCN$	11% (14)
			$GCN \rightarrow DC$	7% (8)
	GCE	(47)	GCN	(122)

Table 4.5: The most likely actions after a Click

As can be seen in Table 4.5, most of the clicks are followed by reading a mention. This action is by far the dominant action performed after clicking on an edge ($\frac{16}{31}$ and $\frac{32}{47}$) and no other frequent pattern starting with a GCE is observed. However, reading the mentions of a node is not as dominant ($\rho > 0.1$) and it is followed closely by exploring an entity by reading its connections to other nodes (i.e., $GCN \rightarrow RF$). The only noticeable difference observed between the two tasks is that reading the mentions of an edge was done more frequently for the Complex task ($\rho < 0.05$).

4.4.4 Starting the Exploration

We also identified the common activities performed by the group who start their exploration from the Graph side (SfG) compared with the group who start from the

Document side (SfT). Table 4.6 lists the most frequent patterns starting with a SfG and the ones starting with a SfT.

	action	%	action	%
SP	$SFG \rightarrow GCN$	63% (49)	$SFT \rightarrow DCN$	43% (13)
	$SFG \rightarrow DC$	9% (7)	$SFT \rightarrow DR$	27% (8)
	$SFG \rightarrow GCE$	4% (3)		
	SFG	(78)	SFT	(30)
CX	$SFG \rightarrow GCN$	60% (34)	$SFT \rightarrow DCN$	33% (14)
	$SFG \rightarrow GR$	14% (8)	$SFT \rightarrow DR$	26% (11)
	$SFG \rightarrow DC$	11% (6)	$SFT \rightarrow CP$	23% (10)
			$SFT \rightarrow DCE$	7% (3)
	SFG	(57)	SFT	(43)

Table 4.6: Starting from the Graph v.s. Starting from the Document

By analyzing these patterns, we observed that clicking on an entity was the very first action taken by participants regardless of their starting point (graph or text) and the task ($\rho < 0.001$). However, for the group who started their exploration from the graph side, clicking on an entity was by far the most dominant action (around 60% for both tasks). On the other hand, for the group who started from the document side, the top two patterns (i.e., clicking on an entity in text (DCN) and reading the text (DR)) were not significant ($\rho > 0.1$).

In fact, "query nodes" was identified as the main starting point for exploring the graph. One should note that since the participants did not submit a query to the system, we refer to the main entities in the task description as "query nodes". While for the complex task "California" and "Stemcell*" nodes were clicked in 72% of the cases, "Lyme*" nodes were selected in 63% of the cases once participants started their search from the graph side.

Another observation for the group who started from the graph was that exploring the graph (corresponding to DC and GR activities) was done more frequently for the complex task (25% compared with 9% for the simple task) and the difference was significant ($\rho < 0.05$).

4.4.5 Common Patterns for Finding Answers

The conducted user study clearly indicated that the current stage of the new interface provided different levels of support for different types of tasks. We conducted a post-task questionnaire to gauge participants' preferences for interacting with knowledge graphs versus documents texts as well as their rationales behind switching from one side to another in the UI provided. Out of 18 participants, 9 found the availability of a knowledge graph very useful for the simple task, while they preferred the text view for the Complex task. However, 4 found the graphs more useful for the complex task and 4 mentioned the graphs were useful for both tasks. One participant preferred the text for both tasks. As commented by most participants, the graphs were the most effective when the searchers were clear about what information they were looking for (e.g., an answer to a specific QA question). However, when the nature of task was complex (e.g., CiQA topics), they were not sure how to navigate in the graphs and would mostly go through the text to find relevant information. Table 4.7 summarizes the participants' preferences in regards to the representation (knowledge graph versus documents) as well as their main motivations for leveraging these representations.

Types of documents the graph is useful for	Longer documents	11
	with a lot of names	10
	more technical	8
Main benefits of using graph	Locate certain pieces of information	12
	Get the big picture / overview of document	11
	Connecting pieces of information	11
Switching from document to graph	To explore related entities	14
	To look at the big picture	8
	To locate a more interesting part of text to jump to	8
Switching from graph to document	To Further read about a fact	11
	To learn more about the current element of graph	11
	To understand a label in graph	8

Table 4.7: Summary of Searchers Preferences

In this experiment we assume the Rationality Principle [418] holds. That is, the searchers' behaviour is purposeful and hence they carry out a sequence of actions to achieve some goal. As mentioned, the participants interacted with our system in order to find a set of "answers" for two different tasks within a 10 minute time limit. These answers were evaluated by using NIST judgements provided by TREC. For the QA task, participants found 2.42 correct answers on average to the 5 factoid question. For the CiQA topic, there was a total of 2 vital and 13 okay nuggets present in the 10 listed documents, out of which the participants were able to retrieve 0.63 vital and 2.0 okay nuggets on average.

One of the most interesting outcomes of analyzing the interaction patterns was to identify the ones that led to locating an "answer". While for the simple task an answer is a known factoid (mostly an entity), for the Complex task an answer is a snippet of text that contains some evidence or support for a given statement. We created state diagrams that illustrate the patterns that led to locating an answer.

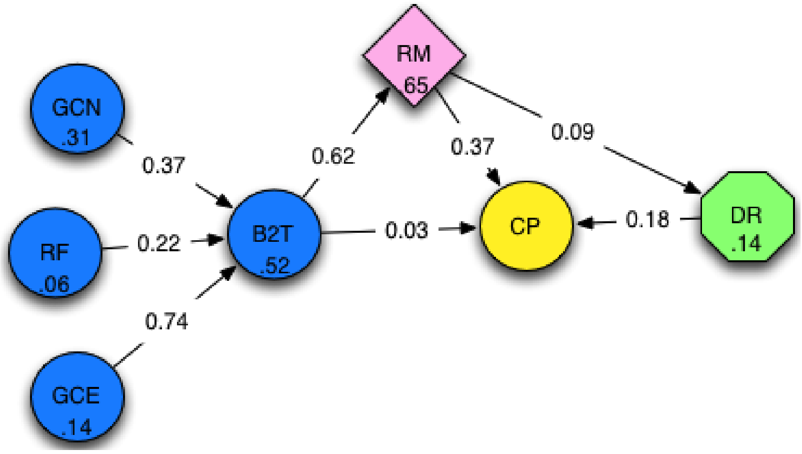


Figure 4.2: State Diagram for Frequent Patterns that Led to an Answer - Simple task

As depicted in Figures 4.2 and 4.3, blue states (circles) correspond to the Graph's contribution and green states (octagons) correspond to the Document's contribution in finding an answer. The state RM could belong to either of these two groups based on the preceding nodes in this state diagram. Links labels indicate the probability of transition from the source state to the destination state regardless of other states in this graph. This is the conditional probability calculated using Equation 4.1. Also, each state contains a weight that indicates the probability of getting to an answer by starting from this state. That is, traversing the state diagram starting from this node and ending at CP. Since for some states there are multiple paths leading to CP, we select the path with the maximum probability and record this path as the best candidate pattern for leading to an answer. We calculated these probabilities using Equation 4.2.

$$P(state_i) = \frac{\max(freq(state_i \rightarrow ... \rightarrow state_{CP}))}{freq(state_{CP})} \tag{4.2}$$

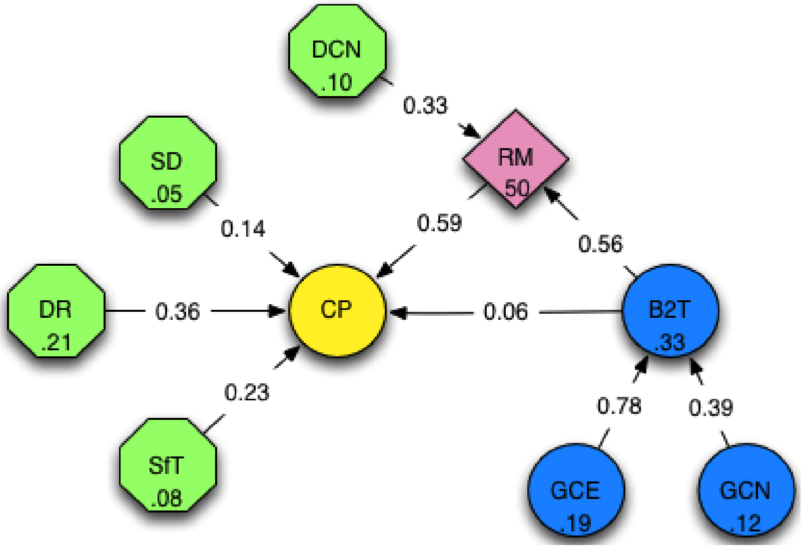


Figure 4.3: State Diagram for Frequent Patterns that Led to an Answer - Complex task

Where max($freq(state_i \rightarrow ... \rightarrow state_{CP})$) indicates the frequency of the most repeating patterns that starts from $state_i$ and ends at CP; $freq(state_{CP})$ indicates the total number of paths that lead to CP. That is, the sum of all maximal patterns ending in CP. For example, in Figure 4.2, starting from B2T there are three paths that lead to CP:

(a) B2T→CP with a probability of 0.08;

(b) B2T→RM→CP with a probability of 0.52;

(c) B2T→RM→DR→CP with a probability of 0.06;

Therefore, we consider path (b) as the most successful path that starts from going back to text and ends at locating an answer. We removed the less likely patterns from these diagrams for the sake of clarity. Therefore, the probabilities of edges exiting from a state do not sum to 1 in these figures.

We made the following key observations:
(1) Overall, there are more distinct paths (i.e. maximal repeating patterns) that lead to an answer for the complex task (129 paths) than the simple task (65 paths). This resulted in a more complex structure depicted in Figure 4.3.

(2) Graphs were more likely to initiate a path to an answer for the simple task than the Complex task (0.52 v.s. 0.33 respectively ; $\rho < 0.05$);
(3) In the cases that using the Graph led to finding an answer, nodes are more likely to be the contributing factor for the simple task (0.31 v.s. 0.14 ; $\rho < 0.05$). However, clicking on an edge was significantly more beneficial for locating an answer in the Complex task than the Simple one (0.14 v.s. 0.19 ; $\rho < 0.05$);
(4) The paths starting from RM were the most likely paths that ended in CP. These patterns correspond to finding the answers by going through the mentions of entities and relations in text.

4.4.6 Summary of Findings

In this section we revisit our research questions from Section 4.2.2 and discuss our findings.

1 (a-b). Overall, the participants clicked on nodes more than the edges in both tasks. This trend was strongly significant for simple task, while edges were clicked more during the complex task.
1 (c-d). Overall, for both simple and complex tasks, the participants started their exploration from the graphs more than the documents. However, this trend was significantly stronger for the simple task. Also, starting from the document side was done more frequently for the complex task than the simple task.
1 (e). While clicking on an entity was the main activity done by the group who started from the graphs, it was strongly significant for the simple task. On the other hand, exploring the graph was a significant pattern for this group during the complex task.
2 (a-b) Figures 4.2 and 4.3 depicted the most frequent patterns that involved finding an answer by the participants. While there are similar patterns observed across two tasks, the set of patterns for the complex task was more diverse.
2 (c-d-e) Overall, the graphs provided more support for the simple task as compared with the complex task. Also, nodes were proved more useful in locating an answer for the simple task, while edges appeared more frequently in that paths that led to an answer during the complex task.

4.5 Discussion

Our findings have implications for designing a search framework that is effective in supporting complex and exploratory search tasks.

Our quantitative results indicated that participants who started their explorations with the graph side engaged in graph exploration activities (e.g. exploring nodes and relations and navigating through the canvas) much more frequently during the complex search task than the simple task. This difference was significant. As well, the more complex structure of the diagram in Figure 4.3 indicates that searchers take a more diverse set of paths to an answer when performing complex search tasks. These observations can provide some ground for the need to offer more advanced mechanisms for interacting with information than what is currently offered by modern search engines. In fact, information seekers are more motivated to explore the information space during complex search tasks. Further, different searchers exhibit different information seeking behaviour in order to locate the relevant pieces of information in retrieved documents. A better understanding of the common interaction patterns can help the search engines to identify and facilitate these search tactics.

We also observed how nodes and edges in our knowledge graphs were utilized differently during simple and complex search tasks. It is not surprising to see that nodes contributed to locating factoid answers while edges were utilized mainly for locating relevant nuggets in the complex task. To elaborate, since the answers for the simple task are entities, nodes should be more helpful to locate the factoid information in text. On the other hand, relevant evidence supporting the "position of California w.r.t Stemcell research" is expressed by sentences / text snippets and more context is required to judge and identify these answers by the searchers. Therefore, edges provide more support for locating more complex information. This finding was also observed in Table 4.2 as the "interaction pattern" of length 4, denoted by [GCE→B2T→RM→CP] was more frequent as compared with its counterpart [GCN→B2T→RM→CP] for the Complex task and the opposite was true for the simple task.

An interesting takeaway is that edges in knowledge graphs incorporate rich semantic information which can assist the sensemaking activities and provide more context for how different concepts are connected in a domain. Other spatial information representations such as hierarchical structures (e.g. faceted browsers or tables of content) may not be as effective as their edges encode simpler Is-a type relationships.

Overall, applying IE techniques to highlight the mentions of key entities and the relations connecting them in text appears to be beneficial for locating relevant information. This is illustrated in the state diagrams as the probability of starting from RM and ending at CP.

4.5.1 Limitations of the Current Framework

While our work provided an initial foray into designing a knowledge graph extension to the documents retrieval paradigm, it is clear that the current search framework has many limitations in supporting complex and exploratory search tasks. In fact, while our motivation for leveraging knowledge graph representations was to assist sensemaking and exploration activities, these graphs were perceived as more beneficial for simple search tasks. That is, our quantitative analysis of search logs indicated that graphs were more likely to initiate a path to an answer in the simple task than in the complex task.

A closer examination of participants' qualitative comments regarding the utility of graphs and documents for completing the designed search tasks helped with identifying the main shortcomings of knowledge graphs extension in our designed framework. These shortcomings can be categorized as limitations in *extraction* and *representation* of relevant information as well as *interaction* with this information. These shortcomings and candidate solutions to address them are discussed next.

Leveraging External Resources. As we monitored the searchers finding relevant information about a topic they were not very familiar with (e.g., "Stemcell research"), we realized they were making use of the graphs to learn basic facts (e.g., "Stemcells are undifferentiated biological cells") about the salient entities or the query terms. However, since documents mostly lack this basic information the corresponding graphs would not contain such nodes or links either. Therefore, while deriving knowledge graphs from the same set of retrieved documents can lead to generating representations that are relevant to a searcher's query, augmenting the generated graphs with information related to the existing entities can be helpful. These basic facts can be extracted and suggested to the searchers by leveraging external knowledge sets such as DBpedia that contain entity-relationship triples.

Providing Overviews. We identified a major barrier to effective application of automatically generated knowledge graphs to complex search scenarios. As noted by many participants, for the larger graphs, it was not clear where to start and where to go next in the graph. This was the main reason mentioned by the participants who preferred the document text where the nature of the search task was complex. Since the users of exploratory search systems are usually engaged in complex search scenarios it is easy for them to get lost or frustrated in the middle of a search session and just abandon their exploration. It is also very difficult for them to keep track what they have browsed so far and what is there to explore further.

Hence, exploring different approaches for providing overviews of generated knowledge graphs seem to benefit the searchers. These high level overviews of underlying knowledge graphs can be utilized as starting points, while they can also assist the searchers with staying oriented during search sessions.

Enabling Alternative Interaction Paradigms. One of the main challenges for conducting an effective exploratory search is to mitigate the information overload. That is, once these graphs become very large, how can the searcher control the scope of the information they are willing to interact with at any given time.

As identified by many participants, the poor visibility of labels for the large graphs was the main barrier for utilizing the graphs for exploration and information finding. Some of the participants mentioned they would like to see only those parts of the graph that are related to the node they are currently viewing. They found partially visible graphs less confusing to explore. Therefore, presenting graphs with different levels of granularity and enabling enhanced interaction mechanisms with dynamic and on-demand views of these graphs should also be explored.

4.6 Chapter Summary and Next Steps

In this chapter we have reported the results of an initial user study conducted to develop a better understanding of searchers during information finding and analysis activities using a new search framework that displays documents and knowledge graphs side by side. We gained valuable insights by observing different information finding patterns and searchers' exploration within a new search paradigm. We conclude that utilizing graphs of concepts and relationships, which are derived from documents, can be effective for finding relevant information when the information need is well defined. Our findings also demonstrate that providing meaningful relations that explain how different entities of a domain are connected are crucial for supporting more complex search task.

Probing the relatively limited support provided by our framework for complex search tasks, we identified three areas of improvement for the current search framework that is designed around coupling knowledge graphs and documents. The first area of improvements involves efforts in augmenting the *extraction* of informative triples through leveraging a variety of existing knowledge bases such as DBPedia. The second and third areas of improvement focus on *representations* of these extracted triples as well as new *interaction paradigms* to interact with them.

Since the main focus of this dissertation is on providing effective means of *representing* these extracted triples as well as new *interaction paradigms* to explore these entities and relationships, we envision two main directions to pursue in this dissertation in order to extend the current framework and provide more support for complex and exploratory tasks: Providing both global and local views of knowledge graphs as well as enabling alternative interaction paradigms to smoothly transition between these two views.

Our findings, in many ways, highlight the tension between two alternative approaches to search: Overview, filter, detail-on demand [495] versus Expand-from-known [562]. Essentially, knowledge graphs, as low-level representations of entities and relationships in a domain of interest, seem to be beneficial for browsing the immediate context-graph around a specific node of interest and then expand the scope of exploration to other regions of the graph. We observed some instances of expand-from-known searching among our participants who started their exploration from query-nodes or the nodes in the graphs that they were the most familiar with. On the other hand, providing high level overviews of the entire graphs can assist the information seekers with obtaining a visual preview of how salient concepts of a domain are laid out and enable a step-by-step plan for exploration.

The next chapter contrasts two different representations of document information, hierarchies, designed to focus specifically on these high-level overviews, versus knowledge graphs, with the goal of quantifying differences in user behavior, performance, and perception.

Chapter 5

Alternative Representations of Search Results

> *Information exploration should be a joyous experience, but many commentators talk of information overload and anxiety.*
>
> – Wurman, 1989

Chapter 4 reported the results of an initial study on the efficacy of search interfaces that combine textual and knowledge graph representations of the search results in supporting information seeking tasks. The findings highlighted the role of spatial representations in structuring the information that can help locate relevant information when they are coupled with textual content of documents. We also observed how different representations of the same underlying information can lead to different information seeking strategies and how the complexity of the search tasks can bias searchers towards utilizing different components of a representation (e.g. starting from documents text versus interacting with nodes in the knowledge graphs versus examining the edges in the graphs corresponding to semantic relationships). Finally, we identified the main shortcomings of knowledge graphs extension in our designed framework and the tension between two alternative approaches to search: Overview, filter, detail-on demand versus Expand-from-known.

In this chapter we directly contrast two different spatial representations of search results, knowledge graphs (corresponding to the expand-from-known paradigm) and hierarchies (corresponding to the overview-first paradigm), in order to provide a broader understanding of how different ways of structuring the search results can impact information seeking behaviors and outcomes. More specifically, we are interested in identifying the unique characteristics of each representation as they both externalize semantic information from textual content and in structuring them in a 2D graphical structure.

The research described in this chapter is the first mixed methods study of multiple representations of search space, which resulted in identifying their relative strengths and weaknesses in supporting look-up and exploratory information seeking tasks. The findings of this work culminated in designing a novel representation of search results, deemed Hierarchical Knowledge Graphs [475], that enables the user to engage in two alternative navigation paradigms: they can exploit overview layers to explore the collection at a higher level followed by targeted immersion in the detailed view (See Chapter 6).

5.1 Motivation

In the domain of on-line search, the output of current search engines is normally sufficient for many well-defined online tasks, including navigational queries, transactional queries, and many types of informational queries. However, as we note in related work, when information is sought to address broad curiosities, e.g. for learning and other complex mental activities, retrieval is necessary but may not be sufficient [597, 21]. Specifically, there are many open research questions about how to design interfaces to support exploratory search using techniques that organize the retrieved information into meaningful structures. Search results presented by modern search engines are an example of an ordered list sorted by relevance, i.e. a *vectorial model* [340]. However, information seekers often express a desire for a user interface that organizes search results into meaningful groups to help make sense of the results, to infer relationships between concepts, and to help decide what to do next [215, 352]. As a result of this desire for organization, *spatial models* [392], i.e. hierarchies and networks, have also been used to organize information and support sensemaking [340]. While both hierarchies and networks have been shown to be useful in the structuring of content (e.g. [87, 385, 90]) little work has explored the similarities and differences between these two representations. This chapter explores how two specific visualizations of information – Knowledge Graphs (or Knowledge Maps) and Hierarchical Trees – support exploratory search tasks.

In this chapter, we present the quantitative and qualitative results of a study contrasting participants' perspectives on the use of knowledge graphs versus hierarchical trees to support exploration of data for the purpose of developing an answer for informational queries. We describe the design of interfaces and our evaluation of the use of network and hierarchical data structures during exploratory search tasks.

5.2 Representing Search Results for Exploratory Search

Marchionini [352] notes that there are interactive aspects to exploratory search, rather than simply viewing the query satisfaction or information retrieval problem as optimally matching documents to a query. Characteristics of interfaces to support exploratory search, drawn from research in human-computer interaction, include the use of high-level overviews and rapid previews to facilitate sensemaking during the exploratory process. The incorporation of overviews argues for some organization of search results that both presents this overview and allows the user to explore the data and its interconnected relationships more fully through filtering and the examination of user-selected details [495]. We review some of the past efforts on providing representations of search results and their evaluations next and highlight differences to our work.

5.2.1 Existing Search Results Representations

Any system that supports information seeking must structure information to make it accessible. The way information is organized and made available affects the strategies used to access this knowledge and thus information-seeking performance [354, 213, 87].

Given the observation that organization of search results benefits users, one might then ask what organizations of search results exist. In Section 2.4.1 we reviewed the existing techniques for organizing search results (e.g. the taxonomy proposed by Wilson et al. [611]). Fully elaborating on all of the organization techniques or visualizations for search results is beyond the scope of this chapter, and the interested reader is referred to the above taxonomies and Sections 2.4.1 and 2.4.2. However, some visualizations of search result data are specifically salient to our research, in particular, Ltifi et al.'s [340] vectorial model, hierarchies and networks. Ltifi et al [340] proposed a classification of visualization techniques for knowledge discovery including visualizations for linear data (e.g. timelines), multi-dimensional data (e.g.

scatterplots), vectorial models (e.g. relevance-ordered results), hierarchies (trees, tables of contents) and networks (e.g. knowledge graphs). Given the non-numerical characteristic of web search results, the latter three types of visualizations (vectorial, hierarchical, and network) are particularly useful for displaying search results. While the vectorial representation presents results as a ranked list, hierarchical and network representations can be used to display grouping, similarity or relationships among search results.

Trees are a common tool for representing hierarchies. A hierarchical structure is mainly made up of organizational links that organize the information into categories (topics) with no or few cross-links between categories. Google's "Knowledge Graph" enhances basic search results with structured data, essentially presenting a network organization of search results [502]. Google claims the knowledge graph enhances search in three main ways: query disambiguation, a summarization of related facts, and exploratory search suggestions (based on what other users explored next).

Most network visualizations tend to provide a global perspective on a graph by attempting to represent an overview of the information space so no information is missing and the data can speak for itself. Most of these techniques are based on Shneiderman's Visual Information Seeking Mantra [495]: "Overview first, zoom and filter, then details on demand". For example Sanchez and Llamas [468] followed this principle to visualize a large combination of concept maps to distinguish between an interface for the author and an interface for the end user that facilitates the exploration tasks. Some of the common techniques for visualizing large network data is discussed in Section 2.4.2.3.

5.2.2 Evaluating Search Results Representations

Novick and Hurley [392], working in the field of education, performed extensive research on the use of spatial models such as networks, hierarchies, and matrices. In particular, they were interested in the properties of these spatial models that were particularly suited to problem solving. Our work differs in its focus on information retrieval and the representation of search results. As well, our work differs in that Novick and Hurley do not develop interfaces that support problem solving; instead, they use questionnaire data to elicit from participants which representations participants think might best support information representation.

More recently, researchers in information retrieval have performed evaluations of techniques for representing search results, examining both hierarchical structures (e.g. [87, 385, 150]) or networks ([478]). However, these results investigate how different properties of one structure may affect users' behaviour, whereas our work aims at understanding the type of support provided by two inter-related structures for different types of search tasks on the Web.

Other recent work in search results representation focuses on a single visualization (e.g. concept maps) that seeks to represent both hierarchies and networks ([23, 22, 90]) to support information seeking and finding. However, the focus of this research was on comprehension of the representations through a quantitative study. Our focus is on understanding how hierarchical versus network representations support different types of search tasks.

Our past work [478] described in Chapter 4 investigated the effects of combining a knowledge graph with textual documents. Our goal was to understand user behaviour with respect to different search tasks. We argued that a hybrid approach that combines the coherent content of text with the organized structure of graphs may better support information finding and sense making. Our main takeaway from this past work was that utilizing graphs of concepts and relationships which are derived from documents can be effective for finding relevant information when the information need is well defined. These findings also demonstrate that providing meaningful relations that explain how different entities of a domain are connected are crucial for supporting more complex search task. This chapter broadens this work by looking specifically at the contrast between hierarchical representations (e.g. trees) and network representations (e.g. knowledge graphs).

5.3 Application Design

One challenge with any application that presents search results from an exploratory query to users is that the goal is rarely a static representation of the content returned by a user's query. Instead, the goal is to develop an interface that allows a user to interact with the content, to filter and select specific content, essentially to explore the information returned. As a result, the representation is linked to the interface that contains it and supports manipulation and exploration of it [215]. In fact, this observation is one of the main takeaways from our past work reported in Chapter 4: while the way that the information is structured and presented to the user (e.g. a knowledge graph versus a textual document) can impact sensemaking and synthesizing new knowledge, the interaction mechanisms that the search UI is providing can significantly constrain information seeking strategies and experiences.

From Section 5.2, we see that there is a need for interaction with representations of search results, and that, following Ltifi et al. [340], alongside vectorial models of search, hierarchies and networks are also viable representations for knowledge discovery. To develop an interface for exploratory search that would allow us to explore the characteristics of hierarchical and network visualizations, we engaged in an iterative process using a series of walkthoughs, thinkalouds, and pilot studies.

5.3.1 Prototype Development

To develop our representations and interface, we began with a low-fidelity design, where paper prototypes were used to explore user perception of representations of data and user interaction with those representations. We initially designed two low fidelity interfaces. The first interface employed a graph structure in which the entities from each article were the nodes and the sentences describing a semantic relation between them were the edges. In order to investigate how users navigate through large graphs to find information, our knowledge-graph-based prototype was designed such that the user would start from the overview page that contained all the nodes that had a high number of connections to other nodes in the graph. These nodes could be considered as representatives of different components of the graph and would help distribute different sub-graphs into different pages. The user was able to expand any of the nodes on the overview page and would proceed to a new page that contained the selected node and all the nodes that had a link to this node. The user could expand a new node on this page or collapse the expanded node and go back to the previous page.

The second interface utilized a hierarchy (or a tree) structure to organize headings and sub-headings of the articles, as observed in each page's table-of-contents. Each tree was in a collapsed format initially and the user would expand and collapse nodes to drill down into document content. We also created an overview node that linked all the trees in our collection. Interfaces were seeded with data gleaned from Wikipedia pages on Canadian capital cities.

We conducted a thinkaloud study to evaluate our paper prototypes with six participants (two female) to gather a set of features required for these interfaces. Data was presented in both graph and tree form to each participant. We asked participants to think aloud about what the data represented and how they would interact with the data. We also collected qualitative data on different use cases of these interfaces with respect to different search tasks. From this initial study, we redesigned our interfaces.

5.3.2 Final Design

The qualitative data and participants' feedback helped refine both the design of our search result presentation and our interface for manipulating the representation of the search results. We used force and pack layouts (as part of the D3 library[1]) to visualize the graphs and trees respectively.

When the user launches the graph-based application (Figure 5.1, top), they are presented with a knowledge graph containing labelled nodes and unlabelled links between nodes. Nodes that represent entities with low frequency are hidden in the initial view, and only appear once a higher-frequency, connected node is clicked. This ensures that the graph does not become too cluttered. Once the user clicks on a node, that node and all connected nodes are highlighted, while the remainder of the graph is alpha-blended into the background. By hovering over any connected node in highlighted portion of the graph, the user can see the relationship(s) between the two nodes in the snippet window located on the left side of the interface (Figure 5.1. top). For each relationship in the snippet region, participants have a link that allows them to view the corresponding Wikipedia article.

The tree interface is shown in Figure 5.1, bottom. When the user launches the application, the user is presented with a fully expanded tree. By clicking on any node within the tree, that portion of the Wikipedia document corresponding to the node is presented in the preview area at the left of the interface. Under the snippet in question, there is a link to view article, allowing users to access the article in question.

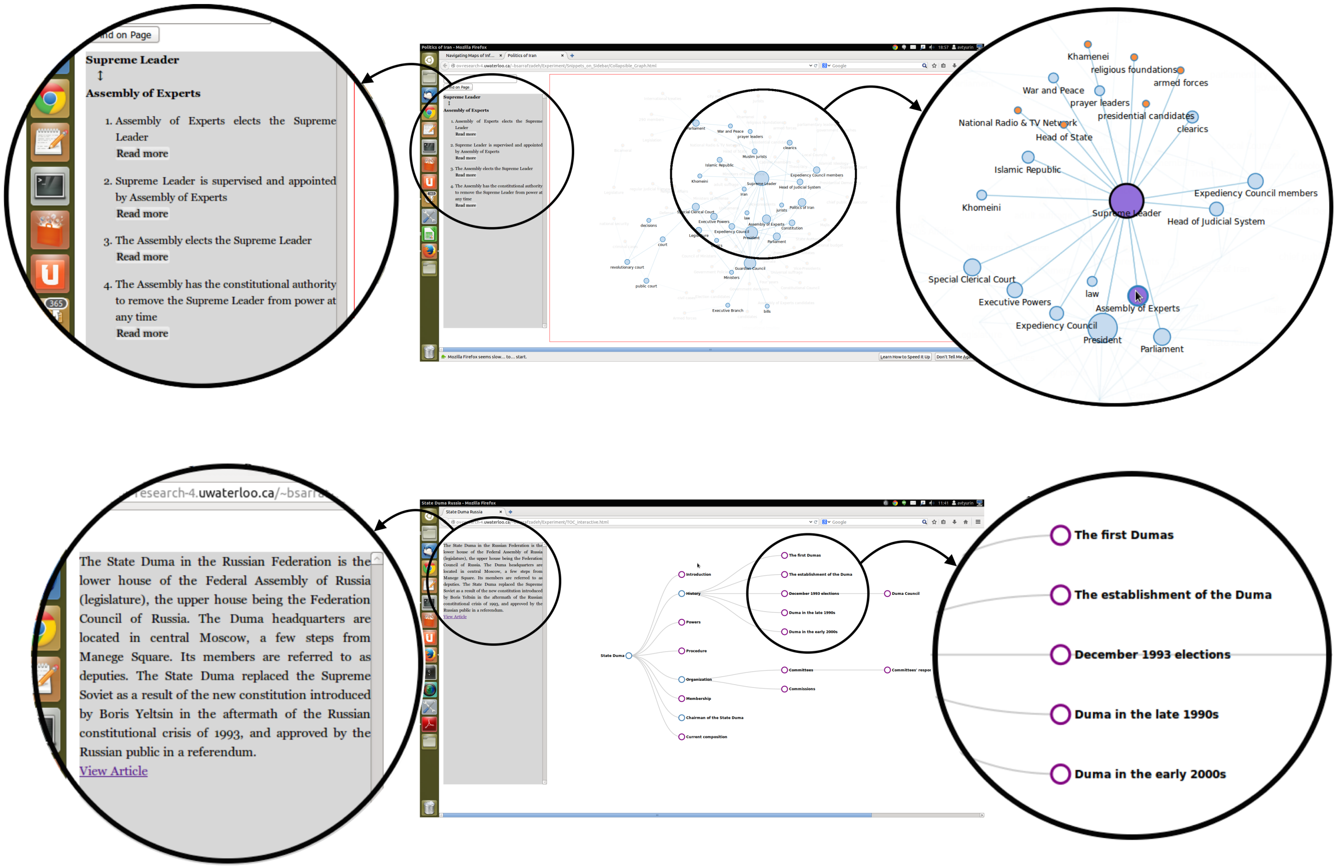
5.4 Experimental Design

To detail our experimental design, we first discuss the data extraction method that we used to populate our interface with data. Next, we present the tasks in our study and describe our participant population. Finally, we describe the data we capture from each participant.

5.4.1 Data Extraction

To populate our interactive applications, we created two distinct data sets: one focusing on history and the second on global politics. For the history data set, we

[1] http://d3js.org/


nd on Page

Supreme Leader
↕
Assembly of Experts

1. Assembly of Experts elects the Supreme Leader
Read more
2. Supreme Leader is supervised and appointed by Assembly of Experts
Read more
3. The Assembly elects the Supreme Leader
Read more
4. The Assembly has the constitutional authority to remove the Supreme Leader from power at any time
Read more

research-4.uwaterloo.ca/~bsarr

The State Duma in the Russian Federation is the lower house of the Federal Assembly of Russia (legislature), the upper house being the Federation Council of Russia. The Duma headquarters are located in central Moscow, a few steps from Manege Square. Its members are referred to as deputies. The State Duma replaced the Supreme Soviet as a result of the new constitution introduced by Boris Yeltsin in the aftermath of the Russian constitutional crisis of 1993, and approved by the Russian public in a referendum.
View Article

Figure 5.1: The graph and tree visualization interfaces. Note the callouts of nodes and document snippets.

used the previous search task exploring former capital cities of Canada. For the politics search task, we created a data set representing governmental structures in Iran and Russia.

To create this data set, we first collected a set of Wikipedia articles by querying the Web using a popular search engine. We retrieved the top 10 articles in Wikipedia based on their relevance to three queries corresponding to three topics: “Former Capital Cities of Canada”, “Political System of Iran” and “Political System of Russia”.

To create our knowledge graph, we employed our Information Retrieval and Extraction (IRE) system that processes a text collection and generates (entity-relation-entity) triples [474]. This module is implemented in four phases. During the first phase we create the input corpus by collecting retrieved documents based on a given query. Next, we extract entities from text using state-of-the-art entity taggers [2]. We then select the sentences that contain at least two entities in them and parse them using Stanford Dependency Parser. For each sentence, we extract meaningful relations between the entities by finding the shortest path in the corresponding parse tree. For example we extract *The Constitutional Act divided the Province of Quebec into Upper and Lower Canada* as a relationship between the entities *Constitutional Act* and *Upper Canada*. We constructed a set of patterns based on dependency triples that lead to semantically meaningful relations. In the final phase we generate labels for the extracted relations and rank them based on relevance to the query and the informativeness of the extraction. Once the knowledge graph is generated, we hand-tune some aspects of the graph by correcting minor errors caused by the extraction of entities and relations. For more details please see Chapter 3.

For the tree based interface, we extracted the Tables of Content (TOCs) embedded in each Wikipedia article. We then manually extended the table-of-contents by adding subheadings to each section in order to provide a richer structure for the trees. Overall, our goal was to create visualizations that could realistically be created by computer algorithms while ensuring equivalent, high-quality for each of the generated visualizations.

5.4.2 Search Tasks

We noted earlier that researchers have defined search queries as simple or complex. With respect to the complexity level, each participant performed one Simple (i.e. question answering) and one Complex (i.e. essay writing) task. We also used two different topics (i.e. History and Politics) to investigate the relation between the topic and content knowledge with the structure used to organize the retrieved information. In addition, for our complex tasks we closely follow Bystroӥ and Hansen's [83] recommendation that three levels of description should be used to specify a search task: a contextual description, a situational description and a topical description. The queries we asked people to find information to satisfy in our study were the following:

[2] https://cogcomp.cs.illinois.edu/page/software_view/NETagger

Simple Politics: What governmental body or bodies are involved in the impeachment of the President of Iran and of Russia?

Complex Politics: Imagine you are a high school student who is going to write an essay on the Political Systems of Iran and Russia. Knowing little about the presidents of these two countries, you wish to determine which president has more power. Find at least 3 arguments to justify your answer."

Simple History: As a result of which act were Upper and Lower Canada formed?

Complex History: Imagine you are a high school student who is going to write an essay on the History of Canada. Knowing little about Canadian History, you wish to know which cities have served as a capital for Canada. You would also like to understand the reasons behind moving the capital from one city to another.

To assess the study design, we piloted with four participants. The pilot ensured that the usability of the system was sufficient to support interaction and provided guidance on the semi-structured interview to collect qualitative data on distinctions and use cases of the designed interfaces.

To limit study length and ensure coverage of simple and complex queries within subjects, our final study design was a 2 × 2 × 2 [interface, interface-topic, topic-complexity] mixed design with interface as a within subjects factor, topic to interface assignment and complexity to topic assignment as between subject factors. This resulted in eight different groupings of participants. Each participant saw both interfaces. In the first interface, they had either politics or history, with the other topic in the second interface. For the two topics, each participant saw a complex query on one topic and a simple query on the other. In order to control for order effects, we rotated the order in which the tasks and the interfaces were assigned to the participants. That is, participants were randomly divided into 8 groups.

5.4.3 Participants

Once the study design was final, we recruited 26 (13 female) participants from different areas of Science, Math and Engineering for this study, all of whom use the Internet on a regular basis to search for information.

5.4.4 Procedure and Data Collection

The study proceeded as follows. After a brief introduction to the study, participants were given an initial questionnaire that evaluated their knowledge of the first query's topic. Participants were then presented with their first interface, were given an introduction to the features of the interface demonstrating how each feature of the interface worked, and were then given some time to manipulate the interface.

Once participants had developed some comfort with the features of the interface (approximately three minutes), participants were given the query and told to manipulate the interface as if they did not know the answer to the query and wished to locate it. To capture data on participants' actions, participants were asked to "think aloud" during each task and share their thoughts and strategies with the researcher. For both tasks, the participants were given 15 minutes and were required to find relevant information by providing a reference sentence or sentences from the interface or document collection to justify their arguments or answers. The need to find specific information ensured that each participant manipulated the interface to find relevant information.

After providing an answer to the query, participants completed a post-task questionnaire that evaluated the experience they just had. We used questionnaires provided by TREC-9 Interactive Searching track [3] and modified them to fit into our experiment design. At the end of each task, via a semi-structured interview participants were asked to reflect on their experience with using the assigned interface for performing the assigned task. They were encouraged to think about the conceptual usability of the type of structure utilized for information organization as well as the technical usability of the application. At the end of the second task a semi-structured interview format was used to elicit comparison between the two interfaces with respect to the different types of search tasks and to reflect on the design of an "ideal" interface that could support them more efficiently. Participants received a $10 incentive for their participation.

Data was captured in a variety of ways. Each interface was instrumented with a logger which monitored participants during the search sessions. Both movement on the computer screen and participants' interactions with the system were captured. Interactions we collected included clicking on nodes or edges, reading snippets, viewing articles, and the time they spent reading the articles. The activity logs for two of the participants were corrupted so we excluded their data from our activity log analysis. Experimental blocks and a post-task semi-structured interview were audio

[3] www-nlpir.nist.gov/projects/t9i/qforms.html

recorded. Finally, two assessors evaluated the quality of answers provided by the participants for each of the search tasks independently. Simple queries were rated as either correct or incorrect. To receive a correct rating, both answer and referenced document section were required to be correct. Complex questions were rated on a scale. Scores for all queries were normalized to reflect a value in the range [0, 1]. Inter-assessor reliability was evaluated using Pearson coefficient and an overall value of 0.8 for simple queries and 0.9 for complex queries was found.

5.5 Main Observations

In this section, we present an analysis of data collected during the study. We first present some numerical data collected from search logs which provides a broad overview of participants' contrasting behaviours given different interfaces and given queries of differing complexity. Next, we present the results of our qualitative analysis, clustered into four broad themes: Biasing Factors, Task Effects, Data Relations, and Problem Solving Approach.

5.5.1 Validating Search Tasks

In any study where the goal is to explore search result representations for exploratory search, one concern is whether or not the search tasks are representative of exploratory search tasks. In our task design, we were guided by Marchionini's work on exploratory search [352]. Leveraging two task domains, politics and history, we created one look-up task and one exploratory search task within each domain using Marchionini's definitions, yielding four tasks overall. The politics tasks asked participants to compare two different governmental structures, Iran and Russia, rationalizing and providing citations for answers they provide. Similarly, the history tasks asked participants to discover something about the history of Canada and, again, rationalize and provide citations for their answers. For our complex tasks, in particular, we argue that the tasks combine aspects of knowledge acquisition or comparison (the *learn* subcategory of exploratory search) with analysis, synthesis, and evaluation (the *investigate* subcategory).

Another concern is whether the actual topics are of sufficiently similar complexity that topic effects do not overwhelm other factors in our results. To address this, beyond ensuring counterbalancing of topics, we analyzed topic effects vis a vis our dependent variables to determine whether either the history or politics task resulted in statistically significantly varying behaviours. Interestingly, our *look-up tasks* in both history and politics, where participants returned a factoid, differed in quality of answers, time reading, and document views ($F_{1,24} = 6.02, p < 0.05$ for quality; $F_{1,24} = 6.00, p < 0.05$ for reading; and $F_{1,24} = 21.22, p < 0.001$ for document views). However, for our *exploratory tasks*, i.e. our complex tasks where participants were asked to learn or investigate, scores did not differ significantly between the two topics areas of history and politics ($p > 0.05$ in all cases). Because our primary interest is supporting exploratory search, we argue that our complex tasks are of sufficiently similar complexity as to limit topic effects.

Finally, alongside care designing our search tasks and a analysis of topic effects on dependent numerical measures, we also examined our qualitative data to determine whether participants found the tasks to be aligned with their conceptualization of exploratory search. The comments made by our participants when they were presented by the tasks descriptions indicated that these tasks were indeed complex, i.e. that they were ambiguous and open ended in nature. As well, different participants interpreted the task descriptions differently and came up with different strategies based upon their interpretation, further validating the open-ended, exploratory nature of the search tasks.

5.5.2 Log Data Analysis

As noted earlier, our data logged all user action with the system. Of particular interest to us was information on the scoring of participant responses, the number of nodes clicked in each interface, the number of documents read, and the amount of time spent reading documents. Table A.1 summarizes this data. The Mark column contains scoring of participant responses. Clicks is a count of the number of nodes clicked on. Views is the number of instances when a participants used the interface to view the actual Wikipedia document (as opposed to relying on the information contained in the interface). Finally, ViewTime is the amount of time in seconds spent reading documents (as opposed to manipulating the interface).

We performed a repeated measures ANOVA with interface (tree versus graph) as a within subject effect and query complexity as a between subjects effect. Dependent variables were scoring of query results, clicks with the interface, number of document views, and time spent viewing documents. Overall, RM-ANOVA indicated that interface had a statistically significant effect on dependent variables ($F_{4,20} = 5.83, p < 0.01, \eta^2 = 0.54$). Query complexity was not significant, nor was there any interaction between complexity and interface. Univariate tests of dependent variables with respect to interface (tree versus graph) show statistically significant effect on number of document views ($F_{1,23} = 26.29, p < 0.001, \eta^2 = 0.53$) and on time spent viewing documents ($F_{1,23} = 6.01, p < 0.05, \eta^2 = 0.21$). Marks and Clicks were not significant.

	Marks	Clicks	Views	ViewTime
Graph	0.74 (0.27)	18.7 (3.2)	1.6 (0.43)	131 (37)
Tree	0.43 (0.04)	17.9 (2.2)	4.9 (0.49)	1228 (444)

Table 5.1: Mean (Standard Deviation) values for marks (average independent evaluator scores), clicks on nodes, document views, and document view time.

.

Overall, our data indicate that the knowledge-graph visualization allows participants to glean more information from the data structure (67% fewer document views, on average) in less time (almost 90% less time reading documents). The knowledge graph is designed to represent the information in the document in a way that obviates the need to read extensively, and it was very successful at accomplishing this. Over half of all participants examined either 0 or 1 documents while using the knowledge graph (mean of 1.6 documents), whereas all except one participant examined at least three documents with the tree structure (mean of 4.9 documents). Qualitatively, we note that the knowledge graph also fared better in score, though not statistically significantly better. As well, the workload in both documents (as measured by node clicks) was very similar (18.7 versus 17.9 clicks per query on average).

5.5.3 Qualitative Data

Given the statistical advantage enjoyed by the graph representation, the next question we wished to explore involved participant perspectives on each of these representations of search results. How did they differ? What were the advantages and disadvantages of each from a user perspective? We present four themes arising from our qualitative data analysis in this section.

5.5.3.1 Biasing Factor: A Willingness to Explore

Exploratory behavior, defined by the National Library of Medicine as “the tendency to explore or investigate a novel environment”, is driven by curiosity and is evident in most exploratory searches. Both lookup and exploratory searches use curiosity in their search models, though the actual curiosity which drives each type of search is slightly different [56]. *Specific curiosity* is the desire for a particular piece of information, as typified by an attempt to solve a problem or puzzle. *Diversive curiosity* is a more general seeking of stimulation or novelty, for example a television viewer flipping between channels. In information seeking, specific curiosity corresponds to well-defined goals and directed searching, while diversive curiosity corresponds to ill-defined goals and exploratory browsing [399].

In our thinkaloud data and in our follow-up interview data, we identified specific versus diversive curiosity as a factor that influenced participants' perceptions of each web interface. Essentially, some participants preferred an interface over the other based on the amount of time they were willing to spend in exploratory browsing. Linking to specific curiosity, if an interface is effective in accomplishing a search task but required extensive time browsing, the participant would rather use a different interface. Participants patience with the search task was influenced by the tension between the drive to solve the problem (specific curiosity) versus the tolerance for browsing (diversive curiosity):

> "*For specific questions, it depends on how much time I'm willing to spend. If I have more time I'd like the tree, because it's more scattered and I can learn more objectively.*" [P4]

> "*So I feel like the Tree would be good if I wanted to sit down and spend time reading about a topic and I wasn't looking for something specific. Whereas if I was looking for something very specific, for that, I think I would like the other one [graph] better. Cause it was already doing the keyword search and it was easier to pick out things.*" [P8]

> "*If I need a fast way, I go to the graph. I use the tree only when I'm learning deep about a new domain.*" [P4]

This observation is in line with the initial work on information foraging; Pirolli and Card [412] defined the profitability of an information source "as the value of information gained per unit cost of processing the source." Cost is defined in terms of time spent, resources utilized and opportunities lost when pursuing a search strategy instead of others [462]. We find that diversive curiosity biases toward the tree structure, whereas specific curiosity biases toward the graph.

5.5.3.2 Task Effects: Finding Versus Learning

The Web has provided the opportunity to browse and navigate through an extensive information space. However, beyond simply finding basic answers, web searchers also engage in learning and discovery [352].

As noted in our study design, we incorporate two tasks with different levels of complexity. Given these two levels of complexity, in post-experiment interviews the participants were able to compare the two interfaces based on the specificity of the information they were looking for. Interestingly, however, participants were divided on which interface was better for simple versus complex search tasks.

Overall, most participants found the graph interface more practical for finding specific information and simple question answering tasks. Both question-answering and keyword-based tasks were typically perceived of as advantaging the graph structure:

> "*For the question-answering task I'd rather use the graph. Because I want to know exactly if this word is linked to that word. If there are two words appear in the same sentence you can quickly find an answer and I don't have to read the whole article.*" [P9]

> "*When I was searching for specific keywords, with the tree interface I actually had to go to the article itself to search. so it wasn't useful. Whereas the other one [graph] actually gives me access to the keywords.*" [P14]

> "*To learn I think the hierarchy interface is good if I want to learn say about history of Canada, because then you start from step 1 and the you go to the next level."* [P10]

This is not to say that our participants were universal in their beliefs about data visualizations. Some participants found that the tree was significantly better for finding a specific piece of information. P2, P3, and P6 all articulated variants of this belief:

> "*But when you are trying to find a specific answer to a question, then the tree structure is good, because it helps you traverse from the root to a leaf node.*" [P2]

> "*I like the tree better for specific questions. It categorizes things better.*"[P3]

> "*Tree is pretty good for finding exact information.*" [P6]

To try to understand this phenomenon better, we looked at other demographic data collected from our participants, and found an intersection between the belief that a tree was better suited to search tasks and our participants self-rated prior knowledge of the topic being examined. Participants were biased toward a tree structure for broad learning of the task domain particularly when they had low prior knowledge. This result seems to replicate findings by Amadieu et al. [23] on the use of network structures versus hierarchical structures in the education domain, i.e., that low knowledge learners benefited from hierarchical structures in free recall performance and exhibited reduced disorientation, whereas high knowledge learners performed better and followed a more coherent reading sequence given a network structure. Participants, too, noted this phenomenon:

> "*So if you are an expert in a domain, you want the view very focused [knowledge graph]. But if you don't know much about a domain, you want to see an overview first [tree].*" [P6]

5.5.3.3 Data Relations: Derivative Versus Multifaceted, Local Versus Global

Visualizations of structures, i.e. of entities and relationships, inherent in large data sets can help users understand the structure of the data and make information more accessible. However, participants may perceive a domain to have a derivative/hierarchical structure or a multi-faceted structure. If the representation of search results mimics that perceived structure, participants prefer that structure:

> "*If you are searching for something that is already structured and we already know the names of these categories, then the tree is helpful.*" [P2]

> "*For the [tree] interface, if I was using it for a topic like Geography, then I'm looking for continents, countries, cities, states, capitals, ... Then I know the headings and then I know which path to take.*"[P3]

> “*I think [tree] would also be useful if you had some sort of notion of how things are laid out, like if there was a chronological order. Yeah, if that was chronological that would be nice cause you could gauge where you needed to click. Like you saw something that was really-really previous and two nodes down to find something more recent, even if you don’t know exactly which one.*”[P8]

To clarify, the data relations in question are those perceived to be salient *by the participant*. If salient relationships are viewed as derivative or hierarchical (e.g. ‘is-a’ relationships), then a tree can best capture this view of data, whereas if salient relationships are more heterogeneous and resist structure as a hierarchy, that advantages the graph-based representation.

