UNIVERSITY OF CALIFORNIA
SANTA CRUZ

REPETITION AND DIVERSIFICATION IN MULTI-SESSION TASK ORIENTED SEARCH

A dissertation submitted in partial satisfaction
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

COMPUTER SCIENCE

by

Sarah K Tyler

December 2013

The Dissertation of Sarah K Tyler
is approved:

Professor Yi Zhang, Chair

Professor David Helmbold

Professor Marilyn Walker

Professor Jamie Callan

Tyrus Miller
Vice Provost and Dean of Graduate Studies

Contents

List of Figures

List of Tables

Abstract

Repetition and Diversification in Multi-Session Task Oriented Search

Sarah K Tyler

As the number of documents and the availability of information online grows, so to can the difficulty in sifting through documents to find what we're searching for. Traditional Information Retrieval (IR) systems consider the query as the representation of the user's needs, and as such are limited to the user's ability to describe the information he or she is seeking. Several challenges can arise within this single query-results paradigm.

In this model the user is required to know enough about the domain to express his or her information need. Constructing a query can be difficult in it's own right. Without being well versed in the jargon of a new domain, it may be difficult to formulate a query that adequately expresses one's information need. It may also be challenging to understand the contents of a document, let alone how a given document relates to other documents in the field. This fact can make it more challenging for the user to identify the relevant documents out of the search results. Thus learning about new ideas and new topics that one is unfamiliar with can be challenging.

In many situations it may not be possible for all the information a user is seeking to be present in a single source, or to be returned within a single query or small set of queries. The knowledge sought by the user may be too broad or too deep. Someone researching medical treatments may not find one single source with enough information to make an adequate decision regarding his or her health. In a literature review, the absence of information on a given idea is itself information. An idea could be so novel that it's never been attempted or discussed. To know that an idea of is novel, however, would require a good understanding of many documents,

not just any single source. The user would require a solid grasp on the entire search space, not just the search results from a single query.

Finally, the information itself may be changing. As the medical community evaluates new drugs and treatments that might outperform current methods, society's understanding of bioscience grows as does the literature and information available to the searcher. Similarly, new research continues to be published at conferences and in journals, by academics and industry alike. A literature review from a few years ago may be considered incomplete and obsolete today. New information is being built upon past information. Thus a user may need to constantly go back and revisit past documents and information already acquired, even as they continue to seek out new information.

In this dissertation we present explore multi-session task oriented search. We begin this dissertation with a study of on-going multi-session search, search that cannot be satisfied within a single query or session. We focus primary on repetitive queries and repeat search result clicks to find patterns in the way individuals search. These actions can be viewed as a simple form of on-going search. When studying repetitive queries we discover that as users continue to issue the same query, they show a desire to return to past content as well as find new information. We find users are selective in their new-findings, and even though users may revisit the same URLs frequently, these new-findings might offer more insight into the user's underlying interest.

We then study repeat findings, when a user revisits a webpage following a search result. We find a distinction between intra and cross session re-findings, where the former suggested a possibly difficulty understanding documents and the latter was indicative of picking up a task. We find that a user's understanding of a URL may evolve over time, and that the query used when finding this URL becomes "better", in that it is shorter, more commonly used, and ranks the revisited URL higher.

From these findings about query reissuing and page revisitation we propose a relationship based framework called ARTEMIS. ARTEMIS stands for Assisted Relationship Tracing for Exploratory Multi-Session Informational Search. ARTEMIS is designed to help users better understand the search space by promoting connections between new URLs that the user may be unfamiliar with to URLs the user has previously visited and shown a strong interest in. By showing how a new document may relate to prior ones the user gains a better understanding of both the content of the new document, and the potential relevance of the document to the user's underlying task.

Finally, we demonstrate the potential of ARTEMIS with a proof of concept implementation accompanied by a task oriented user study. The study is conducted using lay people preforming a skilled task, which is an important and challenging subtask of the general search behaviors explored in the previous sections. The results show novice searches have an increased ability to find documents, and to understand them while using ARTEMIS.

To my husband,

Domingo Colon.

Acknowledgements

I would first like to express my sincerest gratitude to my adviser, Yi Zhang, who guided me on my journey from graduate student to researcher. I would like to thank her for her encouragement and support throughout the past six years, both professionally and personally and for her continued friendship. Without her guidance as well as friendship, this dissertation would not have been possible.

I would also deeply grateful my committee members, Jamie Callan, David Helmbold and Marilyn Walker. I would especially like to thank Jamie Callan who first introduced me to information retrieval in his Introduction to Language Technologies class days at Carnegie Mellon which sparked my interest in this field, for his career advice following undergraduate school and encouraged me down the path of graduate school when I was ready to make the plunge.

I am deeply indebted to the industry internship mentors, but specifically to Jaime Teevan, Susan Dumais, Sebastian De La Chica and Peter Bailey who gave me the opportunity to work with real world, large scale data which helped shaped my dissertation and the way in which I tackle problems. I would also like to thank Brian Carver for his support in running our experiment at the Berkeley Law School, and his advice regarding our legal study proof of concept implementation and study design.

I am also eternally grateful to my countless friends and family members who not only helped me beta test my proof of concept implementation, but offered design feedback to improve the interface. I'm especially thankful to Andy Herman and Brian Railing who uncovered numerous bugs, and for beta testing my multiple pilot studies.

I wouldn't be finishing my PhD without the friendship of my IRKM labmates at the University of California Santa Cruz, Jian Wang, Aaron Michelony, Lanbo Zhang, Qi Zhao, Yize Li, Shawn Wolfe, Jonathan Koren, and Yunfei Chen. Your companionship helped me through

the many trials and tribulations of graduate school. I have greatly enjoyed our conversations in the lab, both research and non-research alike.

I am also grateful for the NSF Graduate Student fellowship, the Cota Robles fellowship, as well as the CITRIS Grant, which funded my research.

Finally, I owe a great deal of my success of my Ph.D. study to the encouragement and support from my family. I am deeply indebted to my wonderful husband Domingo for the tremendous amount of love, support and encouragement he showed me, my parents, and my daughter, Nicole. I dedicate my dissertation to Domingo, for continuing to believe in me even when I doubt myself and to my daughter, Nicole, who reminds me every day that there is life beyond research.

Chapter 1

Introduction

In the early days of information retrieval, queries were often considered out of context. A query issued by the user to a search engine was considered the sole representation of that user's single information need. The results either satisfied or failed to satisfy said need. This approach to information retrieval is overly simplistic.

In some instances the information need may be persistent and may span multiple queries over different sessions, different days and even longer time intervals. A cancer patient likely cannot become sufficiently versed with possible treatment options when planning a course of action in a single query, session or day. Each new query and set of search results can improve his understanding, but may not satisfy his need in isolation. A politician concerned with her image, as another example, may be curious as to how she is being perceived by the community, and may wish to see what new information about her is published online throughout the election cycle. Any new potential scandal or opinion piece is relevant to her, and she may search regularly, clicking any search result whether she's previously visited it or not. While the search results may satisfy the politician's query on any given day, the politician may have the same need and

require a different set of results to satisfy the need on another day. A successful search during one session does not negate the need to search again in the future.

As the user's underlying understanding or interest in a topic changes, his or her specific information need may gradually evolve over time and lead to new topics. Our cancer patient may seek new information to augment what he has previously learned. As he navigates through the search space, he may return to past sites and view information he has previously encountered as his understanding grows, or visit new sites to gain additional information. New search results previously viewed as irrelevant may now be deemed relevant and vise versa. In this way information found on one topic can lead to the discovery of new topics and thus new information needs. This act of piecing together information from various sources is referred to as the *information journey* in the field of digital libraries and *exploratory search* or *information gathering* in the field of Information Retrieval.

As the information need evolves, it may also become more specific. Over time the cancer patient may settle on one particular family of treatments. If he has already decided against a method of treatment, he may be unwilling to consider it further and ignore search results on that topic. Only a subset of search results containing information related to the query may be of interest to the user. While the politician may benefit from a broad diversification of search results, the cancer patient's search results would benefit from a more targeted approach.

In some cases a lack of understanding of how information relates can become an impediment to a successful search. We consider a third example of a novice researcher performing a literature review. Often a paper may be relevant to a literature review in non-obvious ways. A researcher's search results may include papers from different domains, published in unfamiliar venues that don't appear relevant from the paper's title, abstract or snippet alone. These foreign papers may use different jargon and terminology than she is familiar with. Since she may be

unfamiliar with these new domains, she may not have the technical background nor familiarity with the terminology to see how this new paper relates to her literature review. She may discount a paper from a venue she's not familiar with, or from a lesser known author. Academic papers can be dense and time consuming to read. She may be unwilling to read a paper whose relevance is not clear or appears non-existent. Thus she may not realize her mistake. This is sometimes called *biased search for information* in psychology and is a form of *confirmation bias* [117, 106].

All three of our fictional users are engaged in an ongoing information journey that cannot be satisfied in a single session. While they are also seeking new information, the cancer patient and scholar may need to return back to previously visited sources, in order to better understand that new information.

1.1 Limitations of Traditional Search Engines

Prior work has treated the dual need to return to past valuable content, and explore the search space as two separate research areas.

Exploratory Search, *Information Foraging*, *Berry Picking* and *Information Journey* are all terms used by researchers to describe the process of seeking information that generally imply the on-going iterative process. They combine the notion of directed search and exploratory browsing with a goal of achieving *Intelligence Amplification* [167]. Intelligence amplification is the process of using information systems to increase human intelligence. Yet the focus is still on helping users find the right set of documents, (or the "facilitation" stage of the information journey). Little support exists to help in the "interpretation" stage, where users must piece together what they've found [3].

Two approaches designed to help the user browse the document space are clustering, and facets. The cluster hypothesis states "closely associated documents tend to be relevant to the same requests" [153]. The idea behind clustering search results is to group similar results, thus allowing the user to understand the result set space and navigate through subgroups of documents. Faceted search operates under a similar principle, grouping items according to similarity, typically via metadata about the document. Faceted search generally has pre-defined facets, and may allow the searcher to search using these facets whereas clusters are usually not pre-defined. Both approaches offer some support for the interpretation stage, as users can assume grouped documents are somehow related. However, this similarity is usually specific to the current set of search results, not past history or user background.

Another potential issue with clustering and faceted approaches is that when search results are grouped, they can "hide" or bury repeat URLs within clusters. Studies have shown, even just re-ranking can adversely affect the user's ability to re-find a document when explicitly targeting it [138].

Researchers have also explored the need to return to past information in the context of task resumption. The goal of task resumption is to aid the user when they return to a task they previously abandoned. This is usually accomplished through some sort of persistent store, either past history such, queries issued or past visited results or through an annotation capability. This approach places the onus on the user to understand how results may relate to his or her task, and only assists simple cross session re-searches where the information from multiple sources does not need to be aggregated and related in order to be understood fully.

1.2 The Goal of Studying Multi-Session Task Oriented Search

To address these challenges we propose a two step processing. In the first part we explore how users engage in multi-session task oriented search. We explore evidence of repetition seeking and diversification seeking in the user. We use these findings to build a framework (dubbed ARTEMIS) for aiding the user in these searches.

1.2.1 To Goal of Exploring Repetition and Diversification Seeking Behaviors

Up to this point repetition and diversification seeking have been seen as two separate concepts. In terms of repetitive behavior, prior work has explored the behavior of returning to past content via a search engine (re-finding) and re-issuing the same query (re-searching). On the diversification side, researchers have explored novelty detection, and diversification of search results.

1.2.2 To Goal of ARTEMIS

We propose a new framework dubbed ARTEMIS which stands for Assisted Relationship Tracing for Exploratory Multi-Session Informational Search. ARTEMIS is designed to help in on-going search where the user might be exploring a topic. As we will see, it has implications for other types of queries, such as transactional and shopping queries. The goal for ARTEMIS is two fold (1) account for persistent visitation patterns and ephemeral interests simultaneously and (2) provide support to help the user understand how new or possibly overlooked results may relate to the those previously explored.

ARTEMIS identifies highly important URLs based on past history, those that are frequently visited and that the user may be more familiar with, which are referred to in the framework as "starred" results. This is a departure from past systems which seek to identify results which are likely to be revisited or likely to be most relevant. While there is some overlap between the starred results and those more likely to be revisited, and between the starred results and those deemed most relevant, these two sets are not the same.

ARTEMIS then compares starred results to those the user has explored less, and which the user may be unfamiliar with to uncover possible relationships. By showing how highly valued information from a source (such as content on a frequently visited URL) relates to information from another source, the user may be more efficient and gain a better understanding of a topic area. Returning to our literature review example, if the user sees that a new paper shares many of the same citations as a paper she has previously cited, or that if the new paper is often co-cited with the paper she has previously cited, she may be more willing to consider the paper.

In our relationship based model, the rank shows the relevance of the document to the query and the relationships and stars show how a document might relate to the underlying task. This effectively decouples recommendation (relevance of the search result to the query) from justification (relation of the search result to past information) as proposed by Vig et all [155].

1.3 Key Contributions of this Dissertation

In this dissertation we explore two common forms of on-going repetitive information needs, query re-issuing (re-searching) and web page revisitation (re-finding) in depth. We then use these findings to design the ARTEMIS framework. Finally, we implement a proof of concept system to show the validity of this approach.

Key contributions are:

- An in-depth exploration of repetitive queries called re-searches from an on-going information seeking perspective. In this exploration we discovered a dual need to return to valuable content and discover new information when issuing the same query, even for queries typically thought of as navigational. We find evidence that in some cases re-finding may be the first step towards non re-findings, or new findings. We find that users may seek diversification of new-finding along personal interests. They are often discerning in the new-findings, and that such clicks do not happen at random. We also find in some cases users frequently click on more than one URL, indicating a possible connection between the two URLs.

- An in-depth exploration of repeat findings, or re-findings and the underlying user intentions. We find that users often settle on a query when re-finding and that this query tends to be 'better' in terms of being shorter, and the rank of the re-finding result being higher. Thus re-finding and re-searching are tightly coupled behaviors. We also found a distinction between intra and cross session re-findings, where the former suggested a possible difficulty understanding them and the latter was indicative of picking up a task as evidenced by the click trails. Further we find that users are more likely to re-find at the beginning and end of a session.

- We develop the ARTEMIS framework to help users both return to valuable content (re-find) while simultaneously aiding in the discovery of new information. ARTEMIS works by identifying repetitive interest URLs, which are typically the most relevant to the underlying task. URLs that are not identified as repetitive interest URLs are compared against the repetitive interest URLs to uncover possible relationships. These relationships can help a user understand why a document that does not appear relevant may actually be so.