Beyond the specific relationships between entities, another theme that appeared in our data involved the scope of information required to satisfy a query. The tree structure seems to provide a comprehensive overview for the information space. Even if we provide groupings and overviews for our graphs, the graph interface best serves exploration at the entity-relationship level. As a result, several participants liked the tree structure for cases where they needed a comprehensive overview of the domain:

> “*If I’m learning about a new domain, in the case that I want to cover the entire domain and get a general understanding of everything but at the surface, I’d like the tree.*” [P7]

5.5.3.4 Problem Solving Approach: Depth-First Versus Breadth-First

According to Brown [77] information seeking is a goal-driven activity in which needs are satisfied through problem solving. This view is comparable to Wilson’s model of information seeking [612], which considers information seeking as a problem-solving process with the goal of reducing uncertainty about the information being searched.

> “*... if you have a large amount of data then you’re kind of confused, you don’t know which part to look at, which connection to look at. ... I would use hierarchy to get an idea of how everything is organized and then maybe I go and try to dig in*

> *more, find out the relations between terms.*
> *- For digging more would you use the graph?*
> *- Yes. But again, even in the graph, it should be the specific, the focused one. Not the whole thing.*" [P21]

In unpacking this quote, we note that the process of directly addressing a problem is essentially a depth-first process where the knowledge graph allows a focused exploration of a region. On the other hand, with confusion the breadth-first or tree-based exploration is beneficial as it allows the user to iteratively reduce confusion, obtain an overview, and only slowly exploit detail. Many participants indicated similar concepts of confusion, nervousness, or inadequacy as a rationale for their preference for the tree structure:

> "*[The graph interface] is not friendly. Too many things!*" [P2]

> "*Because I get frustrated jumping from one node to the other for a while and don't get any information I want. ... If the graph is too big, I get scared of it! ... Too many things, so I don't know where to look*" [P26]

> "*when the graph is too big, I don't know where to look ... and I don't also know where to start. Because I'm not familiar with most of the information .. the Councils, the positions, the names, ... So I don't know where to start.*" [P26]

More generally, many research domains argue for overview and structuring of content to permit sense-making and reduce confusion. Information visualization is founded on the techniques people use to structure and cluster visual stimuli (see, for example, [583], or Section 2.4.2). Problem solving research in psychology connects aspects of visual perception and structuring of content to comprehension [432]. Designing for visually impaired readers argues for well-structured hierarchical content to allow more rapid sense-making [221], even in the absence of vision. Essentially, overviews are invaluable when people feel a need to orient themselves within data.

Alongside confusion and the need to orient ones-self within a domain, one's problem-solving strategies may bias behaviour. Research into problem solving strategies has a long history in the psychology domain. One well-known characterization of *coping* or problem-solving strategies identifies two groups of individuals: Problem-focused and Emotion-focused [319]. Problem-focused individuals tend to directly address a problem, whereas those with an emotion-focused strategy seek to reduce the effects of the problem. In web search, Kim [290] found that problem-solving style had some impact on navigational patterns. Emotion-focused subjects traversed several layers of nodes before returning to the starting page (i.e. a depth-first navigation), whereas problem-focused subjects spent more time checking nodes available in the same level (i.e. a breadth-first navigation). Acknowledging the lack of personality-testing in our questionnaires [290], the link between confusion, nervousness, or fear and a desire for a hierarchical structure that allows depth-first exploration may merit further inquiry.

5.6 Discussion

5.6.1 Understanding Tree Versus Graph Visualizations

As we note in the motivation for this chapter, our goal with this research was to explore the differences between graph and tree visualizations, specifically to understand their similarities and differences with respect to the search process. Our results explore these differences, triangulating both quantitative data from log files and qualitative data from participant interviews to understand how search result representations influence search behaviour.

From our log data, we note that the hierarchical structures in our study serve as pointers to passages in a document due to their similarity to tables-of-contents. Essentially, they simplify the process of locating topics, but the monotonic relationship that they represent – for example an is-a relationship – limits the information they can represent. The end result is that hierarchies result in a greater need to read the document rather than find the information contained within the visualization, shown, in our log data, by more instances of reading documents, and a longer period of time reading documents. Specifically, participants read documents three times more frequently and spent almost ten times more time reading. Our data also highlights the advantages and disadvantages of gleaning information from a knowledge-graph versus finding the relationship within source material and generating an abstract version of the knowledge graph for oneself. However, it is also clear that one representation is not better than the other in any subjective sense. Many of our participants expressed a need for combining both interfaces into one interface which enables switching between a global and a local view of the information space.

5.6.2 Design Implications

When designing search interfaces, the process of creating a view of search results remains challenging. Information visualization tools such as the InfoVis Toolkit[4], SpotFire[5] and InfoZoom[6] typically support multiple representations of search results. Our work does not dispute the accepted practice of recognizing that heterogeneous, interactive visualizations are the best way to allow exploration of a data set generated by search queries.

Our study highlights the complementary nature of hierarchical structures and knowledge graphs as representations of data. Our data indicates that hierarchies allow a more gradual depiction of and immersion into the domain, essentially fostering sense-making of the overall content (see Section 5.5.3.3). On the other hand, participants note that graphs are "more engaging" (P4), yield "more control over exploration" (P8), or are "similar to my mind" (P16). This, then begs the question of whether hierarchies and knowledge-graphs could be combined, but the challenge with combining hierarchical structures with knowledge-graphs is that hierarchies represent topics within a corpus, whereas nodes in a knowledge graph represent entities and their relations. Any one entity in a knowledge graph can map onto several topics in a hierarchy: For example, a political figure or governmental structure (e.g. the Guardian Council) or a historical event (e.g. the War of 1812) may be mentioned in all topics in a hierarchy, depending on how pervasive that entity is to the overall corpus.

[4] http://sourceforge.net/projects/ivtk/

[5] http://spotfire.tibco.com/

[6] http://www.softlakesolutions.com/

On the other hand, within our knowledge graphs, nodes have different prominence based upon the number of edges they connect to. Entities that are more pervasive in the document have more connections and, hence, can be assumed to be more central to the topic. One alternative to hierarchical structures is to consider central entities within a knowledge graph, those entities that have a higher connectivity to the graph. By setting thresholds, one might be able to structure a multi-level view of a knowledge graph around central entities. In Chapter 6 we are exploring this option as one way to effectively combine the advantages of knowledge graphs and hierarchies into a single view. Rather than breaking down topics or concepts as in our tree view or in concept maps, the multi-level view of knowledge graphs focusing on central entities simply introduces information seekers to those entities or objects most central to a retrieved corpus.

5.6.3 Limitations

In designing any study, compromises must be made. In this section, we discuss three potential confounds: interface effects versus information representations; the effect that hand-tuning may have had on results; and the generalizability of our results given corpus size and topic/task selection. In this section, we address each of these concerns.

Any time one conducts a user study comparing two artifacts, it is always possible to bias the study through selective design. A poor user interface or poor interaction design can disadvantage one experimental option, leading to biased results. To limit this confound, we conducted multiple rounds of pilot studies and made modification to ensure that each representation was sufficiently rich that participants could perform a significant portion of the information seeking task within the visualization. In analyzing our data, we found that participants in our study data indicated no dissatisfaction with the interaction within the visualizations, and, instead, focused on the visualizations themselves. Even on probing during de-briefing interviews, participants would frequently discuss the advantages and disadvantages of knowledge representations (hierarchies versus graphs) when asked to comment on each interface.

A second concern revolves around the ecological validity of our results, particularly in light of hand tuning. As we noted in our experimental design section, we used automated algorithms to generate knowledge graphs [474] and extracted hierarchies from tables-of-contents or headings within documents. However, we then performed some refinement of the hierarchies (adding low-level sectioning to documents) and knowledge graphs (mainly refinement of coreferencing). We address this point in two ways. First, arguably, to ensure that confounds are *not* present in our results, hand-tuning (or at least manual verification) is essential. Otherwise, error-prone algorithms and poorly structured data could influence the effectiveness of any individual representation of search results, focusing the data around the algorithmic failures as opposed to the nature of hierarchies versus graphs (we revisit error prone representations and investigate their impact on information seeking behaviors and outcomes in Chapter 7.) Second, it is important to note that the manual refinement we performed was very limited. In hierarchies, we created a richer set of leaf nodes, but did not modify the overall structured content of the document; in knowledge-graphs, a small set of entities (less than 10%) needed to be combined when coreference resolution failed. As research in automatic summarization and coreference resolution continues, these problems will hopefully be addressed by researchers working in natural language processing.

Finally, task and corpus has been a concern in past iterations of this paper. Our tasks and topic effects are discussed in Section 5.5.1. We argue that the 10 most relevant documents from Wikipedia represents a set of documents similar to the number explored in real-world web searching tasks. First, while web searches return more results, work on information seeking argues that the effective size of a *relevant* document set for web search results is significantly smaller that all documents returned – on the order of six documents – hence the importance of ranking algorithms in information retrieval [279, 252]. Second, not every retrieved document is directly relevant to any specific information seeking task. A user may look within any individual document from a set of top ranked documents, and he or she may also combine information from multiple sources to satisfy his or her information needs.

5.7 Chapter Summary

In this chapter, we presented the results of a study evaluating knowledge graphs and trees as spatial representations of Web search results. Our analysis includes both log data gleaned from participant interactions with data representations and qualitative interview data gleaned from thinkalouds and semi-structured interviews. Overall, we find that knowledge graphs are effective in capturing the entities and relationships in a corpus in a way that reduces participant reliance on actual retrieved documents, i.e. participants viewed significantly fewer documents for significantly less time. As well, the quality of participant responses to pre-specified queries (a measure of how effective visualizations are at representing data) was statistically unaffected by representation. Finally, from the perspective of our participants, we find that tree-based representations are better suited to learning, provide better overviews of a domain, and are more approachable for participants who are confused. Graphs, in contrast, work best for directly seeking answers, and appear to be a more playful mechanism for exploring the details of individual entities and their relationships.

The next Chapter explores our efforts in combining knowledge graph and hierarchical representations of search results into a unified structure where strengths of each representation is preserved and the shortcomings are minimized.

Chapter 6

Hierarchical Knowledge Graphs

The art of information system design (which, I am certain, has a long future) is to find the form and timing of information presentation which will best aid the system user in whatever task he has in hand.

– Oddy, Information Retrieval via Man-Machine Dialogue, 1977

In information retrieval and information visualization, hierarchies are a common tool to structure information into topics or facets, and network visualizations such as knowledge graphs link related concepts within a domain. The research described in Chapter 5 demonstrated the complementary advantages of these two spatial representations of search results in supporting complex and exploratory search tasks. While these findings provided many interesting insights about how each of these representations is perceived and utilized by the information seekers for search tasks of varying complexity, one open question remains: can we design a representation that seamlessly merges these two representations?

In this chapter, we explore a multi-layer extension to knowledge graphs, hierarchical knowledge graphs (HKGs), that combines hierarchical and network visualizations into a unified data representation. Through interaction logs, we show that HKGs preserve the benefits of single-layer knowledge graphs at conveying domain knowledge while incorporating the sense-making advantages of hierarchies for knowledge seeking tasks.

Specially, this chapter describes our algorithm to construct these visualizations, analyzes interaction logs to quantitatively demonstrate performance parity with networks and performance advantages over hierarchies, and synthesizes data from interaction logs, interviews, and thinkalouds on a testbed data set to demonstrate the utility of the unified hierarchy+network structure in our HKGs.

6.1 Motivation

As noted earlier in this thesis, there are two predominant paradigms for finding information on the Web: Searching (i.e, Search by query) and Browsing (i.e, Search by Navigation) [397, 257]. While current search engines, following a "search by query" paradigm, are generally sufficient when the information need is well-defined in the searcher's mind, examining search results remains a necessary step within a larger information seeking process [353, 307]. To elaborate, Searching requires the user to translate an information need into queries, while Browsing accommodates the knowledge gap between what the user is able to communicate and what the system requires to find the desired information. This knowledge gap (also formalized as an 'anomalous state of knowledge' by Belkin [50]) is more evident when information is sought to address broad curiosities, for learning and other complex mental activities [597, 31].

In Chapter 5, this thesis explored the relative benefits of hierarchies and networks and noted that the benefits are largely complementary: hierarchies provide users with some understanding of central topics, allowing them to develop a better overview of information; whereas networks allow people to glean concrete information from the representation rather than needing to extensively read individual documents [479]. Given the complementary advantages of knowledge graphs and hierarchies, our main research question in this work is that whether we can algorithmically generate a seamless data structure that combines the advantages of both hierarchies and networks into a single unified structure.

Given our main goal of designing search interfaces that can assist searchers with learning and investigating tasks, we are interested in solutions that support sensemaking, which is central to all exploratory search activities. Sensemaking is the process of searching for a representation and encoding data in that representation to answer task-specific questions [462]. Russell et al. [462] present four main cognitive stages that are involved in sensemaking: an interaction between a bottom-up 'Search for representation' phase and a top-down representation instantiation phase, a phase of shifting representations to fit newly discovered data into the current representation, and an application of the representation to the user's specific task. The top-level representation that results from sensemaking has many different names: a "holistic cognitive structure" [592], a mind-map, a concept map [391]. Regardless, the construction of this representation must be either provided explicitly by the interface or constructed implicitly by the user [462, 592] before the user can fully "make sense" of the retrieved information.

As a first step towards providing representations that are effective in supporting sensemaking activities, Chapter 2 reviewed the available literature on search user interfaces in order to gain a better understanding of strengths and weaknesses of different spatial representations of search results and whether there have been any efforts to combine multiple structures into one unified representation that preserves the strengths of underlying structures while mitigating the weaknesses. Because of the importance of structure in search, there have been efforts to contrast strengths and weaknesses of different spatial representations and groupings of search results. In Chapter 5 we reviewed the past efforts in evaluating the efficacy of different representations and identified different shortcomings of network representations. Essentially, a main drawback to network structures is that it is hard to both get an overview of an information space and to navigate through the network effectively.

Given our observation of the complementary benefits of hierarchies and networks, one remaining question is whether studies have examined the use of hierarchies and networks as combined – synchronized and simultaneous – representations of search results. We have found little work that explores combined networks+hierarchies. Part of the challenge may arise from the complexity of seamlessly integrating both hierarchies and networks into a single unified structure. For example, hierarchies are typically best when the structure aligns well with the user's task, but, given this alignment, entities in networks may have many multiple 'parents' within the structure, yielding a many-to-many relationship within the hierarchy, i.e. a three-dimensional graph.

In this chapter, we evaluate the efficacy of hierarchical knowledge graphs (HKGs) as a combined representation of low-level entity relationships and high-level central concepts. We generate these knowledge graphs automatically using a simple parsing algorithm [474], then extract hierarchies using a dynamic thresholding approach. We evaluate these HKGs using a mixed methods approach. Quantitative data argues that HKGs preserve the transparency advantages of knowledge graphs and structural advantages of hierarchies. Qualitative data triangulates with quantitative observations and provides additional insight into the advantages and disadvantages of both hierarchical and network visualizations.

6.2 Hierarchical Knowledge Graphs

In this section, we describe hierarchical knowledge graphs, an extension of knowledge graphs that include hierarchical information about the lower level graphical structures. The rationale behind our proposed approach for employing hierarchical knowledge graphs to represent search results is the complementary benefits [479] of hierarchies [215, 619] and network structures [391, 31, 478] to support exploratory browsing of search results. More specifically, hierarchies provide a breadth-first exploration of the information that allows the user to iteratively reduce confusion, obtain an overview, and slowly exploit detail. They thus provide a structured way to navigate from more general concepts to more fine grained data and are valuable when people feel a need to orient themselves. In contrast network structures allow users to glean more information from the representation (document reading time is reduced), are more engaging, yield more control over exploration at the lower level of inter-concept relationships [479], and are more similar to one's mental model [103, 479, 387, 33].

Given the complimentary benefit of networks and hierarchies, the next question is how to design a representation that can seamlessly merge these two representations. We take the approach that a knowledge graph will be an appropriate low-level representation and seek to incorporate a hierarchical view of this low-level representation of corpus content. To incorporate a hierarchical view into a knowledge graph, we need to find answers to the following three design questions (DQs):

1. How do we integrate network and hierarchical views into a single,

seamless data structure?

2. How can both the global and the local view of a knowledge graph be co-visualized?
3. How can transitions between views be designed to maximize visualization stability?

To answer these DQs, we first focus on DQ1 and describe the design of our data structure. Next, to address DQ2 and DQ3 we describe an interface that supports interaction with the data structure. Alongside our DQs, we add one additional constraint to our design. We want to ensure that both the low-level knowledge graph and the hierarchies gleaned from that knowledge graph can be automatically generated from a targeted search performed by the user.

6.2.1 Visualization Design and Creation

As noted above, given that we take the approach that a knowledge graph will constitute the lower-level visualization of our data, the task becomes creating a knowledge graph and creating a hierarchy that is gleaned from and corresponds directly to the underlying knowledge graph.

Figure 6.1 depicts the system architecture that supports the process of automatically generating the hierarchical knowledge graph representation. To simplify hierarchy generation, we create a 3-level hierarchy for any document corpus. Beyond the base layer knowledge graph, there is an intermediate layer of central concepts gleaned from the knowledge graph. Finally, at the top-level, the documents, themselves, represent the top level of the hierarchical knowledge graphs. In Figure 6.1, three main steps are depicted to generate hierarchical knowledge graphs: Document Retrieval (yielding the top-level of the hierarchy – corresponding to the *Collection View*), Knowledge Graph Generation (yielding the bottom level of the hierarchy – corresponding to the *Detailed View*), and Hierarchy-from-graph Generation (yielding an intermediate view of an individual knowledge graph, which we dub a *minimap*[1]).

[1] The term minimap is drawn from the gaming literature. It represents a less detailed overview of a gaming world, allowing the user to orient themselves.

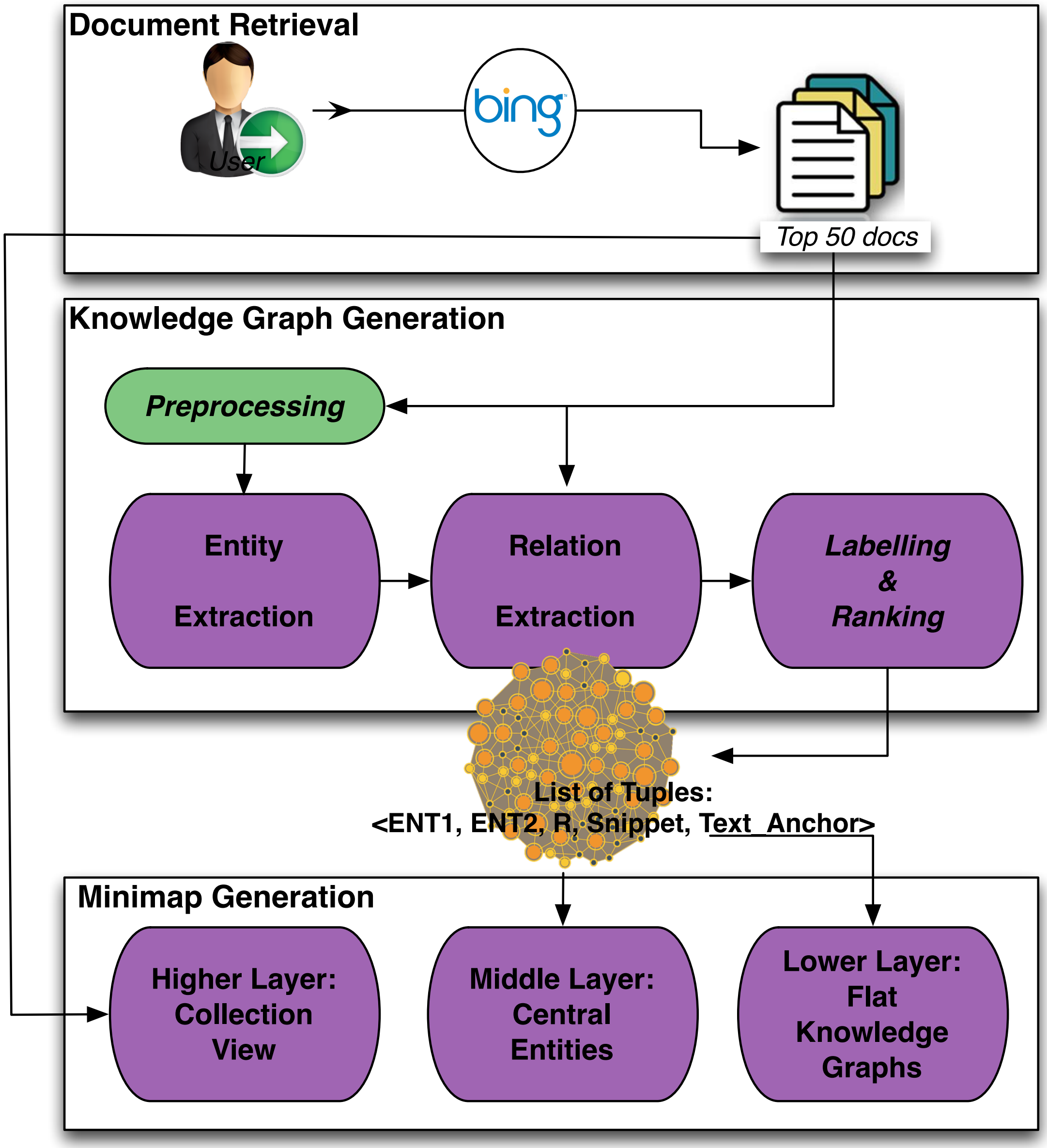
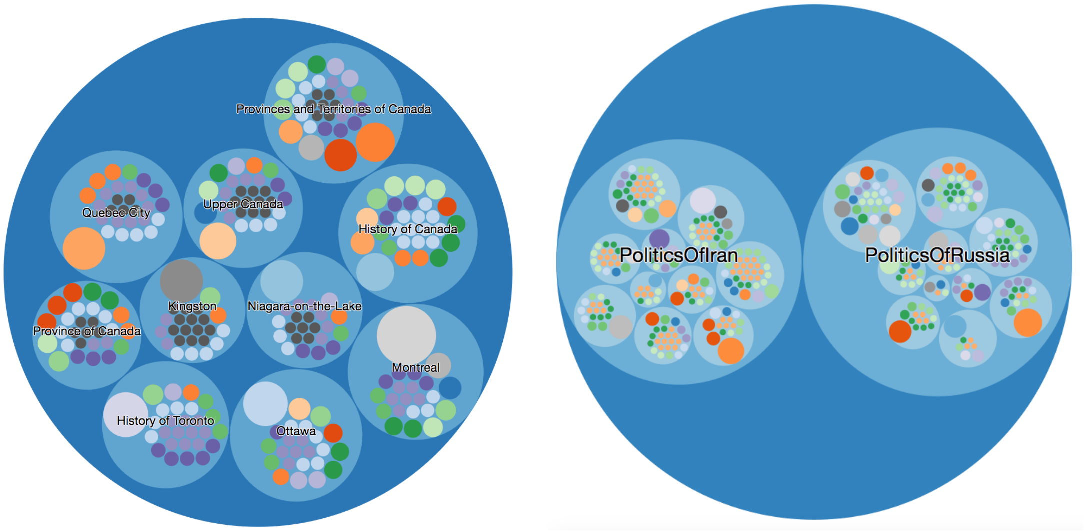

Document Retrieval
User
bing
Top 50 docs
Knowledge Graph Generation
Preprocessing
Entity Extraction
Relation Extraction
Labelling & Ranking
List of Tuples:
<ENT1, ENT2, R, Snippet, Text_Anchor>
Minimap Generation
Higher Layer: Collection View
Middle Layer: Central Entities
Lower Layer: Flat Knowledge Graphs

Figure 6.1: Generating Hierarchical Knowledge Graphs

6.2.1.1 Document Retrieval.

The Document Retrieval component aims at creating an initial document collection based on a user's query. This collection will then be used as an input for the Knowledge Graph Generation component and will represent the top view of the target hierarchy.



(a) Collection View for the History Topic (b) Collection View for the Politics Topic

Figure 6.2: Collection View generated for two different queries.

To generate a document corpus, we configured our IRE system to use the Bing Search engine to retrieve the top n documents for a query while attempting to ensure a reasonable quality of information in the retrieved documents. By default, to ensure that retrieved documents are consistent in their credibility and coverage, we specify Wikipedia as the target domain. Furthermore, because it is known that searchers typically view only a few results [248] and rarely stray past the first page of results [49], we selected n=10 documents to generate collections. The target domain from which to glean documents (e.g. a user might specify WebMD for medical documents, 'gov' for public policy documents, 'bbc' for news) and the size of the initial collection can be specified by the user at the time of query submission.

Finally, since most exploratory search tasks require multiple queries to retrieve documents for different aspects of the information need, this component assigns one partition per query so the user can narrow down the retrieved collection further. For example, considering our History and Politics topics Figure 6.2 demonstrates the Collection Views generated for the set of retrieved documents. While the "Politics of Iran and Russia" query resulted in two partitions for the retrieved documents, i.e. PoliticsOfIran and PoliticsOfRussia (Figure 6.2b), the "History of Canada" query resulted in one partition only (Figure 6.2a). For any query with multiple partitions, the searcher can drill down to any of the partitions which alters the view to a *Partition View*. In this view 10 bigger bubbles corresponding to 10 documents retrieved for that partition are shown. Figure 6.3 demonstrates the PoliticsOfIran partition view for the Politics query. Please note that for queries with one partition the Collection View and the Partition View are identical.

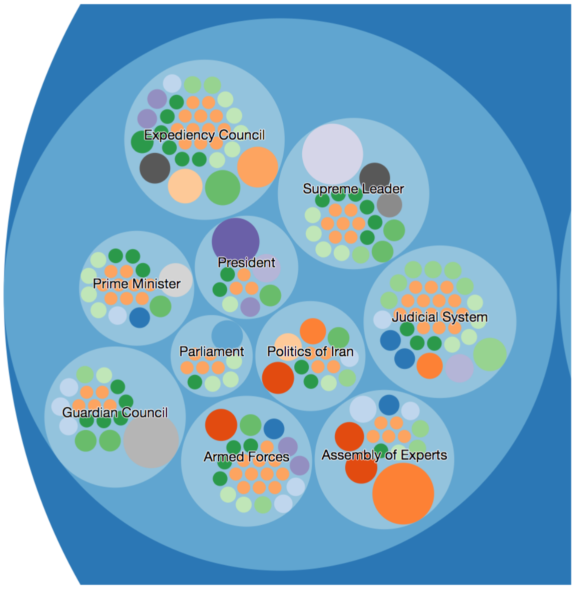


Figure 6.3: PoliticsofIran Partition View for the Politics query.

6.2.1.2 Knowledge Graph Generation.

Leveraging our developed IRE system, we create knowledge graphs for an individual document or set of documents as follows. (1) Entity taggers [2] are used to extract entities from text. (2) Sentences that contain at least two entities are selected and parsed using the Stanford Dependency Parser. For each sentence, we extract meaningful relations between the

[2] https://cogcomp.cs.illinois.edu/page/software_view/NETagger

entities by finding the shortest path in the corresponding parse tree. (3) Finally, labels are automatically generated for the extracted relations. The labeled relations are ranked based on relevance to the query and the informativeness of the extraction [477]. See Chapter 3 for more details.

The outcome is a set of tuples in the form of <entity1, entity2, relation, snippet, document_anchor>. These tuples collectively correspond to a knowledge graph representation of retrieved documents where *entity* is usually a term or a noun phrase in text that corresponds to a concept in the domain, *relation* corresponds to a simplified sentence that is semantically complete and describes how entity1 and entity2 are connected, *snippet* is a short portion of text from which the corresponding entity pair and the relationship is derived, and *text_anchor* is an HTML anchor that links the extracted tuple to the corresponding portion of the source document in the collection. For example, from a paragraph on powers and responsibilities of a president the following tuple can be extracted: <president, parliament, "President nominates the Cabinet members to the Parliament", snippet, [URL][anchor]>.

These tuples are visualized as a knowledge graph where nodes are the entities and edges are the relationships between them. This visualization constitutes the lowest layer of the hierarchy and provides a *Detailed View* of the search space.

6.2.1.3 Minimap Generation.

The final component of this system generates a hierarchical representation of the search results by extracting a middle layer from the input Knowledge graph tuples and provides bidirectional mappings between all three layers. As noted earlier, we call this layer the minimap layer.

A natural result of the entity-relationship tuples extracted above is that some entities have a higher number of edges, i.e. are of higher degree. A higher edge count implies a larger number of connections to other entities in the graph; in other words, those entities with higher edge counts were more frequently linked with other entities in the document. We call these higher degree vertices *central concepts* and hypothesize that one alternative to hierarchical faceted structures is to consider a multi-level view of a knowledge graph around central concepts. The multilevel view focusing on central concepts simply introduces information seekers to those

Algorithm 1 Extracting Central Concepts

Require: *Nodes*: array of nodes in the knowledge graph, *min_degree*: a prespecified threshhold for the minimum degree of node to be considered as a central concept (starting value = 3), *max_count*: an experimentally derived threshold for the maximum number of Central Concepts to be included in the middle layer (default value = 15).

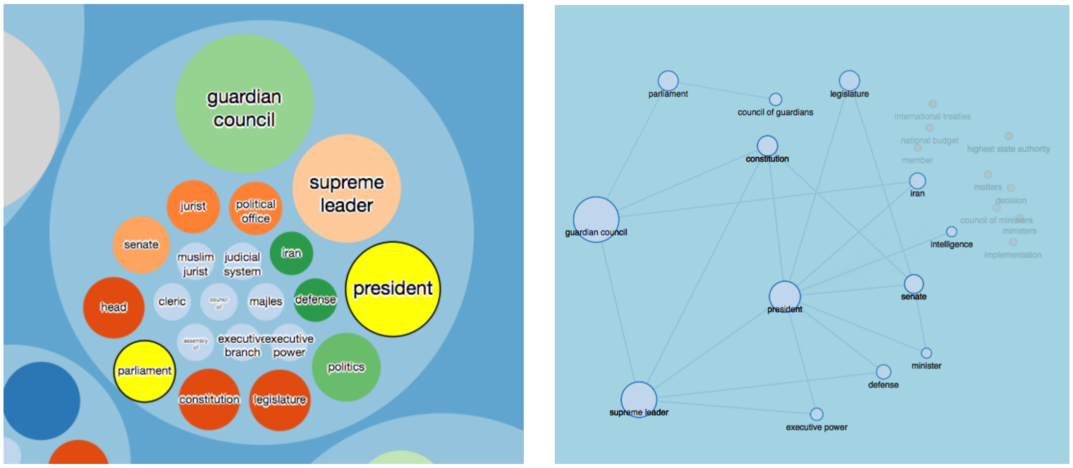
```
function ExtractCC($Nodes, min\_degree, max\_count$)
    while true do
        CentralNodes $\leftarrow$ []
        for all $node$ in $Nodes$ do
            if $node.degree \geq min\_degree$ then
                CentralNodes.add($node$)
        if $CentralNodes.size() \leq max\_count$ then
            return $CentralNodes$
        $min\_degree{+}{+}$
```

entities or objects that are most frequently linked to other entities within the corpus. Generating the hierarchy becomes a thresholding task to appropriately scope the intermediate level of the visualization. Algorithm 1 describes this process more formally.

6.2.2 Prototype Development

Given our hierarchical representation (DQ1), we must support mechanisms for viewing and interacting with the visualization (DQ2 and DQ3). In information retrieval, it is difficult to separate any visualization for representing search results from the interface that contains that visualization [215]. We iteratively designed an interface to support navigation of our hierarchical knowledge graphs via a series of pilot studies.

Based on established literature and pilot studies we found that knowledge graphs can become overwhelming or confusing for participants [119, 299, 397, 479]. The overwhelming nature of the full knowledge graph leads to a need to create filtered views of our graph. These filtered views draw inspiration from the "expand-from-known" paradigm in information visualization [562]. Specifically, at the top level of the full corpus, a user selects a document, then a central concept from the minimap visu-



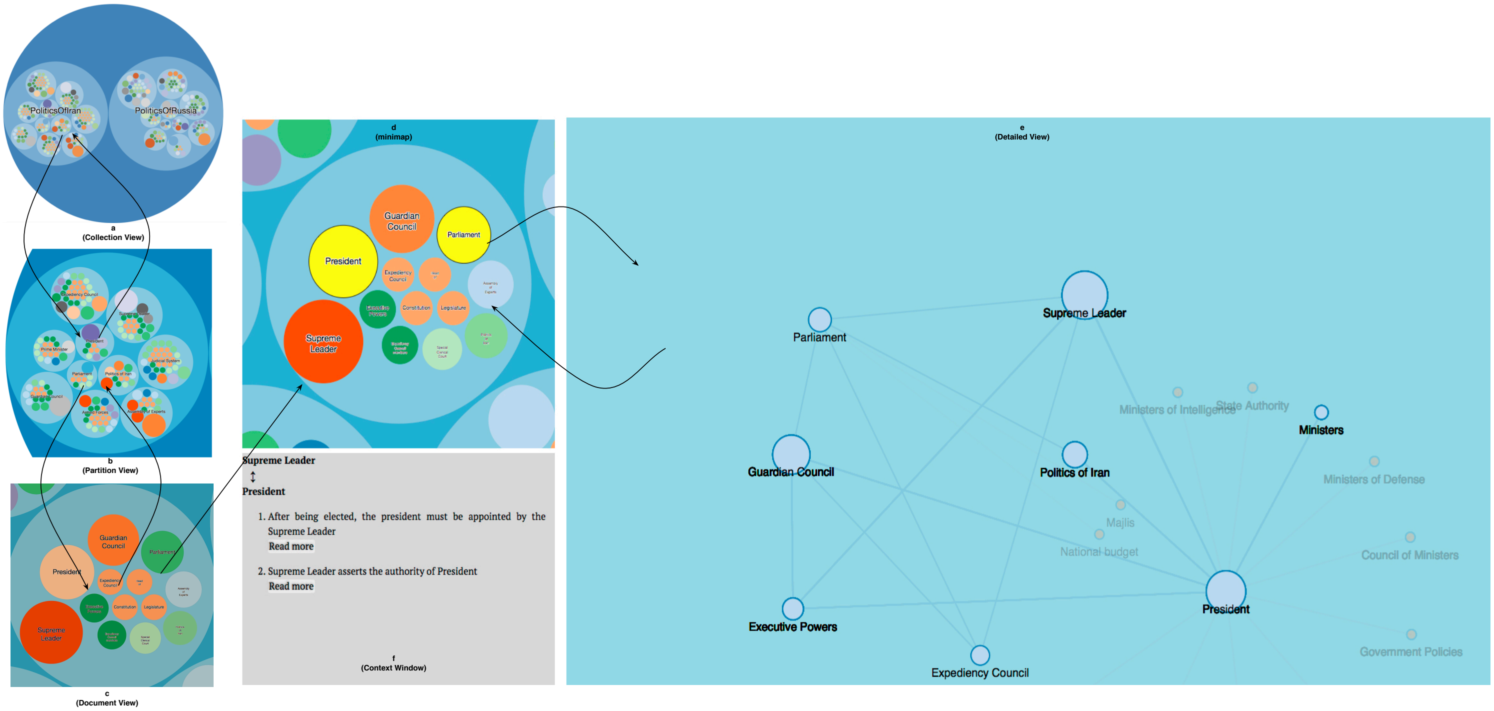
(a) Minimap View

(b) Detailed View

Figure 6.4: A Minimap is generated from the central concepts in the Knowledge Graph. The user can select a subset of central concepts from the Minimap (highlighted in yellow) which will be added to the Detailed View.

alization. While preserving the entire knowledge graph, we alpha-blend all nodes in the knowledge graph *except* those nodes directly related to the central concept from the minimap. Recall that the central concept is simply a high-degree vertex from the knowledge graph; therefore, the central concept and all its linked nodes are shown saturated. As a result, users can identify the central concept, linked entities, and can see closely related additional entities. Together, this focused detailed view seems to effectively support expand-from-known at the knowledge graph level. Figure 6.4 demonstrates a detailed view of the underlying knowledge graph generated based on the central concepts selected in the provided Minimap.

As well, for the Hierarchical View, the biggest challenge to address was the disorientation among the participants during transitions between collection, minimap, and knowledge graph views, a common problem in interfaces that show multiple levels of abstraction. To address this disorientation (DQ3), we maintained the connection between the hierarchical view and the graph view in two ways. First, the user can move between the layers of Collection View and the Document View smoothly through a zooming functionality that changes the focus of the UI (see Figure 6.5).



Supreme Leader
↕
President

1. After being elected, the president must be appointed by the Supreme Leader
   Read more
2. Supreme Leader asserts the authority of President
   Read more

Figure 6.5: MultiLayer Graph Interface: (a) Collection View; (b) Partition View; (c) Document View; (d) Minimap (i.e. Global View); (e) Detailed View (Local View); (f) Snippet Window

Second, the interplay between the Document View and the Detailed View is designed such that the overview of the document is present at all times, in terms of a callout on the left side of the screen, an actual *minimap* as in computer gaming, which allows the user to maintain a sense of where he or she is while manipulating the fine-grained nodes and edges in the Detailed View.

The iterative process culminated in the final prototype shown in Figure 6.5. In this interface, we see an initial overview, the Collection View that presents an overview of the underlying documents' structure in the collection (Figure 6.5-a). The collection view can potentially provide multiple partitions on the documents. Figure 6.5-b illustrates one partition of a collection. As an information seeker drills down on each document, the view is altered (Figure 6.5-c) such that an overview of the document is presented. The Document View provides a Global View of the corresponding document in terms of its central concepts. In this overview, the salient concepts in that article are visualized as circles of different sizes, where size indicates the frequency of occurrence in that article. We used force and pack layouts (as part of the D3 library[3]) to visualize the different layers of the knowledge graph representation.

The lowest layer of our representation is the Detailed View (Figure 6.5-e). This view is a knowledge graph that represents entities and relationships between them. The Detailed View, similar to the graph interface presented in Chapter 5, contains labeled nodes and unlabeled links between nodes. Nodes that represent entities with low frequency are hidden in the initial view, and only appear once a higher-frequency, connected node is clicked, ensuring that the graph does not become too cluttered. Once the user hovers over a node, that node and all connected nodes are highlighted, while the remainder of the graph is alpha-blended into the background. Clicking on a node can expand it by adding in its related nodes. Alternatively, clicking on a node can collapse its neighbours if they are expanded already. Nodes can also be dragged and placed at different parts of the canvas. This functionality can help with organizing the graph structure in a way that is more meaningful to the user and it can help with minimizing label overlap in the graph [479].

Edges can similarly be highlighted by hovering. By clicking on any edge, the user can see the relationship(s) between the two corresponding nodes (linked by this edge) in the context window located on the lower left side of the interface (Figure 6.5-f). For each relationship in the context region, a hyperlink allows users to view the corresponding web page.

6.3 Experimental Design

Given that we have designed hierarchical knowledge graphs, a related question is how hierarchical knowledge graphs compare to hierarchies and/or knowledge graphs with respect to information seeking tasks. To evaluate this question, we need a set of control interfaces (reference interfaces that can be compared to hierarchical knowledge graphs, HKGs) and a reference data set. These can then be leveraged to design an experiment. As well, experimental design should replicate, as closely as

[3] http://d3js.org/

possible, past work to ensure experimental validity.

In our recent work [479], we developed two interfaces for exploratory search: one knowledge graph interface and one hierarchical tree interface. To preserve experimental validity, we use identical interfaces as control interfaces. We also leverage the identical data sets, ensuring that topic is eliminated as a confound. Finally, we use exactly the same experimental task, ensuring that performance numbers are representative between experiments.

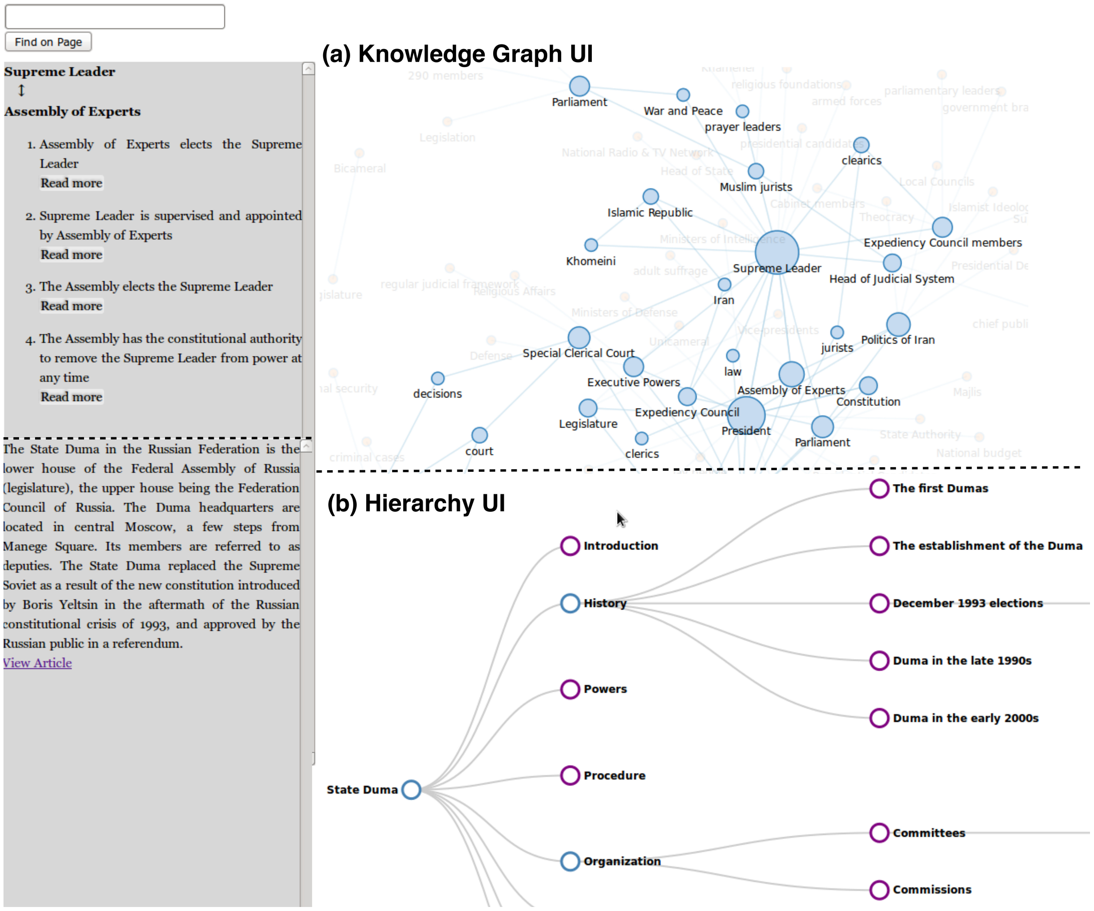
6.3.1 Control Interfaces

- The first interface, a knowledge graph interface, functions as follows: As the interface starts, nodes that represent entities with low frequency are hidden in the initial view, and only appear once a higher-frequency, connected node is clicked. Users can also filter the knowledge graph by clicking on a node; when a user clicks on an edge, snippets and links associated with that edge are shown in a preview pane on the left side of the interface.
- The second interface utilized a hierarchy (or a tree) structure to organize headings and sub-headings of the articles, as observed in each page's table-of-contents. When the user launches the application, the user is presented with a fully expanded tree. By clicking on any node within the tree, that portion of the Wikipedia document corresponding to the node is presented in the preview area at the left of the interface.

Figure 6.6 depicts these two interfaces. Contrasting these interfaces with Figure 6.5 shows a similar preview pane for snippets. Links within the snippets function identically across all three interfaces.

6.3.2 Data Set

We leveraged the two data sets from our previous study: A history data set, specifically a corpus of Wikipedia articles describing the historical locations of the capital city of Canada; And a global politics data set, a Wikipedia corpus representing governmental structures in Iran and Russia. These data sets are described in Section 5.4.1.


Find on Page

Supreme Leader
↕
Assembly of Experts

1. Assembly of Experts elects the Supreme Leader
   Read more
2. Supreme Leader is supervised and appointed by Assembly of Experts
   Read more
3. The Assembly elects the Supreme Leader
   Read more
4. The Assembly has the constitutional authority to remove the Supreme Leader from power at any time
   Read more

The State Duma in the Russian Federation is the lower house of the Federal Assembly of Russia (legislature), the upper house being the Federation Council of Russia. The Duma headquarters are located in central Moscow, a few steps from Manege Square. Its members are referred to as deputies. The State Duma replaced the Supreme Soviet as a result of the new constitution introduced by Boris Yeltsin in the aftermath of the Russian constitutional crisis of 1993, and approved by the Russian public in a referendum.
View Article

Figure 6.6: Control interfaces for Knowledge graph and Hierarchy. More details in Sarrafzadeh et al. [479].

6.3.3 Search Tasks

We used the same two exploratory search tasks [352], a simple and a complex exploratory search task, as follows:

Simple Politics: What governmental body or bodies are involved in the impeachment of the President of Iran and of Russia? (sample question)

Complex Politics: Imagine you are a high school student who is going to write an essay on the Political Systems of Iran and Russia. Knowing little about the presidents of these two countries, you wish to determine which president has more power. Find at least 3 arguments to justify your answer.

Simple History: As a result of which act were Upper and Lower Canada formed? (sample question)

Complex History: Imagine you are a high school student who is going to write an essay on the History of Canada. Knowing little about Canadian History, you wish to know which cities have served as a capital for Canada. You would also like to understand the reasons behind moving the capital from one city to another.

In Section 5.5.1 we validated these search tasks – both quantitatively and qualitatively – and ensured that the complex tasks are representative of exploratory search tasks and that both topics are of sufficiently similar complexity.

6.3.4 Study Design

Our study design was a 3 × 2 × 2 [interface, topic, complexity] mixed design. For Knowledge graph and hierarchy, we leverage the data set from our previous study [479] as control interfaces.We add additional participants for our HKGs to yield our mixed design as follows.

For HKGs, each participant performed two different tasks, one simple and one complex. The topic area (history or politics) differed for each of these tasks. More formally, for these participants, our design was a 2 × 2 full factorial mixed design, with topic and complexity as within subjects factors and complexity to topic assignment as a between subject factor. We counter-balanced the order in which the tasks were assigned to the participants.

Alongside the HKG participants, leveraging data from our previous work [479] adds two additional levels of Interface (hierarchical tree or knowledge graph) as a between subject factor. Combining the data sets yield the 3 × 2 × 2 mixed design [interface, topic, complexity] with interface as a between subjects factor, and topic and complexity as within subjects factors.

6.3.5 Participants

In total we analyze data from forty seven participants. Twenty six participants, thirteen female, used hierarchies and knowledge graphs, the control interfaces. An additional twenty-one participants (4 female) used HKGs, the experimental condition, as a between subjects factor. All participants use the Internet on a regular basis to search for information. Participants were aged between 18 and 45 years old (62% were between 20 and 29 years old). Participants received a $15 incentive for their participation.

6.3.6 Procedure

After introducing the study, participants were presented with an experimental interface (populated with an unrelated data set), and were given time to familiarize themselves with the interface and data structure. Once participants had developed some comfort with the features of the interface (~ 3 minutes), participants completed a questionnaire assessing their familiarity with the topic used for the first task. They were then given the description of their task (see above), and were asked to complete the task using the interface (15 minutes per task). Participants completed a post-task questionnaire that evaluated the experience; we used questionnaires provided by TREC-9 Interactive Searching track [4] modified to fit our experiment. The same process was repeated for the second task.

At the end of the second task, a semi-structured interview explored participants' experience using the interface. Interviews explored the conceptual usability of the visualization, the technical usability of the application and the efficacy of the interface for different types of search tasks. Feedback on competing interfaces was also collected from participants.

6.3.7 Data Collection

Alongside a mixed design of within subject and between subject factors, we perform a mixed methods analysis of both quantitative and qualitative data [127]. Data was captured as follows:

[4] www-nlpir.nist.gov/projects/t9i/qforms.html

(a) The interface was instrumented with a logger which monitored movement on the computer screen and participants' interactions with the system. Interactions collected included node or edge clicks, snippets read, articles viewed, and time spent reading the articles. In HKGs, the transition between the layers and switches between the MiniMap and the knowledge graph were captured.

(b) Two assessors evaluated the quality of answers provided by the participants for each of the search tasks independently. Simple queries were rated as either correct or incorrect. Complex questions were rated on a scale. Scores for all queries were normalized to reflect a value in the range [0, 1]. Inter-assessor reliability was evaluated using Pearson coefficient and an overall value of 0.97 for simple queries and 0.94 for complex queries was found.

(c) We captured field notes during participant interactions, audio recorded all sessions, transcribed final interviews, and collected questionnaire data. This data was analyzed collectively using open coding to extract low-level themes and axial coding to identify thematic connections between elements. Coding was performed incrementally as each participant's data was collected, and saturation was found after coding qualitative data from field notes and transcripts for 15 of our 21 participants.

6.3.8 Hypotheses and Research Questions

Quantitative data allows us to test the following hypotheses:

- Hierarchical knowledge graphs result in fewer document views and less time spent reading documents than do hierarchical trees.
- Hierarchical knowledge graphs exhibit statistically similar behaviors to Knowledge Graphs.

Alongside hypothesis testing, our log data provides insight into whether hierarchies are used in hierarchical knowledge graphs and on whether task complexity affects the use of hierarchies. As well, to triangulate quantitative data, we leverage our qualitative data to compare and contrast the nature of the hierarchies between the tree interface and the hierarchical knowledge graphs and to understand whether the hierarchies provide similar affordances.

6.4 Results

6.4.1 Quantitative Analysis

Scoring of participant responses by independent evaluators and log file analysis produced the quantitative measures in Table A.1 for Hierarchical Knowledge Graphs (H. Graphs), Hierarchical Trees (H. Trees), and Knowledge Graphs (K. Graphs). Rows represent measures for Marks (MK), Nodes clicked (NK), Edges Clicked (EC), Document Views (V) and Document View Time (VT). We break each measurement out by two query levels, Simple and Complex, as described previously.

		H. Graphs	H. Trees	K. Graphs
Simple	MK	0.43 (0.21)	0.32 (0.20)	0.37 (0.14)
	NC	11.4 (8.6)	19.0 (10.04)	11.38 (9.4)
	EC	18.3 (8.9)	NA	27.15 (12.9)
	V	**2.38 (1.61)**	**6.08 (2.49)**	**2.38 (3.00)**
	VT	**145.6 (153.7)**	**1430.9 (2302.8)**	**211.6 (228.0)**
Complex	MK	0.62 (0.18)	0.57 (0.28)	0.58 (0.16)
	NC	13.38 (9.2)	20.09 (17.7)	26.23 (19.12)
	EC	23.09 (12.7)	NA	41.07 (19.4)
	V	**2.15 (2.13)**	**4.38 (2.24)**	**4.38 (2.24)**
	VT	**103.4 (97.6)**	**985.38 (1848.3)**	**78.76 (131.5)**

Table 6.1: Hierarchical (H.) Graphs vs. Hierarchical Trees and Knowledge (K.) Graphs: Mean (Standard Deviation) values for marks (MK - average independent evaluator scores), clicks on nodes (NC) and edges (EC), document views (V), and document view time (VT). Bolded dependent variables exhibited significant differences in post-hoc testing.

6.4.1.1 Hypotheses Testing

Multivariate analysis of variance with respect to interface (tree versus graph versus hierarchical graph), topic (history versus politics), and task (simple versus complex) for Marks (MK), Views (V), and View Time (VT) shows a statistically significant effect of interface ($F_{6,172} = 7.126, p <$

$0.001, \eta^2 = 0.2$) and task ($F_{3,86} = 12.22, p < 0.001, \eta^2 = 0.3$) on dependent variables. Post-hoc factor analysis using Tukey correction indicates that the tree interface exhibited statistically significantly higher numbers of document views than both hierarchical graphs and knowledge graphs. As well, the tree exhibited statistically longer reading times than hierarchical graphs ($p < 0.05$), but not than knowledge graphs (p = 0.064) in our analysis. Hierarchical graphs and knowledge graphs did not differ significantly in their effects on any dependent variables. Task significantly impacted the marks but no other variables.

Clicks are not directly comparable between H. Trees, H.Graphs, and K.Graphs, as edges are not clickable in hierarchies (NA value in Table 1). Performing pairwise comparison between H.Graphs and K.Graphs, our analysis showed no statistically significant effect on dependent variables ($F_{3,30} = 0.752, p > 0.5, \eta^2 = 0.70$), including node click and edge click behavior.

Given the above analyses, we reject both null hypotheses and conclude that our hypotheses are supported by our data set. Hierarchical Knowledge Graphs preserve the advantages of Knowledge graphs over hierarchical trees in both reading time and in document views. Focusing specifically on our hierarchical graph, we find that our hierarchical graph has statistically lower document views (61% fewer document views, on average) and time reading (90% less time reading documents) than does hierarchical trees and that its behavior is statistically indistinguishable from the prior observations of knowledge graph interfaces. Furthermore, the effect size measures, η^2, are significantly above the threshold (0.14) typically considered to be a large effect, lending support to these differences being sufficiently large to be meaningful. In summary, our quantitative results support our hypothesis that our hierarchical knowledge graphs fully preserve the quantitative advantages identified by our prior work [479] (described in Chapter 5) for knowledge graphs over hierarchies.

6.4.1.2 Additional Quantitative Analysis

Given the statistically indistinguishable nature of HKGs and Knowledge Graphs, one question is if (and whether) intermediate hierarchical representations are used. It is possible that Hierarchical Knowledge Graphs are indistinguisable from Knowledge Graphs because users ignore the hierarchy and simply leverage the knowledge graph.

	GlobalView	MiniMap	DetailedView
Simple Task	27.03%	14.61%	58.0%
Complex Task	23.83%	17.24%	58.90%

Table 6.2: Percentage of Time spent on each of Global View, Minimap and Detailed View

.

To specifically explore this question, we looked at how much time users spent on each of the provided views in our HKG interface. Overall, our data indicated that participants took advantage of all three layers relatively similarly across both Simple and Complex tasks. Further, while the the time spent on detailed view dominates other views (58% for the simple task and 59% for the complex task), over 40% of time was spent on additional views in the hierarchy (Table 6.2). Looking specifically at how participants spent their time in different layers of the hierarchy (i.e. utilizing different views of the data) for different tasks we see that the time spent at the detailed view is similar for both levels of complexity. On the other hand, participants seem to spend less time in MiniMap than Global for the simple task (Pairwise t-tests with Tukey correction yields statistical significance, $p < 0.01$). For Complex task, however, time in Global versus mid-level are not statistically different ($p > 0.1$). Essentially, in the complex task, sensemaking is split between global and minimap views of the hierarchy more equitably, i.e., the minimap is particularly useful during our complex tasks.

We also explored usage patterns of views. Figure 6.7 is a heatmap that visualizes use of different views for intervals of 1% of task length. Early in the task, we see frequent use of the global view. While difficult to see, MiniMap usage peaks just after the halfway point in the task, but there is no strong concentration of use. The hierarchy, and particularly the MiniMap visualization, seems to be used throughout the task.

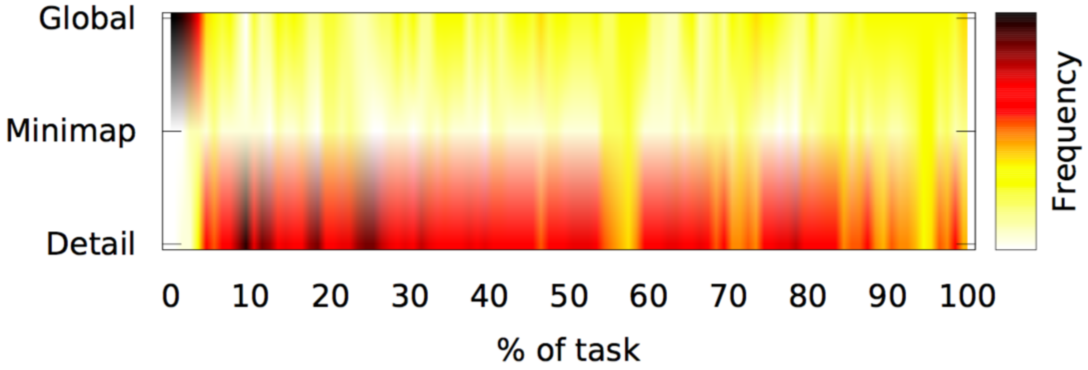


Figure 6.7: Heatmap visualizing the patterns of users navigating views in HKG for intervals of 1% of task length.

6.4.2 Qualitative Analysis

The next question we explore involves participant perspectives on hierarchical knowledge graphs as a representation of search results. We were particularly interested in the overviews knowledge graphs provide for the information space and their contrast with Table-of-content-based hierarchies.

To address these questions, we performed open-coding of observations, transcripts, and questionnaire data. We coded incrementally, and saturation occurred after fifteen participants were coded. We coded all participants for completeness. Once open coding was complete, axial coding and thematic analysis was performed collaboratively by the researchers. We present three themes arising from our qualitative data analysis: Supporting Exploratory Search Tasks, Imposing a Structure versus Open Exploration and the Self-Orienting nature of HKGs.

6.4.2.1 Supporting Exploratory Search Tasks

As noted in our study design, we incorporate two exploratory information seeking tasks with different levels of complexity. In post-experiment interviews the participants were able to compare how different task complexities are supported by the assigned interface.

The hierarchical graph representation was found to provide more support for the Complex Task (i.e., more open ended and exploratory tasks such as essay writing or learning) versus Simple tasks (such as question answering and specific knowledge finding). This observation seems to be true for any multi-level structure which provides an overview and allows a gradual immersion into details: Finding a specific piece of information to satisfy a simple query is best done using a traditional search engine.

Looking specifically at HKGs and complex tasks, the overview allowed participants to identify the central concepts of a domain at a glance and the size of the circles indicates their prominence in the corresponding article. As many participants noted, 'relevance' or 'prominence' of a concept with respect to the main topic or the domain they are exploring is an important asset in Complex Search tasks. This qualitative observation may explain the more equitable use of the MiniMap representation for complex search tasks noted in our quantitative analysis. Complex tasks required synthesizing, rationalizing, and comparing, which seem to require more awareness of the entire data set.

This identification of central concepts was also linked to a perception of value of the MiniMap as a starting or entry point into the topic of the document being examined. Several participants articulated a belief that the overview provided by Central Concepts helped with "going from knowing nothing to having a plan", "learning terminology", "relevance, importance, or prominence", and "objectively learning about a domain". In particular, the *objective* nature of central concepts was cited by many participants as key to their utility.

As White and Roth [597] point out, exploratory search is motivated by complex information problems, poor understanding of terminology and information space structure, and often a 'desire to learn'. Vakkari [557] also argues "more support is needed in the initial stages of a task", when users have an unstructured mental model. Inspired by Kim [290], Sarrafzadeh et al. [479] found that hierarchical trees provide this benefit in unfamiliar domains. A strength of our design of hierarchical knowledge graphs is that it enables the user to engage in two alternative navigation paradigms. Users can exploit overview layers to explore the collection at a higher level followed by targeted immersion in the detailed view.

6.4.2.2 Imposing a Structure versus Open Exploration

While most participants were unanimous that the hierarchical representation imposes a [subjective] [rigid] structure onto the information space, their attitude towards this phenomenon varied. The level of domain knowledge and the complexity of the search tasks were found to be the major factors affecting their attitude.

When the searcher is dealing with a domain where he has limited knowledge, he is more open to accepting the structure that the representation imposes. Both hierarchical trees and hierarchical knowledge graphs incorporate imposed structures. Participants articulated a variety of advantages to structures: it was “easier to follow”, “contained important aspects” that “simplified focus”, and guided participants in “where to go” or “what steps to follow”. With respect to hierarchical trees, some participants simply “trusted” the designer of the hierarchy (e.g. the author of an article) to be “logical” or “rational” in the way he broke down things. This was particularly true for participants with limited knowledge of a topic domain and replicates findings in our prior work [479] and and the work by Amadieu et al. [23] that low knowledge learners benefited from hierarchical structures in free recall performance and exhibited reduced disorientation.

In the case of higher domain knowledge, our participants were split in their preferences and attitudes. Some still trusted the logic behind the layout of a hierarchical trees and the fact that their knowledge of the domain can guide them to find what they want using this hierarchy. They trusted the designer to place items in close proximity to where the item should be. Other participants strongly opposed the rigid structure of a hierarchy, feeling it was “not the way I think”, “based on the mindset of the author”, or “did not match the domain structure”.

One interesting perspective of the multi-layer graph representation which presents central concepts of a domain as an overview for each document is that it reflects the knowledge graph concepts. This reflection made it, for many participants, more flexible and exploratory, a window into the knowledge graph. Many participants commented on this phenomenon, noting it was “guiding but not imposting”, “more open”, “sparked interest” in the lower level structure, or was “visually appealing” and “fun”.

6.4.2.3 Self-Orienting or Relative Positioning

One main advantage of the Hierarchical Tree visualization in the previous chapter was the explicit connections between nodes (categories or headings) in the representation. These edges help in two ways:

1. At a glance, you can tell why a concept appeared in this overview, or in this domain. To whit, the hierarchical structure exists the way it does because of a human author's decision.
2. The Path from the root to each of these nodes in the Tree Layout can provide useful information on where a concept is positioned relative to the topic.

Some of our participants articulated this distinction between these two types of hierarchies particularly well.

> "*This [hierarchical knowledge graph] kind of tells you these are the main things happening here; but I mean it could be a bit deceptive! because when a concept is there, you don't know how it is related to the main topic. you should go there and you find out if it was actually related. For example, it shows you "American Revolution" is there a lot; but I have no idea, how it's related to Upper Canada. Whereas in Tree, there is a clear connection that this is directly related to that; and you can see how it's related. Because this heading is directly related to the main topic. Just by seeing the position of the heading you can understand what the heading is talking about. But in Circles, you don't really know!*"[P3]

> "*The nice thing about the hierarchy [Tree] is that, if we consider history as a Domain ... it starts at one point and ends at another point, like the current date. Having a hierarchy or a timeline makes sense. Whereas [hierarchical knowledge graphs] gives you important concepts and stuff but it doesn't tell you 'where the seventy two resolutions occur' or 'what's the order' and when history is the domain that's important.' For example, if you are looking at First World War, Upper Canada, Lower Canada, and you have a Hierarchy [Tree], you start in the direction of the colonization and go in the order that things happen. Whereas [with Hierarchical Knowledge Graphs], it's just whenever! 'Political crisis' is over there but is that in War of 1812? Or did it happen at World War II? all of them? none of them?*" [P9]

In Chapter 5 we note that participants may perceive a domain to have a derivative/hierarchical structure or a multi-faceted structure. If salient relationships are viewed as derivative or hierarchical (e.g. 'is-a' relationships), then a tree can best capture this view of data, whereas if salient relationships are more heterogeneous and resist structure as a hierarchy, that disadvantages the hierarchies.

This is not the case in our MiniMap, where the connection between each of these main concepts and the main topic is unknown at first glance. Central concepts are simply extracted based on their high connectivity with other concepts within a specific document within a corpus. However, it is also true that it would be quite surprising if highly linked concepts were not, somehow, important components of any individual document. The more pervasively they link, the more they interconnect with other concepts, the more important it is to understand them and their relationship. In this way, HKGs become self-orienting for out participants.

6.5 Limitations

Any study has limitations. Because we leverage the research methodology and the data sets from our prior work [479], we inherit the limitations of that study, including topic and implementation issues which may bias the study. We discussed our approach to addressing these limitations in Section 5.6.3. Despite this, there is also a strength in replication: if interfaces are redesigned, data sets differ, and tasks are unique it becomes difficult to ensure a lack of confound in experimental design. We address this by preserving, to the limit possible, all aspects of a similar study within this space contrasting data structures.

Our mixed design of within and between subject factors is a particular strength to our study design. Because topic (history/politics) and task complexity are within-subject factors, they are controlled across participants. Because we are most interested in interface and it is a between subject factor, to observe statistical significance we need good separation of dependent variables between the two data sets, reducing the likelihood of a type-one error in our analysis.

6.6 Chapter Summary

The primary goal of the research described in this chapter was to explore whether we could combine benefits from both knowledge graphs and hierarchies into one data structure for visualizing search results. We note that our hierarchical graphs significantly reduce documents read and reading time as compared to hierarchical trees and perform on par with knowledge graphs. We also provide evidence that the hierarchy is used by participants via analysis of interaction logs.

Qualitative data from our participants does indicate that hierarchies grounded in tables-of-contents are more familiar, easier to follow, and more focused. This in turn helps users orient themselves in the data. The vetted nature of hierarchical tables-of-contents was also perceived to be an asset absent from our hierarchical knowledge graphs. The hierarchies in our knowledge graph were viewed slightly differently, as noted above, with a more quantitative perspective giving them a certain cachet with respect to the unbiased nature of topic selection.

A final issue to consider is whether any hierarchy might provide benefits. While it may, one advantage of the hierarchy in our HKGs is its tight connection to the entities contained in a knowledge graph and the ease of automatically extracting the hierarchy through thresholding. Another advantage is flexibility: while we currently leverage only three levels – corpus, central concept, and knowledge graph – it is easy to generalize the hierarchy to an arbitrary number of thresholds depending on the complexity of the domain. We do not generalize the hierarchy in this work because, for a first experimental validation, there are a limited number of factors that can be assessed. However, future work can address more detailed inquiries into scalability to larger corpora, scalability to multi-level hierarchies, and contrasts with other hierarchies such as automatic clusters or user-specified facets.

In summary, we find that our hierarchical knowledge graphs preserve many of the previously observed advantages of traditional knowledge graphs, i.e. fewer document views and reduced reading time. Alongside this, HKGs introduce an effective hierarchical representation into knowledge graphs. In the next chapter, we take a step back and ask what happens if there are errors in the extraction algorithms? Essentially, how error-prone HKGs impact the information seeking behaviors and outcomes?

Chapter 7

Error Prone Representations of Search Results

"Knowledge rests not upon truth alone, but upon error also."

– Carl Gustav Jung (1875 - 1961)

In Chapters 4, 5 and 6, we studied representations independently from extraction methods that generate them. This separation was valuable as it provided an opportunity to observe the impact of different representations of the search results on the outcome of complex information seeking scenarios independent of the accuracy of the algorithms that produce the underlying data. In this chapter we take a step back and ask what happens if there are errors in the extraction? How resilient are new representations to these errors? And how do errors impact a user's ability to leverage these representations to acquire knowledge?

To probe these questions, we perform a mixed methods analysis of the effect of precision and recall on the performance of hierarchical knowledge graphs for two different exploratory search tasks. To this end, we leverage the information extraction algorithm that we developed earlier in Chapter 3 and compare users behavior and outcomes with the erroneous output of the IE algorithm to users behavior and outcomes with the corrected outputs of the information extraction algorithm. While the quantitative data shows a limited effect of precision and recall on user performance and user effort, qualitative data provides evidence that the type of exploratory search task (e.g., learning versus investigating) can be impacted by precision and recall. Furthermore, our qualitative analyses find that users are unable to perceive differences in the quality of extracted information. We discuss the implications of our results and analyze other factors that more significantly impact exploratory search performance in our experimental tasks.

7.1 Motivation

While representing information is critical to effective support of exploratory search, the process of extracting and presenting that information to the user is challenging. Information extraction (IE) algorithms are not perfect [516], meaning that the base level information to be presented to a user may contain errors and omissions, and, as such, may reflect compromises in either *precision* or *recall* by the algorithm. The TAC [1] and TREC [2] tracks on evaluating IE and question answering (QA) algorithms report varying levels of precision and recall for these algorithms when compared against manually created ground truth datasets.

To address the extraction and presentation challenge, alongside developing an information extraction algorithm that is capable of automatically extracting semantic information from the textual content of documents (see Chapter 3), in Section 3.5.1.2 we proposed an evaluation framework that decouples the evaluation of the efficacy of representations of the extracted information from the performance of the algorithms that generate these extractions. As a result of this decoupling, we generated knowledge-graph representations of search results that were populated with gold entity-relationship data and focused on improving the interactive and visual aspects of knowledge graphs while the underlying information is as accurate as the textual content of retrieved documents.

In this chapter, we address the second step of our proposed evaluation methodology, essentially, we present a mixed-methods study [126] that examines the impact of imperfect IE on exploratory search tasks. To conduct this evaluation, we leverage an existing system that supports exploratory search tasks via hierarchical knowledge graphs (HKGs) [475]. Users interact with two HKGs, one created manually by human experts

[1] https://tac.nist.gov/

[2] https://trec.nist.gov/

(the typical approach to generating ground truth to benchmark IE systems [516]) and a second graph that was automatically generated and exhibits significantly lower precision and recall compared to the manually generated graph.

While our expectation was that precision and recall would impact user performance (i.e., success in an exploratory search task) or the effort expended during search (e.g., the number of documents viewed), our results indicate that neither performance nor effort was significantly impacted by differing levels of precision and recall. To probe this result in greater detail, we analyze qualitative data collected via observations and interviews. Our qualitative data indicates that task characteristics may be an important factor to consider in exploratory search. Specifically, for investigate-style tasks [352] where there is a defined set of facts to retrieve, recall may impact user behavior because it is necessary to find specific facts within the information presented. In contrast, more open-ended synthesis or comparison tasks, where salient data can be more flexibly applied by the user, seem more resilient to lower recall rates.

7.2 Past Research

As noted in the introduction, our primary interest in this work is to develop an understanding of how errors in information extraction – and consequently error-prone representations – impact exploratory search. In this section, we provide background in IR research on the effect of error on information seeking and examine past research in supporting exploratory search.

7.2.1 Error Effects in Information Seeking

It is often assumed if an evaluation measure coupled with a test collection reveals that system A provides higher quality output than system B, then the user will both prefer system A and that system A will more effectively support the user's information seeking task [20, 559]. However, the IR research on this topic indicates that the relationship between output quality and system efficacy is not clear.

In document retrieval, there is a long history of research that examines how human search performance varies with system effectiveness [222, 550, 20, 549, 114, 15, 504, 469, 506, 578]. As we noted in Chapter 2, the broader goal of this line of work is to understand when system effectiveness improvements are meaningful or useful in improving users' information seeking abilities in practice. These studies, however, have resulted in contradictory findings. Early studies suggest that "better" systems, as measured by system oriented metrics such as precision and recall, do not necessarily translate into better task performance [222, 550, 223, 549, 504]. More recent work has detected potential correlations between various system effectiveness metrics and human preferences [15, 469] and between precision and user performance [510].

Even if one gives credence to more recent work relating effectiveness to user preference and user performance, the reason for inconsistent effects of system performance on human performance in past work is unclear. Differences could lie in the definition of 'relevance' and how it is used as a basis for evaluation of document retrieval systems [148, 529]; they could also lie in the discrepancies between the metrics used for evaluation (e.g., MAP vs P@10) and the type of task the user performs (e.g., recall-based or complex information seeking) [549]; sample size may provide increased power to discriminate effects [15, 469]; or, finally, differences could be a result of the lack of UI support for meaningful user interaction with the retrieved results [332, 504, 550]. Hersh et al. [223] also found that while precision and recall weren't associated with success in medical QA tasks, other factors including experience of the searcher and cognitive abilities in spatial visualization were correlated with the ability to answer questions correctly. Further, in Section 2.5.3.3, we highlighted other factors contributing to observing inconsistent effect of system performance on users success, particularly in the domain of exploratory search tasks.

Possibly due to the ambiguous link between system performance and effectiveness, there have been calls to extend evaluation of IR systems from an analysis of the output of the system to the outcome of the search task [281, 559]. Furthermore, there is also an evolving drive toward evaluations of how effectively IR systems support complex, evolving, long term information seeking goals, such as learning and exploration [129].

7.2.2 Designing for Exploratory Search

While there has been research on understanding complex and exploratory search (see Section 2.2 and [597, 611] for a survey), there are many open questions when it comes to the design and evaluation of IR systems that provide tailored and adaptive support for long-term search needs. We have identified three areas of past research that explore support for users with more complex, exploratory search tasks. First, in Section 2.3 we reviewed a body of work in the IR community that aims to deliver "information" and not documents. Within this body of work, Open Information Extraction (Open IE) [40] techniques have been widely applied to extract semantic information from the text of documents to support a variety of down stream applications such as question answering or to populate knowledge bases (see [384] for a recent survey of these systems).

The second area focuses on investigating ways that search systems can represent and provide the extracted information to help searchers in evaluating and contextualizing search results (see Section 2.4). Within this space of techniques for the organization of search results, researchers have performed evaluations of representations including concept maps [23, 22, 90], hierarchies (e.g. [87, 385, 150]) and networks ([478]). To extend these efforts, in our prior work [479, 475] we designed search interfaces looking specifically at the contrast between hierarchical and network representations.

Finally, developing solutions to support users' exploratory search tasks also includes significant challenges in evaluation. The recent SWIRL Workshop [129] has identified the most relevant research questions to be addressed in order to develop new evaluation models that are suited for complex and exploratory information seeking. We reviewed some of the approaches to evaluating exploratory search systems in Section 2.5. A major step towards this goal is to design and study characteristics of search tasks that elicit exploratory behavior. These studies, in turn, provide data on searchers performing these tasks, specifically focused on task outcomes and searcher behaviors. Designing tasks for exploratory search studies can be especially difficult since inducing exploratory style search requires searcher to individually interpret the tasks, results, and their relevance [308] which is at odds with maintaining some level of experimental control and consistency [307]. The literature [352, 84, 83, 307, 328, 605] suggests a number of desirable characteristics for exploratory search. As

Wildemuth and Freund [605] note, it is crucial to construct tasks having particular attributes, knowing that our study findings can then be generalized to all search tasks having those attributes.

Our primary interest is in how errors in IE impact end-user performance and effort in exploratory search. Within the broad domain of the impact of error in IE, we were able to identify only one work by Chu-Carrol and Prager [112] that examined how user performance degrades in the face of imperfect named entity and relation extraction. Their experimental results demonstrated that significant document retrieval gain can be achieved when state-of-the-art IE systems are used and that recall has more significant impact on document retrieval performance than precision when adopting the MAP metric. Their results focus on assessment of document retrieval, not on assessment of support for exploratory search.

Synthesizing past research, we see ambiguity in the effect of errors in the domain of information retrieval, and an increasing focus on outcome versus output centric assessment. Coupled with this, we note that exploratory search tasks require systems that support browsing, and, within information retrieval, this has given rise to systems that retrieve and present information to users in formats that support browsing. However, absent from past research is any assessment of the effect of information extraction errors on exploratory search interfaces. While we concur that systems would ideally have perfect precision and recall, in the near term it seems unlikely that computational information extraction will be perfected, further motivating exploration of the effect of information extraction errors on interfaces that support exploratory search.

7.3 Evaluation Framework

In order to evaluate how varying levels of precision and recall in IE impact user behavior, it is necessary to embed the IE algorithm in a system that supports the overall task. As we noted earlier, in this chapter, our primary interest is in how varied levels of precision and recall in IE impact exploratory search.

With this goal in mind, in this section we describe a framework for evaluating systems that support exploratory search. We first focus on the necessary system components to support exploratory search, including an IE algorithm and an interface that leverages IE output to support user interaction. We then characterize the behavior of our IE algorithm to demonstrate that our system includes varied levels of precision and recall.

7.3.1 System Support for Exploratory Search

As noted in our thesis statement in Section 1.1, to support exploratory search, a system must have two components: an information extraction system that can identify and extract relevant information from a corpus (e.g., search engine results); and, an interactive UI that presents the information to users and allows users to browse the extracted information for sensemaking [257, 397]. Any arbitrary system designed to support exploratory search will exist on a range from a standard search engine interface (an ordered list where a user can select and browse individual documents) to systems that extract and synthesize information for the user. An example of the latter type of system include those found at the TREC Complex Answer Retrieval (CAR) track [142] which explores the design of systems that apply information extraction to synthesize content and generate an essay on a particular topic. Ultimately, any system on this spectrum requires the two components identified above: an information extraction component and an interactive UI to support browsing.

A common approach to present-day exploratory search systems is to leverage IE to identify relevant entities and their relationships within the retrieved documents. These entities and relationships can then be displayed in graphical form (e.g., as a knowledge graph, or hierarchy, or concept map, etc) which provides the user with a spatial representation of the information space for sensemaking [414]. This representation can then be embedded in an interface that allows a user to interact with the representation, to filter and select specific content, and essentially to explore the information returned [479].

7.3.2 Evaluating Exploratory Search Systems

Given that our goal is to understand how errors in extraction can impact exploratory search tasks, it is necessary to analyze IE algorithms as a component within larger frameworks [516, 112, 111, 39]. There are two aspects to system performance: accuracy and effectiveness. Assessing the accuracy of an algorithm can be performed through benchmarking and/or combined efforts tasks (e.g., TREC or CLEF tasks). System effectiveness for exploratory search, on the other hand, requires evaluating how well the systems aids in the exploratory search tasks it is designed around. In the following two subsections, we describe in more detail, these two types of assessments. Our goal can therefore be stated as an attempt to characterize the impact of accuracy on effectiveness.

7.3.2.1 System Accuracy

Evaluating information extraction is challenging. There are no clear guidelines as to what constitutes a valid proposition to be extracted, and most information extraction evaluations consist of a post-hoc manual evaluation of a small output sample [516]. There is also no agreement on an appropriate data set to use for information extraction [384]. However, Stanovsky and Dagan [516] have developed a methodology that leveraged the recent formulation of QA-SRL [212] to create the first independent and large-scale gold benchmark corpus.[3] Stanovsky and Dagan's benchmark is based on a set of guiding principles that underlie most Open IE approaches. This benchmark has provided an opportunity to evaluate the output of an Open IE system using both precision and recall. While acknowledging that this benchmark may not be perfect, we leverage it as the most up-to-date standard for evaluating IE systems.

7.3.2.2 System Effectiveness

It has long been understood in IR that a system understanding of relevance is not always consistent with what a user desires [473] and so we must also understand how systems impact user performance. To do this, representative search tasks are required. Marchionini [352], referencing Bloom's taxonomy of educational objectives [63], distinguishes three broad categories of search tasks as *Lookup*, *Learn*, and *Investigate*. While these categories are depicted as overlapping activities, exploratory search is more pertinent to the *Learn* and *Investigate* activities. As a

[3] The corpus is available at: https://github.com/gabrielStanovsky/oie-benchmark

result, exploratory search is defined as searching that supports learning, investigating, comparison or discovery [307, 597]. From this understanding, we can distill exploratory search tasks into fitting into one of two themes. The first theme includes those tasks that facilitate learning to achieve knowledge acquisition, comprehension of concepts, interpretation of ideas and comparison or aggregation of concepts. The second theme covers those investigative tasks that involve discovery, analysis, synthesis and evaluation.

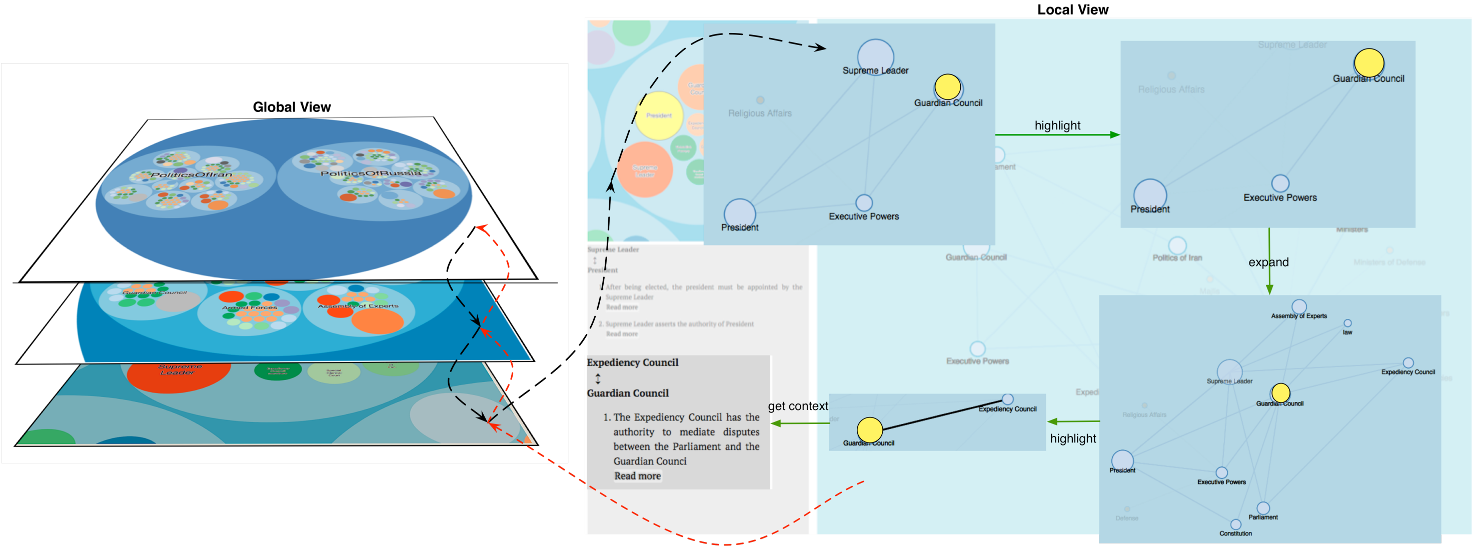
Based upon the aforementioned works and a survey of existing classification by Li and Belkin [328], we believe that exploratory search tasks should: provide uncertainty and ambiguity about the information need and in how to satisfy it; suggest a specific knowledge acquisition, comparison or discovery task; be in an unfamiliar domain for the searcher; represent a situation that a user can relate to and identify with; be of sufficient interest to test users; and, be formulated such that the user has enough imaginative context to facilitate immersion in the task. Any task that meets these criteria provides sufficient complexity that the end-to-end experience with an exploratory search system can be fully and properly assessed.

7.4 Experimental Design

To detail our experimental design, we start with instantiating a system to support exploratory search. Next, we describe the study design, our participants and the experimental procedure. Finally, we describe the data we capture from each participant.

7.4.1 Instantiating Our Evaluation Framework

Our experimental platform to support exploratory search leverages information extraction, information visualization, and an interface to support browsing and sensemaking. First, to examine the effect that different levels of precision and recall have on exploratory search, we use two different information extraction outputs. One set represent the raw, uncorrected output of our IE algorithm, described in Chapter 3; the second represents fully human-corrected output. Next, to visualize information, we



Supreme Leader
↕
President

1. After being elected, the president must be appointed by the Supreme Leader
Read more

2. Supreme Leader asserts the authority of President
Read more

Expediency Council
↕
Guardian Council

1. The Expediency Council has the authority to mediate disputes between the Parliament and the Guardian Counci
Read more

Figure 7.1: The Populated Exploratory Search Interface Leveraged in Our Study. This interface, represents search results as a hierarchical knowledge graph structure which enables the user to engage in two alternative navigation paradigms. While users can frequently exploit the overview layers to explore the collection at a higher level followed by targeted immersion in the detailed view, they can alternatively descend within an area of interest in a graph and continue exploring in the detailed view itself by manipulating nodes and edges and getting more context on relationships.

leverage our hierarchical knowledge graphs (see Chapter 6 and [475] for more details), which is a type of concept map that represents entities hierarchically and the relationships between those entities. Finally, to support interaction, we leverage the interface we developed and evaluated previously as it was experimentally shown to support exploratory search tasks effectively. Overall, to preserve experimental validity, our system is identical in functionality to the hierarchical knowledge graph system described in Chapter 6. However, unlike the study in the previous chapter, in this chapter we use both automatically generated and manually generated IE results to build knowledge graphs, resulting in significantly different levels of precision and recall. Each of these components is described in detail below.

7.4.1.1 Automatic Information Extraction

For our experiment, we generate knowledge graphs for a set of documents using our developed IRE system, which automatically extracts entity-relationship triples as follows: (1) entity taggers [4] are used to extract entities from text; (2) sentences that contain at least two entities are selected and parsed using the Stanford Dependency Parser; (3) for each sentence, we extract meaningful relations between the entities by finding the shortest path in the corresponding parse tree; (4) labels are automatically generated for the extracted relations. The labeled relations are ranked based on relevance to the query and the informativeness of the extraction (See Chapter 3). The results (entities and relations) are emitted along with the snippet from which the entities and relation were selected and an HTML anchor to the snippet in the source text.

To determine the performance of our IE system, we used the Open-IE benchmarking toolkit [516]. As can be seen in Figure 7.2, our system performance is representative of state-of-the art systems; specifically, we have tuned relative precision and recall of our system such that it achieves a precision of 0.65 and recall of 0.24 for the task, the mid-point of the precision and recall curves presented in [516].

Some may question our decision to choose approximately median (as opposed to optimal) performance for our information extraction system. If our goal was to investigate the best possible performance of systems leveraging automatic information extraction, choosing the best system would be justifiable. However, our goal is to examine the effect that errors in information extraction have on performance. Accordingly, selecting the top performing system would yield a biased experiment which would be limited to insights about the best-performing algorithms, whereas more representative performance across a class of algorithms allows us to generalize to better performing systems. That is, such systems should perform at least as well as our system relative to manually tuned extractions.

7.4.1.2 Hierarchical Knowledge Graphs

Once information has been extracted, this information must be presented to the user. For this study, we leverage our proposed data structure

[4] Reference implementation used https://cogcomp.cs.illinois.edu/page/software_view/NETagger.

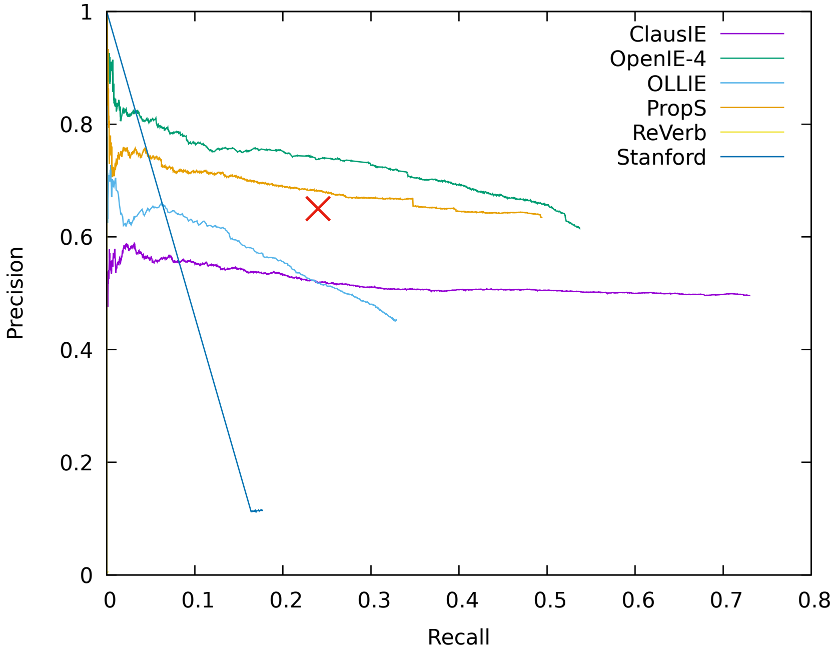


Figure 7.2: Precision-recall curve for the different Open IE systems using Stanovsky et al.'s toolkit [516]. The X represents our system.

called hierarchical knowledge graphs (HKGs) [475], introduced in Chapter 6. Like a typical knowledge graph, HKGs provide entity-relationship depictions of the information contained in a corpus where entities are represented by vertices and relationships by edges connecting entities; as well, like hierarchies, HKGs provide a hierarchical representation of the low-level information in knowledge graphs by leveraging the degree of connectedness of vertices to select a subset of vertices for an overview of key concepts within an information space. The benefit of this organization is that both the low-level knowledge graphs and the hierarchies can be automatically extracted using information extraction. However, given that IE algorithms may be compromised in precision and recall, automatically generated HKGs are more sparse and less precise than manually generated counterparts.

7.4.1.3 Exploratory Search Interface

As our experiments require users to complete exploratory search tasks (outlined in the following section), we leverage the interface we developed in prior work [475], which was previously shown to be effective for exploratory search tasks on HKGs [479, 475]. This interface allows smooth transition between overview and detailed views of HKGs (supporting overview-filter-detail-on-demand search [495] and expand-from-known searching [562]). Figure 7.1 presents this interface for context.

One advantage of leveraging an existing interface is that we eliminate the confound of interface effects on user behaviour. Specifically, because our interface is identical to past work, we can compare human-corrected output to output from past work. Then, if the output is similar, any deviations in results between different error levels are attributable to the effect of extraction errors rather than interface idiosyncrasies.

7.4.2 Exploratory Search Tasks

In order to evaluate the efficacy of error prone knowledge graphs and how they impact the exploratory search performance, we leverage the complex search tasks we previously designed.

As we described in Section 5.5.1, these tasks combine aspects of knowledge acquisition/comparison (Marchionini's *learn* subcategory) with analysis, synthesis, and evaluation (Marchionini's *investigate* subcategory). In addition, the task descriptions closely follow Bystrom̈ and Hansen's [83] recommendation that three levels of description should be used to specify a search task: a contextual description, a situational description and a topical description and query. Finally, in previous studies described in Chapters 5 and 6, we conducted quantitative and qualitative analysis on participants performing these exploratory search tasks and showed that these tasks were indeed complex (i.e., that they were ambiguous, open ended and exploratory in nature) and they were of sufficiently similar complexity as to limit topic effects.

For context, we provide the task descriptions used in our study:

Complex Politics: Imagine you are a high school student who is going to write an essay on the Political Systems of Iran and Russia. Knowing little about the presidents of these two countries, you wish to determine which president has more power. Find at least 3 arguments to justify your answer.

Complex History: Imagine you are a high school student who is going to write an essay on the History of Canada. Knowing little about Canadian History, you wish to know which cities have served as a capital for Canada. You would also like to understand the reasons behind moving the capital from one city to another.

7.4.3 Acquiring Knowledge Graphs at Different Levels of Quality

We generate HKG representations at two levels of quality in two steps; First, from our. previous studies, we leverage the document sets as well as the hand curated set of entity-relationship tuples that were generated by having experts manually refine the output of our IRE system.These documents are Wikipedia articles related to the tasks previously described. The expert crafted tuples form what can be considered the best possible execution of the extraction algorithm and represent a "best effort" for knowledge graph creation and so form our *gold standard*. Accordingly, the second step involved automatically running the extraction algorithm to yield an inferior graph that contains errors that would impact user performance.

7.4.3.1 Characterizing Precision and Recall of Automatic vs Hand-Tuned IE

To test whether precision and recall rates differ between *Auto*matic and *Gold* graphs, we note that the *Gold* generated graph for the History task contains 2,957 entity-relationship tuples and the Politics graph contains 3,231 tuples. In contrast, the *Auto*matically generated graphs contain 1,782 and 2,735 tuples, respectively. We ran the automatic graph through the Open-IE benchmarking toolkit with the manually curated graph as ground truth. For Politics, the automatic graph achieved a Precision of 0.56 and a Recall of 0.33. For History, the automatic graph achieved a Precision of 0.7 and a Recall of 0.31. These results are in line with what we might expect given our earlier results from running our IE algorithm on the Open-IE benchmark (See Section 7.4.1.1).

7.4.4 Study Design

Different components of our evaluation framework were discussed in previous subsections. As a final step, we describe a controlled in-lab experiment to investigate the effect of an error-prone Open IE system's output (i.e., the automatically ran system) on users conducting exploratory search tasks. The main goal of our study is to investigate the effect of errors on user behaviour. To facilitate this, our study design was a 2 x 2 [error-level, topic] mixed design with error-level and topic as within-subject factors and error-level to topic assignment as a between-subject factor. The two levels of error were *Gold* and *Auto*, which correspond to the manual and automatic run of the algorithm previously described. The topics correspond to the History and Politics tasks. The order of error-level to topic assignment is fully counter balanced to mitigate any order or learning effect. This resulted in a full-factorial design with 4 groups. Participants were randomly assigned to groups.

We recruited 25 (11 female) participants from different areas of Science, Arts and Business, Math and Engineering for this study, all of whom use the Internet on a regular basis to search for information. Participants were aged between 20 and 45 years old (80% were between 20 and 29 years old). They received a $15 incentive for their participation.

7.4.5 Procedure

Our study began with a brief introduction and a short familiarization period ($\sim$ 3 minutes) with the experimental interface that was populated with a completely unrelated dataset. The goal of this familiarization period was to allow the participants to "figure out" the nuances of the search interface unconstrained by a particular task. Following this period, participants were given an initial questionnaire intended to gauge their prior knowledge of their first task. The questionnaire combined a self-assessment of their own prior knowledge and a list of three questions, ordered by increasing difficulty, to provide an objective assessment.

Following this questionnaire, participants were presented with the description of the task and were asked to complete the task using the interface (15 minutes per task). To capture data on participants' use of the interface, they were asked to provide reference sentences that indicate the information included in the essays was not purely based on their prior knowledge of the topic. After 15 minutes had elapsed, participants completed a post-task questionnaire that evaluated their experience of the task. We used questionnaires provided by TREC-9 Interactive Searching track [5] modified to fit our experiment. Additionally, we inquired about each participant's perceived assessment of the quality of the knowledge graphs in terms of the information they provided, including whether they noticed errors, inconsistencies, or missing information. The same process was repeated for the second task.

Following the completion of second task, a semi-structured interview was conducted to explore participants' experience of the interface. This interview focused on the conceptual usability and efficacy of the interface for different tasks as manifested in a participant's perception of task complexity, the strategies they employed to complete tasks, and their belief as to whether this type of interface supports these types of tasks. We also inquired about their perceptions of knowledge graph quality, the information they found using the graph, and whether they noted any differences in graph quality between the tasks (see Appendix Section **??**).

As a final step, participants were presented with a list of factoid questions on both topics to collect feedback on how their behaviour may change when performing question answering style tasks. Participants were given a small amount of additional time to familiarize themselves with the interface to complete this type of task. Once they were comfortable, they were asked to find the answer to one final question that was erroneous as a means to determine whether they would notice such an error during their information seeking process. We collected strategies and changes in quality perception when using the interface for question answering tasks (see Appendix **??** for more details on interviews).

[5] Available at www.nlpir.nist.gov/projects/t9i/qforms.html.

7.4.6 Behavioural Data Collection

Following Creswell [127], we collect data to perform a mixed methods analysis of both quantitative and qualitative factors. We log all participant interactions with the system, which includes the node or edge clicked, articles viewed, time spent reading articles, as well as more topological data (e.g., interactions with the UI itself, such as transitioning between layers or using the "minimap"). In addition, we captured field notes and audio during all participant sessions, transcribed the final interviews, and collated the questionnaire data. Quantitative data was analyzed using SPSS. Qualitative data was analysed using open coding to extract low-level themes and axial coding to identify thematic connections. This process was conducted incrementally as each participant completed the study and we attained qualitative saturation after 14 out of 25 participants. Sampling continued to ensure that participant set cardinality was sufficient to support statistical discrimination.

As participants were required to provide essays for each search task they completed, we had two independent assessors evaluate the quality of these answers. To ensure consistency with prior work, we reused the marking scheme provided by Sarrafzadeh et al. [479]. To aid in further analyses we normalized all scores to be in the range [0,1] and inter-assessor reliability was found to be 0.90 using the Pearson coefficient (i.e., reasonable agreement).

7.5 Quantitative Analysis

Because past work suggests that searchers are able to adapt their behavior to the system performance by controlling their effort (e.g. issuing more queries [504], reading more documents [550], allowing more interaction with the search results [510]) in the search process, our interaction logs can be used as a proxy for the effort the searcher is contributing during the search task and whether there are differences when comparing Auto and Gold conditions. Of particular interest to us was clicking on edges to read the relationship label (EdgeClick), viewing snippets to get the context around extracted relation label (Snippet), viewing articles to continue the exploration using the articles themselves (ViewArticle),

	Group 1 (HA - PG)		Group 2 (HG - PA)		Group 3 (PG - HA)		Group 4 (PA - HG)	
	Auto	Gold	Auto	Gold	Auto	Gold	Auto	Gold
Mark	0.52 (0.12)	0.56 (0.14)	0.72 (0.15)	0.66 (0.15)	0.40 (0.16)	0.65 (0.11)	0.65 (0.11)	0.64 (0.16)
Snippet	12.71 (6.7)	16.30 (12.30)	11.50 (5.80)	11.83 (6.77)	14.83 (3.97)	10.0 (2.83)	12.0 (8.55)	13.67 (5.72)
ViewArticle	3.0 (2.08)	4.0 (3.21)	0.67 (0.82)	1.5 (1.38)	3.17 (1.72)	1.0 (1.56)	1.67 (1.50)	1.5 (1.87)
Duration	169.57 (120.76)	189.58 (174.87)	53.83 (58.36)	19.33 (22.85)	163.33 (90.81)	44.0 (63.34)	69.0 (81.03)	78.17 (95.80)
EdgeClicks	20.0 (10.20)	27.86 (27.51)	25.5 (11.04)	21.83 (11.50)	27.0 (14.08)	27.67 (15.02)	30.5 (15.04)	21.67 (10.33)
Prior Knowledge	History	Politics	History	Politics	History	Politics	History	Politics
	1.0 (0.1)	0.14 (0.38)	1.0 (0.89)	0.5 (0.83)	0.67 (0.82)	0.0 (0.0)	1.3 (0.82)	0.83 (0.98)

Table 7.1: Contrasting relative performance and effort effects: Mean (Standard Deviation) values for dependent variables. Each group indicates the assignment of automatic or gold extraction (A or G) as well as the order of the Search Tasks (H for History and P for Politics).

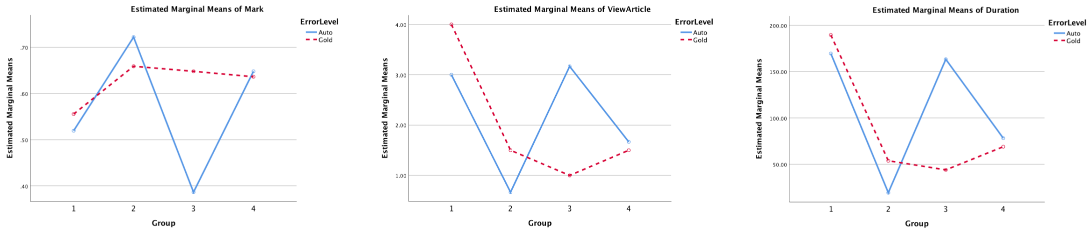
and overall time on task (Duration). Table 7.1 summarizes means (standard deviations) for these quantitative measures as well as the scoring of participant responses by independent evaluators (Mark) and estimates of their prior knowledge for each topic.