- We evaluated a proof of concept implementation with a task oriented user study using unskilled searchers. We implement a proof of concept system tied to a legal search engine and conducted a user study based with lay people. We found that even naive system star strategies were effective in aiding users to find relevant documents, and to better able to understand the documents once found.

1.4 Dissertation Outline

The rest of this dissertation is as follows. Chapter 2 discusses related work in the fields of information retrieval, digital libraries, and human computer interaction. Chapter 3 defines definitions and describes the datasets we use to accomplish this work. In Part II we conduct log analysis and user intent surveys to explore what repetitive and diversification behaviours looks like. Chapter 4 explores how users behave when re-issuing the same query over long periods of time, which are the simplest form of on-going search. We next explicitly look at re-finding in Chapter 5, what it looks like and how we might better support it. In Part III we use our findings from Part II to design a framework to account for these behaviors. Taking what we learned in Chapters 4 and 5 we describe our ARTEMIS framework in Chapter 6. Next in Chapter 7 we present a proof of concept implementation, and present our findings with a user study in the legal domain. Finally, in Chapter 8 we summarize our contributions, and gives some possible directions for the continuation of this work.

Parts of the dissertation include work that has been previously published in past conferences. Chapter 4 is based on a paper published in the Conference of Information Retrieval and Knowledge Management (CIKM) in 2012 [150] Chapter 5 contains details from papers published in both the Web Search and Data Mining Conference (WSDM) in 2010 [148] and again from CIKM in 2010 [149].

Part I

Introduction of the Problem Space

Chapter 2

Background

In this chapter we review the literature related to this dissertation. We begin with a discussion on information seeking behaviors within search, and the challenges that have been identified. Next we discuss a very simple case of ongoing search, namely that of query reissuing and page revisitation. These actions have been studied in the context of task resumption, but not from an on-going information seeking perspective. It is by studying these patterns in Chapters 4 and 5 that new patterns of behavior emerge. Finally, we address how researchers have addressed similar challenges of navigating search results and understanding documents in the past, namely clustering and faceted groupings of search results, better text summaries and document snippets in the context of search and justifications in context of recommender systems.

2.1 Information Seeking & the Search Process

Searchers who gather information related to a given topic are said to be *Information Seeking*. Information seeking behavior is often divided into two categories, goal drive (extrinsic

search [67], diverse curiosity [111]) and undirected or without a goal (intrinsic exploration [67], diverse curiosity [111]). Generally speaking, *Exploratory Search* focuses on the users behavior and interactions with the search engine, where *Information Journey* describes the process of acquiring knowledge.

2.1.1 Exploratory Search and Related Concepts

Exploratory search[97, 166] is on-going, and can span multiple queries, sessions or days with an open-ended, persistent and multi-faceted. The goal is usually to learn or increase understanding. White and Roth define exploratory search as a combination search and browsing behavior that can help lead to deeper understanding [167]. They further characterize users engaging in exploratory search as often (1) unsure of the domain, (2) unsure of their goal, or (3) unsure of how to achieve their goal. Marchionini divided search activities into three parts, lookup, learn and investigate, and defined exploratory search as a combination of the learn and investigate components of search [97]. The lookup phase consists of discrete, analytical type searches, and are currently the best supported. Learning and Investigating are similar in that they are iterative stage that requires cognitive effort on the searcher for interpreting the findings, however Marchionini relates learning more to precision, and Investigating more to recall.

Researchers have also used the terms "berrypicking" [14], "information foraging" [115], and "information scent" [114, 35] to describe a ongoing, roaming, and possibly undirected search, where users gather pieces of related information from multiple sources.

Information Foraging relates seeking of information to foraging for food [115]. This analogy allows researchers to apply optimal foraging theory [134] to search behavior. There are two models in optimal foraging theory: the patch model and the diet model. The patch model describes an environment where food is distributed in patches. The forager may continue to

consume food in one patch or move to another once the food in the current patch is consumed. In the diet model the forager seeks to maximize the calories gained from food consumed relative to the handling or processing time (capturing, consuming, digesting of prey) and search time (finding the prey). The information forager, or *informavores*, makes similar choices whether to continue to consume a source, or move on to new sources [115, 116]. In the Information foraging analogy users follow the trail of information through snippets and thumbnails, referred to as the *information scent* [114, 35].

The notion of berry picking is similar to information foraging in that the searcher is navigating through the information space looking for information to consume. In the berry picking analogy information "berries" are scattered rather than grouped in patches. The emphasis of the berry picking analogy is on the sequence of behaviour, and the evolving nature of search [14]. Berrypicking allows for the flexibility of a changing information need, where information scent (and information foraging) tend to refer to targeting in on increasingly interesting pieces of information with a definitive goal [48].

Many strategies have been suggested to help the user who is engaged in exploratory search. Marchiolonini argued for more interactive systems that allow the user to have more control [97]. Hearst suggested clustering and faceted views to help the user make sense of the search space [62], discussed in detail in Section 2.3.1. Other approaches include keeping track of histories and providing annotation scratch space [102, 158, 96].

2.1.2 The Information Journey

In the field of digital libraries, researchers have studied the way in which users engage in information searches, including as a series of searches related to the same task or need

[83, 121, 29]. Like exploratory search, there is a notion of gathering information for a deeper understanding, however there is usually a defined end goal.

Adams and Blandford [3] studied the *information journey* of users, finding the journey generally included three stages: information initiation, facilitation and interpretation. While most modern day search algorithms focus on facilitation (or the gathering of information stage), Adams and Blandford found support for initiation and interpretation were lacking. One difficulty found by Adams and Blandford was helping the user to interpret the information that was presented to them.

In many cases, users show improvement towards information seeking over time. Each subtask or session that the user is engaged in can also be thought of as a different stage. Liu and Belkin [93] studied multi-session information seeking by creating an information gathering task for users. The task was subdivided into three similar tasks. They found that each task, or "stage", was strongly correlated to the usefulness of documents. However, while user knowledge generally increases over multiple stages, a "ceiling effect" was also encountered [94]. Similar observations have been made by Vakkari [152], who found users adapt and improve over multiple stages. However, in Vakkari et al. study subjects were experienced searchers with domain level expertise. On the other hand, Warwick et al. [161] studied how expertise in information seeking develops and found that users with less prior knowledge were reluctant to change their established search strategies, relying on keyword searches even when such strategies were proving ineffective. Additionally a study by Aula et al [9] on difficult information searches suggested that when expert users begin to have difficulty in a search task, their search behaviour more closely mimics novices. These two studies suggest that difficult subject matter may hinder or even reduce an expert searcher's ability to preform information searches.

Lin and Xie focused on transmuting searches, those where the information need gradually changes over time [92]. In their user study He et al found users were more likely to make direct query modifications in the initial stage, when exploring diverse topics [59].

Dörk et al. [49] proposed a model for information seeking based on urban flaneur, or wander, designed to aid in the interpretation of gathered information. Recently the need to support multi-session search tasks have been of interest to researchers. Bron et al explored multi-session search tasks spread over aggregated search engines [25], finding that users tended to prefer the tabbed displays, where results are separated by source, to blended ones where results are intermingled when the information need remains consistent.

2.2 Repetition in Search History

A user's history is a powerful tool for predicting the user's behavior. According to the polyrepresentation principle, users needs stem from more than just the current need, but the underlying problem space, task or interest. By incorporating the context, the search engine can reduce uncertainty and improve retrieval performance [69, 70]. The polyrepresentation principle also implies a user's need may not be satisfied with a single query. They may continue to search relative to the past information found.

2.2.1 Understanding Repetition in Search

Past work has explored repeated actions by individuals when interacting with a search engine in terms of re-issuing the same query [8, 125, 139], and in the context of clicking on search results they have visited before [108, 137]. Sanderson and Dumais [125] focused on the temporal nature of query re-issuing, finding daily patterns, while Teevan et al. [139] explored repetitive queries in the context of re-finding. Both studies found query re-issuing to be a

common behavior. Repetitive queries have been used to aid in collaborative search [132, 11] as well as clustering [162] for query recommendation [10, 180]. Past research has also shown that half of all webpages a person visits are pages the person has seen before [108, 137, 38].

The user's ability to return to past content can be hindered by many things. As the number of mobile devices grows, the ease at which a user can return to previous content decreases [30]. Search engine algorithms change over time, as does the cyber landscape. In some cases as much as 49% of the top ten search results when issuing the same queries over again in a given month [127]. These changes can also hamper a user's ability to return to previously seen pages [138].

Reasearchers have studied how people keep information encountered on the Web, and found people store Web-based information for future use in many ways, including by doing nothing and relying instead on tools to help them return [26, 74]. In particular, [74] noted strategies such as emailing a web URL to one self. Such methods become cumbersome as the number of pages a user wishes to revisit increase. Browser tools, such as the Web browser back button [125], bookmarks [2], and browser histories [78] have been used to study re-visitation. These tools have been built into the browser and do not attempt to model re-visitation behaviour in the search algorithms. As a result, the system relies more on the user's ability to actively attempt to re-finding. If the user does not recall the circumstance in which they found the original page, this task of re-finding can be quite difficult.

2.2.2 Helping the User Return and Resume

Tools to aid in Re-Finding have been divided into two camps: tools to help remember the URL, and tools to help remember the query. Under the assumption that good queries are hard to formulate. Researchers have explored storing complex queries for future use [118] as well

as using the saved search results list to improve re-finding [140]. The Re:Search Engine [140] and the SearchBar [102], are both designed to primarily support re-visitation. Re:Search uses the past result lists for a given query and merges it with the new list in order to provide stable results for searchers. SearchBar maintains the query histories and browsing histories of the user, displaying the user's common queries to aid in re-finding by grouping queries into topic-centric tasks as well as allowing the user to enter notes about the pages [102]. Morris and Horvitz proposed a persistent store, S^3, to help alleviate the need for users to re-issue queries and to share results in collaborative search settings [103]. Wang et al and MacKay and Walker proposed search bars specifically aimed towards task resumption [158, 96], which allow the user to specify when a page is related to a new task, and thus allow them to re-find it. These strategies usually involve storing either the queries or the search results, and do not support changing needs or a desire for additional search results that an individual might have with the same query.

More recently, researchers have explored re-ranking repeat URLs within the search results based on past clicks, skips (search results not clicked, but higher than a clicked results) and misses (search results deeper than the deepest clicked result)[128] using lambda rank. Shokouhi et al found that all three types as well as features like query similarity were useful in predicting future clicks.

2.2.3 Query Intention Drift

Researchers have also noted that the intention behind queries tend to change over time. Query dynamics, and the temporal nature of queries have been studied across users [156, 15, 44]. The types of temporal patterns associated with various queries are linked to retrieval performance. There is a transient nature to some information needs. For example, a query for "world cup soccer" may be in reference to different events in different years, and may

mean different things during different times of the same year. During the event, a user may wish to see the current standings, or an update on a score, whereas in the off season the user may be seeking more general information about the world cup. With this in mind, temporal aspects in web page content have been used to re-rank search results [179, 141]. In our study of query re-search, we notice intention drifts by the same user.

2.3 Aiding in Organization and Understanding of Results

One way to help users make sense of the result set space is to organize results in some meaningful manner. The organizational view then conveys information to the user. The two common approaches are clustering search results and faceted search.

Another approach to help users understand documents is to summarize them in the form of search result snippets. Such summarization is referred to as query-based summarization when the summary depends, at least in part, on the query [145].

Researchers in recommender systems have also addressed the understanding problem. Approaches include making the recommendation more transparent to the users, or providing additional justifications for the recommendation not based on the underlying recommendation algorithm.

2.3.1 Organizing Search Results

The two common approaches of helping the user navigate the result set are to cluster the results, or provide a faceted view of the results. These two approaches group results according to similarities. In the clustering approach, the relationship between documents is usually drawn from the content of the documents. As typical of clustering algorithms, the relationships may not be defined or known ahead of the query. In the faceted view approach, documents are typically

grouped according to their meta data, but can also include contextual or semantic similarities. Faceted search often allows the user to search by created fielded queries that specify values for each facet. In this approach the labels tend to be known in advance.

Clustering Search Results

Not all search results are equally interesting to each user. Some queries are inherently ambiguous. The canonical example is the query *jaguar*. Searchers interested in high end cars are likely not interested in large cats and vice versa. By grouping (clustering) results by similar content, users can drill down into the group that interests them, and ignore the search results that do not.

According to Carpineto et al [32] the three areas that most benefit from search result clustering are (1) topic exploration, (2) subtopic retrieval and (3) alleviating information overlook. By alleviating information overlook, Carpineto et al is referring to the phenomenon of users to only skim the top results for a search engine and thereby might miss relevant documents. By clustering documents, many documents are summarized into a single view.

One of the earliest document clusters to be built on top of an information retrieval (IR) system was Scatter/Gather. Scatter/Gather was first design to accommodate the browsing behavior [41], creating fast partitions by utilizing randomization, and clustered based on cosign similarity of terms. When the user selected one or more clusters, the documents in those clusters would be reclustered, allowing the user to effectively wonder the corpora. Hearst and Pedersen extended the Scatter/Gather idea to search results [63], clustering results after the initial ranking. They found different clusters may be of use to different users.

These days clustering methods for search results tend to be categorized among two dimensions: single words or phrases [54, 75, 173, 174, 177] vs summaries [110, 63, 89, 88], and

flat [63, 54, 75, 110, 176] vs hierarchical [52, 33, 177, 86, 170, 50]. Flat clusters are discrete, where hierarchical clusters are clusters where one cluster may be subsumed by another in whole or in part. A related concept to hierarchical clustering is soft clustering. A clustering method is said to be hard when each document is assigned membership to one cluster. Membership is boolean; the document belongs in the cluster or not. A soft clustering method, by comparison, may assign degrees of membership for a document in multiple clusters. Approaches for flat clustering includes transactional k-means [54], relational fuzzy clustering [75], singular value decomposition [177, 110], regression [176] and cosine similarity [41] and term frequency, inverse document frequency (tf-idf) [89]. While Scatter/Gather is technically a flat cluster at each iteration, it was specifically designed to allow users to browse the document space. As such, it effectively creates subclusters as the user selects a given cluster. Thus some have considered it a hierarchical approach.

Traditionally clusters have been constructed based on the content of either the full documents or search result snippet, however researchers have proposed other methods, including link structure [159] and named entities present in the document [144] and time [6].

Clusters built on single words attempt to find a single keyword of phrase to describe the cluster where as clusters with summaries are meant to help the user understand the documents in the cluster. One drawback to summaries is that they are often not readable. Therefore Zamir and Etzioni proposed suffix tree clustering to identify common phrases between documents [173, 174].