For each of our measures, we performed a sample size power estimate [234] to ensure that our sample size is sufficient to identify statistically significant differences in dependent variables. We used means, standard deviations and T-statistics at 0.975 in our power estimate to support a two-tailed analysis of effects at the 95% confidence interval. We found that that sample sizes of between 12 and 18 were sufficient to analyze Mark, ViewArticle and SnippetViews. Our sample size of 25 exceeds this threshold.

We performed a repeated measures ANOVA with ErrorLevel (Auto versus Gold) as a within subject effect and Group as a random factor. The group variable encodes error level to topic assignment as well as the ordering of the tasks. For example, in Group 1 (HA - PG), participants performed the **H**istory task with **A**utomatically extracted information first and **P**olitics task with **G**old data next. Dependent variables were Mark, ViewArticle, Duration, Snippet and EdgeClick.

Overall RM-ANOVA indicated that there was no statistically significant effect of ErrorLevel on Mark ($p > 0.1$), as a measure of search performance, nor was there any interaction between Group and ErrorLevel. There was, however, a significant effect of the group as a random factor on Mark ($F_{3,25} = 4.808, p < 0.05, \eta^2 = 0.3$) and Duration

($F_{3,25} = 3.935, p < 0.05, \eta^2 = 0.4$). Post-hoc analysis, using Tukey correction, indicates that Group 1 and 3 had lower marks than Group 2; Group 4 was in the middle, not significantly different than any other group. As well, participants in Group 1 spent more time reading articles than participants in Group 2. In terms of view article, there was no significant effect of error level, but there was a significant Group x ErrorLevel interaction ($F_{3,25} = 4.269, p < 0.05, \eta^2 = 0.4$). Figure 7.3 demonstrates these trends in our data.



(a) Marks

(b) ViewArticle

(c) Duration

Figure 7.3: Effects of Error Level and Group on (a) Marks, (b) ViewArticle and (c) Duration.

7.6 Qualitative Findings

In order to triangulate quantitative data, we leverage our qualitative data to first understand how the designed exploratory search tasks were perceived by the participants and next dive deeper on characterizing exploratory search behavior with error prone knowledge graphs. The analysis of our qualitative interviews and think-alouds indicated that participants varied in their perception of complexity of the tasks while there were reasonable level of agreement on different attributes of each search task. Furthermore, different ways searchers can be impacted by error-prone knowledge graphs and the factors that help or hinder noticing these errors while engaged in complex information seeking activities were identified. Finally, we contrasted different types of search tasks at different levels of complexity and specificity of the information being sought and how they differ in the way they are impacted by errors in retrieved information.

The following subsections elaborate on these themes.

7.6.1 Characterizing the Designed Exploratory Search Tasks

The qualitative analysis of our post task interviews resulted in a rich characterization of the two exploratory search tasks and how they compare with respect to their perceived complexity as well as main distinguishing attributes. We found that our participants were split in their perception of the complexity of these two tasks and that a variety of factors such as prior knowledge of the searcher, interpretation of the task, and the strategies employed impacted their perception.

Participants who found the History task more difficult ([P1-5, P7, P9, P10, P15, P20]) mentioned a variety of reasons including *representation does not match the perceived domain structure* [P3, P10, P15], *Politics is a more familiar domain* [P5, P7, P10], *"needed a deeper exploration"* [P9], *"searching without a keyword"*[P3], *"more detailed fact finding"*[P6, P8, P13].

A different group of participants, however, perceived the Politics tasks to be more difficult [P4, P6, P11-13, P17-18, P21-22, P24-25]. The most common rationales included *"Politics task asks a higher level question and needs a higher level understanding of a domain"* [P8, P10], *"involves learning, complex reasoning and interpretation"* [P6, P8, P11, P13], *"is multi-faceted"* [P13], *"involves objectively learning about a domain that requires more time"* [P11, P18] and *"more unfamiliar and technical terminology"* [P13].

Furthermore, these two search tasks, although both exploratory, were perceived to possess different attributes. Overall, the Politics task was found to be more subjective, required more high level information and belonged to the Learn category of exploratory search tasks [352]. This task was also found to be a precision oriented task as exploration was done mostly at a higher level and learning a few accurate facts about how each president functions locally or globally could suffice for comparison. The History task, on the other hand, was seen as more objective and was characterized as more of a finding task that involves more detailed information and fact retrieval, closer to the Investigate category of exploratory search task [352]. This task also seemed to share attributes of recall oriented tasks where a higher number of relevant facts need to be retrieved to satisfy the search task.

> *"For the Politics task I thought that I had to learn something, whereas for the other one I didn't feel I need to learn anything. Like in order to make any reasonable statement I will have to actually learn something about these countries."* [P8]

One interesting distinguishing attribute was the notion of *subjectivity* that was incorporated into the Politics task description. All participants were unanimous in the observation that the Politics task, as opposed to the History task, is asking about a subjective topic, *the power of a president*, which can be interpreted and rationalized differently. This in turn impacted the perceived complexity of this task based on how participants defined power and whether or not they perceived this task as an "objective learning task" or a "subjective goal oriented search". We observed that participants who started the task without any predetermined outcome of the task (i.e. which president is more powerful) found the task more difficult as they first opted to objectively learn about the Politics of both countries in order to identify the aspects on which compare the presidents as well as find out which president is indeed more powerful. As well, to our surprise, the majority of the participants who found the Politics task more difficult had indeed more prior knowledge of this topic (i.e. were familiar with the politics of at least one of the two countries). One potential reason for this observation was that participants with higher prior knowledge of the topic were aware of more facets that can define power and they had higher expectations of what constitutes a reliable comparison between the two presidents. On the other hand, participants who interpreted this task as a goal oriented search with the mere objective of finding evidence that conforms to the president that they believed had more power, were more satisfied with their search outcome and as a result perceived the task to be easier to perform.

7.6.2 Perceived Quality of Generated Knowledge Graphs

The second theme that arose from our qualitative analysis explored how knowledge graphs generated at different levels of quality, as measured by precision and recall, were perceived by participants performing exploratory search tasks. In our post-tasks interviews the participants enganged in an in-depth discussion of how the knowledge graphs and the information they presented were perceived in terms of their quality, accuracy and whether or not they had a good coverage of the information they were expected to represent. Given that the automatically generated knowledge graphs had a recall as low as 0.32 and a precision of 0.60 (averaged over Politics and History topics), we expected this difference to be noticeable by the participants.

To our surprise, participants did not notice any significant quality differences between the knowledge graphs they interacted with across the two search tasks and they were not able to identify the automatically generated ones when were specifically asked to do so. Participants, when asked specifically about the quality of the information represented by the graphs and whether or not they noticed errors, inconsistencies, incomplete or unreadable sentences, missing information, etc, mentioned they found the graphs to be comparable across both tasks. While a few participants mentioned they noticed minor errors such as typos, or duplicates they didn't think graphs associated with one of the tasks were significantly better or worse than the other.

Next, participants were told that one of their tasks did use automatically generated knowledge graphs while the other used experts-curated data and were asked to guess which task was done with the automatically generated graphs. The majority of participants were still not able to identify the task with lower quality graphs. We received some comments regarding the information that they expected to see or the structure they expected the graphs to be represented with but wasn't the case for them.

> *I think if you come across contradictory knowledge when you're not confident about your prior knowledge then I would maybe try to accept both of them and try to justify them in my mind.* [P5]

7.6.3 Factors impacting the recognition of errors

The final theme from our qualitative interviews explored the factors that may help or hinder noticing of errors in automatically extracted information and contrasted the impact of these errors between lookup and exploratory search tasks. We identified two categories of factors that can influence the likelihood of noticing errors during information seeking activities: (a) Individual Differences; (b) Task Complexity and Scope. We elaborate on these two factors next.

7.6.3.1 Individual Differences or User Effects

The interview continued with discussing the reasons that might have hindered participants from noticing errors or lower quality of graphs for one of the tasks. We found that the intent of the user performing the task (i.e., information seeking versus validating information), their confidence in the task (i.e., their prior knowledge of the topic as well as their language competency) and different types of cognitive biases impact the likelihood of noticing errors.

> *"Now I can see that mostly when I'm reading sentences I look at keywords only. Especially when it's not written in my first language. And sometimes it feels like all English sentences should make sense, which is not the case in my first language, Persian. I do validate sentences as I read them in my first language!"*[P13]
> *"I think there are multiple factors to consider. One thing is about my confidence in some language. When I read sentences in English, I don't question the validity as often as I do when I read text in my first language. So I'm more focused on understanding what the sentence is trying to say and not so much on spotting potential grammar issues, typos or inconsistent information. [...] And for more technical texts it's a mix of unfamiliarity and lack of confidence in my own knowledge of the domain as well as a higher level of trust in the validity of information that makes me accept what I read as facts."*[P13]

> *"I think if you come across contradictory knowledge when you're not confident about your prior knowledge then I would maybe try to accept both of them and try to justify them in my mind. E.g. See "President is the highest authority in Iran" and also see "President answers to Supreme Leader". And think to myself probably President is the highest authority in a different sense and not like being the Head of State! So I would trust both but try to justify them somehow! So I would come up with my own rationale: maybe Supreme Leader's authority is different. E.g. only considering religious matters. So when I don't have any prior knowledge I'm more likely to just trust what I see."*[P5]

7.6.3.2 Task Effects

The final part of our interviews contrasted different characteristics of simple (e.g. LookUp or Factoid Question Answering tasks) against complex and exploratory search tasks (e.g. Learn or Investigate categories) and how they are impacted by errors. In order to enable the participants to compare simple and exploratory search tasks, they were presented with a list of factoid questions on the same History and Politics topics and were asked to use the system for a few minutes to find answers to these questions. Participants were encouraged to think-aloud and share their strategy for locating the specific answers to each question while manipulating the knowledge graphs presented by the UI.

Once participants performed a few question-answering tasks they were asked to compare between the characteristics of this task and the previous essay writing tasks that they had completed and whether this interface supports these two types of tasks differently. Alongside participant comments, data from our field notes and observations indicated that the impact of poor quality of extracted information is higher for lookup and factoid question answering tasks than in the exploratory search tasks. The sensemaking aspect of exploratory search tasks coupled with the time willing to be spent in these tasks seem to play a role in mitigating the impact of errors to some extent.

> *"I'm more likely to notice errors when I dig deeper, I'm investigating further and when I'm inclined to read the articles text. These are the case in the Complex Task."* [P9]
> *"[for a Complex task] I'm trying to get a big picture or form an idea, whereas here [in Simple task] I'm looking for specific information. It's like a 'ray of light' and a small error like that would have completely deflected it! Where as for the complex task I was 'shining a lot of light' so small errors could cast a little shadow but then you'll get lots of other beams of light! "* [P8]

As a final question, participants were given a factoid question to find an answer to which was particularly designed to direct the user towards erroneous information in edge labels that contained the answer to this question. Observing the participants as they came across inaccurate and misleading information, whether or not they would notice it and how it would impact their judgement of utility of the information enabled us to first confirm some of the previous factors that were mentioned as leading to noticing errors (e.g. prior knowledge, expectations, etc) more reliably, and second, better understand whether different search tasks at different levels of complexity are impacted by errors differently.

The question that was chosen for this final stage was from Politics of Iran and asked “Whether there is an authority in Iran which can dismiss the Supreme Leader?”. Since most of our participants, regardless of their initial knowledge of Politics of Iran had learned about the Politics of this country through their attempt at the Exploratory Politics tasks, and that the Supreme Leader of Iran seemed to be a very powerful political entity, they approached this question with higher prior knowledge as well as prior expectations and bias towards what the answer should be. As well, at this points our participants were aware of presence of errors in knowledge graphs.

The relation label that contained the answer to this question was corrupted due to a parser error which resulted in a semi-readable sentence which would misleadingly specify ‘Supreme Leader supervised Assembly of Experts’. While the snippet corresponding to this sentence would contain the accurate information as ”Supreme Leader of Iran is elected and supervised by the Assembly of Experts”.

About half of our participants read the relation label and dismissed it as it was not providing an answer to the question. The other half, however, read the label and then clicked on <View More> to see the snippet. Among this group some still did not notice the inconsistency between the information in the relation label and the snippet provided. A closer speculation of what they read and understood from these sentences led to interesting observations about how different types of cognitive bias had led to accepting erroneous information and not noticing the inconsistencies as a result of extraction errors.

In order to unpack this observation, in the next subsection, we provide some background on cognitive biases and the existing research on how these biases can impact human judgement, and in particular the performance of online information seeking activities. We end this section by reflecting on cognitive biases we observed during the conducted user study.

Bias .

Cognitive biases are defined as a pattern of deviation in judgement occurring in particular situations, where a deviation may refer to a difference from what is normatively expected, either by the judgement of people outside of the situation or by independently verifiable facts [551]. Biases play a central role in human judgement and decision making and span a number of different dimensions. As well, biases can be observed in information retrieval in situations where searchers seek or are presented with information that significantly deviates from the truth [594].

Prior work has identified many different types of bias. Biases are not necessarily impacting the human judgement and their decision making process negatively. For example, they can form information-processing shortcuts leading to more effective actions in a given context or enable faster decisions when timeliness is more valuable than accuracy [551, 594]. One type of bias, however, that can lead to observing irrational search behavior is the *confirmation bias*, which describes people's unconscious tendency to prefer confirmatory information [382]. During information seeking activities, this type of bias can make searchers to seek evidence that supports their hypothesis and disregard evidence that refutes it [295, 594].

As previously discussed, most participants when pressured in time and engaged in information seeking activities, tend to skim sentences and only read the keywords. These participants mentioned their eyes are only focused on the keywords while the rest of the sentence is framed in their mind. We observed that how different types of bias led this framing of the sentences to result in misjudging the utility of the information. While 'confirmation bias' led some of the participants to just accept 'Assembly of Experts' as yet another council that is supervised by the Supreme Leader, to our surprise, even some of the participants with very high prior knowledge of this topic who already knew the answer to this question did not notice the error in the sentence. When they were asked to describe their thought process we learned that higher prior knowledge can bias the framing of the sentences in a way that should make sense to the reader:

> *I actually read the sentence as "Supreme Leader is supervised by"! So I added 'is' and 'by' to the sentence myself. Maybe because I expected this? I mean I knew that Assembly of Experts supervises Supreme Leader. So I already knew the answer and*

> *I was only looking for some evidence to support my answer. So it's like my mind is framing the sentences or reading them in a way that I expect them to be!* [P15]

Research on anchoring and adjustment has shown that people typically perform little revision to their beliefs, especially if those beliefs are strongly held [551, 594]. Furthermore, it is well established that people have constraints on their ability to process information. In fact, information foraging theory [413], inspired by the idea of bounded rationality and satisficing, states that information seekers, will choose behaviors that optimize the utility of information gained as a function of interaction cost. We notice this bias more strongly in the participants who approached the exploratory Politics task with an intention of finding evidence to support the power of the president they believed is more powerful as well as in specific question answering tasks for mid-range and high prior knowledge participants.

7.7 Discussion

In this section, we synthesize results from both quantitative and qualitative data. From our quantitative results, we note that Group, rather than Error condition or Task, resulted in significantly different performance on exploratory search tasks, as highlighted by the dependent measure Mark.

To investigate this further, we looked more closely at any confounds within each group that could potentially impact the variance we see in search performance and behavior. Because of the structure of our experimental task, Group, alongside being a between subject factor (encoding differences between participant) also encodes topic to error assignment (as shown in Table 7.1), groups differ in which topic was assigned to Gold versus Auto in order to balance out topic effects. Given our measure of Group as a significant factor, it is possible that lower quality extractions (i.e., Auto condition) impacted the two search tasks differently. Our qualitative data provides some evidence that task was perceived differently by participants; however, we find that while participants were split in their perception of the complexity of the tasks, one task did not dominate another in terms of complexity.

To further probe task-error effects, we performed a Univariate ANOVA analysis on History and Politics marks separately and noticed a statistically significant effect of group on History marks ($F_{3,25} = 4.373, p < 0.05, \eta^2 = 0.4$) but not on the Politics marks. This distinction indicates that the performance of our participants were impacted by the error condition assigned for the History task but not for the Politics task even though both datasets were generated with comparable levels of precision and recall. Repeating the same analysis on History marks using two fixed factors ErrorEffect and OrderEffect indicated a highly significant effect of ErrorEffect ($F_{1,25} = 10.806, p < 0.005, \eta^2 = 0.34$). OrderEffect, however, had no statistically significant impact on History marks.

Our qualitative data highlights two possible explanations for this group effect. First, prior knowledge of participants could help them recover from the poor quality of the knowledge graphs in terms of the information they represent. Second, the History task was identified as a investigate-style exploratory search task, and this class of search task might be more impacted by recall rates than a Learn/Synthesize task.

To further probe this, we note that our pre-task questionnaires data provided an assessment of participant knowledge. Figure 7.4 shows a potential correlation between prior knowledge and overall performance (marks) on search tasks. To test this observation, we coded prior knowledge from our three questions onto a three-point scale. We performed a Univariate ANOVA with ErrorEffect and OrderEffect, as fixed factors, prior knowledge as a random factor, and History marks as a dependent variable. The results indicated no statistically significant impact of prior knowledge, ErrorEffect, and OrderEffect. There were also no significant interaction between Group factors and prior knowledge.

Synthesizing these observations, while we observe no initial effect of error on performance, combining qualitative data with post-hoc statistical analysis, we find some evidence that precision and recall rates may significantly impact one of our tasks, the history task, more than the politics task. This, potentially makes sense. Because the history task is an investigate task with, as noted by our participants in qualitative results, a set of answers that are targeted rather than open-ended, errors in precision and/or recall might result in concepts useful to the search task being omitted from the data set.

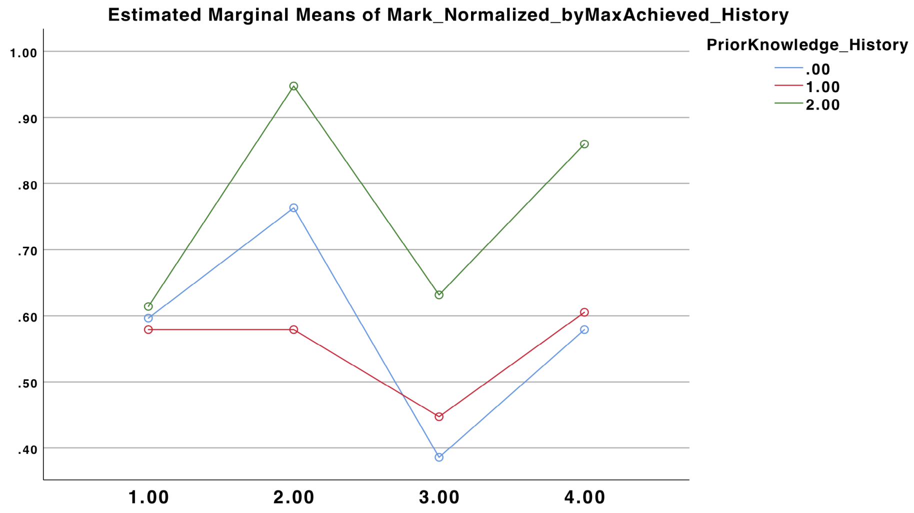


Figure 7.4: Effects of Prior Knowledge on Marks for History Task. x-axis indicates Group as a random factor and y-axis indicates Mark as a dependent variable.

7.8 Limitations and Future Work

We took significant care to address potential limitations in our study, by: performing power estimates in initial quantitative results to ensure a sufficient sample size for statistical effects to become visible; triangulating quantitative and qualitative data to cross-validate results; performing post-hoc quantitative analysis to validate qualitative themes that emerged from participants. However, one challenge with measuring the effect of IE errors is that, due to the need to understand how error impacts outcomes, tasks need both an experimental condition (automatically generated IE) and a control condition (manually corrected IE). This limits the number of tasks to those with manually generated ground truth. We control for this by cross-validating our results in ground truth output with past results and leveraging the same tasks.

However, one obvious area of future work is to add additional exploratory search tasks from Marchionini's taxonomy [352]. Our emergent qualitative results indicate that, within the broad category of exploratory search tasks, different types of exploratory search may be impacted differently by errors. Understanding where and how the current levels of precision for IE algorithms impact each of these different task types will help to clarify where and how useful current levels of IE accuracy are for different types of tasks.

Moreover, we only have investigated this problem in the context of HKGs due to our ability to replicate prior results in a consistent manner. While different UIs may yield different interaction behaviours, HKGs that belong to the class of search systems that directly organize and present search results using spatial representations [611], were designed to combine alternative interaction paradigms [495]: Users can exploit the overview layers to explore the collection at a higher level, which can be followed by targeted immersion in the detailed view.

Accordingly, we believe that the behaviours observed in our study are reflective of similar systems that seek to support *search by navigation* through developing search interfaces that allow the user to interact with and explore the spatial representations of the search results. While we cannot suggest that our results would apply to other classes of search interfaces (e.g., systems that provide a classification of the results using different metadata or facets), our findings encourage further investigation to test this possibility.

Alongside this, anther avenue of future work is to investigate other potential factors that may affect user performance in search tasks. For example, cognitive biases can result in irrational search behavior and influence searchers' relevance judgment of information [594]. As well, bounded rationality impacts the way information seekers optimize their information processing efforts even at the cost of achieving a sub-optimal outcome [413]. Our qualitative data provides some evidence that biases might be salient: participants who approached the exploratory Politics task with an intention of finding evidence to support the power of the president they believed is more powerful limited browsing behaviors because participants felt 'already informed' on the topic.

7.9 Chapter Summary

In the field of information retrieval, precision and recall measures are the gold standard by which algorithms are evaluated. Given that the information retrieval domain has yet to realize perfect precision, this chapter explores the impact of imperfect retrieval on two different information seeking tasks: an investigative search task and a learning oriented task. Through a mixed methods analysis, we find, first, limited statistical impact of error on task outcome measures of exploratory search, despite imperfect information extraction. Second, we find through qualitative and follow-on statistical analysis, potential task type effects. Together these results can inform both the design of user-facing exploratory search tools and provide a roadmap for evaluation of user-facing information retrieval systems.

Chapter 8

Conclusion

What is success? I think it is a mixture of having a flair for the thing that you are doing; knowing that it is not enough, that you have got to have hard work and a certain sense of purpose.

– Margaret Thatcher

In this chapter, we discuss the implications from previous chapters, and revisit the research questions that we proposed in Chapter 1. Finally, avenues for future research are discussed.

8.1 Overview

The research described in this thesis is inspired by the premise that exploratory search tasks require sensemaking and sensemaking involves constructing and interacting with representations of knowledge.

> *Most cognitive scientists believe, learning best begins with a big picture, a schema, a holistic cognitive structure, which should be included in the lesson material-often in the text. If a big picture resides in the text, the designers' task becomes one of emphasizing it. If this big picture does not exist, the designers' task is to develop a big picture and emphasize it. (West, Farmer and Wolff, 1991, p. 5) [592].*

Given the benefits of structuring information in supporting analysis and sensemaking, a main research question we wished to explore in this thesis was that whether spatial representations of information that are generated automatically and correspond to a searcher's information need can enable more effective extraction and assimilation of information?

To explore the impact of alternative representations of search results on user behavior during exploratory search tasks, we stated that:

> *Supporting complex and exploratory search tasks requires designing search systems that move beyond the current query-response paradigm in three main directions: (1) algorithms that move beyond document retrieval and provide information relevant to a user's query; (2) interfaces that move beyond a turn-taking interaction with a ranked list of documents and provide richer representations of the search results, as well as (3) mechanisms for accessing and interacting with them in ways that support exploration and sensemaking.*

To defend this thesis statement, this dissertation presented a series of discrete research endeavours, with each chapter exploring a different facet of the above claim. In Chapter 3 we addressed the problem of extracting information – that is more granular than documents – as a response to a user's query. Through a series of designing and evaluating alternative representations of search results in Chapters 4, 5 and 6, we looked at how this extracted information can be represented such that it extends the document-based search framework's support for exploratory search tasks. Finally, in Chapter 7 we probed the ecological validity of this research by exploring error-prone representations of search results and how they impact a searcher's ability to leverage these representations to acquire knowledge.

In the next section we revisit our primary research questions from Chapter 1 and address them based on the research presented in this dissertation.

8.1.1 Research Questions

[RQ1]. How can semantic information be automatically externalized as a response to a searcher's query?

There are two aspects to this research question: *externalization of the information* and the *connection to a searcher's query*. In Chapter 3 we specified two design requirements for the Information Retrieval and Extraction system we developed based on these two aspects: (1) The extraction tool should be tailored to the search query submitted by the user; and (2) The output of the extraction tool is expected to support locating fragments of information by externalizing semantic relationships between different entities and concepts that are discussed in the textual content of retrieved documents.

To address these two requirements, we developed an Information Extraction (IE) component as an extension to a Search and Document Retrieval (SDR) component such that the IE algorithm is applied directly to the output of the SDR component. Further, this IE component externalizes semantic information from the content of the retrieved documents (i.e. the output of the SDR component) as a series of entity-relationship tuples which can map directly to the semantic space of the documents content.

[RQ2]. How can this extracted semantic information be presented such that it extends the document-based search framework's support for information seeking activities?

In Chapter 4 we investigated this question at three levels:

1. *Design*: How can such a framework, that presents search results as both textual documents and their corresponding knowledge graphs, be envisioned?
2. *Searchers' Behavior*: How is such a framework used by the searchers? What are common behaviors? Are they impacted by the complexity of the search task?
3. *Outcome*: How successful searchers are in locating relevant information and completing search tasks once such a framework is provided?

We designed a search interface that couples documents with their corresponding knowledge graphs through bi-directional mappings between mentions of entities and semantic information in text to their corresponding nodes and edges in the graphs. Given this coupling the interface displays two alternative representations of search results, i.e. documents' text and their corresponding knowledge graphs representation side by side. We demonstrated that utilizing graphs of concepts and relationships that are derived from documents can be effective for finding relevant information, where salient entities and their neighbors in the graphs can assist with well defined goals and labeled edges describing how different concepts are related are used more frequently during complex search tasks. Probing the relatively limited support provided by our framework for complex search tasks, we identified two main directions to pursue in this dissertation in order to extend the current framework and provide more support for complex and exploratory tasks: Providing both global and local views of knowledge graphs as well as enabling alternative interaction paradigms to smoothly transition between these two views.

Our findings in many ways highlighted the tension between two alternative approaches to search: Overview, filter, detail-on demand [495] versus Expand-from-known [562]. Essentially, we found that knowledge graphs, as low-level representations of entities and relationships in a domain of interest, seem to be beneficial for browsing the immediate context-graph around a specific node of interest and then expand the scope of exploration to other regions of the graph. We observed some instances of expand-from-known searching among our participants who started their exploration from query-nodes or the nodes in the graphs that they were the most familiar with. On the other hand, providing high level overviews of the entire graphs can assist the information seekers with obtaining a visual preview of how salient concepts of a domain are laid out and enable a step-by-step plan for exploration. These observations motivated a more in-depth research into the efficacy of other spatial representations of search results and how they compare with knowledge graphs in supporting exploratory search tasks. We address this question next.

[RQ3]. How do alternative spatial representations of information, that externalize semantic information in documents' content, fare in presenting results for exploratory search tasks?

To address this question, in Chapter 5 we contrasted two different representations of document information, hierarchies, designed to focus specifically on high-level overviews that were absent in our previous representation, versus knowledge graphs, with the goal of quantifying differences in user behavior, performance, and perception. Our findings highlighted the complementary nature of hierarchical structures and knowledge graphs as representations of search results. More specifically, hierarchies provide a breadth-first exploration of the information that allows the user to iteratively reduce confusion, obtain an overview, and slowly exploit detail. They thus provide a structured way to navigate from more general concepts to more fine grained data and are valuable when people feel a need to orient themselves. In contrast, network structures allow users to glean more information from the representation (document reading time is reduced), are more engaging, yield more control over exploration at the lower level of inter-concept relationships, and are more similar to one's mental model.

To elaborate, in our experiments hierarchies resulted in a greater need to read the document rather than find the information contained within the visualization, shown, in our log data, by more instances of reading documents, and a longer period of time reading documents. Specifically, participants read documents three times more frequently and spent almost ten times more time reading. While, our quantitative analysis indicated that networks allow people to glean concrete information from the representation rather than needing to extensively read individual documents, which resulted in statistically lower reading times and a better quality of provided responses (although not significantly better), our qualitative observations highlighted the strengths of hierarchical structures in fostering sense-making of the overall content and allowing searchers to develop a better view of the information space. In fact, we found that our participants were biased towards a hierarchical structure for broad learning of the task domain particularly when they had low prior knowledge.

Overall, our quantitative and qualitative findings broadened our understanding of strengths and weaknesses of each of these representations and argued that one representation is not better than the other in any subjective sense. Many of our participants expressed a need for combining both interfaces into one interface which enables switching between a global and a local view of the information space.

[RQ4-a]. How can we combine alternative representations of search results – specifically hierarchical and knowledge graph representations – into a unified structure where local and global views of the data are co-visualized and seamless transitions between these views are enabled?

As a first inquiry into the design of a hybrid structure that combines hierarchical and knowledge graph representations we addressed three design questions in Chapter 6:

1. How do we integrate network and hierarchical views into a single, seamless data structure?
2. How can both the global and the local view of a knowledge graph be co-visualized?
3. How can transitions between views be designed to maximize visualization stability?

To address the challenges of seamlessly integrating both hierarchies and networks into a single unified structure we were inspired by the qualitative benefits of hierarchical representations of search results we observed in Chapter 5. Essentially, hierarchies provide a breadth-first exploration of the information that allows the user to iteratively reduce confusion, obtain an overview, and slowly exploit detail. They thus provide a structured way to navigate from more general concepts to more fine grained data and are valuable when people feel a need to orient themselves. In order to realize this level by level immersion into the entities and relationships represented by the knowledge graphs we proposed an alternative to hierarchical structures through a multi-level view of a knowledge graph around central entities. These central entities are the nodes that have a higher connectivity to the rest of the graph. Given this approach, we designed HKGs such that our knowledge graphs would constitute the lower-level visualization of our data, and a hierarchy was then gleaned from and corresponded directly to the underlying knowledge graph.

[RQ4-b]. Given that we introduced Hierarchical Knowledge Graphs as a new structure that combines the hierarchical overviews with local context of knowledge graphs representations, a related question is whether this new structure preserves the strengths of the underlying representations?

To evaluate this question, we leveraged the hierarchy and knowledge graph representations we developed in Chapter 5 as control interfaces as well as the data sets we used to populate these representations. Next, we performed a mixed methods analysis of both quantitative and qualitative data we collected through our experiment. The quantitative analysis of interaction logs of our participants supported both of the following two hypotheses:

Hypotheses:

1. Hierarchical knowledge graphs result in fewer document views and less time spent reading documents than do hierarchical trees.
2. Hierarchical knowledge graphs exhibit statistically similar behaviors to Knowledge Graphs.

Findings:

1. Hierarchical Knowledge Graphs preserve the advantages of Knowledge graphs over hierarchical trees in both reading time and in document views. Focusing specifically on our hierarchical graph, we find that our hierarchical graph has statistically lower document views (61% fewer document views, on average) and time reading (90% less time reading documents) than does hierarchical trees and that its behavior is statistically indistinguishable from the prior observations of knowledge graph interfaces.
2. Value of hierarchy (i.e. higher levels of the HKG) for complex tasks: We verified the usage of the higher levels of our HKGs to ensure they were indeed utilized during the search tasks. We found that, in the Complex Task, sensemaking was split between global and minimap views of the hierarchy more equitably, i.e., the minimap is particularly useful during our complex tasks.
3. The hierarchical graph representation was found to provide more support for the Complex Task (i.e., more open ended and exploratory tasks such as essay writing or learning) versus Simple tasks (such as question answering and specific knowledge finding). This is promising given our earlier results in Chapter 4 indicating knowledge graphs can support look-up tasks and are not suitable for satisfying complex information needs.
4. Our results also highlighted the value of Minimap and Overviews as a *starting point*. Looking specifically at HKGs and complex tasks, the overview allowed participants to identify the central concepts of a domain at a glance and the size of the circles indicates their prominence in the corresponding article. This identification of central concepts was also linked to a perception of value of the MiniMap

as a starting or entry point into the topic of the document being examined.

5. We identified two main distinctions between the notion of hierarchies in Hierarchical Trees versus Hierarchical Knowledge Graphs and how they were perceived by our participants:
   - Imposing a Structure versus Open Exploration: One interesting perspective of the multi-layer graph representation which presents central concepts of a domain as an overview for each document is that it reflects the knowledge graph concepts. This reflection made it, for many participants, more flexible and exploratory, a window into the knowledge graph.
   - Self-Orienting or Relative Positioning: One main advantage of the Hierarchical Tree visualizations (Discussed in Chapter 5), was the explicit connections between nodes (categories or headings) in the representation. We found that these edges help in two ways: (1) At a glance, you can tell why a concept appeared in this overview, or in this domain. To whit, the hierarchical structure exists the way it does because of a human author's decision. (2) The Path from the root to each of these nodes in the Tree Layout can provide useful information on where a concept is positioned relative to the topic.

In summary, the primary goal of the research described in this chapter was to explore whether we could combine benefits from both knowledge graphs and hierarchies into one data structure for visualizing search results. We note that our hierarchical graphs significantly reduce documents read and reading time as compared to hierarchical trees and perform on par with knowledge graphs. We also provide evidence that the hierarchy is used by participants via analysis of interaction logs. Qualitative data from our participants does indicate that hierarchies grounded in tables-of-contents are more familiar, easier to follow, and more focused. This in turn helps users orient themselves in the data. The vetted nature of hierarchical tables-of-contents was also perceived to be an asset absent from our hierarchical knowledge graphs. The hierarchies in our knowledge graph were viewed slightly differently, as noted above, with a more quantitative perspective giving them a certain cachet with respect to the unbiased nature of topic selection.

A final issue to consider is whether any hierarchy might provide benefits.

While it may, one advantage of the hierarchy in our HKGs is its tight connection to the entities contained in a knowledge graph and the ease of automatically extracting the hierarchy through thresholding. Another advantage is flexibility: while we currently leverage only three levels – corpus, central concept, and knowledge graph – it is easy to generalize the hierarchy to an arbitrary number of thresholds depending on the complexity of the domain. We do not generalize the hierarchy in this work because, for a first experimental validation, there are a limited number of factors that can be assessed. However, future work can address more detailed inquiries into scalability to larger corpora, scalability to multilevel hierarchies, and contrasts with other hierarchies such as automatic clusters or user-specified facets.

To summarize, we find that our hierarchical knowledge graphs preserve many of the previously observed advantages of traditional knowledge graphs, i.e. fewer document views and reduced reading time. Alongside this, hierarchical knowledge graphs introduce an effective hierarchical representation into knowledge graphs.

[RQ5]. Given that IE algorithms are not perfect in the real world, how resilient are HKGs to these errors? And how do error-prone HKGs impact a searcher's ability to leverage these representations to perform exploratory search tasks?

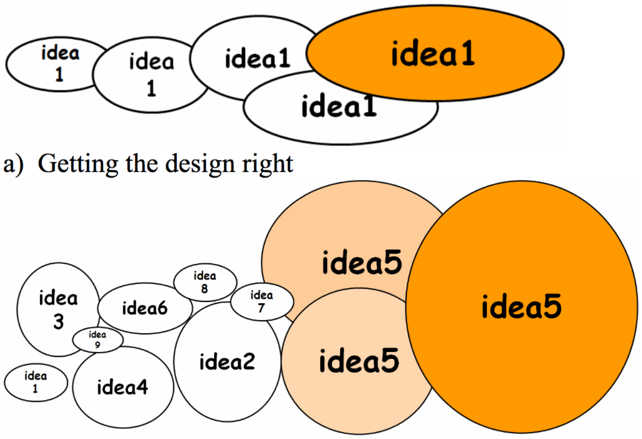
In Chapter 7 we took a step back and started investigating the question that given that IE algorithms are not perfect how practical HKGs are in the real world? And how do automatically generated HKGs impact a searcher's ability to leverage these representations to perform exploratory search tasks? To probe these questions, we performed a mixed methods analysis of the effect of precision and recall on the performance of hierarchical knowledge graphs for two different exploratory search tasks [476]. To this end, we leveraged the information extraction algorithm that we developed earlier in Chapter 3 and compared users behavior and outcomes with the erroneous output of the IE algorithm to users behavior and outcomes with the corrected outputs of the information extraction algorithm. While the quantitative data showed a limited effect of precision and recall on user performance and user effort, qualitative data triangulated with follow-on univariate analysis provided evidence that the type of exploratory search task (e.g., learning versus investigating) can be impacted by precision and recall. Furthermore, our qualitative analyses found that users were unable to perceive differences in the quality of extracted information. A more in-depth inquiry into the ways our participants perceived, verified and interacted with presented information resulted in an identification of different factors that can influence the likelihood of noticing errors during information seeking activities; Among them individual differences such as the level of prior knowledge or the cognitive biases as well as the complexity and scope of tasks impacted the searchers ability to recognize errors in representations the most. Together these results can inform both the design of user-facing exploratory search tools and provide a roadmap for evaluation of user-facing information retrieval systems.

8.2 Discussion, Limitations, and Future Work

We started this research with a vision of how search systems could enable searchers to learn, investigate, and make sense of the search results, essentially to perform exploratory search tasks more effectively. This vision foresaw the creation of novel, richer representations of search results that are supported by advanced interaction capabilities of the search UIs that contain them. As a result, these new representations and user interaction with these representations could fundamentally change the ways searchers explore information and synthesize new knowledge. In this section, we discuss our approach to system design and evaluation, limitations, and future work that arises from this thesis.

8.2.1 Understanding Alternative Representations of Search Results via Research Through Design

Our approach to realizing the vision of search systems that better enable exploratory search was inspired by two complementary established paradigms: 1) The Research Through Design Model [649] where the researcher focuses on making the right thing; artifacts intended to transform the world (of online information seeking) from the current state to a preferred state; and 2) Buxton's premise of "*getting the right design versus getting the design right*" [80, 543]. Essentially, the research



a) Getting the design right

b) Generating designs, choosing the right design, and then getting the design right

Figure 8.1: Getting the Right Design vs. Getting the Design Right; Figure adapted from Greenberg and Buxton [190]

through design model integrates the existing models and theories from the behavioral science domain ("true knowledge") with the technical opportunities demonstrated by engineers ("how knowledge") [190]. As such, we grounded our initial choice of knowledge graphs as a suitable representation in the large body of prior work in the fields of education and learning that demonstrated the efficacy of concept maps and knowledge graphs in supporting learning, comprehension of content, and locating relevant information [390, 246, 90, 391, 22, 185]. As well, through an active process of ideating, iterating, and critiquing potential designs for extending the textual representation of documents, we continually reframed the problem as an attempt to "make the right thing".

To elaborate, our first study, reported in Chapter 4, presented counter-intuitive results regarding the limited utility of knowledge graphs for supporting complex search tasks despite what prior work would suggest. Our usability evaluations indicated two main shortcomings of our knowledge graphs visualizations: lack of overviews and lack of support for expand-from-known mechanisms to interact with local, more focused views of the graphs. To address this counter-intuitive result, in Chapter 5, rather than correcting shortcomings of knowledge graphs, we, instead, explored alternative representations and contrasted differences in user behavior. This direction is inline with Greenberg and Buxton [190]'s view on the utility of exploring many possible early designs for illustrating the essence of an idea to “get the right design”, and, only afterwards, refining the design for a particular idea through iterative testing and development. In this regard, our characterization of complementary strengths of network and hierarchies resulted in realizing and evaluating a novel representation of search results that combines aspects of knowledge graphs representations (that externalize semantic relationships between concepts) as well as hierarchical structures (that provide overviews of the space and mechanisms for exploiting details in steps).

8.2.2 Evaluating Exploratory Search Systems

Another important aspect of the research that was conducted in this thesis lies in the evaluation methodologies that were employed to characterize different aspects of supporting exploratory search activities through alternative representations of search results. Essentially, we leverage mixed methods research to understand whether the exploratory search systems we design satisfy the ultimate goal of enabling searchers to obtain knowledge more effectively. By mixing both quantitative and qualitative analyses and data, we gain in breadth and depth of understanding the efficacy of different representations and how they are perceived by information seekers engaged in simple and exploratory search tasks. For example, in contrasting characteristics of network and hierarchical representations for supporting exploratory search tasks, while our knowledge graph representations exhibited statistical advantages over hierarchies resulting the reduction in time spent on reading document and the quality of answers provided, our qualitative findings highlighted the efficacy of hierarchical structures in fostering sensemaking and their utility for supporting broad learning tasks. Further, while our quantitative analysis supported our hypotheses that HKGs are statistically similar to knowledge graphs in their ability to reduce the need to read documents and that relevant information can be gleaned from the representation itself, our qualitative findings provided some insights into how the hierarchical views of HKGs are viewed differently by the searchers compared with hierarchical tree structures and and what factors might bias a searcher towards one representation versus the other.

Alongside applying Mixed Methods approaches for evaluating the search UIs we developed, we also contributed to the existing body of work that argues that a true evaluation of search systems needs to consider both the accuracy of outputs as well as the success of the search outcomes. Essentially, this research argues that in exploratory search tasks, both information retrieval and information representation are important factors in sensemaking. While retrieval is well-supported by modern search engines, the representation of retrieved information requires information extraction, and the efficacy of information extraction algorithms – as measured by precision and recall – is believed to be one of the most crucial aspects in creation of suitable information representation. While boosting the precision and/or recall of information extraction systems is valuable, it is unclear whether IE algorithms will attain human-like accuracy in the near term. Accordingly, given that IE algorithms are not perfect, how do current precision and recall levels in IE algorithms impact user performance in exploratory search tasks?

While our motivation for asking the above question was essentially an overall assessment of how HKGs can work in the real world given imperfect IE algorithms, there are three primary reasons that we feel the above is an open and significant question in the broader scope of information retrieval. First, in the document retrieval community, assessments of human performance in question answering tasks given varied precision and recall has produced mixed results, with some assessments finding that information retrieval precision and recall can significantly impact human performance [510, 469] and others finding no significant impact [222, 223, 549, 504]. Second, if, in exploratory search tasks, there is a need to support both query and browsing [397, 257], there is some evidence that interaction with search results (i.e., browsing) may limit the negative impacts of compromised precision and recall [332, 559]. Finally, while it would seem obvious that the goal of boosting precision and recall is primarily to satisfy user requirements with respect to IE – is the IE algorithm accurate enough for the user's task? – we have found only one examination of how error-prone IE impacts user performance [112], and the goal of this earlier evaluation was to assess the impact on document retrieval, not on exploratory search.

Overall, our evaluation approach described in chapter 7 makes two contributions to IE. First, it demonstrates an approach to evaluating IE that embeds IE in a task flow and evaluates IE performance in situ using a balanced mix of quantitative and qualitative (i.e., mixed) methods. While the approach to in situ IE evaluation is straight forward and both qualitative and quantitative analyses are mature research methods, we argue that this in situ evaluation is essential in ensuring that improvements in IE performance are meaningful. In particular, if one accepts that the goal of improved IE is to substantively improve task outcomes, then the fact that we have found only one instance of IE evaluation that considers user performance [112] is a significant concern, one that this thesis provides a roadmap for addressing in future work. Second, through our quantitative and, in particular, our qualitative results, this work highlights how different exploratory search task types (e.g., investigate vs compare/synthesize) may be more or less resilient to varied precision and recall. Understanding this is a first step toward addressing the contradictory effects that variations in precision and recall have had on information retrieval outcomes when those outcomes are measured via human performance metrics [222, 223, 332, 549, 504, 510, 469, 559].

8.2.3 Limitations and Future Work

There are two main aspects to the limitations of this research:

1. How generalizable our findings are given our experimental design and the evaluation methodologies we used to test our developed solutions?
2. How practical HKGs are to be deployed in real settings?

We expand on these two aspects in the next two subsections.

8.2.3.1 Generalizability of Findings

Regarding the validity and generalizability of the findings reported in this dissertation, we took significant care to control the impact of different confounds, such as characteristics of the search tasks, quality of extracted information, and idiosyncrasies of the search UI design through leveraging consistent tasks and datasets as well as cross-validating results across a set of reference interfaces. We also triangulated qualitative and quantitative data through the application of mixed methods approaches to better understand users' behavior and minimize bias. Additionally, we administered the same style of controlled experimentation using a tiered, 2-step evaluation methodology that decoupled the contributions of the accuracy of extraction systems from the efficacy of representations of this extracted information and the search UIs that contained these representations to the final outcomes of exploratory search tasks. We believe that this tiered model of evaluation is essential in the assessment of search solutions that support exploratory search as it incorporates the performance of two main components of such systems: Information retrieval and extraction module and the Search UI module.

This level of control, while necessary to address our research questions specified in Chapter 1, led to some limitations of our experiments. In particular, one challenge with any methodology that considers both an experimental condition (e.g. automatically generated knowledge graphs) and a control condition (manually refined data) is that the number of tasks and topics for which ground truth entity-relationship triples can be extracted is limited. Therefore, one obvious area of future work is to add additional types of search tasks and test our system with more topics. To elaborate, Marchionini [352] broadly categorizes different types of exploratory search tasks under two classes of Learn and Investigate (see Figure 2.1 for a list of all of these tasks). In particular, searches that support learning aim to achieve knowledge acquisition, comprehension of concepts, interpretation of ideas and comparisons or aggregation of concepts. Searches that support investigation, on the other hand, aim to achieve Bloom's [63] highest-level objectives such as analysis, synthesis, and evaluation and require substantial topical knowledge. These tasks involve elements of discovery, synthesis and evaluation.

In our experiments, we designed an instance of Learn category on the topic of Politics of Iran and Russia as well as an instance of Investigate category on the topic of History of Canada. Our post-task evaluations confirmed that our Politics task combined aspects of knowledge acquisition (e.g. learning about the political system of foreign countries), comprehension of concepts (e.g. different political entities) and comparison (of presidential powers), while our History task was more focused on discovery and finding reasons behind moving the capitals of Canada in line with Marchionini's characterization of Investigate style tasks. Our quantitative findings coupled with our emergent qualitative results, reported in Chapter 7, indicated that within the two broad categories of exploratory search tasks, different types of tasks may be impacted differently by errors. Understanding whether and how the current levels of precision and recall for IE algorithms impact each of these different task types will help to clarify how predictive the assessments of outputs of IE systems are for judging the success of the search tasks and what other factor might be also taken into account when evaluating the performance of the exploratory search systems. For example, in Chapter 7 we found some evidence that other characteristics of users including their prior knowledge and cognitive biases can help or hinder their ability in leveraging error-prone representations of search results. Coupling the searchers characteristics with the nature of the search tasks, our results provided some evidence that it is entirely possible that changes in precision and recall might have no effect on people's ability to learn or investigate topics beyond a certain level. Overall, these results encourages more research in this area and the need for developing better metrics that can measure real improvements to search interfaces that support exploratory search tasks.

8.2.3.2 Practicality of our Proposed Exploratory Search Framework

One open question that this thesis hasn't fully addressed is whether HKGs are now considered working prototypes. That is, can these representations be generated on the fly and deployed in the wild such that searchers can leverage them to satisfy a range of exploratory information needs.

To answer this question we need to reflect on the steps that were taken to reach this current stage of development and what is left to be addressed. As noted earlier in this chapter, we adapted a research through design [649] approach in order to address our second research question, i.e, *how changes in the ways that extracted semantic information is represented is going to impact the ways that searchers acquire knowledge, learn or investigate?* To this end, based on established theories of learning and education, we started with knowledge graphs as a promising representation to support learning and comprehension of topics and observed how they can be incorporated by search UIs in order to enable searchers perform information seeking tasks of varying complexity. Next, in keeping up with the research through design approach, as well as an attempt to "*choosing the right design first and then getting the design right*" [80], our earlier research questions gave rise to additional areas of exploration, to characterize the efficacy of competing representations of search results in supporting exploratory search tasks, and culminated in the design and evaluation of a novel representations, HKG, that combines and preserves the strengths of previously evaluated representations knowledge graphs and hierarchies.

In developing our HKGs, through a discovery-focused approach, we broadened our understanding of how search systems can be designed to incorporate alternative representations of search results and how they can be leveraged by searchers to interact with the information space in ways that were not possible before. We are now at the stage where we can transition from *discovery* to *invention*[1], as a process behind engineering and design [190], and can begin to take techniques – existing or new – that work in the theory or in a lab setting, and extend them to work in the complexity of the real world.

Our work in Chapter 7 is the first step towards this goal. By understanding how error-prone representations, as an inevitable outcome of imperfect IE algorithms, impact HKG's efficacy in supporting exploratory search goals we get one step closer to deploying these systems in the real world. Given our findings regarding the relative resilience of these representations to errors in the output of IE systems, we can begin generating these representations automatically, and for additional tasks and topics. Using these new types of exploratory search tasks we can conduct controlled evaluations of automatically generated HKGs and measure how successful searchers are in completing exploratory search tasks when compared against a standard search engine.