Faceted Search

In faceted search specifics about the documents, (properties, characteristics or meta data) referred to as facets are created for each document. These facets allow documents to be

grouped into multiple overlapping taxonomies in a meaningful manner. A user can then drill down into the search results via the facets. For example, in the academic papers domain facets could include author, conference, and keyword. A searcher may issue a query for a general topic. Upon seeing the search results and facets, she may select a "conference:sigir" facet to limit her search results to those published in that venue. Faceted search provides a means for users to effectively build complex queries through use of an interface. Amazon[1], Ebay[2] and Google Shopping[3] are all examples of search engines with a faceted component.

A key advantage to faceted search is that they are often intuitive and easy for users to use [61, 171, 131] and explore results more deeply [84]. Additionally, searchers often feel more organized and in control during their search session [84]. Other studies have found, however, too many facets can overwhelm the user and degrade preformance [131] which has lead towards to research into which and how many facets to display [80]. Conventional wisdom of facet selection includes result set coverage, those that have have high entropy (and thus partition the result set roughly evenly) and those whose partitions have low overlap with other facets chosen (i.e. are not superseded by another facet) [79, 151, 147].

The most wildly cited academic faceted search engines include the Flamenco (FLexible information Access using MEtadata in Novel COmbinations) Project, Relation Browser, and mSpace. Both Relation Browser and mSpace provided support for discovery of facet based information. Relation Browser [31] was designed to allow users to gain insight into the search result space through it's preview-oriented interface. The interface included bars that represented counts for each facet values. Thus after entering a query a user could get a visual snapshot of what the document space entailed. mSpace [126] utilized a technique known as Backward Highlighting [169]. When an item was selected, the facets for which that item appeared where

[1]http://amazon.com
[2]http://ebay.com
[3]http://google.com/shopping

highlighted. This allows the user to get a visual representation of the facet values for that item. Flamanco was used to study interface design [60] and automated metadata creation [135, 136] and usability [62] in faceted search.

While we focus on search in this dissertation, it is worth mentioning that facets have also been used for *Faceted Navigation*, which is a form of intrinsic navigation, a special case of exploratory search.

2.3.2 Query Based Summerization

Perhaps the most common approach to document summerization is sentence ranking [175, 145, 163, 168]. While most approaches use a form of extraction, or selecting content from the text to serve as the summary, OCELOT and Varadarajan and Hristidis attempted to synthesize a summary [16]. OCELOT used statistical modeling to identify which terms should be in the summary, and would iteratively pick and replace similar words from outside sources in an attempt to improve readability. Varadarajan and Hristidis identified query related fragments and combined them by converting the document into a graph and using proximity search and minimum spanning trees over the fragments to construct a summary [154]. One issue with synthesized summaries is readability. Kanungo and Orr explored predicting the readability of search result snippets [77].

Researchers have also explored the role of entities in query based document summaries. One potential goal of document summerization is to improve novelty. Li and Croft used named entity recognition to find entity patterns within sentences. Sentences were then re-ranked [91]. *Entity Ranking* is the task of finding and ranking entities with respect to a query. It is an area of research that has it's own track at TREC [12, 13] and INEX [45, 47, 46]. Blanco and Zaragoza suggested ranking of entities was not enough and explored finding support sentences for entity

ranking to explain the relationship between the entity and the query [22], akin to selecting snippet selection in search. In their analysis they found context, especially the sentence before and after, helped improve coverage over bag-of-words approaches.

2.3.3 Explanations in Recommender Systems

The approach we will employee to aid in the interpretation stage is to provide explanations to the user. This is a popular technique in recommendation systems that has also been applied to the news domain by Blanco et all [21]. It has been shown that users tend to trust recommendations more with explanations [130] and are more likely to accept recommendations [64] that are transparent.

Recommendations can be complex and difficult for users to understand when based on how the item was recommended, leading some researchers to differentiate between (1) why the item was recommended (how the underlying algorithm choose the item) and (2) justification for the recommendation (based on something the user might understand) [155]. Tintarev and Masthoff proposed a taxtonomy of the seven beneficial aims of explanations in recommender systems [143]: transparency [27, 101, 130], scrutability [20, 43], trust [7, 64, 130, 160], effectiveness [7, 19, 20, 27, 64, 99, 142], persuasiveness [39, 64, 160], efficiency [39, 43, 64, 99, 100, 101, 142], and satisfaction [64]. Increasing transparency has been shown to increase system usability through the principle of User Control [107]. Although Tintarev and Masthoff [143] describe scrutiny as the acceptance of explicit feedback from the system to the user, Herlocker et al [64] have also pointed out when a system makes an erroneous recommendation, explanations can help a user discount the bad recommendation, without the faulty recommendation effecting their belief or trust in the other recommendations.

Identifying good recommendations is easier in a closed domain where something is known about the items being compared (e.g. movies, books, etc). Additionally, as the user's search history grows, the number of items pairs for possible explanations grows exponentially.

Chapter 3

Definitions and Data Sets

In this Chapter we review definitions used throughout this dissertation, and describe the datasets we use in our analysis.

3.1 Basic Definitions

In this section we give basic definitions used both in this dissertation and commonly found in the literature. Readers familiar with this domain will likely already be familiar with these terms.

3.1.1 Query instance

A *Query Instance* is a specific query issued at a specific time by a specific user, noted as q_*. A *Query String* is the string supplied to the search engine by the user. The same query string can be issued at two different times creating two different query instances.

3.1.2 Finding

In the literature a *Finding* occurs when the user clicks on a search result after issuing a query. The term is used to indicate that the user found the URL. In this dissertation we primarily use the term *search result click*, however we provide this definition as it may be beneficial in understanding future terminology.

3.1.3 Session Boundary

Sessions are typically thought of as groups of related activities in the search engine. A *Session Boundary* marks the end of one session and the beginning of another. In this dissertation, we define a session boundary as a time interval of inactivity of more than 30 minutes, as motivated by prior work [34]. While different researchers may use different time boundaries, we note one key finding from Jones and Klinker was that regardless of the time interval, there are often some tasks that will span time-based session boundaries [72]. We find this definition of session boundary acceptable for our purposes, as our intent in exploring multi-session on-going search is to explore the notion of evolving, incomplete or ongoing tasks, such as those often used to acquire knowledge. We wish to exclude from our analysis short tasks where the query is re-issued, or the task is completed, during only a brief period of time. Thus we expect this definition of session is adequate for our study.

3.1.4 Task

We use *tasks* in this dissertation to indicate a collection of on-going related searches. This is a departure from the literature where tasks typically have a clear achievable end goal or state and are often completed in one session (e.g. purchase movie tickets.) Our tasks, in

contrast, cannot be completed within a single session, are broader and may not have a clear end (e.g. research cancer treatments.)

3.2 Key Terminology

In this section we describe the key terminology used frequently throughout this dissertation. To aid in understanding, Table 3.1 gives a brief overview of key terms, and can be used as a quick reference.

Re-Search (§3.2.1) A re-search occurs when the user issues a query that is similar to previously one.

Same Query Re-Search (§3.2.1) A same query re-search occurs when a user issues a query that is identical to a query previously issued.

Minimal-Change Re-Search (§3.2.1) A minimal change re-search occurs when a user issues a query that contains only superficial changes to a query previously issued.

Term Overlap Re-Search (§3.2.1) A term overlap query re-searching occurs when the user issues a query that contains some minimal number of terms in common with a query previously issued.

Substantial Change vs. **Non-Substantial Change** differences in queries (§3.2.1) A query represents a substantial change from another if it differs by at least one non-stop word term. It is a non-substantial change otherwise.

Continued on next page

Table 3.1 – continued from previous page

Previous Query When discussing two query instances, the previous query is the one that was issued first.
(a) Query Based Termonology
Re-Finding (§3.2.2) A re-finding occurs when the user clicks on a search result that is the same URL as another previously clicked search result for some subset of user history. **New-Finding** (§3.2.3) If a search result click is not a re-finding, it is considered a new-finding
(b) Search Result Click Based Termonology
Trail A trail is a sequence of related, not necessarily consecutive, behaviour. **Query Trail** (§3.2.4) A query trail is a subset of query history issued by the user. A re-search query trail is a trail where every query is a re-searching of a previous query in the trail. **Click Trail** (§3.2.5)
(c) Termonology used when Aggregating Behaviours

Table 3.1: Summary of key terminology used in this dissertation.

3.2.1 Re-Search

Re-searching occurs when a query issued by the user is the same or very similar to a previously issued query. In the strictest sense, a query instance, q_j, is an example of *re-search*

if there exists another a query instance, q_i, with an identical query string previously issued by the same user. More broadly, q_j can also be considered to be an example of *re-search* if q_j's query string is similar to, but not necessary identical to, the query string of q_i. This definition of re-search does not include URL(s) of search results, thus abandoned queries can also be re-searchings and are included in our analysis.

The pair of queries is called a *re-search pair*. The second query issued is referred to as the *Re-search* query instance and the first query instance is referred to as the *previous* query instance. There may be many query re-searches for the same query string issued by a single user. In general we only consider one *re-searching pair* for each *re-search* query instance: the one where the previous query instance is the one that most immediately precedes the re-search query instance. We note these two query instances need not be consecutive. We use "previous query" as short hand notation to mean the previous query instance.

Types of Re-Search

Based on the literature, we focus on three types of re-search in this dissertation, *same query*, $\sim_=$, *minimal change query*, $\sim_M$, and *term overlap query*, $\sim_T$. Based on our definitions below, *same query* $\subseteq$ *minimal change query* $\subseteq$ *term overlap query*.

Same Query ($\sim_=$). Two queries are an example of *same query* re-search if the queries are identical. In Table 3.2, Q_5 is a same query re-search of Q_3, and Q_7 is a same query re-search of Q_6.

Minimal Change Query ($\sim_M$). In keeping with previous work [139, 148, 149], queries are considered to be similar according to *minimal change* if the types of differences between the query strings, in general, do not affect the meaning of the queries. Such differences can include a combination of stop words, white spaces, non alpha-numerics, URL identifiers

Day	Label		Query	Click
Mon.	Q_1		swine flu incidence	
		$C_{1,1}$		healthmap.org/swineflu
		$C_{1,2}$		www.swine-flu-map-animation.com
		$C_{1,3}$		www.cdc.gov/H1N1Flu
	Q_2		swine flu deaths	
	Q_3		h1n1	
		$C_{3,1}$		en.wikipedia.org/wiki/H1N1
		$C_{3,2}$		www.cdc.gov/H1N1Flu
Tues.	Q_4		swine flu	
		$C_{4,1}$		www.cdc.gov/H1N1Flu
		$C_{4,2}$		h1n1.nejm.org
Wed.	Q_5		h1n1	
	Q_6		cdc swine flu	
		$C_{6,1}$		www.cdc.gov/H1N1Flu
Sat.	Q_7		cdc swine flu	
		$C_{7,1}$		www.cdc.gov/H1N1Flu

Table 3.2: A hypothetical search history.

such as "www." and ".com" (called *domain variant* in the literature [139, 148]), casing, or term ordering. Thus the strings "Church of Latter Day Saints" and "Latter Day Saints church" are considered to be a minimal change of each other. Such changes may be unnoticeable to the user; the user may not even realize he or she is issuing a query different than one previously issued. While these types of differences typically do not alter meaning, in some specific examples they can. For example, "summit" may refer to any mountain peek where "the summit" is the title of a book about climbing Mount Everest, "White House" refers to a specific building, where "white house" does not. Similarly "Department chair couches offers" and "chair department offers couches" contain the same terms, though they have very different meanings.

Some prior work has also considered query strings that are spelling mistakes, acronyms, synonyms and merging of terms to be minimally different. For example "fb", "facebok" and "face book" would all be considered minimal change queries. For the majority of this dissertation we omit these from our definition of minimal change as a query term may be an acronym or a misspelling of two distinct phrases with different meanings. This omission also preserves the transitivity of the similarity relation. Thus groups of minimally changed query instances are all minimal changes of each other, which better fits the notion of a query being changed in undetectable ways to the user with little, if any, effect the meaning. Preserving transitivity also ensures minimal change re-searching is a subset of term overlap re-searching.

In Chapter 5 spelling mistakes and merging of terms is included in the definition of minimal change. We discuss this design choice more in Section 3.2.1.

Term Overlap Query ($\sim_T$). In some instances a user may wish to find additional search results to ones previously seen, and may alter the query by adding or removing terms. Such query modifications have been well studied [85, 73, 65] including in terms of retrieval of the same URL [65]. To study re-search with this type of query modification, two queries

are considered to be similar according to *term overlap* if at least half their terms are shared. That is, if the Jaccard similarity between the two query strings (equation 3.1) is greater than a threshold. In this dissertation we set the threshold to 0.5. This strictly greater than rule forces at least two terms in common between q_i and q_j when $q_i \neq q_j$.[1] As with minimal change query similarity, case, stop words, non-alphanumerics, and URL identifiers are ignored. Such terms, however, account for a small number of overall terms in the query and they often do not affect the similarity[2].

$$Jaccard(q_i, q_j) = \frac{|q_i \cap q_j|}{|q_i \cup q_j|} \tag{3.1}$$

As an example from Table 3.2, Q_2 is a term overlap re-searching of Q_1, and Q_4 is a term overlap re-searching of Q_2. Again, we don't consider Q_4 to be a term overlap re-search of Q_1, since Q_2 more immediately precedes Q_4.

The Jaccard similarity was chosen because it is intuitive in terms of what the similarity means, and captures the notion of one query being modified into another. Prior research also suggests that other commonly used methods of similarity, such as the vector space measures, are not well suited for finding similar queries [118].

We note the transitive property no longer holds for *term overlap* similarity. However, unlike the *minimal change* similarity metric, this similarity metric better captures the concept of an evolving query, fitting the notion of an evolving information need during the information journey.

[1] In the AOL search log, discussed in detail in Section 3.3.1, 39.6% of all queries consisted of a single term, 25.1% have two terms and 35.3% of queries had more than two terms.

[2] In the AOL dataset, 8.6% (approximately 123 thousand) of the non-same term overlap re-searchings would have had a $Jaccard \leq .5$ if non alphanumerics, stop words and URL identifiers were not removed. An additional 113 thousand query pairs would have $Jaccard > .5$ with the inclusion of such terms.

Substantial vs. Non-Substantial Query Changes

In some instances we may wish to consider a pairing of queries that do not have string similarities in common, for example if we wish to say that have semantic similarity in the query strings or based on click behaviour. In this case we differentiate between *Non-Substantial Change* and *Substantial Change* query. As we will see in Chapter 4 *Minimal Change* and *Same Query* Re-searchings have very similar use patterns. In the context of later chapters we group these two types of changes into *Non-Substantial Change*. In later chapters we also including spelling variants and word merge as examples of non-substantial change in keeping with prior work. In this dissertation we considered queries with a normalized edit distance of less than 0.05 or an absolute edit distance less than 2 as misspellings or spelling variants of each other. This choice was motivated by exploration of our logs.