8.2.4 Synthesis

The work presented in this thesis exists at the intersection of a number of fields: natural language processing for information extraction; web-based information retrieval; information seeking; information visualization; and interaction design. Our primary focus, as noted in our thesis statement, was to explore how information extraction and representation of search results could support exploratory search tasks. This thesis

[1] Scott Hudson at the ACM UIST 2007 Panel on Evaluating Interface Systems Research

advances the field through a characterization of the benefits of hierarchical versus network visualizations of search results, through the design of hierarchical knowledge graphs, and through the careful design and evaluation of systems for supporting exploratory search. Our hope is that contributes to the goal of supporting ever-more-complex information seeking behaviours.

References

[1] Android Studio.

[2] Smart Profile, 2010.

[3] Education and marriage, 2011.

[4] MotoX, 2016.

[5] Distribution of employment income of individuals by sex and work activity, Canada, provinces and selected census metropolitan areas, 2017.

[6] Flicktek, 2018.

[7] Leap Motion, 2018.

[8] Myo arm band, 2018.

[9] TinyMCE, 2018.

[10] Johnny Accot and Shumin Zhai. More Than Dotting the I's — Foundations for Crossing-based Interfaces. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, CHI '02, pages 73–80, New York, NY, USA, 2002. ACM.

[11] Naveed Afzal, Ruslan Mitkov, and Atefeh Farzindar. Unsupervised relation extraction using dependency trees for automatic generation of multiple-choice questions. In *Advances in Artificial Intelligence*, pages 32–43. Springer, 2011.

[12] David Ahlström, Khalad Hasan, and Pourang Irani. Are You Comfortable Doing That?: Acceptance Studies of Around-device Gestures in and for Public Settings. In *Proceedings of the 16th International Conference on Human-computer Interaction with Mobile Devices &#38; Services*, MobileHCI '14, pages 193–202, New York, NY, USA, 2014. ACM.

[13] Jae-wook Ahn and Peter Brusilovsky. Adaptive visualization for exploratory information retrieval. *Information Processing & Management*, 49(5):1139–1164, 2013.

[14] Roland Aigner, Daniel Wigdor, Hrvoje Benko, Michael Haller, David Lindbauer, Alexandra Ion, Shengdong Zhao, and Jeffrey Tzu Kwan Valino Koh. Understanding Mid-Air Hand Gestures: A Study of Human Preferences in Usage of Gesture Types for HCI. Technical report, 11 2012.

[15] Azzah Al-Maskari, Mark Sanderson, Paul Clough, and Eija Airio. The good and the bad system: does the test collection predict users' effectiveness? In *Proceedings of SIGIR*, pages 59–66. ACM, 2008.

[16] Harini Alagarai Sampath, Rajeev Rajeshuni, and Bipin Indurkhya. Cognitively inspired task design to improve user performance on crowdsourcing platforms. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, pages 3665–3674, New York, NY, USA, 2014. ACM.

[17] Daniel C Alexander and James C Gee. Elastic matching of diffusion tensor images. *Computer Vision and Image Understanding*, 77(2):233–250, 2000.

[18] Anwar Alhenshiri, Carolyn Watters, Michael Shepherd, and Jack Duffy. Building support for web information gathering tasks. In *System Science (HICSS), 2012 45th Hawaii International Conference on*, pages 1687–1696, 2012.

[19] Ali Mohamed Nabil Allam and Mohamed Hassan Haggag. The question answering systems: A survey. *International Journal of Research and Reviews in Information Sciences (IJRRIS)*, 2(3), 2012.

[20] James Allan, Ben Carterette, and Joshua Lewis. When will information retrieval be "good enough"? In *Proceedings of SIGIR '05*, pages 433–440. ACM, 2005.

[21] James Allan, Bruce Croft, Alistair Moffat, and Mark Sanderson. Frontiers, challenges, and opportunities for information retrieval: Report from swirl 2012. In *ACM SIGIR Forum*, volume 46, pages 2–32. ACM, 2012.

[22] Franck Amadieu and Ladislao Salmerón. Concept maps for comprehension and navigation of hypertexts. In *Digital Knowledge Maps in Education*, pages 41–59. Springer, 2014.

[23] Franck Amadieu, André Tricot, and Claudette Mariné. Interaction between prior knowledge and concept-map structure on hypertext comprehension, coherence of reading orders and disorientation. *Interacting with computers*, 22(2):88–97, 2010.

[24] John Robert Anderson. *Learning and memory: An integrated approach.* John Wiley & Sons Inc, 2000.

[25] Lisa Anthony and Jacob O Wobbrock. $N-protractor: A Fast and Accurate Multistroke Recognizer. In *Proceedings of Graphics Interface 2012*, GI '12, pages 117–120, Toronto, Ont., Canada, Canada, 2012. Canadian Information Processing Society.

[26] Shaikh Shawon Arefin Shimon, Courtney Lutton, Zichun Xu, Sarah Morrison-Smith, Christina Boucher, and Jaime Ruiz. Exploring Non-touchscreen Gestures for Smartwatches. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, CHI '16, pages 3822–3833, New York, NY, USA, 2016. ACM.

[27] David Armstrong, Ann Gosling, John Weinman, and Theresa Marteau. The place of inter-rater reliability in qualitative research: an empirical study. *Sociology*, 31(3):597–606, 1997.

[28] Daniel Ashbrook, Patrick Baudisch, and Sean White. Nenya: Subtle and Eyes-free Mobile Input with a Magnetically-tracked Finger Ring. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 2043–2046, New York, NY, USA, 2011. ACM.

[29] Daniel Ashbrook and Thad Starner. MAGIC: A Motion Gesture Design Tool. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, CHI '10, pages 2159–2168. ACM, 2010.

[30] Kumaripaba Athukorala, Dorota Głowacka, Giulio Jacucci, Antti Oulasvirta, and Jilles Vreeken. Is exploratory search different? a comparison of information search behavior for exploratory and lookup tasks. *Journal of the Association for Information Science and Technology*, 67(11):2635–2651, 2016.

[31] Anne Aula and Daniel M Russell. Complex and exploratory web search. In *Information Seeking Support Systems Workshop (ISSS 2008), Chapel Hill, NC, USA*, 2008.

[32] David P Ausubel. The psychology of meaningful verbal learning. 1963.

[33] David Paul Ausubel, Joseph Donald Novak, Helen Hanesian, et al. Educational psychology: A cognitive view. 1968.

[34] Alan Baddeley. Working memory. *Science*, 255(5044):556–559, 1992.

[35] Ricardo Baeza-Yates, Berthier Ribeiro-Neto, et al. *Modern information retrieval*, volume 463. ACM press New York, 1999.

[36] Peter Bailey, Nick Craswell, Ian Soboroff, Paul Thomas, Arjen P de Vries, and Emine Yilmaz. Relevance assessment: are judges exchangeable and does it matter. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 667–674. ACM, 2008.

[37] Gilles Bailly, Jörg Müller, Michael Rohs, Daniel Wigdor, and Sven Kratz. ShoeSense: A New Perspective on Gestural Interaction and Wearable Applications. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, CHI '12, pages 1239–1248, New York, NY, USA, 2012. ACM.

[38] Rachel Bainbridge and Joseph A Paradiso. Wireless Hand Gesture Capture through Wearable Passive Tag Sensing. In *2011 International Conference on Body Sensor Networks*, pages 200–204, 5 2011.

[39] Niranjan Balasubramanian, Stephen Soderland, Oren Etzioni, et al. Generating coherent event schemas at scale. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1721–1731, 2013.

[40] Michele Banko, Michael J Cafarella, Stephen Soderland, Matthew Broadhead, and Oren Etzioni. Open information extraction from the web. In *IJCAI*, volume 7, pages 2670–2676, 2007.

[41] Leon Barnarda, Ji Soo Yia, Julie A Jackoa, and Andrew Sears. An empirical comparison of use-in-motion evaluation scenarios for mobile computing devices. *International Journal of Human-Computer Studies*, 62(4):487–520, 2005.

[42] Joel F Bartlett. Rock 'N' Scroll Is Here to Stay. *IEEE Computer Graphics and Applications*, 20(3):40–45, 5 2000.

[43] Feza Baskaya, Heikki Keskustalo, and Kalervo Järvelin. Time drives interaction: simulating sessions in diverse searching environments. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 105–114, 2012.

[44] Marcia J Bates. The design of browsing and berrypicking techniques for the online search interface. *Online Information Review*, 13(5):407–424, 1989.

[45] Marcia J Bates. The design of browsing and berrypicking techniques for the online search interface. *Online review*, 13(5):407–424, 1989.

[46] Marcia J Bates. What is browsing-really? a model drawing from behavioural science research, 2007.

[47] Giuseppe Di Battista, Peter Eades, Roberto Tamassia, and Ioannis G Tollis. *Graph drawing: algorithms for the visualization of graphs*. Prentice Hall PTR, 1998.

[48] Asker M Bazen and Sabih H Gerez. Fingerprint matching by thin-plate spline modelling of elastic deformations. *Pattern Recognition*, 36(8):1859–1867, 2003.

[49] Steven M Beitzel, Eric C Jensen, Abdur Chowdhury, David Grossman, and Ophir Frieder. Hourly analysis of a very large topically categorized web query log. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 321–328. ACM, 2004.

[50] Nicholas J Belkin. Anomalous states of knowledge as a basis for information-retrieval. *Canadian Journal of Information Science-Revue Canadienne Des Sciences De L Information*, 5(MAY):133–143, 1980.

[51] Nicholas J Belkin. Some (what) grand challenges for information retrieval. In *ACM SIGIR Forum*, volume 42, pages 47–54. ACM New York, NY, USA, 2008.

[52] Nicholas J Belkin et al. Interaction with texts: Information retrieval as information seeking behavior. *Information retrieval*, 93:55–66, 1993.

[53] Janine Berg. Income support for the unemployed and the poor. In *Labour Markets, Institutions and Inequality*, Chapters, chapter 10, pages 263–286. Edward Elgar Publishing, 2015.

[54] Joanna Bergstrom-Lehtovirta, Antti Oulasvirta, and Stephen Brewster. The Effects of Walking Speed on Target Acquisition on a Touchscreen Interface. In *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services*, MobileHCI '11, pages 143–146, New York, NY, USA, 2011. ACM.

[55] Eugen Berlin, Jun Liu, Kristof van Laerhoven, and Bernt Schiele. Coming to Grips with the Objects We Grasp: Detecting Interactions with Efficient Wrist-worn Sensors. In *Proceedings of the Fourth International Conference on Tangible, Embedded, and Embodied Interaction*, TEI '10, pages 57–64, New York, NY, USA, 2010. ACM.

[56] Daniel E Berlyne. Conflict, arousal, and curiosity. 1960.

[57] Ana M Bernardos, David Gómez, and José R Casar. A Comparison of Head Pose and Deictic Pointing Interaction Methods for Smart Environments. *International Journal of Human–Computer Interaction*, 32(4):325–351, 2016.

[58] Donald J Berndt and James Clifford. Using Dynamic Time Warping to Find Patterns in Time Series. In *KDD workshop*, volume 10, pages 359–370. Seattle, WA, 1994.

[59] Roman Bertolami, Matthias Zimmermann, and Horst Bunke. Rejection Strategies for Offline Handwritten Text Line Recognition. *Pattern Recognition Letters*, 27(16):2005–2012, 12 2006.

[60] Hugh Beyer and Karen Holtzblatt. *Contextual design: defining customer-centered systems*. Elsevier, 1997.

[61] Jeffrey P Bigham, Michael S Bernstein, and Eytan Adar. Human-computer interaction and collective intelligence. *Handbook of Collective Intelligence*, 2015.

[62] Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked data-the story so far. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 5(3):1–22, 2009.

[63] Benjamin S Bloom. Taxonomy of educational objectives: Handbook i: Cognitive domain. *New York: David McKay*, 1956.

[64] Francoise Boch and Annie Piolat. Note taking and learning: A summary of research. *The WAC Journal*, 16:101–113, 2005.

[65] Richard Bolt. "Put-that-there": Voice and gesture at the graphics interface. In *Proceedings of the 7th annual conference on Computer graphics and interactive techniques - SIGGRAPH '80*, pages 262–270, 1980.

[66] Pia Borlund. The iir evaluation model: a framework for evaluation of interactive information retrieval systems. *Information research*, 8(3):8–3, 2003.

[67] Ria Mae Borromeo and Motomichi Toyama. An Investigation of Unpaid Crowdsourcing. *Hum.-centric Comput. Inf. Sci.*, 6(1):68:1–68:19, 12 2016.

[68] Florian Boudin. A comparison of centrality measures for graph-based keyphrase extraction. In *International Joint Conference on Natural Language Processing (IJCNLP)*, pages 834–838, 2013.

[69] Gerlof Bouma. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL*, pages 31–40, 2009.

[70] Jonathan Bragg and Daniel S Weld. Sprout: Crowd-powered task design for crowdsourcing. In *ACM Symposium on User Interface Software and Technology (UIST'18)*, 2018.

[71] T T Brewer, D S Hoeger, L K McCambridge, T L Kelsey, A R Claflin, K R Robertson, and M W Van Flandern. Method and system for activating double click applications with a single click, 1997.

[72] Stephen Brewster. Overcoming the Lack of Screen Space on Mobile Computers. *Personal Ubiquitous Comput.*, 6(3):188–205, 1 2002.

[73] Stephen Brewster and Lorna M Brown. Tactons: Structured Tactile Messages for Non-visual Information Display. In *Proceedings of the Fifth Conference on Australasian User Interface - Volume 28*, AUIC '04, pages 15–23, Darlinghurst, Australia, Australia, 2004. Australian Computer Society, Inc.

[74] Stephen Brewster, Joanna Lumsden, Marek Bell, Malcolm Hall, and Stuart Tasker. Multimodal 'Eyes-free' Interaction Techniques for Wearable Devices. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, CHI '03, pages 473–480, New York, NY, USA, 2003. ACM.

[75] Andrei Broder. A taxonomy of web search. In *ACM Sigir forum*, volume 36, pages 3–10. ACM, 2002.

[76] Marc Bron, Jasmijn Van Gorp, Frank Nack, Maarten de Rijke, Andrei Vishneuski, and Sonja de Leeuw. A subjunctive exploratory search interface to support media studies researchers. In *Proc. of SIGIR*, pages 425–434, 2012.

[77] Mary E Brown. A general model of information-seeking behavior. In *Proceedings of the ASIS Annual Meeting*, volume 28, pages 9–14. ERIC, 1991.

[78] John M Budd. Relevance: Language, semantics, philosophy. 2004.

[79] Michael Buhrmester, Tracy Kwang, and Samuel D. Gosling. Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data? *Perspectives on Psychological Science*, 6(1):3–5, 1 2011.

[80] Bill Buxton. *Sketching user experiences: getting the design right and the right design.* Morgan kaufmann, 2010.

[81] Tony Buzan and Barry Buzan. How to use radiant thinking to maximize your brain's untapped potential, 1996.

[82] Katriina Bystr`Om. Information and information sources in tasks of varying complexity. *JASIST*, 53(7):581–591, 2002.

[83] Katriina Byström and Preben Hansen. Conceptual framework for tasks in information studies. *JASIST*, 56(10):1050–1061, 2005.

[84] Katriina Byström and Kalervo Järvelin. Task complexity affects information seeking and use. *Information processing & management*, 31(2):191–213, 1995.

[85] Carrie J. Cai, Shamsi T. Iqbal, and Jaime Teevan. Chain reactions: The impact of order on microtask chains. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 3143–3154, New York, NY, USA, 2016. ACM.

[86] Donald J Campbell. Task complexity: A review and analysis. *Academy of management review*, 13(1):40–52, 1988.

[87] Robert Capra, Gary Marchionini, Jung Sun Oh, Fred Stutzman, and Yan Zhang. Effects of structure and interaction style on distinct search tasks. In *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, pages 442–451. ACM, 2007.

[88] Stuart K Card, Jock Mackinlay, and Ben Shneiderman. *Readings in information visualization: using vision to think.* Morgan Kaufmann, 1999.

[89] Stuart K Card, Peter Pirolli, Mija Van Der Wege, Julie B Morrison, Robert W Reeder, Pamela K Schraedley, and Jenea Boshart. Information scent as a driver of web behavior graphs: results of a protocol analysis method for web usability. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 498–505, 2001.

[90] Mary Carnot, Paul Feltovich, Robert Hoffman, Joan Feltovich, and Joseph Novak. A summary of literature pertaining to the use of concept mapping techniques and technologies for education and performance support. 2003.

[91] Mary Jo Carnot, Bruce Dunn, and Alberto J Cañas. Concept map-based vs web page-based interfaces in search and browsing. In *ICTE Tallahassee 2001. International Conference on Technology and Education. Proceedings*, page 183, 2001.

[92] Peter J Carrington, John Scott, and Stanley Wasserman. *Models and methods in social network analysis*, volume 28. Cambridge university press, 2005.

[93] Tom Carter, Sue Ann Seah, Benjamin Long, Bruce Drinkwater, and Sriram Subramanian. UltraHaptics: Multi-point Mid-air Haptic Feedback for Touch Surfaces. In *Proceedings of the 26th Annual ACM Symposium on User Interface Software and Technology*, UIST '13, pages 505–514, New York, NY, USA, 2013. ACM.

[94] Lina M Castano and Alison B Flatau. Smart fabric sensors and e-textile technologies: a review. *Smart Materials and Structures*, 23(5):053001, 5 2014.

[95] Jessica R Cauchard, Janette L Cheng, Thomas Pietrzak, and James A Landay. ActiVibe: Design and Evaluation of Vibrations for Progress Monitoring. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, CHI '16, pages 3261–3271, New York, NY, USA, 2016. ACM.

[96] Edwin Chan, Teddy Seyed, Wolfgang Stuerzlinger, Xing-Dong Yang, and Frank Maurer. User Elicitation on Single-hand Microgestures. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, CHI '16, pages 3403–3414, New York, NY, USA, 2016. ACM.

[97] Liwei Chan, Yi-Ling Chen, Chi-Hao Hsieh, Rong-Hao Liang, and Bing-Yu Chen. CyclopsRing: Enabling Whole-Hand and Context-Aware Interactions Through a Fisheye Ring. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software and Technology*, UIST '15, pages 549–556, New York, NY, USA, 2015. ACM.

[98] Liwei Chan, Rong-Hao Liang, Ming-Chang Tsai, Kai-Yin Cheng, Chao-Huai Su, Mike Y Chen, Wen-Huang Cheng, and Bing-Yu Chen. FingerPad: Private and Subtle Interaction Using Fingertips. In *Proceedings of the 26th Annual ACM Symposium on User Interface Software and Technology*, UIST '13, pages 255–260, New York, NY, USA, 2013. ACM.

[99] Dana Chandler and Adam Kapelner. Breaking monotony with meaning: Motivation in crowdsourcing markets. *Journal of Economic Behavior & Organization*, 90:123–133, 6 2013.

[100] Joseph Chee Chang, Saleema Amershi, and Ece Kamar. Revolt: Collaborative crowdsourcing for labeling machine learning datasets. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 2334–2346. ACM, 2017.

[101] SJ Chang and MH Ko. Behaviors of pim in context of thesis and dissertation research. In *CHI 2008 workshop*, 2008.

[102] Duen Horng Chau, Aniket Kittur, Jason I Hong, and Christos Faloutsos. Apolo: interactive large graph sensemaking by combining machine learning and visualization. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 739–742, 2011.

[103] Hsinchun Chen, Andrea L Houston, Robin R Sewell, and Bruce R Schatz. Internet browsing and searching: User evaluation of category map and concept space techniques. *Journal of the American Society for Information Science, Special Issue on AI Techniques for Emerging Information Systems Applications*, 1998.

[104] Ke-Yu Chen, Kent Lyons, Sean White, and Shwetak Patel. uTrack: 3D Input Using Two Magnetic Sensors. In *Proceedings of the 26th Annual ACM Symposium on User Interface Software and Technology*, UIST '13, pages 237–244, New York, NY, USA, 2013. ACM.

[105] Xiang 'Anthony' Chen, Julia Schwarz, Chris Harrison, Jennifer Mankoff, and Scott E Hudson. Air+Touch: Interweaving Touch &#38; In-air Gestures. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology*, UIST '14, pages 519–525, New York, NY, USA, 2014. ACM.

[106] Justin Cheng, Jaime Teevan, and Michael S Bernstein. Measuring crowdsourcing effort with error-time curves. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 1365–1374. ACM, 2015.

[107] Justin Cheng, Jaime Teevan, and Michael S. Bernstein. Measuring crowdsourcing effort with error-time curves. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 1365–1374, New York, NY, USA, 2015. ACM.

[108] Ed Huai-hsin Chi. A taxonomy of visualization techniques using the data state reference model. In *IEEE Symposium on Information Visualization 2000. INFOVIS 2000. Proceedings*, pages 69–75. IEEE, 2000.

[109] Yejin Choi, Eric Breck, and Claire Cardie. Joint extraction of entities and relations for opinion recognition. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 431–439. Association for Computational Linguistics, 2006.

[110] Chun Wei Choo, Brian Detlor, and Don Turnbull. Information seeking on the web–an integrated model of browsing and searching. 1999.

[111] Janara Christensen, Stephen Soderland, Oren Etzioni, et al. Towards coherent multi-document summarization. In *Proceedings of NAACL*, pages 1163–1173, 2013.

[112] Jennifer Chu-Carroll and John Prager. An experimental study of the impact of information extraction accuracy on semantic search performance. In *Proceedings of CIKM*, pages 505–514. ACM, 2007.

[113] Kenneth W Church and William A Gale. Enhanced good-turing and cat-cal: two new methods for estimating probabilities of english bigrams. In *Proceedings of the workshop on Speech and Natural*

Language, pages 82–91. Association for Computational Linguistics, 1989.

[114] Kenneth Ward Church and Patrick Hanks. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29, 1990.

[115] Charles LA Clarke, Maheedhar Kolla, Gordon V Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, and Ian MacKinnon. Novelty and diversity in information retrieval evaluation. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 659–666, 2008.

[116] James Clarke, Vivek Srikumar, Mark Sammons, and Dan Roth. An nlp curator (or: How i learned to stop worrying and love nlp pipelines). In *LREC*, pages 3276–3283, 2012.

[117] Aaron Clauset, Cristopher Moore, and Mark EJ Newman. Hierarchical structure and the prediction of missing links in networks. *Nature*, 453(7191):98–101, 2008.

[118] Cyril W Cleverdon, Jack Mills, and E Michael Keen. Factors determining the performance of indexing systems,(volume 1: Design). *Cranfield: College of Aeronautics*, page 28, 1966.

[119] Andy Cockburn and Steve Jones. Which way now? analysing and easing inadequacies in www navigation. *International Journal of Human-Computer Studies*, 45(1):105–129, 1996.

[120] Simon Colton, Alan Bundy, and Toby Walsh. On the notion of interestingness in automated mathematical discovery. *International Journal of Human-Computer Studies*, 53(3):351–375, 2000.

[121] Julie Cool. Wage Gap Between Women and Men. Technical report, 2010.

[122] Shelley J Correll, Stephen Benard, and In Paik. Getting a Job: Is There a Motherhood Penalty? *American Journal of Sociology*, 112(5):1297–1339, 2007.

[123] Nelson Cowan, C Morey, and Zhijian Chen. The legend of the magical number seven. *Tall tales about the brain: Things we think we know about the mind, but ain't so, ed. S. Della Sala*, pages 45–59, 2007.

[124] Jim Cowie, Kavi Mahesh, Sergei Nirenburg, and R Zajaz. Minds-multilingual interactive document summarization. In *Working Notes of the AAAI Spring Symposium on Intelligent Text Summarization*, pages 131–132, 1998.

[125] Richard Cox. Representation construction, externalised cognition and individual differences. *Learning and instruction*, 9(4):343–363, 1999.

[126] John W Creswell. *Research design: Qualitative, quantitative, and mixed methods approaches.* Sage publications, 2013.

[127] John W Creswell. *Research design: Qualitative, quantitative, and mixed methods approaches.* Sage publications, 2013.

[128] Lei Cui, Furu Wei, and Ming Zhou. Neural open information extraction. *arXiv preprint arXiv:1805.04270*, 2018.

[129] J Shane Culpepper, Fernando Diaz, and Mark D Smucker. Research frontiers in information retrieval: Report from swirl 2018. In *ACM SIGIR Forum*, volume 52, pages 34–90. ACM, 2018.

[130] Chad Cumby, Katharina Probst, and Rayid Ghani. Retrieval and ranking of semantic entities for enterprise knowledge management tasks. In *Proceedings of the Workshop on Semantic Search (SemSearch'09)*, 2009.

[131] Douglass R Cutting, David R Karger, Jan O Pedersen, and John W Tukey. Scatter/gather: A cluster-based approach to browsing large document collections. In *ACM SIGIR Forum*, volume 51, pages 148–159. ACM New York, NY, USA, 2017.

[132] Liwei Dai, Andrew Sears, and Rich Goldman. Shifting the Focus from Accuracy to Recallability: A Study of Informal Note-taking on Mobile Information Technologies. *ACM Transactions on Computer-Human Interaction*, 16(1):4:1–4:46, 4 2009.

[133] Hoa Trang Dang, Diane Kelly, and Jimmy J Lin. Overview of the trec 2007 question answering track. In *TREC*, volume 7, page 63. Citeseer, 2007.

[134] Florian Daniel, Pavel Kucherbaev, Cinzia Cappiello, Boualem Benatallah, and Mohammad Allahbakhsh. Quality control in crowdsourcing: A survey of quality attributes, assessment techniques, and assurance actions. *ACM Comput. Surv.*, 51(1):7:1–7:40, January 2018.

[135] Alexander Philip Dawid and Allan M Skene. Maximum likelihood estimation of observer error-rates using the em algorithm. *Applied statistics*, pages 20–28, 1979.

[136] Marie-Catherine De Marneffe, Bill MacCartney, Christopher D Manning, et al. Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC*, volume 6, pages 449–454, 2006.

[137] Gianluca Demartini, Djellel Eddine Difallah, and Philippe Cudré-Mauroux. Zencrowd: leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking. In *Proceedings of the 21st international conference on World Wide Web*, pages 469–478. ACM, 2012.

[138] Artem Dementyev and Joseph A Paradiso. WristFlex: Low-power Gesture Input with Wrist-worn Pressure Sensors. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology*, UIST '14, pages 161–166, New York, NY, USA, 2014. ACM.

[139] Brenda Dervin. Useful theory for librarianship: Communication, not information. *Drexel library quarterly*, 13(3):16–32, 1977.

[140] Brenda Dervin. Sense-making theory and practice: an overview of user interests in knowledge seeking and use. *Journal of knowledge management*, 2(2):36–46, 1998.

[141] Niloofar Dezfuli, Mohammadreza Khalilbeigi, Jochen Huber, Florian Müller, and Max Mühlhäuser. PalmRC: Imaginary Palm-based Remote Control for Eyes-free Television Interaction. In *Proceedings of the 10th European Conference on Interactive TV and Video*, EuroITV '12, pages 27–34, New York, NY, USA, 2012. ACM.

[142] Laura Dietz, Manisha Verma, Filip Radlinski, and Nick Craswell. Trec complex answer retrieval overview.

[143] Vania Dimitrova, Lydia Lau, Dhavalkumar Thakker, Fan Yang-Turner, and Dimoklis Despotakis. Exploring exploratory search: a user study with linked semantic data. In *Proceedings of the 2nd IESD*, page 2. ACM, 2013.

[144] Abdigani Diriye, Ann Blandford, and Anastasios Tombros. Exploring the impact of search interface features on search tasks. In *ECDL*, pages 184–195. Springer, 2010.

[145] AnHai Doan, Jeff Naughton, Akanksha Baid, Xiaoyong Chai, Fei Chen, Ting Chen, Eric Chu, Pedro DeRose, Byron Gao, Chaitanya Gokhale, et al. The case for a structured approach to managing unstructured data. *arXiv preprint arXiv:0909.1783*, 2009.

[146] Haiwei Dong, Ali Danesh, Nadia Figueroa, and Abdulmotaleb El Saddik. An Elicitation Study on Gesture Preferences and Memorability Toward a Practical Hand-Gesture Vocabulary for Smart Televisions. *IEEE Access*, 3:543–555, 2015.

[147] Steven Dow, Anand Kulkarni, Scott Klemmer, and Björn Hartmann. Shepherding the Crowd Yields Better Work. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, CSCW '12, pages 1013–1022, New York, NY, USA, 2012. ACM.

[148] Lauren B Doyle. Is relevance an adequate criterion in retrieval system evaluation. Technical report, SYSTEM DEVELOPMENT CORP CALIF, 1963.

[149] D. Christopher Dryer. Wizards, guides, and beyond: Rational and empirical methods for selecting optimal intelligent user interface agents. In *Proceedings of the 2Nd International Conference on Intelligent User Interfaces*, IUI '97, pages 265–268, New York, NY, USA, 1997. ACM.

[150] Nicolas Ducheneaut and Victoria Bellotti. E-mail as habitat: an exploration of embedded personal information management. *interactions*, 8(5):30–38, 2001.

[151] Richard O Duda, Peter E Hart, and David G Stork. *Pattern classification*. John Wiley & Sons, 2012.

[152] Susan Dumais, Edward Cutrell, and Hao Chen. Optimizing search by showing results in context. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 277–284. ACM, 2001.

[153] Susan Dumais, Robin Jeffries, Daniel M Russell, Diane Tang, and Jaime Teevan. Understanding user behavior through log data and analysis. In *Ways of Knowing in HCI*, pages 349–372. Springer, 2014.

[154] Gloria A Dye. Graphic organizers to the rescue! *Teaching Exceptional Children*, 32(3):72–78, 2000.

[155] Georg Ebersbach, Milan R Dimitrijevic, and Werner Poewe. Influence of Concurrent Tasks On Gait: A Dual-Task Approach. *Percept Mot Skills*, 1995.

[156] Pradheep Elango. Coreference resolution: A survey. *University of Wisconsin, Madison, WI*, 2005.

[157] David Ellis. A behavioural model for information retrieval system design. *Journal of information science*, 15(4-5):237–247, 1989.

[158] David Ellis and Merete Haugan. Modelling the information seeking patterns of engineers and research scientists in an industrial environment. *Journal of documentation*, 53(4):384–403, 1997.

[159] David Elsweiler and Ian Ruthven. Towards task-based personal information management evaluations. In *Proc. of SIGIR*, pages 23–30. ACM, 2007.

[160] Johan Engstrom, Emma Johansson, and Joakim Ostlund. Effects of visual and cognitive load in real and simulated motorway driving. *Transportation Research Part F: Traffic Psychology and Behaviour*, 8(2):97–120, 2005.

[161] Martin J Eppler. A comparison between concept maps, mind maps, conceptual diagrams, and visual metaphors as complementary tools for knowledge construction and sharing. *Information visualization*, 5(3):202–210, 2006.

[162] Oren Etzioni, Michael Cafarella, Doug Downey, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S Weld, and Alexander Yates. Unsupervised named-entity extraction from the web: An experimental study. *Artificial Intelligence*, 165(1):91–134, 2005.

[163] Anthony Fader, Stephen Soderland, and Oren Etzioni. Identifying relations for open information extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1535–1545. Association for Computational Linguistics, 2011.

[164] Jerry Fails and Dan Olsen. A Design Tool for Camera-based Interaction. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, CHI '03, pages 449–456, New York, NY, USA, 2003. ACM.

[165] Tom Fawcett. An Introduction to ROC Analysis. *Pattern Recognition Letters*, 27(8):861–874, 6 2006.

[166] Andrea Ferrone, X Jiang, L Maiolo, A Pecora, L Colace, and Carlo Menon. A fabric-based wearable band for hand gesture recognition based on filament strain sensors: A preliminary investigation. In *2016 IEEE Healthcare Innovation Point-Of-Care Technologies Conference*, HI-POCT, pages 113–116, 11 2016.

[167] Debra A Fischer, Megan E Schwamb, Kevin Schawinski, Chris Lintott, John Brewer, Matt Giguere, Stuart Lynn, Michael Parrish, Thibault Sartori, Robert Simpson, Arfon Smith, Julien Spronck, Natalie Batalha, Jason Rowe, Jon Jenkins, Steve Bryson, Andrej Prsa, Peter Tenenbaum, Justin Crepp, Tim Morton, Andrew Howard, Michele Beleu, Zachary Kaplan, Nick VanNispen, Charlie Sharzer, Justin DeFouw, Agnieszka Hajduk, Joe P Neal, Adam Nemec, Nadine Schuepbach, and Valerij Zimmermann. Planet Hunters: the first two planet candidates identified by the public using the Kepler public archive data. *Monthly Notices of the Royal Astronomical Society*, 419(4):2900–2911, 2012.

[168] Tamar Flash and Neville Hogans. The Coordination of Arm Movements: An Experimentally Confirmed Mathematical Model. *Journal of neuroscience*, 5:1688–1703, 1985.

[169] Trevor Fountain and Mirella Lapata. Taxonomy induction using hierarchical random graphs. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 466–476. Association for Computational Linguistics, 2012.

[170] Euan Freeman, Stephen Brewster, and Vuokko Lantz. Tactile Feedback for Above-Device Gesture Interfaces: Adding Touch to Touchless Interactions. In *Proceedings of the 16th International Conference on Multimodal Interaction*, ICMI '14, pages 419–426, New York, NY, USA, 2014. ACM.

[171] Euan Freeman, Stephen Brewster, and Vuokko Lantz. Towards Usable and Acceptable Above-device Interactions. In *Proceedings of the 16th International Conference on Human-computer Interaction with Mobile Devices &#38; Services*, MobileHCI '14, pages 459–464, New York, NY, USA, 2014. ACM.

[172] Rui Fukui, Masahiko Watanabe, Tomoaki Gyota, Masamichi Shimosaka, and Tomomasa Sato. Hand Shape Classification with a

Wrist Contour Sensor: Development of a Prototype Device. In *Proceedings of the 13th International Conference on Ubiquitous Computing*, UbiComp '11, pages 311–314, New York, NY, USA, 2011. ACM.

[173] Masaaki Fukumoto and Yasuhito Suenaga. "FingeRing": A Full-time Wearable Interface. In *Conference Companion on Human Factors in Computing Systems*, CHI '94, pages 81–82, New York, NY, USA, 1994. ACM.

[174] Masaaki Fukumoto and Yoshinobu Tonomura. Body Coupled FingerRing: Wireless Wearable Keyboard. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, CHI '97, pages 147–154, New York, NY, USA, 1997. ACM.

[175] George W Furnas. Generalized fisheye views. *ACM SIGCHI Bulletin*, 17(4):16–23, 1986.

[176] George W Furnas. Effective view navigation. In *Proceedings of the ACM SIGCHI Conference on Human factors in computing systems*, pages 367–374, 1997.

[177] Ujwal Gadiraju, Jie Yang, and Alessandro Bozzon. Clarity is a worthwhile quality: On the role of task clarity in microtask crowdsourcing. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media*, HT '17, pages 5–14, New York, NY, USA, 2017. ACM.

[178] Snehalkumar Gaikwad, Nalin Chhibber, Vibhor Sehgal, Alipta Ballav, Catherine Mullings, Ahmed Nasser, Angela Richmond-Fuller, Aaron Gilbee, Dilrukshi Gamage, Mark Whiting, et al. Prototype tasks: Improving crowdsourcing results through rapid, iterative task design. *arXiv preprint arXiv:1707.05645*, 2017.

[179] Julia Rose Galliers and K Sparck Jones. Evaluating natural language processing systems. 1993.

[180] Maribeth Gandy, Thad Starner, Jake Auxier, and Daniel Ashbrook. The Gesture Pendant: A Self-illuminating, Wearable, Infrared Computer Vision System for Home Automation Control and Medical Monitoring. In *Proceedings of the 4th IEEE International Symposium on Wearable Computers*, ISWC '00, pages 87–, Washington, DC, USA, 2000. IEEE Computer Society.

[181] Gene Golovchinsky, Abdigani Diriye, and Tony Dunnigan. The future is in the past: Designing for exploratory search. In *Proceedings of the 4th Information Interaction in Context Symposium*, pages 52–61. ACM, 2012.

[182] Jorge Goncalves, Denzil Ferreira, Simo Hosio, Yong Liu, Jakob Rogstadius, Hannu Kukka, and Vassilis Kostakos. Crowdsourcing on the Spot: Altruistic Use of Public Displays, Feasibility, Performance, and Behaviours. In *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '13, pages 753–762, New York, NY, USA, 2013. ACM.

[183] Jun Gong, Xing-Dong Yang, and Pourang Irani. WristWhirl: One-handed Continuous Smartwatch Input Using Wrist Gestures. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, UIST '16, pages 861–872, New York, NY, USA, 2016. ACM.

[184] Jonathan Gordon and Lenhart K Schubert. Using textual patterns to learn expected event frequencies. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction*, pages 122–127. Association for Computational Linguistics, 2012.

[185] Nitesh Goyal, Gilly Leshed, and Susan R Fussell. Effects of visualization and note-taking on sensemaking and analysis. In *Proc. of CHI*, pages 2721–2724. ACM, 2013.

[186] Laura A Granka, Thorsten Joachims, and Geri Gay. Eye-tracking analysis of user behavior in www search. In *Proc. of SigIR*, pages 478–479. ACM, 2004.

[187] Wayne D Gray and Deborah A Boehm-Davis. Milliseconds matter: An introduction to microstrategies and to their use in describing and predicting interactive behavior. *Journal of experimental psychology: applied*, 6(4):322, 2000.

[188] Wayne D Gray and Deborah A Boehm-Davis. Milliseconds matter: An introduction to microstrategies and to their use in describing and predicting interactive behavior. *Journal of Experimental Psychology: Applied*, 6(4):322, 2000.

[189] Wayne D Gray and Wai-Tat Fu. Soft constraints in interactive behavior: The case of ignoring perfect knowledge in-the-world for

imperfect knowledge in-the-head. *Cognitive Science*, 28(3):359–382, 2004.

[190] Saul Greenberg and Bill Buxton. Usability evaluation considered harmful (some of the time). In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 111–120, 2008.

[191] Saul Greenberg and Chester Fitchett. Phidgets: easy development of physical interfaces through physical widgets. In *Proceedings of the 14th annual ACM symposium on User interface software and technology*, pages 209–218. ACM, 2001.

[192] Carrie Grimes, Diane Tang, and Daniel Russell. Query logs alone are not enough. 2007.

[193] Vishal Gupta and Gurpreet Singh Lehal. A survey of text summarization extractive techniques. *Journal of Emerging Technologies in Web Intelligence*, 2(3), 2010.

[194] Sean Gustafson, Christian Holz, and Patrick Baudisch. Imaginary Phone: Learning Imaginary Interfaces by Transferring Spatial Memory from a Familiar Device. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology*, UIST '11, pages 283–292, New York, NY, USA, 2011. ACM.

[195] Sean G Gustafson, Bernhard Rabe, and Patrick M Baudisch. Understanding Palm-based Imaginary Interfaces: The Role of Visual and Tactile Cues when Browsing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pages 889–898, New York, NY, USA, 2013. ACM.

[196] Tianyong Hao, Dawei Hu, Liu Wenyin, and Qingtian Zeng. Semantic patterns for user-interactive question answering. *Concurrency and Computation: Practice and Experience*, 20(7):783–799, 2008.

[197] Faizan Haque, Mathieu Nancel, and Daniel Vogel. Myopoint: Pointing and Clicking Using Forearm Mounted Electromyography and Inertial Motion Sensors. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 3653–3656. ACM, 2015.

[198] Sanda Harabagiu, Andrew Hickl, John Lehmann, and Dan Moldovan. Experiments with interactive question-answering. In

Proceedings of the 43rd annual meeting on Association for Computational Linguistics, pages 205–214. Association for Computational Linguistics, 2005.

[199] Donna Harman. Information retrieval evaluation. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 3(2):1–119, 2011.

[200] Chris Harrison, Hrvoje Benko, and Andrew D Wilson. OmniTouch: Wearable Multitouch Interaction Everywhere. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology*, UIST '11, pages 441–450, New York, NY, USA, 2011. ACM.

[201] Chris Harrison and Scott E Hudson. Abracadabra: Wireless, High-precision, and Unpowered Finger Input for Very Small Mobile Devices. In *Proceedings of the 22nd Annual ACM Symposium on User Interface Software and Technology*, UIST '09, pages 121–124, New York, NY, USA, 2009. ACM.

[202] Chris Harrison, Desney Tan, and Dan Morris. Skinput: Appropriating the Body As an Input Surface. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, CHI '10, pages 453–462, New York, NY, USA, 2010. ACM.

[203] G Sandra Hart and E Lowell Staveland. Development of NASA-TLX (Task Load Index): results of empirical and theoretical research. *Human Mental Workload*, 52:139–183, 1988.

[204] Björn Hartmann, Leith Abdulla, Manas Mittal, and Scott R Klemmer. Authoring Sensor-based Interactions by Demonstration with Direct Manipulation and Pattern Recognition. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, CHI '07, pages 145–154. ACM, 2007.

[205] Ahmed Hassan, Yang Song, and Li-wei He. A task level metric for measuring web search satisfaction and its application on improving relevance estimation. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 125–134, 2011.

[206] Ahmed Hassan, Ryen W White, Susan T Dumais, and Yi-Min Wang. Struggling or exploring? disambiguating long search ses-

sions. In *Proceedings of the 7th ACM international conference on Web search and data mining*, pages 53–62, 2014.

[207] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning.* 2 edition, 2009.

[208] Kenji Hata, Ranjay Krishna, Li Fei-Fei, and Michael S. Bernstein. A glimpse far into the future: Understanding long-term crowd worker quality. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17, pages 889–901, New York, NY, USA, 2017. ACM.

[209] Donald T Hawkins, Louise R Levy, and K Leon Montgomery. Knowledge gateways: the building blocks. *Information processing & management*, 24(4):459–468, 1988.

[210] Timothy J. Hazen, Stephanie Seneff, and Joseph Polifroni. Recognition confidence scoring and its use in speech understanding systems. *Computer Speech & Language*, 16(1):49–67, 1 2002.

[211] Luheng He, Mike Lewis, and Luke Zettlemoyer. Question-answer driven semantic role labeling: Using natural language to annotate natural language. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 643–653, 2015.

[212] Zengyou He, Xiaofei Xu, and Shengchun Deng. Data mining for actionable knowledge: A survey. *arXiv preprint cs/0501079*, 2005.

[213] Marti Hearst. *Search user interfaces.* Cambridge University Press, 2009.

[214] Marti A Hearst. The use of categories and clusters for organizing retrieval results. In *Natural language information retrieval*, pages 333–374. Springer, 1999.

[215] Marti A Hearst. Clustering versus faceted categories for information exploration. *CACM*, 49(4):59–61, 2006.

[216] Marti A Hearst and Jan O Pedersen. Reexamining the cluster hypothesis: scatter/gather on retrieval results. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 76–84, 1996.

[217] J Heer and D Boyd. Vizster: Visualizing online social networks.(2005), 2005.

[218] Jessica Heinzelman and Carol Waters. *Crowdsourcing Crisis Information in Disaster-affected Haiti.* Special report (United States Institute of Peace). U.S. Institute of Peace, 2010.

[219] Christian Heipke. Crowdsourcing geospatial data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 65(6):550–557, 11 2010.

[220] Jane Henderson, Shaishav Siddhpuria, Keiko Katsuragawa, and Edward Lank. Fostering Large Display Engagement Through Playful Interactions. In *Proceedings of the 6th ACM International Symposium on Pervasive Displays*, PerDis '17, pages 20:1–20:8, New York, NY, USA, 2017. ACM.

[221] Marion Hersh and Michael A Johnson. *Assistive technology for visually impaired and blind people.* Springer Science & Business Media, 2010.

[222] William Hersh, Andrew Turpin, Susan Price, Benjamin Chan, Dale Kramer, Lynetta Sacherek, and Daniel Olson. Do batch and user evaluations give the same results? In *Proceedings of SIGIR*, pages 17–24. ACM, 2000.

[223] William R Hersh, M Katherine Crabtree, David H Hickam, Lynetta Sacherek, Charles P Friedman, Patricia Tidmarsh, Craig Mosbaek, and Dale Kraemer. Factors associated with success in searching medline and applying evidence to answer clinical questions. *JAMIA*, 9(3):283–293, 2002.

[224] Luke Hespanhol, Martin Tomitsch, Kazjon Grace, Anthony Collins, and Judy Kay. Investigating Intuitiveness and Effectiveness of Gestures for Free Spatial Interaction with Large Displays. In *Proceedings of the 2012 International Symposium on Pervasive Displays*, PerDis '12, pages 6:1–6:6, New York, NY, USA, 2012. ACM.

[225] Paul Heymann and Hector Garcia-Molina. Collaborative creation of communal hierarchical taxonomies in social tagging systems. 2006.

[226] Juan David Hincapié-Ramos, Xiang Guo, Paymahn Moghadasian, and Pourang Irani. Consumed Endurance: A Metric to Quantify Arm Fatigue of Mid-air Interactions. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, CHI '14, pages 1063–1072, New York, NY, USA, 2014. ACM.

[227] Ken Hinckley, Jeff Pierce, Mike Sinclair, and Eric Horvitz. Sensing Techniques for Mobile Interaction. In *Proceedings of the 13th An-*

nual ACM Symposium on User Interface Software and Technology, UIST '00, pages 91–100. ACM, 2000.

[228] Lynette Hirschman and Robert Gaizauskas. Natural language question answering: The view from here. *Natural Language Engineering*, 7(4):275–300, 2001.

[229] Chien-Ju Ho, Aleksandrs Slivkins, Siddharth Suri, and Jennifer Wortman Vaughan. Incentivizing High Quality Crowdwork. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, pages 419–429, Republic and Canton of Geneva, Switzerland, 2015. International World Wide Web Conferences Steering Committee.

[230] Tim Horberry, Janet Anderson, Michael A Regan, Thomas J Triggs, and John Brown. Driver distraction: The effects of concurrent in-vehicle tasks, road environment complexity and age on driving performance. *Accident Analysis and Prevention*, 38(1):185–191, 2006.

[231] John Joseph Horton and Lydia B Chilton. The Labor Economics of Paid Crowdsourcing. In *Proceedings of the 11th ACM Conference on Electronic Commerce*, EC '10, pages 209–218, New York, NY, USA, 2010. ACM.

[232] Takayuki Hoshi, Masafumi Takahashi, Takayuki Iwamoto, and Hiroyuki Shinoda. Noncontact Tactile Display Based on Radiation Pressure of Airborne Ultrasound. *IEEE Transactions on Haptics*, 3(3):155–165, 7 2010.

[233] Mokter Hossain. Crowdsourcing: Activities, incentives and users' motivations to participate. In *2012 International Conference on Innovation Management and Technology Research*, pages 501–506, 5 2012.

[234] David C Howell. *Statistical methods for psychology.* Cengage Learning, 2009.

[235] Yi-Ta Hsieh, Antti Jylhä, Valeria Orso, Luciano Gamberini, and Giulio Jacucci. Designing a Willing-to-Use-in-Public Hand Gestural Interaction Technique for Smart Glasses. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pages 4203–4215, New York, NY, USA, 2016. ACM.

[236] Da-Yuan Huang, Liwei Chan, Shuo Yang, Fan Wang, Rong-Hao Liang, De-Nian Yang, Yi-Ping Hung, and Bing-Yu Chen. Dig-

itSpace: Designing Thumb-to-Fingers Touch Interfaces for One-Handed and Eyes-Free Interactions. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, CHI '16, pages 1526–1537, New York, NY, USA, 2016. ACM.

[237] Scott B Huffman and Michael Hochster. How well does result relevance predict session satisfaction? In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 567–574, 2007.

[238] John E Hummel and Keith J Holyoak. A symbolic-connectionist theory of relational inference and generalization. *Psychological review*, 110(2):220, 2003.

[239] Hugo Hutt, Richard Everson, Murray Grant, John Love, and George Littlejohn. How clumpy is my image? evaluating crowd-sourced annotation tasks. In *2013 13th UK Workshop on Computational Intelligence (UKCI)*, pages 136–143. IEEE, 2013.

[240] J Indratmo and Julita Vassileva. A review of organizational structures of personal information management. *Journal of digital information*, 9(1), 2008.

[241] Peter Ingwersen and Kalervo Järvelin. Information retrieval in context: Irix. In *ACM Sigir Forum*, volume 39, pages 31–39. ACM New York, NY, USA, 2005.

[242] Panagiotis G Ipeirotis, Foster Provost, and Jing Wang. Quality management on amazon mechanical turk. In *Proceedings of the ACM SIGKDD workshop on human computation*, pages 64–67. ACM, 2010.

[243] T Iwamoto and H Shinoda. Finger Ring Tactile Interface Based on Propagating Elastic Waves on Human Fingers. In *Second Joint EuroHaptics Conference and Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems (WHC'07)*, pages 145–150, 3 2007.

[244] Richard J. Jagacinski and Donald L. Monk. Fitts' Law in Two Dimensions with Hand and Head Movements Movements. *Journal of motor behavior*, 17:77–95, 1985.

[245] Ayush Jain, Akash Das Sarma, Aditya Parameswaran, and Jennifer Widom. Understanding workers, developing effective tasks, and

enhancing marketplace dynamics: a study of a large crowdsourcing marketplace. *Proceedings of the VLDB Endowment*, 10(7):829–840, 2017.

[246] P James. Knowledge graphs. *Order*, 501:6439, 1991.

[247] Bernard J Jansen and Udo Pooch. A review of web searching studies and a framework for future research. *Journal of the American Society for Information Science and Technology*, 52(3):235–246, 2001.

[248] Bernard J Jansen and Amanda Spink. An analysis of web documents retrieved and viewed. In *International Conference on Internet Computing*, pages 65–69. Citeseer, 2003.

[249] Kalervo Järvelin, Susan L Price, Lois ML Delcambre, and Marianne Lykke Nielsen. Discounted cumulated gain based evaluation of multiple-query ir sessions. In *European Conference on Information Retrieval*, pages 4–15. Springer, 2008.

[250] Xiaoran Jin, Marc Sloan, and Jun Wang. Interactive exploratory search for multi page search results. In *Proc. of the 22nd international conference on WWW*, pages 655–666. International World Wide Web Conferences Steering Committee, 2013.

[251] Yingzi Jin, Yutaka Matsuo, and Mitsuru Ishizuka. Ranking entities on the web using social network mining and ranking learning. In *WWW 2008 Workshop on Social Web Search and Mining*, 2008.

[252] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, and Geri Gay. Accurately interpreting clickthrough data as implicit feedback. In *Proc. of SIGIR*, pages 154–161. ACM, 2005.

[253] Eleanor Jones, Jason Alexander, Andreas Andreou, Pourang Irani, and Sriram Subramanian. GesText: Accelerometer-based Gestural Text-entry Systems. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, CHI '10, pages 2173–2182. ACM, 2010.

[254] Steve Jones, Stephen Lundy, and Gordon W Paynter. Interactive document summarisation using automatically extracted keyphrases. In *System Sciences, 2002. HICSS. Proceedings of the 35th Annual Hawaii International Conference on*, pages 1160–1169. IEEE, 2002.

[255] Tricia Jones. Incidental learning during information retrieval: A hypertext experiment. In *International Conference on Computer Assisted Learning*, pages 235–253. Springer, 1989.

[256] Ricardo Jota, Miguel A Nacenta, Joaquim A Jorge, Sheelagh Carpendale, and Saul Greenberg. A Comparison of Ray Pointing Techniques for Very Large Displays. In *Proceedings of Graphics Interface 2010*, GI '10, pages 269–276, Toronto, Ont., Canada, Canada, 2010. Canadian Information Processing Society.

[257] Sussane Jul and George W Furnas. Navigation in electronic worlds: a chi 97 workshop. *SIGCHI bulletin*, 29:44–49, 1997.

[258] Hyun Joon Jung and Matthew Lease. Improving consensus accuracy via z-score and weighted voting. In *Proceedings of the 2011 AAAI Workshop on Human Computation*, 2011.

[259] Pyeong-Gook Jung, Gukchan Lim, Seonghyok Kim, and Kyoungchul Kong. A Wearable Gesture Recognition Device for Detecting Muscular Activities Based on Air-Pressure Sensors. *IEEE Transactions on Industrial Informatics*, 11(2):485–494, 4 2015.

[260] Daniel Jurafsky and James H. Martin. Speech and language processing. 2009.

[261] Fiscella K, Franks P, Gold MR, and Clancy CM. Inequality in quality: Addressing socioeconomic, racial, and ethnic disparities in health care. *JAMA*, 283(19):2579–2584, 2000.

[262] Ankit Kamal, Yang Li, and Edward Lank. Teaching Motion Gestures via Recognizer Feedback. In *Proceedings of the 19th International Conference on Intelligent User Interfaces*, IUI '14, pages 73–82. ACM, 2014.

[263] Hsin-Liu (Cindy) Kao, Artem Dementyev, Joseph A Paradiso, and Chris Schmandt. NailO: Fingernails As an Input Surface. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 3015–3018, New York, NY, USA, 2015. ACM.

[264] Hsin-Liu (Cindy) Kao, Christian Holz, Asta Roseway, Andres Calvo, and Chris Schmandt. DuoSkin: Rapidly Prototyping On-skin User Interfaces Using Skin-friendly Materials. In *Proceedings of the 2016 ACM International Symposium on Wearable Computers*, ISWC '16, pages 16–23, New York, NY, USA, 2016. ACM.

[265] Günter Karjoth and Paul A Moskowitz. Disabling RFID Tags with Visible Confirmation: Clipped Tags Are Silenced. In *Proceedings of the 2005 ACM Workshop on Privacy in the Electronic Society*, WPES '05, pages 27–30, New York, NY, USA, 2005. ACM.

[266] Günter Karjoth and Paul A Moskowitz. omid. In *Proceedings of the 2005 ACM Workshop on Privacy in the Electronic Society*, WPES '05, pages 27–30, New York, NY, USA, 2005. ACM.

[267] Gjergji Kasneci, Shady Elbassuoni, and Gerhard Weikum. Ming: mining informative entity relationship subgraphs. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 1653–1656. ACM, 2009.

[268] Gjergji Kasneci, Fabian M Suchanek, Georgiana Ifrim, Maya Ramanath, and Gerhard Weikum. Naga: Searching and ranking knowledge. In *Data Engineering, 2008. ICDE 2008. IEEE 24th International Conference on*, pages 953–962. IEEE, 2008.

[269] Rohit J Kate and Raymond J Mooney. Joint entity and relation extraction using card-pyramid parsing. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, pages 203–212. Association for Computational Linguistics, 2010.

[270] Keiko Katsuragawa. Speach recognition correction with standby-word dictionary.

[271] Keiko Katsuragawa, Ankit Kamal, and Edward Lank. Effect of Motion-Gesture Recognizer Error Pattern on User Workload and Behavior. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces*, IUI '17, pages 439–449, New York, NY, USA, 2017. ACM.

[272] Keiko Katsuragawa, Krzysztof Pietroszek, James R Wallace, and Edward Lank. Watchpoint: Freehand Pointing with a Smartwatch in a Ubiquitous Display Environment. In *Proceedings of the International Working Conference on Advanced Visual Interfaces*, AVI '16, pages 128–135. ACM, 2016.

[273] Keiko Katsuragawa, James R Wallace, and Edward Lank. Gestural Text Input Using a Smartwatch. In *Proceedings of the International Working Conference on Advanced Visual Interfaces*, AVI '16, pages 220–223. ACM, 2016.

[274] Nicolas Kaufmann, Thimo Schulze, and Daniel Veit. More than fun and money. Worker Motivation in Crowdsourcing-A Study on Mechanical Turk. In *Americas Conference on Information Systems*, 2011.

[275] Frank C Keil. *Concepts, kinds, and cognitive development.* mit Press, 1992.

[276] Frank C Keil. *Concepts, kinds, and cognitive development.* mit Press, 1992.

[277] Daniel A Keim. Visual exploration of large data sets. *Communications of the ACM*, 44(8):38–44, 2001.

[278] Daniel A Keim. Information visualization and visual data mining. *IEEE transactions on Visualization and Computer Graphics*, 8(1):1–8, 2002.

[279] Diane Kelly and Leif Azzopardi. How many results per page? a study of serp size, search behavior and user experience. 2015.

[280] Diane Kelly and Colleen Cool. The effects of topic familiarity on information search behavior. In *Proceedings of the 2Nd ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '02, pages 74–75, New York, NY, USA, 2002. ACM.

[281] Diane Kelly, Susan Dumais, and Jan Pedersen. Evaluation challenges and directions for information-seeking support systems. *Computer*, 42(3):60–66, 2009.

[282] Diane Kelly, David J Harper, and Brian Landau. Questionnaire mode effects in interactive information retrieval experiments. *Information processing & management*, 44(1):122–141, 2008.

[283] Diane Kelly, Paul B Kantor, Emile L Morse, Jean Scholtz, and Ying Sun. User-centered evaluation of interactive question answering systems. In *Proceedings of the Interactive Question Answering Workshop at HLT-NAACL 2006*, pages 49–56. Association for Computational Linguistics, 2006.

[284] Atif Khan and Naomie Salim. A review on abstractive summarization methods. *Journal of Theoretical and Applied Information Technology*, 59(1), 2014.

[285] Firas Khatib, Seth Cooper, Michael D Tyka, Kefan Xu, Ilya Makedon, Zoran Popović, David Baker, and Foldit Players. Algorithm

discovery by protein folding game players. *Proceedings of the National Academy of Sciences*, 108(47):18949–18953, 2011.

[286] Davis E Kieras and Davis E Meyer. An overview of the epic architecture for cognition and performance with application to human-computer interaction. *Human–Computer Interaction*, 12(4):391–438, 1997.

[287] David Kim, Otmar Hilliges, Shahram Izadi, Alex D Butler, Jiawen Chen, Iason Oikonomidis, and Patrick Olivier. Digits: Freehand 3D Interactions Anywhere Using a Wrist-worn Gloveless Sensor. In *Proceedings of the 25th Annual ACM Symposium on User Interface Software and Technology*, UIST '12, pages 167–176. ACM, 2012.

[288] Jonghwa Kim, Stephan Mastnik, and Elisabeth André. EMG-based Hand Gesture Recognition for Realtime Biosignal Interfacing. In *Proceedings of the 13th International Conference on Intelligent User Interfaces*, IUI '08, pages 30–39, New York, NY, USA, 2008. ACM.

[289] Jungsoo Kim, Jiasheng He, Kent Lyons, and Thad Starner. The Gesture Watch: A Wireless Contact-free Gesture Based Wrist Interface. In *Proceedings of the 2007 11th IEEE International Symposium on Wearable Computers*, ISWC '07, pages 1–8, Washington, DC, USA, 2007. IEEE Computer Society.

[290] Kyung-Sun Kim. Searching the web: Effects of problem solving style on information-seeking behavior. In *World Conference on Educational Multimedia, Hypermedia and Telecommunications*, volume 1999, pages 1541–1542, 1999.

[291] Kyung-Sun Kim and Bryce Allen. Cognitive and task influences on web searching behavior. *JASIST*, 53(2):109–119, 2002.

[292] Aniket Kittur, Ed H Chi, and Bongwon Suh. Crowdsourcing user studies with mechanical turk. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 453–456, 2008.

[293] Aniket Kittur, Ed H. Chi, and Bongwon Suh. Crowdsourcing user studies with Mechanical Turk. In *Proceeding of the twenty-sixth annual CHI conference on Human factors in computing systems - CHI '08*, CHI '08, page 453, New York, New York, USA, 2008. ACM Press.

[294] Aniket Kittur, Andrew M Peters, Abdigani Diriye, Trupti Telang, and Michael R Bove. Costs and benefits of structured information foraging. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2989–2998. ACM, 2013.

[295] Joshua Klayman and Young-Won Ha. Confirmation, disconfirmation, and information in hypothesis testing. *Psychological review*, 94(2):211, 1987.

[296] Jon Kleinberg. Complex networks and decentralized search algorithms. In *Proceedings of the International Congress of Mathematicians (ICM)*, volume 3, pages 1019–1044, 2006.

[297] Alessandro L Koerich. Rejection strategies for handwritten word recognition. In *Ninth International Workshop on Frontiers in Handwriting Recognition*, pages 479–484, 10 2004.

[298] Oleksandr Kolomiyets and Marie-Francine Moens. A survey on question answering technology from an information retrieval perspective. *Information Sciences*, 181(24):5412–5434, 2011.

[299] Anita Komlodi, Gary Marchionini, and Dagobert Soergel. Search history support for finding and using information: User interface design recommendations from a user study. *Information processing & management*, 43(1):10–29, 2007.

[300] Miriam Konkel, Vivian Leung, Brygg Ullmer, and Catherine Hu. Tagaboo: A Collaborative Children's Game Based Upon Wearable RFID Technology. *Personal Ubiquitous Comput.*, 8(5):382–384, 9 2004.

[301] Natalia Konstantinova and Constantin Orasan. Interactive question answering. *Emerging Applications of Natural Language Processing: Concepts and New Research*, pages 149–169, 2013.

[302] Wessel Kraaij and Wilfried Post. Task based evaluation of exploratory search systems. In *Proc. of SIGIR 2006 Workshop, Evaluation Exploratory Search Systems, Seattle, USA*, pages 24–27, 2006.

[303] Sven Kratz and Michael Rohs. A $3 gesture recognizer: simple gesture recognition for devices equipped with 3D acceleration sensors. In *Proceedings of the 15th international conference on Intelligent user interfaces*, pages 341–344. ACM, 2010.

[304] Kory Kroft, Fabian Lange, and Matthew J Notowidigdo. Duration Dependence and Labor Market Conditions: Evidence from a Field

Experiment*. *The Quarterly Journal of Economics*, 128(3):1123, 2013.

[305] Carol Collier Kuhlthau. Perceptions of the information search process in libraries: A study of changes from high school through college. *Information Processing & Management*, 24(4):419–427, 1988.

[306] Carol Collier Kuhlthau. Seeking meaning. *Norwood, NJ: Ablex*, 1993.

[307] Bill Kules and Robert Capra. Creating exploratory tasks for a faceted search interface. *Proc. of HCIR 2008*, pages 18–21, 2008.

[308] William Kules, Max L Wilson, Ben Shneiderman, et al. From keyword search to exploration: How result visualization aids discovery on the web. 2008.

[309] Todd Kulesza, Saleema Amershi, Rich Caruana, Danyel Fisher, and Denis Charles. Structured labeling for facilitating concept evolution in machine learning. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 3075–3084. ACM, 2014.

[310] Todd Kulesza, Saleema Amershi, Rich Caruana, Danyel Fisher, and Denis Charles. Structured labeling for facilitating concept evolution in machine learning. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 3075–3084. ACM, 2014.

[311] Anand Kulkarni, Matthew Can, and Björn Hartmann. Collaboratively crowdsourcing workflows with turkomatic. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, CSCW '12, pages 1003–1012, New York, NY, USA, 2012. ACM.

[312] Barbara H Kwasnik. A descriptive study of the functional components of browsing. In *Proceedings of the IFIP TC2/WG2. 7 Working conference on Engineering for Human Computer Interaction*, page 191, 1992.

[313] Mikko Kytö, Barrett Ens, Thammathip Piumsomboon, Gun A Lee, and Mark Billinghurst. Pinpointing: Precise Head- and Eye-Based Target Selection for Augmented Reality. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, CHI '18, pages 81:1–81:14, New York, NY, USA, 2018. ACM.

[314] Thomas K Landauer. *The trouble with computers: Usefulness, usability, and productivity.* MIT press, 1995.

[315] J Richard Landis and Gary G Koch. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174, 1977.

[316] Gierad Laput, Robert Xiao, Xiang 'Anthony' Chen, Scott E Hudson, and Chris Harrison. Skin Buttons: Cheap, Small, Low-powered and Clickable Fixed-icon Laser Projectors. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology*, UIST '14, pages 389–394, New York, NY, USA, 2014. ACM.

[317] Gierad Laput, Robert Xiao, and Chris Harrison. ViBand: High-Fidelity Bio-Acoustic Sensing Using Commodity Smartwatch Accelerometers. In *Proceedings of the 29th Annual ACM Symposium on User Interface Software and Technology*, UIST '16, pages 321–333, New York, NY, USA, 2016. ACM.

[318] Edith Law, Burr Settles, and Tom Mitchell. Learning to tag using noisy labels. In *Proc. ECML*, pages 1–29, 2010.

[319] Richard S Lazarus and Susan Folkman. Stress. *Appraisal, and coping*, 725, 1984.

[320] Bongshin Lee, Cynthia Sims Parr, Catherine Plaisant, Benjamin B Bederson, Vladislav Daniel Veksler, Wayne D Gray, and Christopher Kotfila. Treeplus: Interactive exploration of networks with enhanced tree layouts. *IEEE Transactions on Visualization and Computer Graphics*, 12(6):1414–1426, 2006.

[321] Seungyon Claire Lee, BoHao Li, and Thad Starner. AirTouch: Synchronizing In-air Hand Gesture and On-body Tactile Feedback to Augment Mobile Gesture Interaction. In *International Symposium on Wearable Computers*, 2011.

[322] Jens Lehmann, Tim Furche, Giovanni Grasso, Axel-Cyrille Ngonga Ngomo, Christian Schallhart, Andrew Sellers, Christina Unger, Lorenz B̀uhmann, Daniel Gerber, Konrad H̀offner, et al. Deqa: deep web extraction for question answering. In *The Semantic Web–ISWC 2012*, pages 131–147. Springer, 2012.

[323] Luis A Leiva, Alireza Sahami, Alejandro Catala, Niels Henze, and Albrecht Schmidt. Text Entry on Tiny QWERTY Soft Keyboards.

In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 669–678, New York, NY, USA, 2015. ACM.

[324] Jure Leskovec, Kevin J Lang, Anirban Dasgupta, and Michael W Mahoney. Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. *Internet Mathematics*, 6(1):29–123, 2009.

[325] Frank Chun Yat Li, David Dearman, and Khai N Truong. Virtual Shelves: Interactions with Orientation Aware Devices. In *Proceedings of the 22nd Annual ACM Symposium on User Interface Software and Technology*, UIST '09, pages 125–128. ACM, 2009.

[326] Hanchuan Li, Eric Brockmeyer, Elizabeth J Carter, Josh Fromm, Scott E Hudson, Shwetak N Patel, and Alanson Sample. PaperID: A Technique for Drawing Functional Battery-Free Wireless Interfaces on Paper. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, CHI '16, pages 5885–5896, New York, NY, USA, 2016. ACM.

[327] Yang Li. Protractor: A Fast and Accurate Gesture Recognizer. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, CHI '10, pages 2169–2172. ACM, 2010.

[328] Yuelin Li and Nicholas J Belkin. A faceted approach to conceptualizing tasks in information seeking. *Information Processing & Management*, 44(6):1822–1837, 2008.

[329] Rong-Hao Liang, Han-Chih Kuo, and Bing-Yu Chen. GaussRFID: Reinventing Physical Toys Using Magnetic RFID Development Kits. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, CHI '16, pages 4233–4237, New York, NY, USA, 2016. ACM.

[330] Jhe-Wei Lin, Chiuan Wang, Yi Yao Huang, Kuan-Ting Chou, Hsuan-Yu Chen, Wei-Luan Tseng, and Mike Y Chen. BackHand: Sensing Hand Gestures via Back of the Hand. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software and Technology*, UIST '15, pages 557–564, New York, NY, USA, 2015. ACM.

[331] Jimmy Lin, Dennis Quan, Vineet Sinha, Karun Bakshi, David Huynh, Boris Katz, and David R Karger. The role of context in

question answering systems. In *CHI'03 extended abstracts on Human factors in computing systems*, pages 1006–1007. ACM, 2003.

[332] Jimmy Lin and Mark D Smucker. How do users find things with pubmed?: towards automatic utility evaluation with user simulations. In *Proceedings of SIGIR*, pages 19–26. ACM, 2008.

[333] Shu-Yang Lin, Chao-Huai Su, Kai-Yin Cheng, Rong-Hao Liang, Tzu-Hao Kuo, and Bing-Yu Chen. Pub - Point Upon Body: Exploring Eyes-free Interaction and Methods on an Arm. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology*, UIST '11, pages 481–488, New York, NY, USA, 2011. ACM.

[334] Thomas Lin, Oren Etzioni, and James Fogarty. Identifying interesting assertions from the web. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 1787–1790. ACM, 2009.

[335] Angli Liu, Stephen Soderland, Jonathan Bragg, Christopher H Lin, Xiao Ling, and Daniel S Weld. Effective crowd annotation for relation extraction. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 897–906, 2016.

[336] Bing Liu, Yiming Ma, and Philip S Yu. Discovering unexpected information from your competitors' web sites. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 144–153. ACM, 2001.

[337] Jiayang Liu, Lin Zhong, Jehan Wickramasuriya, and Venu Vasudevan. User Evaluation of Lightweight User Authentication with a Single Tri-axis Accelerometer. In *Proceedings of the 11th International Conference on Human-Computer Interaction with Mobile Devices and Services*, MobileHCI '09, pages 15:1–15:10. ACM, 2009.

[338] Mingyu Liu, Mathieu Nancel, and Daniel Vogel. Gunslinger: Subtle Arms-down Mid-air Interaction. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software and Technology*, UIST '15, pages 63–71. ACM, 2015.

[339] Joanne Lo, Doris Jung Lin Lee, Nathan Wong, David Bui, and Eric Paulos. Skintillates: Designing and Creating Epidermal Interactions. In *Proceedings of the 2016 ACM Conference on Designing*

Interactive Systems, DIS '16, pages 853–864, New York, NY, USA, 2016. ACM.

[340] Hela Ltifi, Mounir Ben Ayed, Adel M Alimi, and Sophie Lepreux. Survey of information visualization techniques for exploitation in kdd. In *2009 IEEE/ACS International Conference on Computer Systems and Applications*, pages 218–225. IEEE, 2009.

[341] Hao Lu and Yang Li. Gesture On: Enabling Always-On Touch Gestures for Fast Mobile Access from the Device Standby Mode. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 3355–3364. ACM, 2015.

[342] Zhiyuan Lu, Xiang Chen, Qiang Li, Xu Zhang, and Ping Zhou. A Hand Gesture Recognition Framework and Wearable Gesture-Based Interaction Prototype for Mobile Devices. *IEEE Transactions on Human-Machine Systems*, 44(2):293–299, 4 2014.

[343] Kevin Lynch. The city image and its elements. *MIT Press, Cambridge*, 41:73, 1960.

[344] Kathleen M MacQueen, Eleanor McLellan, Kelly Kay, and Bobby Milstein. Codebook development for team-based qualitative analysis. *CAM Journal*, 10(2):31–36, 1998.

[345] Thomas W Malone. How do people organize their desks?: Implications for the design of office information systems. *ACM Transactions on Information Systems (TOIS)*, 1(1):99–112, 1983.

[346] Inderjeet Mani. *Automatic summarization*, volume 3. John Benjamins Publishing, 2001.

[347] William C Mann and Sandra A Thompson. Rhetorical structure theory: Toward a functional theory of text organization. *Text-interdisciplinary Journal for the Study of Discourse*, 8(3):243–281, 1988.

[348] William C Mann and Sandra A Thompson. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281, 1988.

[349] Christopher D Manning and Hinrich Sch`utze. *Foundations of statistical natural language processing*. MIT press, 1999.

[350] Andrew Mao, Ece Kamar, Yiling Chen, Eric Horvitz, Megan E Schwamb, Chris J Lintott, and Arfon M Smith. Volunteering versus

work for pay: Incentives and tradeoffs in crowdsourcing. In *In Proceedings of the First AAAI Conference on Human Computation and Crowdsourcing (HCOMP '13*, 2013.

[351] Gary Marchionini. *Information seeking in electronic environments.* Number 9. Cambridge university press, 1997.

[352] Gary Marchionini. Exploratory search: from finding to understanding. *CACM*, 49(4):41–46, 2006.

[353] Gary Marchionini and Ben Shneiderman. Finding facts vs. browsing knowledge in hypertext systems. *Computer*, 21(1):70–80, 1988.

[354] Gary Marchionini and Ben Shneiderman. 3.1 finding facts vs. browsing knowledge in hypertext systems. *Sparks of innovation in human-computer interaction*, page 103, 1993.

[355] Gary Marchionini and Ryen White. Find what you need, understand what you find. *International Journal of Human [# x02013] Computer Interaction*, 23(3):205–237, 2007.

[356] G Marin, F Dominio, and P Zanuttigh. Hand gesture recognition with leap motion and kinect devices. In *2014 IEEE International Conference on Image Processing*, ICIP, pages 1565–1569, 10 2014.

[357] Nicolai Marquardt, Alex S Taylor, Nicolas Villar, and Saul Greenberg. Rethinking RFID: Awareness and Control for Interaction with RFID Systems. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, CHI '10, pages 2307–2316, New York, NY, USA, 2010. ACM.

[358] S A Mascaro and H H Asada. Photoplethysmograph fingernail sensors for measuring finger forces without haptic obstruction. *IEEE Transactions on Robotics and Automation*, 17(5):698–708, 10 2001.

[359] Winter Mason and Duncan J Watts. Financial Incentives and the "Performance of Crowds". In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, HCOMP '09, pages 77–85, New York, NY, USA, 2009. ACM.

[360] Julio C Mateo, Javier San Agustin, and John Paulin Hansen. Gaze Beats Mouse: Hands-free Selection by Combining Gaze and Emg. In *CHI '08 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '08, pages 3039–3044, New York, NY, USA, 2008. ACM.

[361] Jess McIntosh, Charlie McNeill, Mike Fraser, Frederic Kerber, Markus Löchtefeld, and Antonio Krüger. EMPress: Practical Hand Gesture Classification with Wrist-Mounted EMG and Pressure Sensing. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, CHI '16, pages 2332–2342, New York, NY, USA, 2016. ACM.

[362] Charles T Meadow. Relevance? *Journal of the American Society for Information Science*, 36(5):354–355, 1985.

[363] Christophe Mignot, Claude Valot, and Noëlle Carbonell. An Experimental Study of Future "Natural" Multimodal Human-computer Interaction. In *INTERACT '93 and CHI '93 Conference Companion on Human Factors in Computing Systems*, CHI '93, pages 67–68. ACM, 1993.

[364] Craig S Miller and Roger W Remington. Effects of structure and label ambiguity on information navigation. In *CHI'02 Extended Abstracts on Human Factors in Computing Systems*, pages 630–631. ACM, 2002.

[365] George A Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, 63(2):81, 1956.

[366] Roberto Mirizzi and Tommaso Di Noia. From exploratory search to web search and back. In *Proceedings of the 3rd workshop on Ph. D. students in information and knowledge management*, pages 39–46. ACM, 2010.

[367] Aditi Misra, Aaron Gooze, Kari Watkins, Mariam Asad, and Christopher A. Le Dantec. Crowdsourcing and Its Application to Transportation Data Collection and Management. *Transportation Research Record: Journal of the Transportation Research Board*, 2414(1):1–8, 1 2014.

[368] Tanushree Mitra, C.J. Hutto, and Eric Gilbert. Comparing person- and process-centric strategies for obtaining quality data on amazon mechanical turk. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 1345–1354, New York, NY, USA, 2015. ACM.

[369] Chihiro Mizuike, Shohei Ohgi, and Satoru Morita. Analysis of stroke patient walking dynamics using a tri-axial accelerometer. *Gait Posture*, 30(1):60–64, 2009.

[370] Michael F Mohageg. The influence of hypertext linking structures on the efficiency of information retrieval. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 34(3):351–367, 1992.

[371] Jane Morris and Graeme Hirst. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational linguistics*, 17(1):21–48, 1991.

[372] Meredith Ringel Morris, Andreea Danielescu, Steven Drucker, Danyel Fisher, Bongshin Lee, m. c. schraefel, and Jacob O Wobbrock. Reducing Legacy Bias in Gesture Elicitation Studies. *interactions*, 21(3):40–45, 5 2014.

[373] Sarah Morrison-Smith, Megan Hofmann, Yang Li, and Jaime Ruiz. Using Audio Cues to Support Motion Gesture Interaction on Mobile Devices. *ACM Transactions on Applied Perception*, 13(3):16:1–16:19, 5 2016.

[374] Mehran Moshfeghi. Elastic matching of multimodality medical images. *CVGIP: Graphical Models and Image Processing*, 53(3):271–282, 1991.

[375] Melissa Moyser. Women in Canada: A Gender-based Statistical Report(89-503-X), Women and Paid Work, 2017.

[376] Lev Muchnik, Royi Itzhack, Sorin Solomon, and Yoram Louzoun. Self-emergence of knowledge trees: Extraction of the wikipedia hierarchies. *Physical Review E*, 76(1):016106, 2007.

[377] George Nagy. 29 Optical character recognition—Theory and practice. *Handbook of statistics*, 2:621–649, 1982.

[378] Mathieu Nancel, Emmanuel Pietriga, Olivier Chapuis, and Michel Beaudouin-Lafon. Mid-Air Pointing on Ultra-Walls. *ACM Transactions on Computer-Human Interaction*, 22(5):21:1–21:62, 8 2015.

[379] Matei Negulescu, Jaime Ruiz, and Edward Lank. A Recognition Safety Net: Bi-level Threshold Recognition for Mobile Motion Gestures. In *Proceedings of the 14th International Conference on Human-computer Interaction with Mobile Devices and Services*, MobileHCI '12, pages 147–150. ACM, 2012.

[380] Matei Negulescu, Jaime Ruiz, Yang Li, and Edward Lank. Tap, Swipe, or Move: Attentional Demands for Distracted Smartphone Input. In *Proceedings of the International Working Conference on Advanced Visual Interfaces*, AVI '12, pages 173–180. ACM, 2012.

[381] Ani Nenkova, Sameer Maskey, and Yang Liu. Automatic summarization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts of ACL 2011*, page 3. Association for Computational Linguistics, 2011.

[382] Raymond S Nickerson. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, 2(2):175, 1998.

[383] Jakob Nielsen and Thomas K. Landauer. A mathematical model of the finding of usability problems. In *Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems*, CHI '93, pages 206–213, New York, NY, USA, 1993. ACM.

[384] Christina Niklaus, Matthias Cetto, André Freitas, and Siegfried Handschuh. A survey on open information extraction. *arXiv preprint arXiv:1806.05599*, 2018.

[385] Kent L Norman and John P Chin. The effect of tree structure on search in a hierarchical menu selection system. *Behaviour & Information Technology*, 7(1):51–65, 1988.

[386] Oded Nov. What Motivates Wikipedians? *Commun. ACM*, 50(11):60–64, 11 2007.

[387] Joseph D Novak. Concept mapping: A useful tool for science education. *Journal of research in science teaching*, 27(10):937–949, 1990.

[388] Joseph D Novak. *Learning, creating, and using knowledge*. Mahwah, NJ: Erlbaum, 1998.

[389] Joseph D Novak. *Learning, creating, and using knowledge: Concept maps as facilitative tools in schools and corporations*. Routledge, 2010.

[390] Joseph D Novak, D Bob Gowin, and Gerard T Johansen. The use of concept mapping and knowledge vee mapping with junior high school science students. *Science Education*, 67(5):625–645, 1983.

[391] Joseph D Novak and Alberto J Cañas. The theory underlying concept maps and how to construct and use them. *FIHM Fl*, 284, 2008.

[392] Laura R Novick and Sean M Hurley. To matrix, network, or hierarchy: That is the question. *Cognitive Psychology*, 42(2):158–216, 2001.

[393] Ian Oakley and Junseok Park. Motion marking menus: An eyes-free approach to motion input for handheld devices. *International Journal of Human-Computer Studies*, 67(6):515–532, 2009.

[394] Vicki L O'Day and Robin Jeffries. Orienteering in an information landscape: how information seekers get from here to there. In *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems*, pages 438–445. ACM, 1993.

[395] Robert N Oddy, NJ Belkin, and Helen M Brooks. Ask for information retrieval: Part i. background and theory. *Emerald: Journal of Documentation.*, page 61, 1982.

[396] Masa Ogata, Yuta Sugiura, Hirotaka Osawa, and Michita Imai. iRing: Intelligent Ring Using Infrared Reflection. In *Proceedings of the 25th Annual ACM Symposium on User Interface Software and Technology*, UIST '12, pages 131–136, New York, NY, USA, 2012. ACM.

[397] Christopher Olston and Ed H Chi. Scenttrails: Integrating browsing and searching on the web. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 10(3):177–197, 2003.

[398] Constantin Orasan, Viktor Pekar, and Laura Hasler. A comparison of summarisation methods based on term specificity estimation. In *LREC*, 2004.

[399] Steven Pace. A grounded theory of the flow experiences of web users. *International journal of human-computer studies*, 60(3):327–363, 2004.

[400] Gabriele Paolacci, Jesse Chandler, and Panagiotis G Ipeirotis. Running experiments on amazon mechanical turk. *Judgment and Decision making*, 5(5):411–419, 2010.

[401] Gabriele Paolacci, Jesse Chandler, and Panagiotis G Ipeirotis. Running Experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, 5(5):411–419, 2010.

[402] Gabriel Parent and Maxine Eskenazi. Clustering dictionary definitions using amazon mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 21–29. Association for Computational Linguistics, 2010.

[403] Kurt Partridge, Saurav Chatterjee, Vibha Sazawal, Gaetano Borriello, and Roy Want. TiltType: Accelerometer-supported Text Entry for Very Small Devices. In *Proceedings of the 15th Annual ACM Symposium on User Interface Software and Technology*, UIST '02, pages 201–204. ACM, 2002.

[404] Jerome Pasquero, Scott J Stobbe, and Noel Stonehouse. A Haptic Wristwatch for Eyes-free Interactions. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, CHI '11, pages 3257–3266, New York, NY, USA, 2011. ACM.

[405] Prajwal Paudyal, Junghyo Lee, Ayan Banerjee, and Sandeep K S Gupta. DyFAV: Dynamic Feature Selection and Voting for Real-time Recognition of Fingerspelled Alphabet Using Wearables. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces*, IUI '17, pages 457–467, New York, NY, USA, 2017. ACM.

[406] David S Pedulla. Penalized or Protected? Gender and the Consequences of Nonstandard and Mismatched Employment Histories. *American Sociological Review*, 81(2):262–289, 2016.

[407] Krzysztof Pietroszek, Anastasia Kuzminykh, James R Wallace, and Edward Lank. Smartcasting: A Discount 3D Interaction Technique for Public Displays. In *Proceedings of the 26th Australian Computer-Human Interaction Conference on Designing Futures: The Future of Design*, OzCHI '14, pages 119–128, New York, NY, USA, 2014. ACM.

[408] Annie Piolat, Thierry Olive, and Ronald T Kellogg. Cognitive effort during note taking. *Applied Cognitive Psychology*, 19(3):291–312, 2005.

[409] Peter Pirolli. Computational models of information scent-following in a very large browsable text collection. In *Proceedings of the ACM SIGCHI Conference on Human factors in computing systems*, pages 3–10, 1997.

[410] Peter Pirolli. *Information foraging theory: Adaptive interaction with information*. Oxford University Press, 2007.

[411] Peter Pirolli. Powers of 10: Modeling complex information-seeking systems at multiple scales. *Computer*, 42(3):33–40, 2009.

[412] Peter Pirolli and Stuart Card. Information foraging. *Psychological review*, 106(4):643, 1999.

[413] Peter Pirolli and Stuart Card. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Proceedings of International Conference on Intelligence Analysis*, volume 5, pages 2–4, 2005.

[414] Peter Pirolli, Stuart K Card, and Mija M Van Der Wege. The effect of information scent on searching information: visualizations of large tree structures. In *Proc. of the working conference on Advanced visual interfaces*, pages 161–172. ACM, 2000.

[415] Peter Pirolli, Patricia Schank, Marti Hearst, and Christine Diehl. Scatter/gather browsing communicates the topic structure of a very large text collection. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 213–220, 1996.

[416] Thammathip Piumsomboon, Adrian Clark, Mark Billinghurst, and Andy Cockburn. User-defined Gestures for Augmented Reality. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '13, pages 955–960, New York, NY, USA, 2013. ACM.

[417] Borislav Popov, Atanas Kiryakov, Ilian Kitchukov, and Krasimir Angelov. Co-occurrence and ranking of entities based on semantic annotation. *International Journal of Metadata, Semantics and Ontologies*, 3(1):21–36, 2008.

[418] Karl R Popper. The rationality principle. *Popper selections*, pages 357–365, 1985.

[419] Martin Potthast, Matthias Hagen, Michael V̀‘olske, and Benno Stein. Exploratory Search Missions for TREC Topics. In Max L. Wilson, Tony Russell-Rose, Birger Larsen, Preben Hansen, and Kristian Norling, editors, *3rd European Workshop on Human-Computer Interaction and Information Retrieval (EuroHCIR 2013)*, pages 11–14. CEUR-WS.org, August 2013.

[420] Swadhin Pradhan, Eugene Chai, Karthikeyan Sundaresan, Lili Qiu, Mohammad A Khojastepour, and Sampath Rangarajan. RIO: A Pervasive RFID-based Touch Gesture Interface. In *Proceedings of the 23rd Annual International Conference on Mobile Computing and Networking*, MobiCom '17, pages 261–274, New York, NY, USA, 2017. ACM.

[421] Manuel Prätorius, Dimitar Valkov, Ulrich Burgbacher, and Klaus Hinrichs. DigiTap: An Eyes-free VR/AR Symbolic Input Device. In *Proceedings of the 20th ACM Symposium on Virtual Reality Software and Technology*, VRST '14, pages 9–18, New York, NY, USA, 2014. ACM.

[422] Wanda Pratt, Marti A Hearst, and Lawrence M Fagan. A knowledge-based approach to organizing retrieved documents. In *AAAI/IAAI*, pages 80–85, 1999.

[423] Yan Qu and George W Furnas. Model-driven formative evaluation of exploratory search: A study under a sensemaking framework. *Information Processing & Management*, 44(2):534–555, 2008.

[424] Dragomir R Radev, Sasha Blair-Goldensohn, Zhu Zhang, and Revathi Sundara Raghavan. Interactive, domain-independent identification and summarization of topically related news articles. In *Research and Advanced Technology for Digital Libraries*, pages 225–238. Springer, 2001.

[425] Dragomir R Radev, Sasha Blair-Goldensohn, Zhu Zhang, and Revathi Sundara Raghavan. Interactive, domain-independent identification and summarization of topically related news articles. In *Research and Advanced Technology for Digital Libraries*, pages 225–238. Springer, 2001.

[426] Ashwin Ram. Interest-based information filtering and extraction in natural language understanding systems. In *Proceedings of the Bellcore Workshop on High Performance Information Filtering*, 1991.

[427] Cartic Ramakrishnan, Pablo N Mendes, Rodrigo ATS da Gama, Guilherme CN Ferreira, and Amit P Sheth. Joint extraction of compound entities and relationships from biomedical literature. In *Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology-Volume 01*, pages 398–401. IEEE Computer Society, 2008.

[428] Adrian Ramcharitar and Robert J Teather. A Fitts' Law Evaluation of Video Game Controllers: Thumbstick, Touchpad and Gyrosensor. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA '17, pages 2860–2866, New York, NY, USA, 2017. ACM.

[429] Arti Ramesh, Shachi H Kumar, James Foulds, and Lise Getoor. Weakly supervised models of aspect-sentiment for online course discussion forums. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 74–83, 2015.

[430] Lev Ratinov and Dan Roth. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, pages 147–155. Association for Computational Linguistics, 2009.

[431] Lev Ratinov, Dan Roth, Doug Downey, and Mike Anderson. Local and global algorithms for disambiguation to wikipedia. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 1375–1384. Association for Computational Linguistics, 2011.

[432] Raj M Ratwani, J Gregory Trafton, and Deborah A Boehm-Davis. Thinking graphically: Connecting vision and cognition during graph comprehension. *Journal of Experimental Psychology: Applied*, 14(1):36, 2008.

[433] Vikas C Raykar and Shipeng Yu. Eliminating spammers and ranking annotators for crowdsourced labeling tasks. *Journal of Machine Learning Research*, 13(Feb):491–518, 2012.

[434] Jun Rekimoto. Tilting Operations for Small Screen Interfaces. In *Proceedings of the 9th Annual ACM Symposium on User Interface Software and Technology*, UIST '96, pages 167–168. ACM, 1996.

[435] Zhou Ren, Jingjing Meng, Junsong Yuan, and Zhengyou Zhang. Robust Hand Gesture Recognition with Kinect Sensor. In *Proceedings of the 19th ACM International Conference on Multimedia*, MM '11, pages 759–760, New York, NY, USA, 2011. ACM.

[436] Jason DM Rennie and Tommi Jaakkola. Using term informativeness for named entity detection. In *Proc. of SIGIR*, pages 353–360. ACM, 2005.

[437] Philip Resnik. Wsd in nlp applications. In *Word Sense Disambiguation*, pages 299–337. Springer, 2007.

[438] Philip Resnik and Jimmy Lin. 11 evaluation of nlp systems. *The handbook of computational linguistics and natural language processing*, 57, 2010.

[439] Daniela Retelny, Michael S. Bernstein, and Melissa A. Valentine. No workflow can ever be enough: How crowdsourcing workflows constrain complex work. *Proc. ACM Hum.-Comput. Interact.*, 1(CSCW):89:1–89:23, December 2017.

[440] DANIELA RETELNY, MICHAEL S BERNSTEIN, and MELISSA A VALENTINE. No workflow can ever be enough: How crowdsourcing workflows constrain complex work. 2017.

[441] Flávio Ribeiro, Dinei Florêncio, Cha Zhang, and Michael Seltzer. Crowdmos: An approach for crowdsourcing mean opinion score studies. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 2416–2419. IEEE, 2011.

[442] Julie Rico and Stephen Brewster. Gestures All Around Us: User Differences in Social Acceptability Perceptions of Gesture Based Interfaces. In *Proceedings of the 11th International Conference on Human-Computer Interaction with Mobile Devices and Services*, MobileHCI '09, pages 64:1–64:2, New York, NY, USA, 2009. ACM.

[443] Gerhard Rigoll, Andreas Kosmala, and Stefan Eickeler. High Performance Real-Time Gesture Recognition Using Hidden Markov Models. In *Proceedings of International Gesture Workshop*, pages 69–80. Springer, 1998.

[444] C J Van Rijsbergen. *Information Retrieval*. Butterworth-Heinemann, Newton, MA, USA, 2nd edition, 1979.

[445] Felix Ritter, Christian Hansen, Volker Dicken, Olaf Konrad, Bernhard Preim, and Heinz-Otto Peitgen. Real-Time Illustration of Vascular Structures. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):877–884, 9 2006.

[446] Walky Rivadeneira and Benjamin B Bederson. A study of search result clustering interfaces: Comparing textual and zoomable user interfaces. *Studies*, 21(5), 2003.

[447] Stephen Robertson. On the history of evaluation in ir. *Journal of Information Science*, 34(4):439–456, 2008.

[448] Kerry Rodden, Wojciech Basalaj, David Sinclair, and Kenneth Wood. Does organisation by similarity assist image browsing? In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 190–197. ACM, 2001.

[449] Simon Rogers, John Williamson, Craig Stewart, and Roderick Murray-Smith. AnglePose: Robust, Precise Capacitive Touch Tracking via 3D Orientation Estimation. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, CHI '11, pages 2575–2584, New York, NY, USA, 2011. ACM.

[450] Jakob Rogstadius, Vassilis Kostakos, Aniket Kittur, Boris Smus, Jim Laredo, and Maja Vukovic. An Assessment of Intrinsic and Extrinsic Motivation on Task Performance in Crowdsourcing Markets. In *ICWSM*, 2011.

[451] Daniel E Rose and Danny Levinson. Understanding user goals in web search. In *Proceedings of the 13th international conference on World Wide Web*, pages 13–19. ACM, 2004.

[452] Louis Rosenfeld and Peter Morville. *Information architecture for the world wide web.* " O'Reilly Media, Inc.", 2002.

[453] Joel Ross, Lilly Irani, M Six Silberman, Andrew Zaldivar, and Bill Tomlinson. Who Are the Crowdworkers?: Shifting Demographics in Mechanical Turk. In *CHI '10 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '10, pages 2863–2872, New York, NY, USA, 2010. ACM.

[454] PH Rossi, MW Lipsey, and HE Freeman. Evaluation: a systematic approach. sage publications. *Thousand Oaks, CA*, 2004.

[455] Anne Roudaut, Eric Lecolinet, and Yves Guiard. MicroRolls: Expanding Touch-screen Input Vocabulary by Distinguishing Rolls vs. Slides of the Thumb. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, pages 927–936, New York, NY, USA, 2009. ACM.

[456] Gustavo Alberto Rovelo Ruiz, Davy Vanacken, Kris Luyten, Francisco Abad, and Emilio Camahort. Multi-viewer Gesture-based Interaction for Omni-directional Video. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, CHI '14, pages 4077–4086, New York, NY, USA, 2014. ACM.

[457] Dean Rubine. Specifying Gestures by Example. *Proceedings of the 18th annual conference on Computer graphics and interactive techniques*, 25(4):329–337, 7 1991.

[458] Jaime Ruiz and Yang Li. DoubleFlip: A Motion Gesture Delimiter for Mobile Interaction. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, CHI '11, pages 2717–2720. ACM, 2011.

[459] Jaime Ruiz, Yang Li, and Edward Lank. User-defined motion gestures for mobile interaction. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, CHI '11, pages 197–206. ACM, 2011.

[460] Jaime Ruiz and Daniel Vogel. Soft-constraints to reduce legacy and performance bias to elicit whole-body gestures with low arm fatigue. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 3347–3350. ACM, 2015.

[461] Tuukka Ruotsalo, Kumaripaba Athukorala, Dorota Głowacka, Ksenia Konyushkova, Antti Oulasvirta, Samuli Kaipiainen, Samuel Kaski, and Giulio Jacucci. Supporting exploratory search tasks with interactive user modeling. *Proc. ASIS&T'13*, 2013.

[462] Daniel Russell, Mark Stefik, Peter Pirolli, and Stuart Card. The cost structure of sensemaking. In *Proc. of INTERACT'93 and CHI'93*, pages 269–276. ACM, 1993.

[463] Daniel M Russell, Malcolm Slaney, Yan Qu, and Mave Houston. Being literate with large document collections: Observational studies and cost structure tradeoffs. In *Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS'06)*, volume 3, pages 55–55. IEEE, 2006.

[464] Tony Russell-Rose. Designing the search experience. In *Human-Computer Interaction–INTERACT 2011*, pages 702–703. Springer, 2011.

[465] Jeffrey M. Rzeszotarski and Aniket Kittur. Instrumenting the crowd: Using implicit behavioral measures to predict task performance. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology*, UIST '11, pages 13–22, New York, NY, USA, 2011. ACM.

[466] Justin Salamon, Alex Williams, Stefanie Mikloska, Duncan Macconnell, Juan P Bello, Mark Cartwright, Ayanna Seals, Edith Law, Oded Nov, and ; Edith Law. Seeing Sound: Investigating the Effects of Visualizations and Complexity on Crowdsourced Audio Annotations. *Proc. ACM Hum.-Comput. Interact. 1, CSCW, Article*, 1, 2017.

[467] Gerard Salton and Michael J McGill. Introduction to modern information retrieval. 1986.

[468] F Sanchez-Zamora and Martín Llamas-Nistal. Adaptive concept maps: Issues on design and navigation. In *Concept Mapping: Connecting Educators. Proceedings of the Third International Conference on Concept Mapping. Tallinn, Estonia & Helsinki, Finland*, 2008.

[469] Mark Sanderson, Monica Lestari Paramita, Paul Clough, and Evangelos Kanoulas. Do user preferences and evaluation measures line up? In *Proceedings of SigIR*, pages 555–562. ACM, 2010.

[470] T Scott Saponas, Jonathan Lester, Jon Froehlich, James Fogarty, and James Landay. Ilearn on the Iphone: Real-time Human Activity Classification on Commodity Mobile Phones. 2008.

[471] T Scott Saponas, Desney S Tan, Dan Morris, Ravin Balakrishnan, Jim Turner, and James A Landay. Enabling Always-available Input with Muscle-computer Interfaces. In *Proceedings of the 22nd Annual ACM Symposium on User Interface Software and Technology*, UIST '09, pages 167–176, New York, NY, USA, 2009. ACM.

[472] T Scott Saponas, Desney S Tan, Dan Morris, Jim Turner, and James A Landay. Making Muscle-computer Interfaces More Practical. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, CHI '10, pages 851–854, New York, NY, USA, 2010. ACM.

[473] Tefko Saracevic. Relevance: A review of and a framework for the thinking on the notion in information science. *J. Amer. Soc. Info. Sci.*, 26(6), 1975.

[474] Bahareh Sarrafzadeh, Rakesh Guttikonda, Kaheer Suleman, Jack Thomas, and Olga Vechtomova. Automatic discovery of related concepts. Technical report, March 2013.

[475] Bahareh Sarrafzadeh and Edward Lank. Improving exploratory search experience through hierarchical knowledge graphs. In *Proceedings of SIGIR*, pages 145–154. ACM, 2017.

[476] Bahareh Sarrafzadeh, Adam Roegiest, and Edward Lank. Hierarchical knowledge graphs: A novel information representation for exploratory search tasks. *arXiv preprint arXiv:2005.01716*, 2020.

[477] Bahareh Sarrafzadeh and Olga Vechtomova. Automatic discovery of related concepts. Technical report, Technical report, University of Waterloo, 2013.

[478] Bahareh Sarrafzadeh, Olga Vechtomova, and Vlado Jokic. Exploring knowledge graphs for exploratory search. In *Proc. of IIiX*, pages 135–144. ACM, 2014.

[479] Bahareh Sarrafzadeh, Alexandra Vtyurina, Edward Lank, and Olga Vechtomova. Knowledge graphs versus hierarchies: An analysis of user behaviours and perspectives in information seeking. In *Proc. of CHIIR*, pages 91–100. ACM, 2016.

[480] Reijo Savolainen. Berrypicking and information foraging: Comparison of two theoretical frameworks for studying exploratory search. *Journal of Information Science*, 44(5):580–593, 2018.

[481] Mike Scaife and Yvonne Rogers. External cognition: how do graphical representations work? *International journal of human-computer studies*, 45(2):185–213, 1996.

[482] Mike Scaife and Yvonne Rogers. External cognition, innovative technologies, and effective learning. *Cognition, education and communication technology*, pages 181–202, 2005.

[483] Michael Schmitz, Robert Bart, Stephen Soderland, Oren Etzioni, et al. Open language learning for information extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language*

Learning, pages 523–534. Association for Computational Linguistics, 2012.

[484] MC Schraefel. Building knowledge: What's beyond keyword search? *Computer*, 42(3):52–59, 2009.

[485] Masaki Sekine, Toshiyo Tamura, Masaki Yoshida, Yuki Suda, Yuichi Kimura, Hiroaki Miyoshi, Yoshifumi Kijima, Yuji Higashi, and Toshiro Fujimoto. A gait abnormality measure based on root mean square of trunk acceleration. *Journal of NeuroEngineering and Rehabilitation*, 10(1):118, 2013.

[486] R Senden, HHCM Savelberg, B Grimm, I C Heyligers, and K Meijer. Accelerometry-based gait analysis, an additional objective approach to screen subjects at risk for falling. *Gait and Posture*, 36(2):296–300, 2012.