Any query change that is not a non-substantial change is considered a *Substantial Change*. *Substantial Change* queries includes *Term Overlap*, as well as instances where the queries have no terms in common.

3.2.2 Re-Finding

Re-finding occurs when an individual clicks a URL from the search results of query instance q_i, and then later clicks on the same URL via the search engine results of another query instance, q_j. The two query instances make up the *re-finding pair*. The user may click more than one search result for any given query instance, and the re-found URL may not be the only clicked result for either query instance. In keeping with previous work [139], we consider the instance to be a re-finding if there is any click overlap.

Not that the query used to re-find a URL may differ substantially from the query used to previously find it. Thus a query instance could be a re-finding of one query instance,

and a re-searching of another. From Table 3.2, Q_4 is a same query re-finding of Q_3 and Q_6 is a substantial change re-finding of Q_4. Queries with substantial changes are interesting in the context of re-finding because they often reflect the fact that the searcher has developed a significantly different way of expressing their information target.

We refer to second query instance to as the *Re-finding* query instance and the first query instance is referred to as the *Previous* query instance. Thus "previous" query refers to the first query instance in either a re-finding or a re-searching pair. As before, we only consider the query instance that most immediately precedes the re-finding query to be the previous query for the re-finding pair.

For most of this dissertation we refer to re-findings in the context of the user's entire search history. In other words, if the URL was clicked in the result set of any previous query, it is a re-finding. In Chapter 4, however, we focus on re-search trails (defined below). Therefore, we consider re-findings relative to the same re-search trail.

3.2.3 New-Finding

A result click that is not a re-finding is referred to as a *new-finding*. Note that a new-finding may not be the first time a user has visited the URL. A web user may visit a URL through other means, such as clicking on a link propagated through social media, clicking on a link in an email, or typing the address in the address bar.

3.2.4 Query Trails

A *query trail* is a sequential subset of queries from a single user's search history, $\{q_0, q_1, ...q_n\}$. When a query is re-searched multiple times, the re-searchings along with the initial instance of the query (q_i) can be grouped into a *trail* of re-search. A query, q_j, is added

to a trail, T, if $\exists q_i \in T$ where $q_i \simeq q_j$. *Re-finding trails* are defined similarly. A *Re-search Trail* and *Re-finding Trail* is a trail where each additional query instance $\{q_1, q_2, ...q_n\}$ in the trail is a re-search or re-finding of some previous instance in the trail.

We consider a re-search to be similar according to the most specific similarity metric, and the full trail to be similar according to the most specific similarity of all re-searchings within the trail. For example, for a *minimal change re-search trail* T, $\forall q_i, q_j \in T, q_i \sim_M q_j$ and $\exists q_k, q_l \in T, q_k \not\sim_= q_l$. Throughout this dissertation we only refer to complete trails, trails where the addition of another query from the user's prior search history would alter the trail's status as a re-finding or re-search trial. Thus we consider $\{Q_3, Q_5\}$ and $\{Q_6, Q_7\}$ from Table 3.2 minimal change re-search trails, and $\{Q_1, Q_2, Q_4, Q_6, Q_7\}$ a term overlap re-search trail.

3.2.5 Click Trail, Hop

After an individual has clicked on a URL from a search result page, he or she may choose to follow the links on the page before moving on to his or her next action with the search engine. This series of clicks on links is called a click trail, and each link in the trail a hop. A trail starts at a search result click, and ends when the user does not click on a link for 30 minutes, or preforms an action that takes the browser away from the current page other than clicking a link such as using a bookmark, closing their browser, entering an address on the address bar, or entering a new query in the search engine [165]. Note that if a trail is longer than 30 minutes, subsequent queries will be considered part of a new session, even if very little time elapses between end of the trail and the query.

When a trail is followed from a URL found via a previous query, we call it the previous trail. When it is followed from a URL re-found via a re-finding query, or followed from a URL

found via a re-search query we call it a re-finding trail or re-search trail respectively. The re-finding trail and re-search trail may involve different hops than the previous trails.

3.3 Datasets

The analysis in this dissertations is comprised of user surveys, log analysis from both publicly available and proprietary sources, and user studies.

3.3.1 Search Logs

We use search logs from two search engines to gain a broader perspective of re-search across millions of users: the AOL search engine and the Sogou search engine.

AOL Search Logs

The AOL search log is a primarily English search log. The AOL set contains 27M distinct queries that were submitted by 650k users between March 2006 to May 2006. Twenty million of these queries have clicks.

One weakness of this dataset is that it only contains the domains of clicked URLs, instead of the full URLs. Nevertheless, analyzing the domain information is still useful for our purposes. Domains often group similar information, either by topic (i.e. the CDC website) or task (i.e. Wikipedia or a news site.) In both cases if the click on a repeated domain is intentional, we assume the user may be able to find the intended information by navigating the resulting website. Additionally, knowing which domain a user will click would still be useful for re-ranking to improve the click through rates. Finally, as we will also see in our analysis of the second search log, experiments with just the domains of the URL in the search results typically yield similar findings as those done with the full URL of the search results.

A user may click the same search result for the same query more than once, for example by opening a new tab, double click on a search result or use the back button to return to the same search instance. To canonicalize our data, we consider a query instance to be unique by query date, query string and user. (The date is provided up to the seconds in the log.) Multiple clicks on the same URL for the same query instance at a given rank are considered the same click.

Sogou Search Log

The Sougou search engine is a search engine based in China that primarily deals with Chinese language queries. The log contains 10.8 million queries with nearly 5M users from August 2006. Since most Chinese words are made up of one or two characters, and there is no spacing between words, it is not straightforward to use the concept of term overlap or minimal change queries on this data set. Therefore we consider only the same-query re-search in the Sogou log.

The log contains the full URLs of search results that were clicked on by the user. Therefore, we can conduct our analysis twice, once with only the domain and once with the full URL of a search result click. When using just domains the results are comparable to that of the AOL log. We also observe that the results are often similar when using the full URL to those from the domain only in the Sogou dataset. This supports the notion that using only the domain information, as with the AOL Log, may be sufficient for our purposes.

One limitation of the Sogou is that it is ordered by time, without time stamps. We cannot define a session based on the time gap. Instead, we approximately define cross-session re-searching by measuring the number of queries that occurred between a candidate query pair, which we call query delta ($\Delta_{queries}$). If the query delta is large, we are reasonably sure that

the previous session has ended and the user is expressing a renewed interest in the same/similar query. Although the median number of queries per session in the AOL dataset was 2, we choose $\Delta_{queries} = 5$ as our session boundary. The focus of our dissertation is ongoing search for tasks that cannot be completed in a single session where the information need may be evolving. A longer gap increases the likelihood of correctly labelling two queries cross session queries.

Unless otherwise noted, all search log findings are on the AOL dataset.

3.3.2 Proprietary Data

During the course of our study we had temporary access to some proprietary data sets for use in the re-finding analysis: 1) one which gives insight into the search engine-related behavior (search engine query logs), 2) another which gives insight into a searcher's behavior after leaving the search engine (Web browser logs), and 3) one which gives insight into the content of the found pages (a large-scale, daily Web crawl). All of these data sets were collected during the month of January 2009. This data is used in Chapter 5.