[487] Tevfik Metin Sezgin and Randall Davis. HMM-based Efficient Sketch Recognition. In *Proceedings of the 10th International Conference on Intelligent User Interfaces*, IUI '05, pages 281–283. ACM, 2005.

[488] Dafna Shahaf and Carlos Guestrin. Connecting the dots between news articles. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 623–632. ACM, 2010.

[489] Dafna Shahaf, Carlos Guestrin, and Eric Horvitz. Metro maps of science. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1122–1130. ACM, 2012.

[490] Dafna Shahaf, Carlos Guestrin, and Eric Horvitz. Trains of thought: Generating information maps. In *Proceedings of the 21st international conference on World Wide Web*, pages 899–908. ACM, 2012.

[491] Aaron D Shaw, John J Horton, and Daniel L Chen. Designing Incentives for Inexpert Human Raters. In *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work*, CSCW '11, pages 275–284, New York, NY, USA, 2011. ACM.

[492] Victor S Sheng, Foster Provost, and Panagiotis G Ipeirotis. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD*

international conference on Knowledge discovery and data mining, pages 614–622. ACM, 2008.

[493] Aashish Sheshadri and Matthew Lease. Square: A benchmark for research on computing crowd consensus. In *Proceedings of the First AAAI Conference on Human Computation and Crowdsourcing*, 2013.

[494] Roy Shilkrot, Jochen Huber, Jürgen Steimle, Suranga Nanayakkara, and Pattie Maes. Digital Digits: A Comprehensive Survey of Finger Augmentation Devices. *ACM Computing Surveys*, 48(2):30:1–30:29, 11 2015.

[495] Ben Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *Visual Languages, 1996. Proceedings., IEEE Symposium on*, pages 336–343. IEEE, 1996.

[496] Ben Shneiderman. *Designing the User Interface: Strategies for Effective Human-Computer Interaction*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 3 edition, 1997.

[497] Jamie Shotton, Andrew Fitzgibbon, Mat Cook, Toby Sharp, Mark Finocchio, Richard Moore, Alex Kipman, and Andrew Blake. Real-time Human Pose Recognition in Parts from Single Depth Images. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '11, pages 1297–1304, Washington, DC, USA, 2011. IEEE Computer Society.

[498] Yedendra Babu Shrinivasan and Jarke J van Wijk. Supporting the analytical reasoning process in information visualization. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1237–1246. ACM, 2008.

[499] Nachum Sicherman. Gender Differences in Departure from a Large Firm. Technical Report 4279, National Bureau of Economic Research, 2 1993.

[500] Shaishav Siddhpuria, Keiko Katsuragawa, James R Wallace, and Edward Lank. Exploring At-Your-Side Gestural Interaction for Ubiquitous Environments. In *Proceedings of the 2017 Conference on Designing Interactive Systems*, DIS '17, pages 1111–1122, New York, NY, USA, 2017. ACM.

[501] Robert Simpson, Kevin R Page, and David De Roure. Zooniverse: Observing the World's Largest Citizen Science Platform. In *Pro-*

ceedings of the 23rd International Conference on World Wide Web, WWW '14 Companion, pages 1049–1054, New York, NY, USA, 2014. ACM.

[502] A Singhal. Introducing the knowledge graph: things, not strings, 2012. *Official Blog (of Google)*.

[503] David Small and Hiroshi Ishii. Design of Spatially Aware Graspable Displays. In *CHI '97 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '97, pages 367–368. ACM, 1997.

[504] Catherine L Smith and Paul B Kantor. User adaptation: good results from poor systems. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 147–154. ACM, 2008.

[505] P Smith, M Shah, and N da Vitoria Lobo. Determining driver visual attention with one camera. *IEEE Transactions on Intelligent Transportation Systems*, 4(4):205–218, 12 2003.

[506] Mark D Smucker, James Allan, and Blagovest Dachev. Human question answering performance using an interactive information retrieval system. *Center for Intelligent Information Retrieval Technical Report IR-655, University of Massachusetts*, 2008.

[507] Mark D Smucker, James Allan, and Blagovest Dachev. Human question answering performance using an interactive document retrieval system. In *Proceedings of the 4th Information Interaction in Context Symposium*, pages 35–44, 2012.

[508] Mark D Smucker and Charles LA Clarke. The fault, dear researchers, is not in cranfield, but in our metrics, that they are unrealistic. In *EuroHCIR*, pages 11–12, 2012.

[509] Mark D Smucker and Charles LA Clarke. Modeling user variance in time-biased gain. In *Proceedings of the Symposium on Human-Computer Interaction and Information Retrieval*, pages 1–10, 2012.

[510] Mark D Smucker and Chandra Prakash Jethani. Human performance and retrieval precision revisited. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 595–602. ACM, 2010.

[511] Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y Ng. Cheap and fast—but is it good?: evaluating non-expert annotations

for natural language tasks. In *Proceedings of the conference on empirical methods in natural language processing*, pages 254–263. Association for Computational Linguistics, 2008.

[512] Rajinder Sodhi, Ivan Poupyrev, Matthew Glisson, and Ali Israr. AIREAL: Interactive Tactile Experiences in Free Air. *ACM Transactions on Graphics*, 32(4):134:1–134:10, 7 2013.

[513] Jean Y. Song, Raymond Fok, Alan Lundgard, Fan Yang, Juho Kim, and Walter S. Lasecki. Two tools are better than one: Tool diversity as a means of improving aggregate crowd performance. In *23rd International Conference on Intelligent User Interfaces*, IUI '18, pages 559–570, New York, NY, USA, 2018. ACM.

[514] Amanda Spink, Judy Bateman, and Bernard J Jansen. Searching the web: A survey of excite users. *Internet research*, 9(2):117–128, 1999.

[515] J Sprinks, R Houghton, S Bamford, and JG Morley. Planet four: Craters—optimizing task workflow to improve volunteer engagement and crater counting performance. *Meteoritics & Planetary Science*, 2019.

[516] Gabriel Stanovsky and Ido Dagan. Creating a large benchmark for open information extraction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2300–2305, 2016.

[517] John Stasko, Carsten G`org, and Zhicheng Liu. Jigsaw: supporting investigative analysis through interactive visualization. *Information visualization*, 7(2):118–132, 2008.

[518] Statista. Statistics & Facts on Wearable Technology.

[519] Ryan Stedman, Michael Terry, and Edward Lank. Aiding human discovery of handwriting recognition errors. In *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*, ICMI '13, pages 295–302, New York, NY, USA, 2013. ACM.

[520] Pamela Stone. *Opting Out?: Why Women Really Quit Careers and Head Home.* University of California Press, 1 edition, 2007.

[521] Anselm Strauss and Juliet M Corbin. *Basics of qualitative research: Grounded theory procedures and techniques.* Sage Publications, Inc, 1990.

[522] David L Strayer, Frank A Drews, and William A Johnston. Cell phone-induced failures of visual attention during simulated driving. *Journal of Experimental Psychology: Applied*, 9(1):23–32, 3 2003.

[523] David J Sturman and David Zeltzer. A Survey of Glove-based Input. *IEEE Comput. Graph. Appl.*, 14(1):30–39, 1 1994.

[524] Chao-Huai Su, Liwei Chan, Chien-Ting Weng, Rong-Hao Liang, Kai-Yin Cheng, and Bing-Yu Chen. NailDisplay: Bringing an Always Available Visual Display to Fingertips. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, CHI '13, pages 1461–1464, New York, NY, USA, 2013. ACM.

[525] Ke Sun, Yuntao Wang, Chun Yu, Yukang Yan, Hongyi Wen, and Yuanchun Shi. Float: One-Handed and Touch-Free Target Selection on Smartwatches. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, CHI '17, pages 692–704, New York, NY, USA, 2017. ACM.

[526] Maoyuan Sun, Lauren Bradel, Chris L North, and Naren Ramakrishnan. The role of interactive biclusters in sensemaking. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*, pages 1559–1562. ACM, 2014.

[527] Peng Sun and Kathryn T. Stolee. Exploring crowd consistency in a mechanical turk survey. In *Proceedings of the 3rd International Workshop on CrowdSourcing in Software Engineering*, CSI-SE '16, pages 8–14, New York, NY, USA, 2016. ACM.

[528] Yunjia Sun, Edward Lank, and Michael Terry. Label-and-learn: Visualizing the likelihood of machine learning classifier's success during data labeling. In *Proceedings of the 22Nd International Conference on Intelligent User Interfaces*, IUI '17, pages 523–534, New York, NY, USA, 2017. ACM.

[529] Don R Swanson. Information retrieval as a trial-and-error process. *The Library Quarterly*, 47(2):128–148, 1977.

[530] Zareen Syed, Evelyne Viegas, and Savas Parastatidis. Automatic discovery of semantic relations using mindnet. In *LREC*, 2010.

[531] Xiaohui Tao, Yuefeng Li, and Ning Zhong. A knowledge-based model using ontologies for personalized web information gathering. *Web Intelligence and Agent Systems*, 8(3):235–254, 2010.

[532] Charles C Tappert. Cursive script recognition by elastic matching. *IBM Journal of Research and development*, 26(6):765–771, 1982.

[533] Valerie Tarasuk and Joan M Eakin. Food assistance through "surplus" food: Insights from an ethnographic study of food bank work. *Agriculture and Human Values*, 22(2):177–186, 6 2005.

[534] Ann Taylor, Mitchell Marcus, and Beatrice Santorini. The penn treebank: an overview. In *Treebanks*, pages 5–22. Springer, 2003.

[535] Jaime Teevan, Christine Alvarado, Mark S Ackerman, and David R Karger. The perfect search engine is not enough: a study of orienteering behavior in directed search. In *Proc. of CHI*, pages 415–422. ACM, 2004.

[536] Gineke A ten Holt, Marcel J T Reinders, and Emile A Hendriks. Multi-Dimensional Dynamic Time Warping for Gesture Recognition. In *Thirteenth annual conference of the Advanced School for Computing and Imaging*, 6 2007.

[537] Rannie Teodoro, Pinar Ozturk, Mor Naaman, Winter Mason, and Janne Lindqvist. The Motivations and Experiences of the On-demand Mobile Workforce. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work &#38; Social Computing*, CSCW '14, pages 236–247, New York, NY, USA, 2014. ACM.

[538] Sigmar-Olaf Tergan and Tanja Keller. *Knowledge and information visualization: Searching for synergies*, volume 3426. Springer, 2005.

[539] Marian Theiss, Philipp M Scholl, and Kristof Van Laerhoven. Predicting Grasps with a Wearable Inertial and EMG Sensing Unit for Low-Power Detection of In-Hand Objects. In *Proceedings of the 7th Augmented Human International Conference 2016*, AH '16, pages 4:1–4:8, New York, NY, USA, 2016. ACM.

[540] Perry W Thorndyke and Sarah E Goldin. Spatial learning and reasoning skill. In *Spatial orientation*, pages 195–217. Springer, 1983.

[541] Robert Tidwell, Robert Akl, Sarath Akumalla, Sarada Karlaputi, David Struble, and Krishna Kavi. Evaluating the Feasibility of EMG and Bend Sensors for Classifying Hand Gestures. In *Proceedings of the International Conference on Multimedia and Human*

Computer Interaction, MHCI '13, pages 1–8, Toronto, ON, Canada, 2013. ASET.

[542] Ramine Tinati, Max Van Kleek, Elena Simperl, Markus Luczak-Rösch, Robert Simpson, and Nigel Shadbolt. Designing for Citizen Data Analysis: A Cross-Sectional Case Study of a Multi-Domain Citizen Science Platform. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 4069–4078, New York, NY, USA, 2015. ACM.

[543] Maryam Tohidi, William Buxton, Ronald Baecker, and Abigail Sellen. Getting the right design and the design right. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 1243–1252, 2006.

[544] Christoph Trattner, Philipp Singer, Denis Helic, and Markus Strohmaier. Exploring the differences and similarities between hierarchical decentralized search and human navigation in information networks. In *Proc. of I-KNOW'12*, page 14. ACM, 2012.

[545] Hsin-Ruey Tsai, Min-Chieh Hsiu, Jui-Chun Hsiao, Lee-Ting Huang, Mike Chen, and Yi-Ping Hung. TouchRing: Subtle and Always-available Input Using a Multi-touch Ring. In *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct*, MobileHCI '16, pages 891–898, New York, NY, USA, 2016. ACM.

[546] Hsin-Ruey Tsai, Cheng-Yuan Wu, Lee-Ting Huang, and Yi-Ping Hung. ThumbRing: Private Interactions Using One-handed Thumb Motion Input on Finger Segments. In *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct*, MobileHCI '16, pages 791–798, New York, NY, USA, 2016. ACM.

[547] Koji Tsukada and Michiaki Yasumura. Ubi-Finger : Gesture Input Device for Mobile Use. 2002.

[548] John W Tukey. *Exploratory data analysis*, volume 2. Reading, Mass., 1977.

[549] Andrew Turpin and Falk Scholer. User performance versus precision measures for simple search tasks. In *Proceedings of SIGIR*, pages 11–18. ACM, 2006.

[550] Andrew H Turpin and William Hersh. Why batch and user evaluations do not give the same results. In *Proceedings of SIGIR*, pages 225–231. ACM, 2001.

[551] Amos Tversky and Daniel Kahneman. Judgment under uncertainty: Heuristics and biases. *science*, 185(4157):1124–1131, 1974.

[552] Seiichi Uchida and Hiroaki Sakoe. A Survey of Elastic Matching Techniques for Handwritten Character Recognition. *IEICE - Transactions on Information and Systems*, E88-D(8):1781–1790, 8 2005.

[553] Geoffrey Underwood, Peter Chapman, Neil Brocklehurst, Jean Underwood, and David Crundall. Visual attention while driving: sequences of eye fixations made by experienced and novice drivers. *Ergonomics*, 46(6):629–646, 2003.

[554] Christina Unger, Lorenz B̀uhmann, Jens Lehmann, Axel-Cyrille Ngonga Ngomo, Daniel Gerber, and Philipp Cimiano. Template-based question answering over rdf data. In *Proceedings of the 21st international conference on World Wide Web*, pages 639–648. ACM, 2012.

[555] Sharanjit Uppal. Employment patterns of families with children. Technical report, 2015.

[556] Pertti Vakkari. Task complexity, problem structure and information actions: Integrating studies on information seeking and retrieval. *Information processing & management*, 35(6):819–837, 1999.

[557] Pertti Vakkari. Relevance and contributing information types of searched documents in task performance. In *Proc. of SIGIR*, pages 2–9. ACM, 2000.

[558] Pertti Vakkari. A theory of the task-based information retrieval process: a summary and generalisation of a longitudinal study. *Journal of documentation*, 57(1):44–60, 2001.

[559] Pertti Vakkari and Saila Huuskonen. Search effort degrades search output but improves task outcome. *JASIST*, 63(4):657–670, 2012.

[560] Alejandro Valerio, David Leake, and Alberto J Canas. Automatically associating documents with concept map knowledge models. In *XXXIII Conferencia Latinoamericana de Informática (CLEI 2007), San José, Costa Rica*. Citeseer, 2007.

[561] Maaike Van Den Haak, Menno De Jong, and Peter Jan Schellens. Retrospective vs. concurrent think-aloud protocols: testing the usability of an online library catalogue. *Behaviour & information technology*, 22(5):339–351, 2003.

[562] Frank Van Ham and Adam Perer. "search, show context, expand on demand": Supporting large graph exploration with degree-of-interest. *Visualization and Computer Graphics, IEEE Transactions on*, 15(6):953–960, 2009.

[563] Radu-Daniel Vatavu and Jacob O Wobbrock. Formalizing Agreement Analysis for Elicitation Studies: New Measures, Significance Test, and Toolkit. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 1325–1334, New York, NY, USA, 2015. ACM.

[564] Radu-Daniel Vatavu and Ionut-Alexandru Zaiti. Leap Gestures for TV: Insights from an Elicitation Study. In *Proceedings of the 2014 ACM International Conference on Interactive Experiences for TV and Online Video*, TVX '14, pages 131–138, New York, NY, USA, 2014. ACM.

[565] Robert Villa, Iván Cantador, Hideo Joho, and Joemon M Jose. An aspectual interface for supporting complex search tasks. In *Proc. of SIGIR*, pages 379–386. ACM, 2009.

[566] Taras K Vintsyuk. Speech discrimination by dynamic programming. *Kybernetika*, 1968.

[567] Paul Viola, John C Platt, and Cha Zhang. Multiple Instance Boosting for Object Detection. In *Proceedings of the 18th International Conference on Neural Information Processing Systems*, NIPS'05, pages 1417–1424, Cambridge, MA, USA, 2005. MIT Press.

[568] Daniel Vogel and Ravin Balakrishnan. Distant Freehand Pointing and Clicking on Very Large, High Resolution Displays. In *Proceedings of the 18th Annual ACM Symposium on User Interface Software and Technology*, UIST '05, pages 33–42. ACM, 2005.

[569] Stephen Voida, Mark Podlaseck, Rick Kjeldsen, and Claudio Pinhanez. A Study on the Manipulation of 2D Objects in a Projector/Camera-based Augmented Reality Environment. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, CHI '05, pages 611–620. ACM, 2005.

[570] Ellen M Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. *Information processing & management*, 36(5):697–716, 2000.

[571] Ellen M Voorhees et al. The trec-8 question answering track report. In *Trec*, volume 99, pages 77–82, 1999.

[572] J̀org Waitelonis, Magnus Knuth, Lina Wolf, Johannes Hercher, and Harald Sack. The path is the destination–enabling a new search paradigm with linked data. *Linked Data in the Future Internet at the Future Internet Assembly, Ghent*, page 8, 2010.

[573] Cheng-Yao Wang, Wei-Chen Chu, Po-Tsung Chiu, Min-Chieh Hsiu, Yih-Harn Chiang, and Mike Y Chen. PalmType: Using Palms As Keyboards for Smart Glasses. In *Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services*, MobileHCI '15, pages 153–160, New York, NY, USA, 2015. ACM.

[574] Ju Wang, Omid Abari, and Srinivasan Keshav. Challenge: RFID Hacking for Fun and Profit. 2018.

[575] Jue Wang, Deepak Vasisht, and Dina Katabi. RF-IDraw: Virtual Touch Screen in the Air Using RF Signals. In *Proceedings of the 2014 ACM Conference on SIGCOMM*, SIGCOMM '14, pages 235–246, New York, NY, USA, 2014. ACM.

[576] R Wang, Hong-Jiang Zhang, and Ya-Qin Zhang. A confidence measure based moving object extraction system built for compressed domain. In *2000 IEEE International Symposium on Circuits and Systems. Emerging Technologies for the 21st Century. Proceedings (IEEE Cat No.00CH36353)*, volume 5, pages 21–24, 2000.

[577] Saiwen Wang, Jie Song, Jaime Lien, Ivan Poupyrev, and Otmar Hilliges. Interacting with Soli: Exploring Fine-Grained Dynamic Gesture Recognition in the Radio-Frequency Spectrum. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, UIST '16, pages 851–860, New York, NY, USA, 2016. ACM.

[578] Yulu Wang, Garrick Sherman, Jimmy Lin, and Miles Efron. Assessor differences and user preferences in tweet timeline generation. In *Proceedings of SIGIR*, pages 615–624. ACM, 2015.

[579] Roy Want. Enabling Ubiquitous Sensing with RFID. *Computer*, 37(4):84–86, 4 2004.

[580] Roy Want. The Magic of RFID. *Queue*, 2(7):40–48, 10 2004.

[581] Roy Want, Kenneth P Fishkin, Anuj Gujar, and Beverly L Harrison. Bridging Physical and Virtual Worlds with Electronic Tags. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '99, pages 370–377, New York, NY, USA, 1999. ACM.

[582] Colin Ware. Visual queries: The foundation of visual thinking. In *Knowledge and information visualization*, pages 27–35. Springer, 2005.

[583] Colin Ware. *Information visualization: perception for design*. Elsevier, 2012.

[584] David Way and Joseph Paradiso. A Usability User Study Concerning Free-Hand Microgesture and Wrist-Worn Sensors. In *Proceedings of the 2014 11th International Conference on Wearable and Implantable Body Sensor Networks*, BSN '14, pages 138–142, Washington, DC, USA, 2014. IEEE Computer Society.

[585] Lars Weberg, Torbjörn Brange, and A Wendelbo Hansson. A Piece of Butter on the PDA Display. In *CHI '01 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '01, pages 435–436. ACM, 2001.

[586] Martin Weigel, Tong Lu, Gilles Bailly, Antti Oulasvirta, Carmel Majidi, and Jürgen Steimle. iSkin: Flexible, Stretchable and Visually Customizable On-Body Touch Sensors for Mobile Computing. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 2991–3000, New York, NY, USA, 2015. ACM.

[587] Martin Weigel, Aditya Shekhar Nittala, Alex Olwal, and Jürgen Steimle. SkinMarks: Enabling Interactions on Body Landmarks Using Conformal Skin Electronics. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, CHI '17, pages 3095–3105, New York, NY, USA, 2017. ACM.

[588] Martin Weigel and Jürgen Steimle. DeformWear: Deformation Input on Tiny Wearable Devices. *Proceedings of the ACM on Inter-*

active, Mobile, Wearable and Ubiquitous Technologies, 1(2):28:1–28:23, 6 2017.

[589] Peter Welinder, Steve Branson, Serge J Belongie, and Pietro Perona. The multidimensional wisdom of crowds. In *Advances in Neural Information Processing Systems*, volume 23, pages 2424–2432, 2010.

[590] Hongyi Wen, Julian Ramos Rojas, and Anind K Dey. Serendipity: Finger Gesture Recognition Using an Off-the-Shelf Smartwatch. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, CHI '16, pages 3847–3851, New York, NY, USA, 2016. ACM.

[591] Matthias Wendt, Martin Gerlach, and Holger D`uwiger. Linguistic modeling of linked open data for question answering. *Proceedings of Interacting with Linked Data (ILD 2012)[37]*, pages 75–86, 2012.

[592] Charles K West, James A Farmer, and Phillip M Wolff. *Instructional design: Implications from cognitive science.* Prentice Hall Englewood Cliffs, NJ, 1991.

[593] Cynthia Weston, Terry Gandell, Jacinthe Beauchamp, Lynn McAlpine, Carol Wiseman, and Cathy Beauchamp. Analyzing interview data: The development and evolution of a coding system. *Qualitative sociology*, 24(3):381–400, 2001.

[594] Ryen White. Beliefs and biases in web search. In *Proc. of SIGIR*, pages 3–12. ACM, 2013.

[595] Ryen W White. *Interactions with search systems.* Cambridge University Press, 2016.

[596] Ryen W White, Bill Kules, Steven M Drucker, et al. Supporting exploratory search, introduction, special issue. *CACM*, 49(4):36–39, 2006.

[597] Ryen W White and Resa A Roth. Exploratory search: Beyond the query-response paradigm. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 1(1):1–98, 2009.

[598] Ryen W White and Ian Ruthven. A study of interface support mechanisms for interactive information retrieval. *JASIST*, 57(7):933–948, 2006.

[599] Jacob Whitehill, Ting-fan Wu, Jacob Bergsma, Javier R Movellan, and Paul L Ruvolo. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Advances in Neural Information Processing Systems*, pages 2035–2043, 2009.

[600] Eric Whitmire, Mohit Jain, Divye Jain, Greg Nelson, Ravi Karkar, Shwetak Patel, and Mayank Goel. DigiTouch: Reconfigurable Thumb-to-Finger Input and Text Entry on Head-mounted Displays. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(3):113:1–113:21, 9 2017.

[601] Edward Whittaker, Sadaoki Furui, and Dietrich Klakow. A statistical classification approach to question answering using web data. In *Cyberworlds, 2005. International Conference on*, pages 8–pp. IEEE, 2005.

[602] Daniel Wigdor and Ravin Balakrishnan. TiltText: Using Tilt for Text Input to Mobile Phones. In *Proceedings of the 16th Annual ACM Symposium on User Interface Software and Technology*, UIST '03, pages 81–90. ACM, 2003.

[603] Daniel Wigdor and Dennis Wixon. *Brave NUI World: Designing Natural User Interfaces for Touch and Gesture*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1st edition, 2011.

[604] Barbara M Wildemuth and Luanne Freund. Search tasks and their role in studies of search behaviors. In *Third Annual Workshop on Human Computer Interaction and Information Retrieval, Washington DC*. Citeseer, 2009.

[605] Barbara M Wildemuth and Luanne Freund. Assigning search tasks designed to elicit exploratory search behaviors. In *Proc. of HCIR 2012*, page 4. ACM, 2012.

[606] Kyle W. Willett, Chris J. Lintott, Steven P. Bamford, Karen L. Masters, Brooke D. Simmons, Kevin R. V. Casteels, Edward M. Edmondson, Lucy F. Fortson, Sugata Kaviraj, William C. Keel, Thomas Melvin, Robert C. Nichol, M. Jordan Raddick, Kevin Schawinski, Robert J. Simpson, Ramin A. Skibba, Arfon M. Smith, and Daniel Thomas. Galaxy Zoo 2: detailed morphological classifications for 304 122 galaxies from the Sloan Digital Sky Survey. *Monthly Notices of the Royal Astronomical Society*, 435(4):2835–2860, 11 2013.

[607] Alex C Williams, Joslin Goh, Charlie G Willis, Aaron M Ellison, James H Brusuelas, Charles C Davis, and Edith Law. Deja vu: Characterizing worker reliability using task consistency. In *Proceedings of the Fifth AAAI Conference on Human Computation (HCOMP)*, pages 197–205. AAAI, 2017.

[608] Jay G Wilpon, Lawrence R Rabiner, Chin-Hui Lee, and E R Goldman. Automatic recognition of keywords in unconstrained speech using hidden Markov models. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 38(11):1870–1878, 11 1990.

[609] Graham Wilson, Thomas Carter, Sriram Subramanian, and Stephen A Brewster. Perception of Ultrasonic Haptic Feedback on the Hand: Localisation and Apparent Motion. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, CHI '14, pages 1133–1142, New York, NY, USA, 2014. ACM.

[610] Max L Wilson et al. The importance of conveying inter-facet relationships for making sense of unfamiliar domains. 2009.

[611] Max L Wilson, Bill Kules, Ben Shneiderman, et al. From keyword search to exploration: Designing future search interfaces for the web. *Foundations and Trends in Web Science*, 2(1):1–97, 2010.

[612] Tom D Wilson. Models in information behaviour research. *Journal of documentation*, 55(3):249–270, 1999.

[613] Raphael Wimmer and Florian Echtler. Exploring the Benefits of Fingernail Displays. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '13, pages 937–942, New York, NY, USA, 2013. ACM.

[614] Jacob O Wobbrock, James Fogarty, Shih-Yen (Sean) Liu, Shunichi Kimuro, and Susumu Harada. The Angle Mouse: Target-agnostic Dynamic Gain Adjustment Based on Angular Deviation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, pages 1401–1410, New York, NY, USA, 2009. ACM.

[615] Jacob O Wobbrock, Meredith Ringel Morris, and Andrew D Wilson. User-defined Gestures for Surface Computing. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, CHI '09, pages 1083–1092. ACM, 2009.

[616] Jacob O Wobbrock, Andrew D Wilson, and Yang Li. Gestures Without Libraries, Toolkits or Training: A $1 Recognizer for User Interface Prototypes. In *Proceedings of the 20th Annual ACM Symposium on User Interface Software and Technology*, UIST '07, pages 159–168. ACM, 2007.

[617] Katrin Wolf, Anja Naumann, Michael Rohs, and Jörg Müller. Taxonomy of Microinteractions: Defining Microgestures Based on Ergonomic and Scenario-dependent Requirements. In *Proceedings of the 13th IFIP TC 13 International Conference on Human-computer Interaction - Volume Part I*, INTERACT'11, pages 559–575, Berlin, Heidelberg, 2011. Springer-Verlag.

[618] Meng-Han Wu and Alexander J. Quinn. Confusing the crowd: Task instruction quality on amazon mechanical turk. In *In Proceedings of the Fifth AAAI Conference on Human Computation and Crowdsourcing (HCOMP '17*, 2017.

[619] Yi-fang Brook Wu, Latha Shankar, and Xin Chen. Finding more useful information faster from web search results. In *Proceedings of the twelfth international conference on Information and knowledge management*, pages 568–571. ACM, 2003.

[620] Fu Xiao, Zhongqin Wang, Ning Ye, Ruchuan Wang, and Xiang-Yang Li. One More Tag Enables Fine-Grained RFID Localization and Tracking. *IEEE/ACM Trans. Netw.*, 26(1):161–174, 2 2018.

[621] Chao Xu, Parth H Pathak, and Prasant Mohapatra. Finger-writing with Smartwatch: A Case for Finger and Hand Gesture Recognition Using Smartwatch. In *Proceedings of the 16th International Workshop on Mobile Computing Systems and Applications*, HotMobile '15, pages 9–14, New York, NY, USA, 2015. ACM.

[622] Xiaobing Xue, Jiwoon Jeon, and W Bruce Croft. Retrieval models for question and answer archives. In *Proc. of SIGIR*, pages 475–482. ACM, 2008.

[623] Ning Yan. Entity-centric information exploration.

[624] Lei Yang, Yekui Chen, Xiang-Yang Li, Chaowei Xiao, Mo Li, and Yunhao Liu. Tagoram: Real-time Tracking of Mobile RFID Tags to High Precision Using COTS Devices. In *Proceedings of the 20th Annual International Conference on Mobile Computing and Net-*

working, MobiCom '14, pages 237–248, New York, NY, USA, 2014. ACM.

[625] Xing-Dong Yang, Tovi Grossman, Daniel Wigdor, and George Fitzmaurice. Magic Finger: Always-available Input Through Finger Instrumentation. In *Proceedings of the 25th Annual ACM Symposium on User Interface Software and Technology*, UIST '12, pages 147–156, New York, NY, USA, 2012. ACM.

[626] Yoonsik Yang, Seungho Chae, Jinwook Shim, and Tack-Don Han. EMG Sensor-based Two-Hand Smart Watch Interaction. In *Adjunct Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology*, UIST '15 Adjunct, pages 73–74, New York, NY, USA, 2015. ACM.

[627] Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. A New Dataset and Method for Automatically Grading ESOL Texts. In *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, HLT '11, pages 180–189. Association for Computational Linguistics, 2011.

[628] Alexander Yates, Michael Cafarella, Michele Banko, Oren Etzioni, Matthew Broadhead, and Stephen Soderland. Textrunner: open information extraction on the web. In *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 25–26. Association for Computational Linguistics, 2007.

[629] Ka-Ping Yee, Kirsten Swearingen, Kevin Li, and Marti Hearst. Faceted metadata for image search and browsing. In *Proc. of CHI*, pages 401–408. ACM, 2003.

[630] Yeliz Yesilada, Robert Stevens, Simon Harper, and Carole Goble. Evaluating DANTE: Semantic Transcoding for Visually Disabled Users. *ACM Transactions on Computer-Human Interaction*, 14(3), 9 2007.

[631] Xing Yi and James Allan. A comparative study of utilizing topic models for information retrieval. In *ECIR*, pages 29–41. Springer, 2009.

[632] Sivan Yogev, Haggai Roitman, David Carmel, and Naama Zwerdling. Towards expressive exploratory search over entity relationship data. In *Proc. of WWW*, pages 83–92, 2012.

[633] Sang Ho Yoon, Ke Huo, Vinh P Nguyen, and Karthik Ramani. TIMMi: Finger-worn Textile Input Device with Multimodal Sensing in Mobile Interaction. In *Proceedings of the Ninth International Conference on Tangible, Embedded, and Embodied Interaction*, TEI '15, pages 269–272, New York, NY, USA, 2015. ACM.

[634] Sang Ho Yoon, Ke Huo, and Karthik Ramani. Wearable textile input device with multimodal sensing for eyes-free mobile interaction during daily activities. *Pervasive and Mobile Computing*, 33:17–31, 12 2016.

[635] Jeffrey M Zacks, Barbara Tversky, and Gowri Iyer. Perceiving, remembering, and communicating structure in events. *Journal of experimental psychology: General*, 130(1):29, 2001.

[636] Mohamed H Zaki and Tarek Sayed. Exploring walking gait features for the automated recognition of distracted pedestrians. *IET Intelligent Transport Systems*, 10(2):106–113, 2015.

[637] Tauhid R Zaman. *Information extraction with network centralities: finding rumor sources, measuring influence, and learning community structure.* PhD thesis, Massachusetts Institute of Technology, 2011.

[638] Oren Zamir and Oren Etzioni. Grouper: a dynamic clustering interface to web search results. *Computer Networks*, 31(11):1361–1374, 1999.

[639] Jing Zhang, Victor S Sheng, Tao Li, and Xindong Wu. Improving crowdsourced label quality using noise correction. *IEEE transactions on neural networks and learning systems*, 29(5):1675–1688, 2018.

[640] Xiaolong Zhang, Yan Qu, C Lee Giles, and Piyou Song. Citesense: supporting sensemaking of research literature. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 677–680. ACM, 2008.

[641] Yang Zhang and Chris Harrison. Tomo: Wearable, Low-Cost Electrical Impedance Tomography for Hand Gesture Recognition. In

Proceedings of the 28th Annual ACM Symposium on User Interface Software and Technology, UIST '15, pages 167–173, New York, NY, USA, 2015. ACM.

[642] Yang Zhang, Robert Xiao, and Chris Harrison. Advancing Hand Gesture Recognition with High Resolution Electrical Impedance Tomography. In *Proceedings of the 29th Annual ACM Symposium on User Interface Software and Technology*, UIST '16, pages 843–850, New York, NY, USA, 2016. ACM.

[643] Yang Zhang, Junhan Zhou, Gierad Laput, and Chris Harrison. SkinTrack: Using the Body As an Electrical Waveguide for Continuous Finger Tracking on the Skin. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, CHI '16, pages 1491–1503, New York, NY, USA, 2016. ACM.

[644] Yi Zhang, Dingding Wang, and Tao Li. idvs: an interactive multi-document visual summarization system. In *Machine Learning and Knowledge Discovery in Databases*, pages 569–584. Springer, 2011.

[645] Yudian Zheng, Guoliang Li, Yuanbing Li, Caihua Shan, and Reynold Cheng. Truth inference in crowdsourcing: Is the problem solved? *Proceedings of the VLDB Endowment*, 10(5):541–552, 2017.

[646] Wang Zhongqin, Ye Ning, Reza Malekian, Fu Xiao, and Wang Ruchuan. TrackT: Accurate tracking of RFID tags with mm-level accuracy using first-order taylor series approximation. *Ad Hoc Networks*, 53, 2016.

[647] Junhan Zhou, Yang Zhang, Gierad Laput, and Chris Harrison. AuraSense: Enabling Expressive Around-Smartwatch Interactions with Electric Field Sensing. In *Proceedings of the 29th Annual ACM Symposium on User Interface Software and Technology*, UIST '16, pages 81–86, New York, NY, USA, 2016. ACM.

[648] Jun Zhu, Zaiqing Nie, Xiaojiang Liu, Bo Zhang, and Ji-Rong Wen. Statsnowball: a statistical approach to extracting entity relationships. In *Proceedings of the 18th international conference on World wide web*, pages 101–110. ACM, 2009.

[649] John Zimmerman, Jodi Forlizzi, and Shelley Evenson. Research through design as a method for interaction design research in hci.

In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 493–502, 2007.

[650] Amal Zouaq, Dragan Gasevic, and Marek Hatala. Ontologizing concept maps using graph theory. In *Proceedings of the 2011 ACM Symposium on applied computing*, pages 1687–1692. ACM, 2011.

APPENDICES

Appendix A

Fostering Non-expert Annotators Reliability for Evaluating Information Extraction Outputs

As we explored different approaches to evaluating the output of Information Extraction systems in Chapter 3 (in particular Section 3.5.1.1) we highlighted different challenges to address. In particular, evaluating the output of IE systems is commonly done by matching the extracted entity-relationship triples with an available gold standard dataset where the mentions of entities and relationships are manually annotated by experts. While these reference datasets are essential for evaluation, once we move away from traditional IE to Open IE systems, gold standard datasets are not readily available, nor are there standard guidelines to construct the ground truth data to evaluate a new dataset. As a result, the construction of these datasets is extremely expensive in annotator-hours (and, as a result financially). One promising alternative is the use of non-expert annotations to judge the quality of system extractions, specifically to identify and categorize errors made by IE systems. Yet, ensuring high quality label assignment by non-domain-expert labelers remains an important challenge.

In this work, we contribute to enabling mechanisms for directly evaluating the output of IE systems for domains where no ground truth labels is available. To this end, we generalize the task of assessing semantic relationships between a series of entity pairs that are extracted by our IE system as an Error Categorization task to be conducted by non-expert annotators recruited from the Amazon Mechanical Turk platform. In order to guide the annotators through this task, two experts go through a vetting process to iteratively categorize errors in a sample of extractions and refine a set of guidelines to describe each of these error categories. The outcome is an annotation codebook that could be used in lieu of annotation task guidelines for the annotators. As a next step, we developed an interactive workflow that guides the annotators through the labeling task and ensures the guidelines that are derived by the experts are closely and consistently followed.

We evaluate this workflow using a subset of the extractions by our IE system that is run on both Politics and History topics, and demonstrate that our codebook-based interaction design significantly enhances labelers' agreement with expert annotators. Overall, this research describes the process of generating guidelines for producing ground truth labels in IE and demonstrates how a novel interaction design can be effectively leveraged to support non-experts' performance of an error categorization task [1].

A.1 Motivation

Labeling instances in a dataset is essential for tasks ranging from training machine learning models to assessing the quality of a variety of NLP system outputs. Techniques for collecting labeled data include recruiting experts for manual annotation [534], extracting relations from readily available sources (e.g. forums) [429], and automatically generating labels based on user behaviors [92].

Alongside the above methods for producing labeled datasets, researchers have also turned to crowdsourcing to generate labeled data as these platforms provide a scalable and efficient way to construct labeled datasets. Successful crowdsourced data labeling typically requires experts to communicate their desired definition of target labels to non-expert annotators

[1] The rest of this appendix is formatted to follow our story for the paper we are submitting to CHI 2021 based on this work and may not be consistent with the main theme of this dissertation.

through a set of guidelines explaining how instances should be labeled without leaving room for interpretation [100].

There are, however, challenges with producing labeling guidelines. Once we move from familiar categories – images of cats or traffic lights, for example – to more abstract labels, guidelines can grow in size and complexity which makes it very difficult for non-experts to follow without training. However, abstract or synthetic labels are common for a variety of task domains – error categorization, sentiment identification, citizen scientist tasks – and in these tasks labeling quality is affected by factors such as labelers' expertise or familiarity with the concept or data [318]. As a result, it can be difficult to train labelers [318], and labelers frequently go through a concept evolution phase as they progress in the task [309] which can lead to inconsistent labeling of similar instances by the same annotator over time.

One highlighted area of interest for the Human-Computation community in this year's call for papers is "interface techniques for augmenting human abilities to perform tasks". While past crowdsourcing research has explored a number of techniques for fostering labeling quality, other domains also explore issues of task design to enhance labeler reliability. The premise of our work is that the application of synthetic labels to data has much in common with the domain of open coding in qualitative research [593], specifically the idea that data must be analyzed, understood, and then assigned to a set of categories based on guidelines developed from the data set. We look specifically at how *annotation codebooks* are developed [344] because the goal of codebook design is to guide non-expert labelers in complex qualitative labeling tasks.

To explore the applicability of codebook design to synthetic labeling by crowdworkers, we present a 3-phase design process to devise an annotation codebook workflow for crowdworkers. Specifically, we elicit labels and processes from domain experts and synthesize these in codebook form, we tune the codebook through two studies: a pilot study with crowdworkers and an in-lab restrospective thinkaloud with non-expert labelers, and we assess our codebook workflow via a summative experiment. Our results indicate that our codebook-centric interaction design leads to statistically signficant increase in agreement with experts' annotations once aspects of workflow are tuned to refine structure and balance effort of label categories.

The remainder of this paper is organized as follows. We present related work on crowdsource reliability and on code book design in the next section. We then describe the design and refinement of our codebook. Finally, we evaluate our codebook workflow and discuss avenues of future work.

A.2 Related Work

A.2.1 Crowdwork Research on Worker Accuracy

Approaches to improving workers' accuracy include Micro-task Decomposition and Consensus Based Aggregation. Researchers have also started to explore novel interfaces to reduce label noise. For example, research has shown that open-ended questions yield higher reliability than Likert questions [293], that crowdsourcing crowdwork task design can yield better crowdworker performance [70, 178, 311], and that clear instructions and appropriate guidance can improve completion rates [177, 309, 335, 528].

Alongside the above mechanisms for improving accuracy, metrics have also been proposed to measure worker reliability, e.g., consensus and consistency. Consensus measures a worker's agreement with other workers [27, 493]. Consensus can be calculated for individual workers based on some metric that assesses proximity to consensus [441], based on some analysis of typical worker error [135, 137, 242, 433, 511, 597], or by leveraging statistical methods [258, 589, 639, 645]. The important aspect of this measure is that all variants represent some estimate of agreement between different between crowdworkers doing the same labeling task.

In contrast, consistency measures a worker's agreement with themselves [106, 134, 527]. The premise of this measure of reliability is based on a crowdworker's ability to yield the same output when repeatedly given a particular input. Crowdsourcing research has only begun to explore, at depth, how consistency should be used in practice for measuring of reliability for crowdworkers [134, 106, 208, 527]. Most closely related to the work in this paper, Williams et al. [607] studied workers' consistency in the context of crowdsourced object detection tasks in two different domains and found that workers' ability to yield consistent responses varied between domains, and was explainable by the duplicate task's difficulty and its position in the worker's task queue. As Williams et al. note [607], consistency is a valuable metric to explore given that most research in quality control in crowdsourcing seeks ways to make workers perform better as an individual. If workers are performing well – i.e. providing correct data – then they should continue to be correct when seeing a repeated task, i.e. they should have high consistency in their labeling [79, 229, 359].

A.2.2 Research on Non-Expert Labeling

Alongside the above crowdwork research, there exists a long history of research and practice, particularly in fields such as qualitative research [344, 521], that seeks to maximize the accuracy of non-expert labelers. In qualitative research, there is a necessity to *code* data, i.e. to attach synthetic labels to data such that data can be clustered according to its meaning and relevance for the problem domain [521]. This process of applying labels typically requires careful interpretation of the underlying content and judicious application of – sometimes ambiguously segmented – labeling categories.

In past work, Mitra et al. [368] contrasted training of labelers (by providing them with feedback on accuracy) with Bayesian Truth Serum, BTS, (where crowdworkers guess others' responses to the task and are rewarded for accuracy). Training was shown to better foster worker accuracy than BTS. Beyond the work of Mitra et al. [368], it seems that no work has explored, in depth, the benefit of more structured application of non-expert labeler guidance and from qualitative research. Labeler guidance typically involves two main steps: (1) Developing category labels to assign to data, i.e. a labeling 'codebook' [344]; (2) Designing an effective annotation workflow (analytical process) that guides non-expert labelers through the experts' codebook [344].

A.2.2.1 Developing an Annotation Codebook

Particularly when the process of data labeling, i.e. of *qualitative data coding* will be outsourced to non-expert labelers, clear guidelines are required to describe the labeling categories. These guidelines are frequently contained in an artifact known as an annotation codebook.

The procedure used to develop codebooks for qualitative analysis involves having experts independently engage in an inductive, iterative process to examine a sample of data, develop codes that describe aspects of that instance, and gradually increase their understanding of common groupings and patterns that emerge from the dataset [100]. The resultant ‘codebook’ of category labels is an important means for documenting the labels, the shared understanding of what each label means and the procedures for applying them [593]. A standard coding manual (developed by O’Connell (1989), later revised by [344]) includes six basic components: (1) the code (i.e. category label name), (2) a brief description, (3) a full description, (4) inclusion criteria, (5) exclusion criteria and (6) clarifying examples. Once the annotation codebook is developed by the experts, it can serve to train new annotators or to outsource annotation, thus freeing domain experts from the tedious coding task [593].

A.2.2.2 Designing Annotation Workflows for Non Experts

While labeling data is a seemingly simple task, past research shows that labeling is challenging, especially for complex domains (e.g. [318]). Labels reflect a labeler’s mapping between data and their underlying concept and is impacted by a labeler’s expertise or familiarity with the concept or data, their judgement, and the ambiguity of the data [309]. Given that codebooks are developed by a group of experts, it is not clear how a different group of annotators with limited expertise and who do not share the same mindset as the original expert assessors can perform complex annotation tasks consistently and efficiently.

To promote consistency in behavior among non-experts (where consistency is defined as similar actions being performed in similar situations [309]), past research has shown that interface design can play a central role in consistency of users interaction who are completing different tasks using the UI. For example, UI agents such as Wizards and Guides [149] are used to help prevent users from making mistakes by guiding them through different information tasks. Specifically in crowdwork, the design of crowdworker tasks has also been shown to enhance the quality of crowdsourced data. Researchers have argued that workflows that leverage multiple tools for the same task can improve worker output accuracy and minimize systematic errors [513]. Alongside this, UIs with less autonomy or task diversity may lead to a better performance (but a worse user experience)[515]. Data format may also influence crowdworker accuracy; as one example of this, Hutt et al. [239] noted that asking crowdworkers to input ranked evaluations versus Likert or binary categories results in higher accuracy. Finally, crowdsource workflows may break down a task into simpler sub-tasks and present them to the crowdworker one at a time to coordinate workers through completing a complex task [439].

In order to develop a task workflow that can support non-expert annotators throughout the annotation process, we utilize three design principles inspired by past research findings on task design and crowdworker training [368]:

1. *Incorporating guidelines as Hard Constraints.* For many interactive user interfaces, the sequence and methods of interaction is determined by hard constraints [149, 188]. Hard constraints are built into the interface by the designer and determine what patterns of interactive behavior are possible. In a labeling task, hard constraints can incorporate the desired labeling guidelines into the UI design, similar to a Wizard, to guide the worker through the decision making steps for each instance.
2. *Divide and Conquer.* Breaking down a complex task into simpler subtasks can reduce the cognitive effort of completing a task. There is evidence that information workers already implicitly break larger tasks down and that people perceive tasks in segments [635]. Microtasking is prevalent in crowdsourcing, where a number of workflows have been developed that decompose large, seemingly complex tasks into microtasks for goals such as as taxonomy creation and copy editing [106].
3. *Mimicking Expert's Categorization Approach.* During the coding phase, experts go through an iterative process of starting from broader categories and, as their understanding of underlying data evolves, begin to see finer discriminations within these broader categories [593]. As a result, the initial categories are broken down into subcategories that provide finer distinctions of initial higher-order categories. One implication of this observation is that categories can be grouped such that they have super-categories (i.e. parent categories) that encompass common attributes among them.

These design principles hint at a workflow design that mimics a decision tree, where leaves are the target category labels to apply and intermediate nodes are the sub-tasks.

A.3 Problem Domain

In this section we describe a problem domain that we use to investigate the efficacy of designing codebook-style workflows for non-expert annotators performing a complex labeling task. Because our goal is to explore codebook-style interaction support for *complex* labeling tasks, we wish to identify a synthetic labeling task where the application of labels requires judicious application of labeling categories. We leverage the domain of error categorization from Natural Language Processing (NLP), and in particular the categorization of errors from NLP-based Information Extraction (IE) algorithms.

A.3.1 Task Description and Dataset

A.3.1.1 Task Description

To explore codebook design to support crowdsourced labeling, we leverage an error categorization task where annotators are asked to evaluate relations that were automatically extracted by an Open Information Extraction (IE) system [474]. An IE system is a Natural Language Processing (NLP) System that extracts entities and their relationship from natural language text. Evaluating the output of IE systems is commonly done by matching the extracted entity-relationship triples with an available gold standard dataset where the mentions of entities and relationships are manually annotated by experts. While these reference datasets are essential for evaluation, once we move away from traditional IE to Open IE systems, gold standard datasets are not readily available, nor are there standard guidelines to construct the ground truth data to evaluate a new dataset. As a result, the construction of these datasets is extremely expensive in annotator-hours (and, as a result financially). One promising alternative is the use of non-expert annotations to judge the quality of system extractions, specifically to identify and categorize errors made by IE systems.

Our motivation for choosing this task as a representative complex labeling task is twofold: (1) high quality IE is essential for supporting a variety of downstream applications such as Question Answering, Generating Knowledge Graphs and Semantic Search; (2) Despite the widespread use of Open IE systems, there are no clear guidelines as to what constitutes a valid proposition to be extracted, and most OpenIE evaluations usually consist of a post-hoc manual evaluation of a small output sample by experts [516]. If we can support – through interaction design – non-experts (e.g. crowdworkers) in effectively perform this labeling task, then this can significantly expand the ability of NLP researchers to produce high quality IE datasets.

A.3.1.2 Dataset

We leveraged the document collection from Sarrafzadeh et al [479] that contains Wikipedia articles on two different topics: history of Canada and politics of Iran and Russia. We ran an available Open IE system described in [474] on this collection to generate a dataset of entity pair, relation label and reference sentence tuples: `entity` is usually a term or a noun phrase in text that corresponds to a concept in the domain, and `relation` corresponds to a simplified sentence that is semantically complete and describes how entity1 and entity2 are connected. For example, from the sentence “President of Iran who took office in August 2013 nominated his coalition cabinet members to the parliament.” the following tuple can be extracted: <president, parliament, “President nominates the Cabinet members to the Parliament”>.

A.4 Creating the Annotation Codebook

A.4.1 Codebook Design

Because there are no widely accepted annotation guidelines for assessing entity-relationship triples generated by Open IE systems, we recruited two experts to evaluate each extracted tuple by listing all syntactic and semantic errors as compared with the reference sentence each tuple was extracted from.

First, to develop initial codes, a subset of all system extractions were reviewed by the two experts to identify common error types observed in the output. The experts engaged in an iterative vetting process [2] to group similar errors and refine the formed groups until no more groups were found. As a result, a taxonomy of seven error categories was created (Table A.1). These categories cover different aspects of an extracted relation label between two entities including whether or not the entity pair are related and whether the extracted label is readable, meaningful and/or informative.

Category	Description
No Relation	E not related
Indirect Relation	E related through a third entity
Wrong RL	E related, but RL not readable
Incomplete RL	E are related, but RL incomplete
Misleading RL	E are related, but RL inconsistent with S
P. Unreadable RL	E related, RL has readability issues
Correct RL	E are related, and RL is accurate

Table A.1: A Taxonomy of Error Categories (Relation Label (RL), Sentence (S), Partially (P), Entities (E))

.

Following the standard codebook template described in above, each of these categories is accompanied with a description, the main criteria for instances that belong to this category, and a few clarifying examples, i.e., Table A.1 plus carifying instances define the codebook. Once the annotation codebook was finalized, we asked two expert annotators (two of this paper's authors) to label the data set independently. Given the 7-category labeling exercise, inter-rater agreement was measured via Cohen's Kappa as 0.509, highly significantly different from chance ($p < 0.001$).

Resolving Expert Disagreement.

Experts then engaged in a deliberation process to resolve disagreements on assigned labels in order to generate a closer proxy of ground truth. A more unified set of experts labels is helpful when analyzing similarities and differences between workers and experts and can also act as a metric

[2] Iterative vetting involves using the output of one stage to determine the next stage, incrementally reaching a final objective.

for evaluating the reliability of the workers accuracy, and whether or not they improve over time.

To this end, experts discussed their rationale for assigning their selected label for every instance where there was disagreement. The goal was to identify the cases where disagreement was caused by either misunderstanding the category definition or a lack of clarity in the guidelines. For example, considering the categories as outlined in Table A.1, *Incomplete RL* was most frequently confused with *Unreadable RL* and *Wrong RL*. The main reason behind this confusion was the lack of clear guidelines for distinguishing between readable and unreadable relation labels given that the label has omitted words.

While the above disambiguation resolved some ambiguous cases, there exist cases where further analysis indicated true ambiguity, i.e. that there could be more than one valid label. For ambiguous cases where there could be more than one valid category label, experts preserved their disagreement as an indication of the ambiguity of an instance (Entity - Relation - Sentence tuple). The assumption is that, if ambiguity exists for experts between two different categories, then an assignment to either category by non-experts demonstrates consensus with experts.

After the resolution process was completed, the inter-rater agreement was re-measured via Cohen's Kappa as 0.728, highly significantly different from chance ($p < 0.001$). Our motivation for measuring inter-rater agreement between the 2 experts before and after the deliberation process is twofold. While the resolved set of labels provide a closer proxy of the ground truth data, the Kappa scores indicate that (1) this labeling task is complex with an initial agreement rate of 51% between the experts; and (2) while the deliberation process improved agreement between the experts and helped refine the annotation codebook, for only 70% of instances both experts agreed on a single category label.

A.4.2 Designing the Codebook Workflow

We developed a workflow for the labeling task that uses hard constraints for guided task support, i.e. a 'wizard' that guides the user through the steps required to complete a task [149]. The workflow structure decomposes goals into small independent tasks. Through iterative design, we structured our workflow as a binary Decision Tree that guides assessors via yes/no decisions. The decision tree is shown in Figure A.1 alongside the interface that supports traversal of the decision tree. The crowd-worker, by answering these questions, is guided to a leaf node (label) that represents one of the error categories.

The decision tree was constructed by leveraging commonalities between different error categories. For example, both 'No Relation' and 'Indirect Relation' have, as their basis, the fact that there is no direct relationship between the two entities: they may have no relation or they may be connected by a third entity.

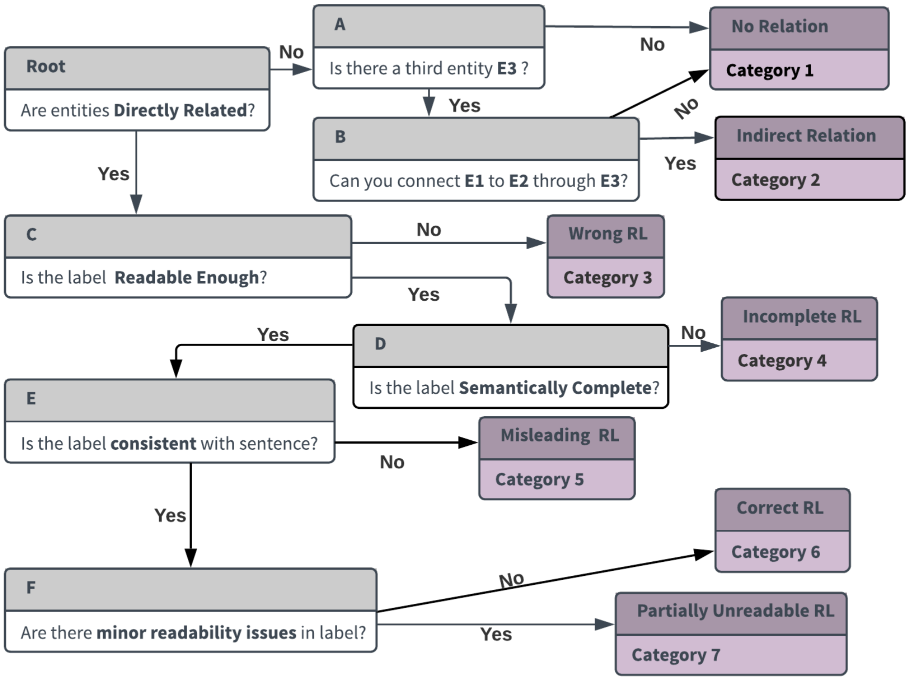

Root
Are entities Directly Related?
No
A
Is there a third entity E3 ?
No
No Relation
Category 1
Yes
B
Can you connect E1 to E2 through E3?
No
Yes
Indirect Relation
Category 2
Yes
C
Is the label Readable Enough?
No
Wrong RL
Category 3
Yes
D
Is the label Semantically Complete?
No
Incomplete RL
Category 4
Yes
E
Is the label consistent with sentence?
No
Misleading RL
Category 5
Yes
F
Are there minor readability issues in label?
No
Correct RL
Category 6
Yes
Partially Unreadable RL
Category 7

Figure A.1: Decision Tree for the Workflow Task Design

We performed two parallel studies to iterate on the codebook and the codebook workflow design. The first assesses the codebook via amazon mechanical turk, and the second uses a retrospective thinkaloud in lab.

A.4.3 Refinement Via Crowdworker Assessment

A.4.3.1 Procedure

We performed an initial evaluation of our codebook and workflow on Amazon Mechanical Turk[3] (AMT) with a small worker pool. A total of 26 workers were recruited through AMT as annotators for the study. To normalize crowdworker quality and efficiency, we ensured our workers had completed at least 500 tasks on AMT with a 90% acceptance rate.

Each annotator was asked to categorize 15 instances. A dataset of 135 instances was used as the main pool to generate task queues for each annotator. The task queues were generated such that (a) each queue contains a duplicate instance that is placed at a random position; (b) duplicate instances are never placed back to back since it could be perceived as a system error by the participants; (c) no queue contains more than 3 instances of the same category and hence each worker sees at least 5 categories); and (d) each instance is categorized by at least 3 annotators.

Annotators were paid $3.00 for completing the task. Before beginning the task, they were required to read a tutorial explaining the task and all seven categories. Additionally, participants completed three practice tasks before beginning the main task where they were given feedback on the correctness of their label.

A.4.3.2 Results and Observations

Our primary interest in this initial tuning experiment was to understand crowdworker error versus our expert labelers. We examined which category labels were assigned in error. The workers were biased towards the label *Unrelated Entities*, which appears to lead to higher consistency scores while hurting agreement scores. Overall, Unrelated Entities accounts for 35% of all category labels assigned. This is significantly different from the frequency of this label assigned by Experts (9%). While we cannot provide data that explains this bias, we hypothesize that the bias toward the Unrelated Entities label may be associated with the effort required to reach this category. This label has the shortest path from the root; therefore, it is the most convenient category to assign to an instance if the Worker wishes to minimize task completion time.

[3] https://www.mturk.com

A.4.4 Refinement Via Retrospective Thinkalound

Alongside assessing crowdworker behavior, recall that the format of the workflow was elicited from expert labeler behavior during the labeling task. However, non-expert labelers may not interact in the same fashion as expert labelers: they may misunderstand terms used by experts; their decision making process may differ; or other aspects of the codebook may be unclear. To assess these factors, we performed a lab-based study where seven participants used the workflow interface to perform the categorization task and were then taken through a retrospective thinkaloud [561] to explore how and why discrepancies occur between an individual and herself and between individuals and experts. Recall that, in a retrospective thinkaloud, participants perform a task in the traditional way, but the task is video recorded. Participants then re-visit their task and thinkaloud while watching their interaction [561]. Retrospective thinkalouds are one way to capture thinkaloud data from a cognitively demanding task where concurrent thinkalouds might interfere with task performance.

The retrospective thinkaloud task proceeded as follows:

- Participants labeled 20 instances (10 unique + 10 duplicates) using the workflow interface.
- Once done, they were asked if they noticed any duplicates (to test whether the duplicates were too easily detectable).
- Next, three of the assigned instances were automatically selected by the system based on ordered criteria as follows:
    1. Instance labels are inconsistent and in complete disagreement with the experts labels;
    2. Instance labels are inconsistent and in partial agreement with the experts (i.e. agrees with at least one expert's label);
    3. Instance labels are consistent but are in complete disagreement with the experts labels;
- These instances are presented to the participant and they were asked to complete the task one more time while thinking out aloud. Note that, for these three instances we collected a set of three (possibly overlapping) categories where each assigned category corresponds to a path taken from the root of the Workflow tree to the leaf category.

- All three paths were then visualized and presented to the participant as a means of reflecting back on the thought process that occurred during the categorization process.
- Finally, the participants were presented with a static interface that simply listed the categories and were asked to comment on how this new interface could help or hurt their experience.

We collected both descriptive statistics and qualitative data from 7 participants (3 female).

A.4.4.1 Statistics

One question we asked participants was whether they noted repeated instances of labeling. Interestingly while all 10 instances were duplicated in this experiment, the majority of participants did not notice that instances were presented twice.

Analyzing the assigned category labels confirmed a very high consistency rate of 87% (for the last 2 category labels assigned to each instance) as well as a high level of observed agreement with experts (average agreement: 46% compared with the initial agreement between the experts of 58%).

Finally we observed that for the majority of cases where the participant was inconsistent in their assigned labels they were actively trying to perform better at the task. In fact, out of all inconsistent category labels (corresponding to 13% of all cases), 65% resulted in a correct category label as opposed to 18% for the opposite case.

A.4.4.2 Qualitative Findings

We performed a standard process of open coding on transcriptions of the think-aloud data, and identified three key themes: (1) the impact of Category Names on the categorization outcome; (2) consistency as a questionable quality reliability metric; and (3) complementarity in the strengths of the task design.

Category Names Participants commented on the choice of category labels and if they matched their understanding of what that category entails. The category names can be interpreted and internalized differently from the intermediate questions that lead to that category and the choice of words in those questions (e.g. ‘consistency’, ‘sufficient information’, etc). For example [P1] was surprised every time she arrived at the ”Wrong Relation Label” category for an instance as they perceived the entities to be unrelated and that’s what she expected this category meant. In the new version of the Interface we renamed this category as “Unreadable Relation Label”.

Alongside Category names, different wordings of guidelines and intermediate questions were refined to improve the clarity and usability of the task interface, thus enhancing participants understanding of categories.

Consistency as a Quality Metric The retrospective walkthroughs provided an opportunity for participants to reflect on the causes of inconsistencies within their own labels and whether consistency was a reliable quality metric. Consistency was perceived to be a questionable metric: while highly consistent annotators lead to more predictability of annotations outcome they can simply be applying the same category over and over without any effort in following the guidelines and improving their understanding of the task. On the other hand, we can have highly inconsistent annotators who seek to improve their annotation accuracy as their knowledge of the task evolves. P3 notes they were “the most inconsistent in the cases I spent much more time investigating.”

Contrasting Workflow and Static Interfaces for Categorization Because participants were presented with a static interface and were asked to contrast it with the Workflow interface for error categorization, we collected strengths and weaknesses of structured workflow. The static interface was deemed to be much easier to use and needed less time to complete the task. However, different participants mentioned that having too many options at once (i.e. 7 error categories to choose from) can be very overwhelming when learning the task. We leverage this data to design a final labeling interaction workflow that balances effort across conditions and allows multiple decisions at leaf nodes.

A.4.5 Final Codebook and Workflow Design

We created a refined workflow design shown in Figure A.2. Contrasting Figures A.1 and A.2, the Hyrbid structure of the new workflow combines

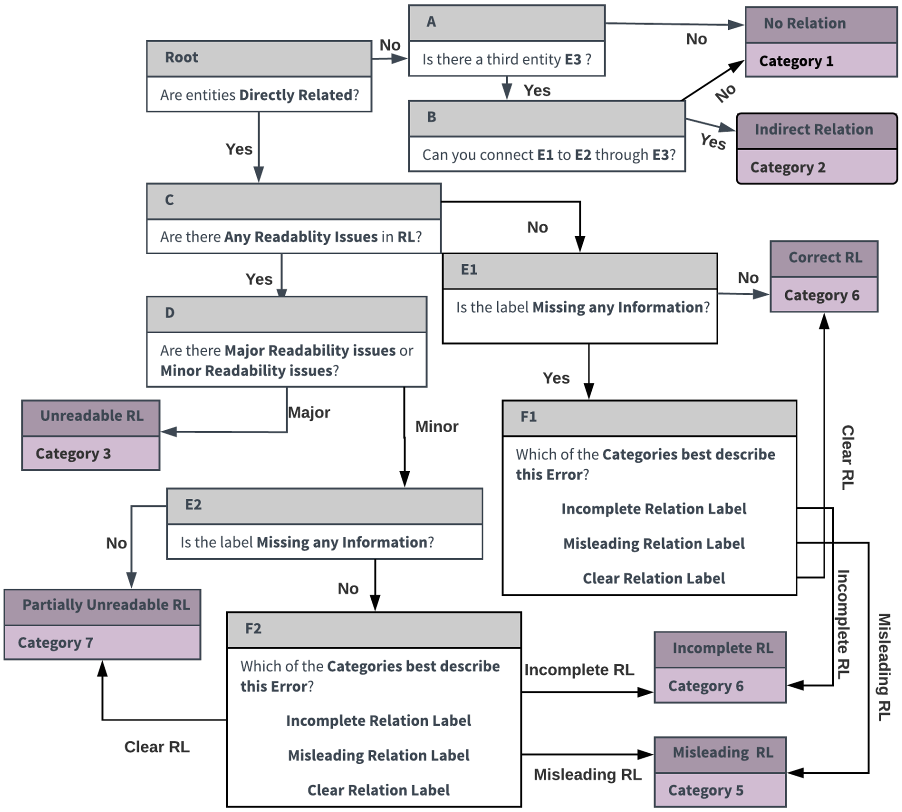

Root
Are entities Directly Related?
No
A
Is there a third entity E3 ?
No
No Relation
Category 1
Yes
B
Can you connect E1 to E2 through E3?
No
Yes
Indirect Relation
Category 2
Yes
C
Are there Any Readablity Issues in RL?
No
E1
Is the label Missing any Information?
No
Correct RL
Category 6
Yes
D
Are there Major Readability issues or Minor Readability issues?
Yes
Major
Minor
Unreadable RL
Category 3
F1
Which of the Categories best describe this Error?
Incomplete Relation Label
Misleading Relation Label
Clear Relation Label
Clear RL
E2
Is the label Missing any Information?
No
Partially Unreadable RL
Category 7
No
F2
Which of the Categories best describe this Error?
Incomplete Relation Label
Misleading Relation Label
Clear Relation Label
Clear RL
Incomplete RL
Incomplete RL
Category 6
Incomplete RL
Misleading RL
Misleading RL
Category 5
Misleading RL

Figure A.2: Decision graph to support the hybrid interface

aspects of a static design with the workflow interface. Modifications are three-fold: (a) the hybrid structure includes a mix of binary and multiple choice at each level to more efficiently guide the decision making process; (b) the new structure is not a tree structure; there is redundancy in the paths to a final category label which helps workers recover from decision making errors earlier in the categorization process; (c) the overall cost of reaching different category nodes in the new structure is more uniformly distributed when compared with the workflow structure. Because this workflow contains aspects of both static and workflow conditions from our earlier study, we label it a hybrid design.

To balance workload for different labels, alongside the more structured workflow we required workers to provide a rationale any time they answered “No” to the question “Are these entities directly related?” For this response, they need to provide two other entities from the same sen-

	Agreement		Agreement w/ Consensus		Consistency		Consensus	Experts
	BA	MA	Exp1-c	Exp2-c	BC	MC	Rate	Agreement
Static-just-off	0.26	0.36	0.18	0.17	0.56	0.64	0.58	0.77
Static-just-on	0.28	0.38	0.06	0.12	0.51	0.61	0.60	0.68
Workflow-just-off	0.26	0.35	0.10	0.18	0.49	0.55	0.41	0.64
Workflow-just-on	0.34	0.43	0.36	0.37	0.53	0.62	0.53	0.71

Table A.2: Contrasting Binary Agreement (BA), Micro Agreement (MA), Inter-Experts Agreement (Experts), Agreement between Consensus for each expert individually (Exp1-c, Exp2-c), Binary Consistency (BC), Micro Consistency (MC), Consensus Rate, and Expert Agreement for Static and Workflow Conditions with and without justification.

tence which were directly related. This requirement was added for two reasons: First, it eliminates the low effort of labeling entities as unrelated; and it also provides valuable information to the NLP analysis such that errors can be corrected.

A.5 Evaluating the Codebook Design

While the overall principle of a qualitative codebook (labels for categories, category desciptions, and examples) is a valuable design template for structuring labeling task descriptions, our codebook workflow includes two additional design elements: a divide and conquer decision workflow that structures decision making, and a justification requirement for one condition to ensure balanced work across conditions. We *hypothesize* that the *combination of these elements* are needed in codebook workflows to *foster enhanced accuracy.*

To evaluate our codebook workflow, i.e. to test our hypothesis that dynamics workflow elements improve accuracy in labeling, we conducted a 2X2 between subjects experimental evaluation with two different workflow conditions (static versus workflow) and two different justification conditions (justification versus no justification). This yielded four different experimental conditions representing combinations of dependent variable: static-just-off; static-just-on; workflow-just-off; and workflow-just-on.

A.5.0.1 Participants

We evaluate our codebook workflow using Amazon Mechanical Turk[4] (AMT) with 200 crowdworkers performing an Entity-Relationship Error labeling task. Workers are split equally between the 4 experimental conditions, yielding 50 workers per condition. As in our earlier pilot study, to normalize crowdworker quality and efficiency, we ensured our workers had completed at least 500 tasks on AMT with a 90% acceptance rate.

A.5.0.2 Method

Each annotator was asked to categorize 15 instances of entity-relationship tuples. A dataset of 135 instances was used as the main pool to generate task queues for each annotator. The task queues were generated such that (a) each queue contains one duplicate instance that is placed at a random position; (b) duplicate instances are never placed back to back since it could be perceived as a system error by the participants; (c) no queue contains more than 3 instances of the same category and hence each worker sees at least 5 categories); and (d) each instance is categorized by at least 3 annotators.

Annotators were paid $3.00 for completing the task. Before beginning the task, they were required to read a tutorial explaining the task and all seven categories. Additionally, participants completed three practice tasks before beginning the main task where they were given feedback on the correctness of their label.

Prior work has shown that various characteristics of a task (e.g. difficulty, sequence) can affect the way workers perform tasks [85]. In order to eliminate task difficulty and ordering as a confound between conditions, we assigned the same set of generated task queues to both static and workflow conditions.

A.5.0.3 Measures

To analyze the effect of workflow on user labeling behavior, our primary measure is agreement with experts, which we measure in two ways: Basic

[4] https://www.mturk.com

Agreement, and Agreement with Consensus. Basic Agreement is further subdivided into binary agreement and micro-agreement.

Basic Agreement reports the average observed agreement for all instances labeled by annotators in the same condition, i.e. the percentage of labels that agree with expert annotators. We analyze basic agreement in two ways:

1. *Binary Agreement*: For each annotator's label, if a worker agreed with at least one of the experts, we assigned them a score of 1. Otherwise, we assign them a score of 0. Binary Agreement scores are in the range [0, 1] and represent the fraction of crowdworker labels that agree with at least one expert labeler.
2. *Micro Agreement*: One problem with binary agreement is that some label categories are more similar than others. Micro Agreement penalizes unrelated labels in the codebook more than labels that are more similar by leveraging the structure of the workflow. Specifically, we group the 7 categories into 2 classes (i.e., similar and not-similar) based on whether the 2 categories share a common parent. For every label, we assign a score of 1 if the worker label agrees with at least one expert label; we assign a score of 0.5 if labels do not agree with at least one expert label but the label is similar to an expert label, i.e. the label and an expert label have a common parent in the workflow; and a score of 0 is assigned otherwise. Micro Agreement is measured on a [0, 1] scale.

Agreement with Consensus, instead, considers the percentage of time that the expert label agrees with the majority vote for a label. Recall that in the experimental method, we ensured that each instance is categorized by at least 3 crowdworkers, meaning that, even with errors, a plurality of crowdworkers may select one label. Agreement with Consensus measures the fraction of instances in the range [0, 1] where the consensus label agrees with the consensus label. We show agreement with consensus scores for each individual expert.

Finally, to ensure that measures such as worker consistency and overall consensus between workers are not impacted, we report these values as well. Again, we break down consistency into binary consistency (where a plurality of workers agree with each other), and micro-consistency, where they receive partial agreement (0.5) if a plurality select from classes with a common ancestor.

A.5.1 Results

Table A.2 shows results of Agreement, Agreement with Consensus, Consistency, and Consensus for each of our four experimental conditions. We also show expert agreement with each other, a theoretical upper bound on performance for consensus measures.

Examining agreement scores, an overall CHI square test reveals statistically significant differences within the tabular data ($\chi^2(df = 6, N = 200) = 27.11$, $p < 0.001$). Bonferroni corrected post-hoc tests involving individual measures shows that, for Agreement with Consensus workflow-just-on is highly statistically significantly ($p < 0.001$) different than other conditions. All other measures are not significant, and other experimental conditions (Workflow-just-off and Static-just-on/off) also do not differ significantly. Finally, measures of consistency or consensus also do not vary by statistically significant margins.

A.6 Discussion

We began the previous section by noting that, alongside a regimented structure for presenting synthetic labels to non-expert labelers, the application of codebook principles to labeling required two things: a workflow to guide labelers through the labeling task; and balanced workload to discourage crowdworkers from "gaming the system", i.e. from choosing a low-effort label, thus simplifying the labeling task.

Our data supports the hypothesis that both aspects of codebook implementation are necessary to increase crowdworker accuracy. Without justification, we see a repeat of earlier crowdworker behavior, where workers select easily reachable categories more consistently to maximize their throughput. We assume that this behavior results from the low-wage environment typical of AMT (median wages as low as $1.38 per hour [231]). Furthermore, simply turning on justification does little to promote crowdworker agreement with experts.

We believe that the primary outcome of this work is an overall guideline for developing and applying codebook interaction to complex labeling tasks for non-expert, crowdworker labelers. The design of the codebook-based workflow can be summarized as a three-step process, where the task requester designs the task as follows:

1. Capture expert labelers' codes and the decision-making process for assigning labels. This step typically involves small scale labeling by experts to develop code labels, independent labeling applying the codes to measure clarity of the guidelines and to allow experts to begin to develop analytical approaches to label assignments, and a resolution step where experts compare errors to refine label and analytical approaches to create a clear methodological description of their work process.
2. Pilot the codebook via retrospective thinkalouds to tune decision tree and place error recovery. While expert labelers can develop label categories and can describe their analytical process, non-expert labelers may struggle to behave in an identical fashion to expert labelers. While this step may seem costly, we posit that retrospective thinkalouds with even a handful of participants can provide valuable insight that can serve to refine the codebook workflow and category labels [383].
3. Pilot the codebook on mechanical turk to analyze crowdworker error and refine the workflow to prevent these errors.

While our codebook workflow did increase agreement with experts, we did not see any changes in crowdworker consensus or self-consistency. Ideally, it would be desirable to see measures such as consensus and consistency correlate with increased agreement with experts, primarily because consensus and consistency can be measured directly from crowdworker data. Potential problems with consistency as a measure of crowdworker accuracy were also noted by participants in our qualitative data collected during our retrospective thinkaloud design study. We believe that one important area of future work is to more fully probe when and how consensus and consistency measures can be used as a proxy of crowdworker accuracy, and, in particular, when and how they fail to correlate with crowdworker accuracy.

A.7 Conclusion

Researchers in the human computation domain are well aware that interface design – and by extension interaction design – is an important aspect in enhancing human abilities when performing complex human-intelligence tasks (aka HITs). In this work, we specifically draw inspiration from the domain of qualitative research, and in particular, the design of codebook-based workflows for qualitative coding. Our results highlight how design around these traditional codebook-style data labeling practices can serve to enhance worker accuracy for labeling tasks where the labeling categories – as in qualitative research – require more judicious application of analytical reasoning to the labeling process.

Appendix B

Study Materials

In this appendix we provide different materials used for conducting the user studies that were described in Chapters 5, 6 and 7.

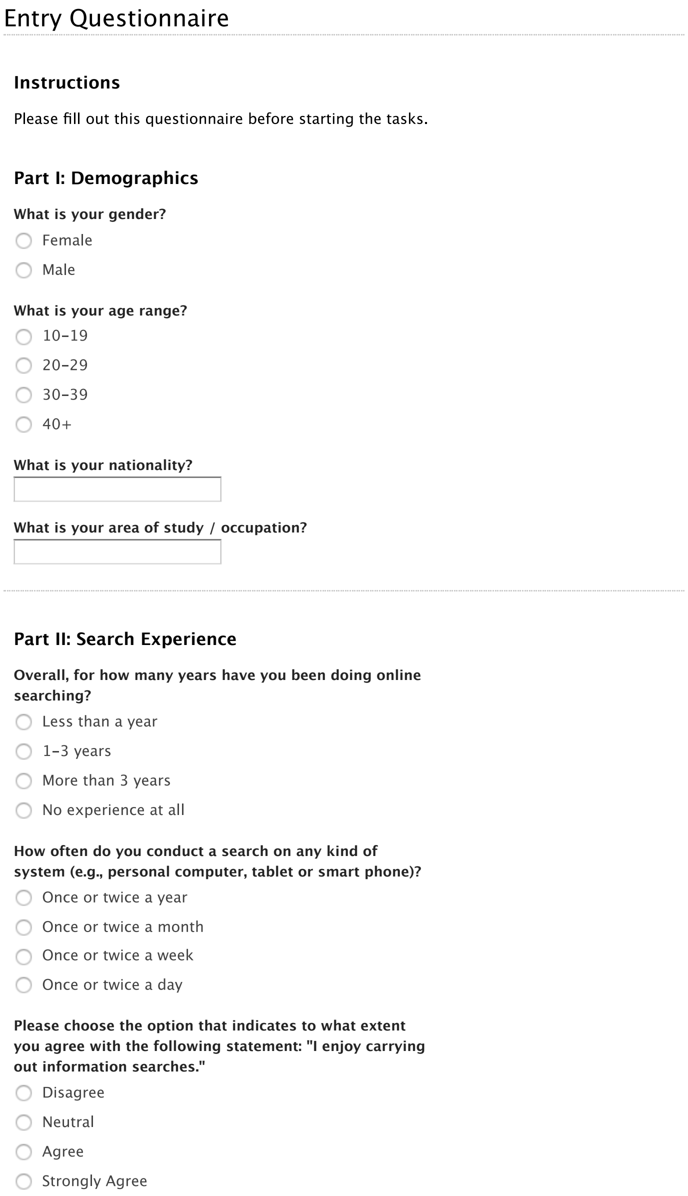
B.1 Questionnaires

All demographics, pre and post task questionnaires that we used for our in lab studies are available online at `https://cs.uwaterloo.ca/~bsarrafz/HKG/Experiment/Forms`. Figure B.1 indicates the entry questionnaire designed to collect some demographics information from the participants.

A set of pre-task questionnaires were designed for each search task in order to gauge participants' interest in and the prior knowledge of each assigned topic. Participants prior knowledge of the topic was collected both as a self-rated score as well as their responses to a set of pre-defined questions. Figures B.2 and B.3 show these questionnaires for Politics and History tasks respectively. After each task is completed participants were directed to fill out a post-task questionnaire as shown in Figure B.4.

B.2 Evaluating Search Tasks Outcomes

In Chapters 5, 6 and 7 we used two main topics – Politics and History – to design search tasks at two levels of complexity. For the *Simple Tasks* a



Entry Questionnaire

Instructions

Please fill out this questionnaire before starting the tasks.

Part I: Demographics

What is your gender?

○ Female

○ Male

What is your age range?

○ 10–19

○ 20–29

○ 30–39

○ 40+

What is your nationality?

What is your area of study / occupation?

Part II: Search Experience

Overall, for how many years have you been doing online searching?

○ Less than a year

○ 1–3 years

○ More than 3 years

○ No experience at all

How often do you conduct a search on any kind of system (e.g., personal computer, tablet or smart phone)?

○ Once or twice a year

○ Once or twice a month

○ Once or twice a week

○ Once or twice a day

Please choose the option that indicates to what extent you agree with the following statement: "I enjoy carrying out information searches."

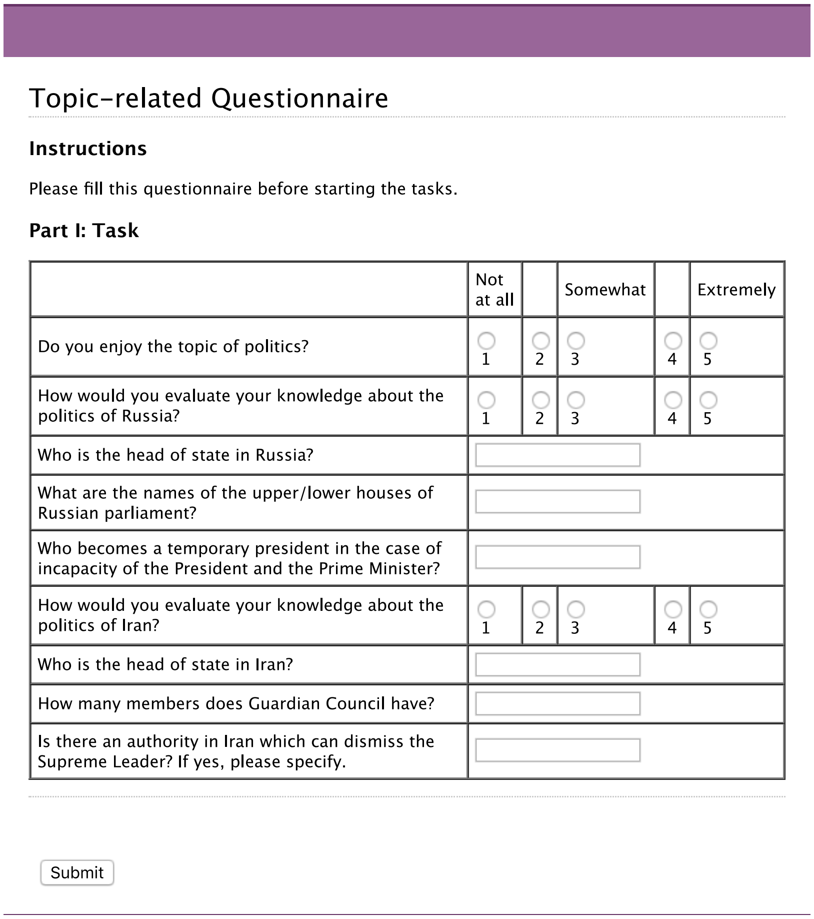
○ Disagree

○ Neutral

○ Agree

○ Strongly Agree

Figure B.1: Demographics Questionnaire



Topic-related Questionnaire

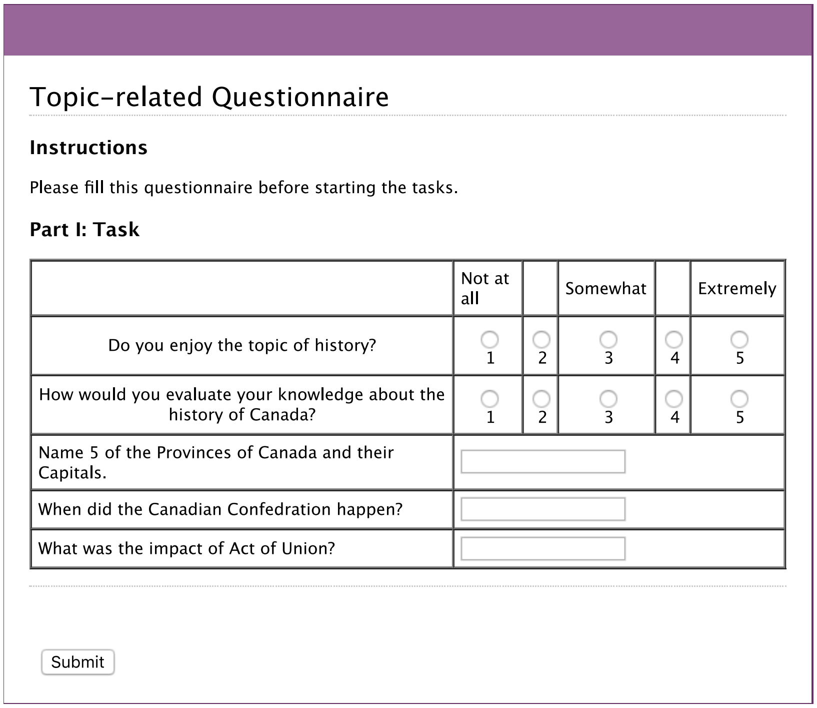
Instructions

Please fill this questionnaire before starting the tasks.

Part I: Task

	Not at all		Somewhat		Extremely
Do you enjoy the topic of politics?	○ 1	○ 2	○ 3	○ 4	○ 5
How would you evaluate your knowledge about the politics of Russia?	○ 1	○ 2	○ 3	○ 4	○ 5
Who is the head of state in Russia?					
What are the names of the upper/lower houses of Russian parliament?					
Who becomes a temporary president in the case of incapacity of the President and the Prime Minister?					
How would you evaluate your knowledge about the politics of Iran?	○ 1	○ 2	○ 3	○ 4	○ 5
Who is the head of state in Iran?					
How many members does Guardian Council have?					
Is there an authority in Iran which can dismiss the Supreme Leader? If yes, please specify.					

Figure B.2: Pre Task Questionnaire for Politics



Topic-related Questionnaire

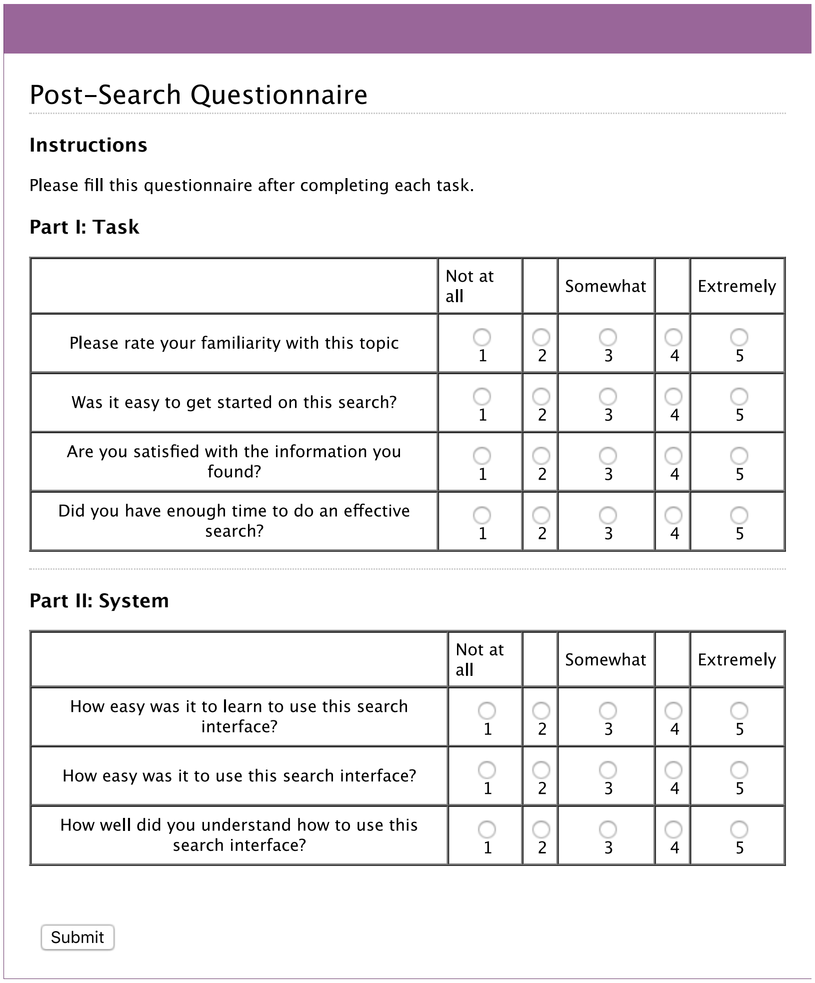
Instructions

Please fill this questionnaire before starting the tasks.

Part I: Task

	Not at all		Somewhat		Extremely
Do you enjoy the topic of history?	○ 1	○ 2	○ 3	○ 4	○ 5
How would you evaluate your knowledge about the history of Canada?	○ 1	○ 2	○ 3	○ 4	○ 5
Name 5 of the Provinces of Canada and their Capitals.					
When did the Canadian Confedration happen?					
What was the impact of Act of Union?					

Figure B.3: Pre Task Questionnaire for History



Post-Search Questionnaire

Instructions

Please fill this questionnaire after completing each task.

Part I: Task

	Not at all		Somewhat		Extremely
Please rate your familiarity with this topic	○ 1	○ 2	○ 3	○ 4	○ 5
Was it easy to get started on this search?	○ 1	○ 2	○ 3	○ 4	○ 5
Are you satisfied with the information you found?	○ 1	○ 2	○ 3	○ 4	○ 5
Did you have enough time to do an effective search?	○ 1	○ 2	○ 3	○ 4	○ 5

Part II: System

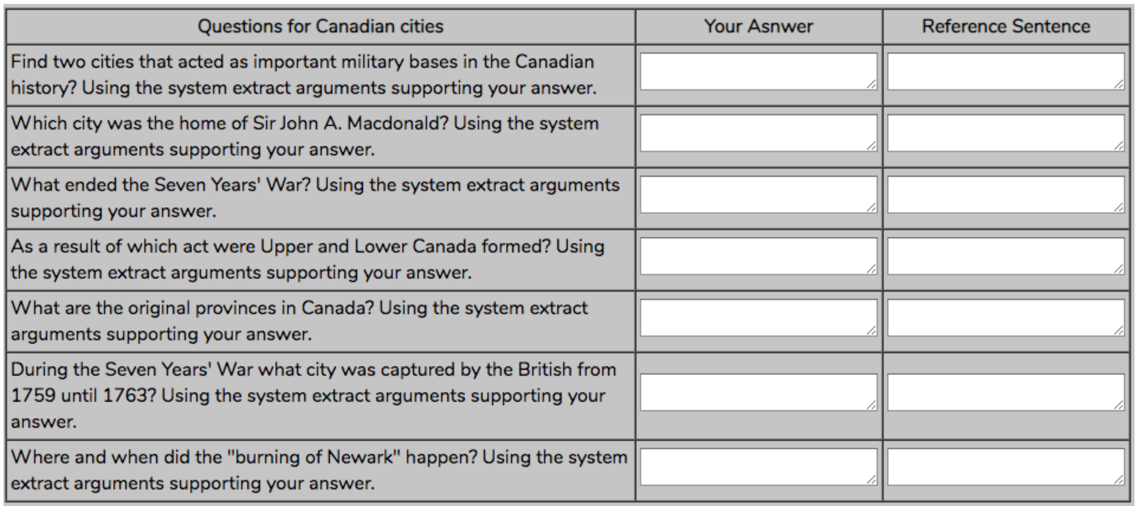
	Not at all		Somewhat		Extremely
How easy was it to learn to use this search interface?	○ 1	○ 2	○ 3	○ 4	○ 5
How easy was it to use this search interface?	○ 1	○ 2	○ 3	○ 4	○ 5
How well did you understand how to use this search interface?	○ 1	○ 2	○ 3	○ 4	○ 5

Figure B.4: Post Task Questionnaire - Same questionnaire was used for both tasks.

series of factoid questions were created where the answers could be one or more entities. We framed the *Complex Tasks* as essay writing tasks which would ask the participants to collect enough information to be able to write an essay on the assigned topic with at least three main arguments. The next two subsections present the questions and the task descriptions used for these two types of tasks.

B.2.1 Simple Tasks

Figures B.5 and B.6 indicate the factoid questions used for History and Politics tasks respectively. To assess participants' performance for the simple tasks, responses were graded as 0.5 mark(s) per correct answer and 0.5 mark(s) per correct citation of reference sentence, for a total of 1 mark per question.

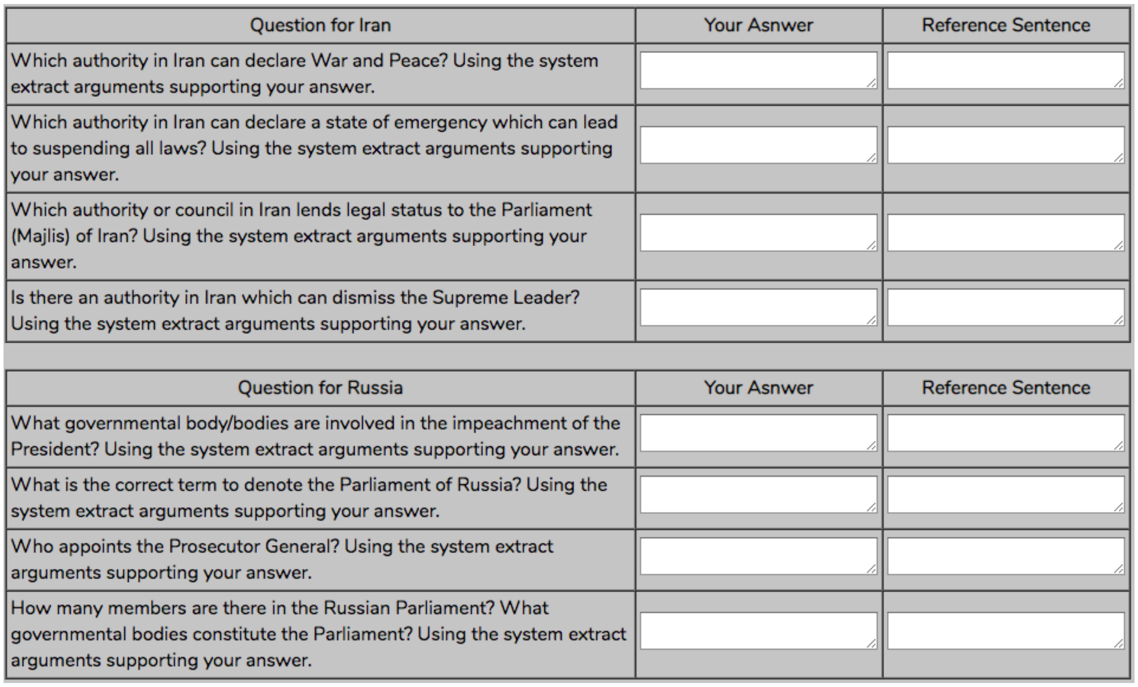


Questions for Canadian cities	Your Asnwer	Reference Sentence
Find two cities that acted as important military bases in the Canadian history? Using the system extract arguments supporting your answer.		
Which city was the home of Sir John A. Macdonald? Using the system extract arguments supporting your answer.		
What ended the Seven Years' War? Using the system extract arguments supporting your answer.		
As a result of which act were Upper and Lower Canada formed? Using the system extract arguments supporting your answer.		
What are the original provinces in Canada? Using the system extract arguments supporting your answer.		
During the Seven Years' War what city was captured by the British from 1759 until 1763? Using the system extract arguments supporting your answer.		
Where and when did the "burning of Newark" happen? Using the system extract arguments supporting your answer.		

Figure B.5: List of Factoid Questions for the Simple Task on History of Canada

B.2.2 Complex Tasks

In all of our experiments for at least one of the assigned tasks the participants would engage in an exploratory information seeking task in order to write a short essay articulating their responses for the given task. To assess participants performance for the complex tasks we designed



Question for Iran	Your Asnwer	Reference Sentence
Which authority in Iran can declare War and Peace? Using the system extract arguments supporting your answer.		
Which authority in Iran can declare a state of emergency which can lead to suspending all laws? Using the system extract arguments supporting your answer.		
Which authority or council in Iran lends legal status to the Parliament (Majlis) of Iran? Using the system extract arguments supporting your answer.		
Is there an authority in Iran which can dismiss the Supreme Leader? Using the system extract arguments supporting your answer.		

Question for Russia	Your Asnwer	Reference Sentence
What governmental body/bodies are involved in the impeachment of the President? Using the system extract arguments supporting your answer.		
What is the correct term to denote the Parliament of Russia? Using the system extract arguments supporting your answer.		
Who appoints the Prosecutor General? Using the system extract arguments supporting your answer.		
How many members are there in the Russian Parliament? What governmental bodies constitute the Parliament? Using the system extract arguments supporting your answer.		

Figure B.6: List of Factoid Questions for the Simple Task on Politics of Iran and Russia

marking schemes for evaluating the quality of the essays provided by our participants.

The marking scheme for the Politics task requires 3 main arguments to be included to support one of the presidents to be more powerful than the other. In order to ensure these arguments are based on the information that was retrieved using the system (and not based on the participant's prior knowledge) each argument needs to have references to the source document. Additionally, each argument is graded as follows:

- 1 point: an argument only refers to one president;
- 2 points: an argument compares both presidents on the same aspect (e.g. military power, rank in the political system, etc.);
- 3 points: same as previous, but the argument also contains information about the authority that limits the president's power (requires a broader understanding of the political system and the power relationships between different political entities);

For the History task, the marking scheme indicates 6 cities as previous capitals of Canada and specifies the marking criteria as:

- 1 point: 3 or fewer correct capitals are listed;
- 2 points: at least 4 correct capitals are listed;
- 2 points: for every well described reason behind changing a capital that is explicitly mentioned in a reference document;
- 1 point: for every questionable or subjective reason provided which includes a reference document;

B.3 Semi-Structured Interviews

In this section we provide a list of questions that guided our semi-structured interviews for studies presented in Chapters 6 and 7.

For the study presented in Chapter 6 each participant used the HKG UI to complete one simple and one complex task (with a randomized order). Once both tasks were completed we conducted semi-structured interviews to gauge their experience with the interface, whether HKGs are more suitable for one task than the other and how they fare against our previously designed Hierarchical Tree UI (described in Chapter 5).

1. Part 1: Contrasting the utility of the HKG interface for simple versus complex tasks.
   (a) Now that you used the same interface for two types of tasks; for which type do you find it more suitable? for question answering task or more open ended one?
   (b) How about your usual Web search experience? How would you go about performing these two tasks using a search engine? How is that different from using this new interface?
   (c) Would your domain knowledge change your preference about the type of task you would do with this interface?
2. Part 2: Demonstrating the Hierarchical Tree based UI as a reference interface, allowing the participant to explore the features and get a *feel* of how completing search tasks using the reference interface would look like.

(a) Now that you did a question-answering task and an essay writing task, you can imagine you were given the other interface [Tree-based UI] instead. How would you like it? How does it compare with the [HKG] UI you used for these tasks?

(b) What are other factors that might affect your choice of interface? (e.g. the search domain, task type, prior knowledge, etc.)

(c) Are there specific scenarios where you prefer a hierarchical tree UI over the HKG interface?

In the followup study presented in Chapter 7 participants used the same HKG UI to complete two complex search tasks. This provided an opportunity to gauge participants' perception of the exploratory search tasks we designed and better characterize their main attributes. As well, we investigated how HKGs of varying quality were perceived by participants and how information seeking can proceed in presence of errors. What follows is the main steps that were generally followed in the interviews:

1. Perception of Task Complexity: how did you feel about the complexity of these tasks? how were they different?
2. Perceived Quality of Generated Knowledge Graphs: how did you find the quality of the information represented by the graphs? was one better than the other? did you notice any errors, inconsistencies, incomplete or unreadable sentences, missing information, etc?
3. Notifying the participant about the experimental conditions: one of your tasks used automatically generated knowledge graphs while the other used experts-curated data. Can you guess which task was done with the automatically generated graphs?
4. Discussing the reasons that might have hindered participants from noticing errors or lower quality of graphs.
5. Contrasting different characteristics of simple versus exploratory search tasks and how they might be impacted by errors. This step involved using the UI for a few simple queries in order to familiarize the participant with factoid type queries.
6. Inquiry regarding a factoid question where the relevant edge in the graphs was erroneous by design.