Search Engine Query Logs - Live Search Engine

To understand the search engines view of re-finding behavior, we studied the query logs from Live Search (now Bing), a major internet search engine. From the logs, we sampled information related to approximately 900 million search result clicks gathered from 106 million users. Similar to the example shown in Table 3.2, the sample included queries and clicked results, as well as time stamp information and the rank position of the clicked results. The sample was filtered to remove spam and processed so that pagination and back button clicks were treated as the same query.

Users were identified by an anonymous ID associated with a user account on a particular computer. As is the case with most log analyses, if a user has more than one computer, that user will have multiple IDs. Conversely, if more than one person uses the same account on a computer, they are amalgamated into a single user.

Web Browser Logs - Windows Live Toolbar

Information about the click trails people followed after running a search was collected via Web browser logs gathered from opt-in users of the Windows Live Toolbar. The toolbar provides augmented search features and reports anonymous Web usage behavior to a central server. Our analysis of the Web browser logs makes use of data from a sample of 4 million users and includes hundreds of millions of pages visits. In addition to containing other URLs, the browser logs contain query URLs associated with multiple search engines, including Live Search, Google, and Yahoo. We used these search engine URLs to identify search trails by extracting the queries from the URLs and analyzing where people went following a result click. We also used the toolbar data to confirm that our findings using the Live Search query logs were consistent across a variety of different Web search engines. However, the analysis reported in Chapter 5 uses the query logs instead of the toolbar logs whenever possible because that data is cleaner and more plentiful, applies to a broader number of users, and contains information about the results presented (order, etc.) in addition to just clicks.

Large Scale Web Crawl

To better understand the result pages people re-found, we also looked at the text content of the pages, captured via a large scale crawl of a sample of Web pages. To understand how the page content changed during the study periods, we crawled each page in our sample

daily. At the onset of the data collection period, we did not yet know which Web pages would be re-found. Instead, we crawled pages that were sampled based on three different visitation-based attributes: the number of unique visitors to the page, the median time between users visits, and the median number of visits per user. In total, 55,000 different pages were sampled. Additional information about the sampling process is described in earlier work by Adar et al [4].

Relating the Three Datasets

The three priority datasets were related in that they covered the same time and referred to, in many cases, the same queries and URLs. Two URLs were considered the same if they appeared likely to refer to the same page. For example, it is common practice (although not always the case) for a primary domain and the subdomain of www. to point to the same content, and thus the initial www. was ignored. Additionally, a trailing slash usually does not alter the page content, and was thus ignored. We did not remove URL parameters as they can often lead to different page content.

3.3.3 User Surveys and Studies

Although the log data described above give a realistic picture of real world behavior, they do not provide insight into what the individual's intention is when conducting a search. In order to explore intentions behind their behaviours, we conducted several Mechanical Turk surveys. One of primary benefits of Mechanical Turk is diversity [28]. Buhrmester et al found Mechanical Turk surveys were typically slightly more demographically diverse, and were of quality at least as good as typical internet surveys. We note that we found reported search behavior through our Mechanical Turk survey is similar to the observed search behavior in our logs, suggesting the validity of the survey findings.

On major limitation of survey taking, however, is satisficing: spending minimal cognitive effort that is sufficient to satisfy the task. In other words, participants may choose the first minimally acceptable answer, rather than the best or most correct answer [82], which can lead to inaccurate or noisy results. This is especially true in the Mechanical Turk setting where rewards are often small, and participants are trying to collect as many possible rewards in a short period of time. Thus we ordered our question responses where satisficing would lead towards a bias of underreporting the behaviors we're interested in.

An additional problem arises when users may be engaged in the survey, but may not understand the questions. To elevate this problem, our survey included several control questions, and questions that were worded slightly differently whose answers should have been related. Participants whose responses were incongruent, for example indicating two very different frequencies for the same behavior, were automatically discarded. Finally, it has been noted that reporting frequency of activity is typically less reliable than explanatory questions [17]. Thus, whenever possible we ask users to describe how and why they engage in certain behaviors, and not just whether they engage in them.

Intentions behind Re-Searching Survey

Our first survey was designed to illicit the frequency and reasons why users re-search. The full survey is listed in appendix A.1. The survey listed several *intentions* a user may have behind re-issuing a query to a search engine that had previously been issued, including

- (1) to revisit a website you have seen before
- (2) to find new websites you have not seen before
- (3) to find new information, regardless of whether it is on a website you have seen before

- (4) to see if or how the search results have changed

Examples were also included to help reduce confusion. Only one participant marked "other", stating that he also re-searched when seeking information for technical standards, specifically to determine when and if the standards had changed.

In our analysis, we refer to the first intention as "re-finding", the second as "new finding" the third as "new content seeking" and the fourth as "result list change."

The full survey is included in the appendix, Section A.1. The participants ranged in age from 12 to 67 years old, consisted of 23 males and 39 females, and were located across the United States.

Re-Finding User Study

In order to get a better picture of whether a re-finding query was actually intended to re-find a particular URL, we conducted a small-scale critical incident user study of 9 individuals (7 males, 2 females).

Participants in this study were volunteers responding to an email solicitation. They installed a Web browser plug-in on their primary work computer, and ran the plug-in for several weeks. The plug-in logged the subject's search engine queries and result clicks, and occasionally popped up a survey following a result click to ask whether the subject had intended to find that particular URL with the issued query. The survey appeared following all re-finding clicks, and following 12.5% of all new-finding clicks. In total, we collected 159 responses from the 9 participants.

Proof of Concept Study and Survey

To show the validity of our framework we built a proof of concept implementation and conducted a user study based on this implementation. The study was conducted over the course of several weeks requiring the users to return to the task after time away. The participants in this study ranged in age from 21 to 61 years old, with a roughly equal gender split with 12 males, and 16 females. The full details of how our study was conducted are described in Chapter 7.

After each session, participants were asked to complete a survey based on their interactions with the search engine. Only the relevant parts of the survey were shown to the user. For example, one section of the survey addresses the relationships presented to the user. Participants in the control group who did not see the explanations also did not see this part of the survey. The survey can be found in the appendix, Section A.2.

Part II

Understanding Human Behavior

Chapter 4

Observations of Re-Search

Many information needs cannot be satisfied in a single session. A politician concerned with her image may be curious as to how she is being perceived by the community, and may wish to see what new information about her is published online throughout the election cycle. Any new potential scandal or opinion piece is relevant to her, and she may search regularly, clicking any search result whether she's previously visited it or not. Alternatively, a cancer patient may want to investigate what kinds of new treatments are available. The patient may seek new information to augment what she has previously learned. As she navigates through the search space, she may return to past sites as her understanding grows, or visit new sites to gain additional information. In both cases, the interest in the given query is persistent and ongoing. A successful search during any given session does not negate the need to search again in the future. We begin our investigating by studying ongoing re-search.

In this chapter we take an information seeking perspective, examining on-going related searches as a series and their effects on the information need. To accomplish this, we use multiple data sources: a user survey (described in Section 3.3.3) to understand the user's stated

intentions, and the AOL and Sogou search logs (described in Section 3.3.1) to compare how common the observed actions are in practice. Prior work on query re-issuing has focused on each query individually [108, 137]. By studying the long term re-searchings, new patterns of behaviour emerge.

The key findings in this chapter include:

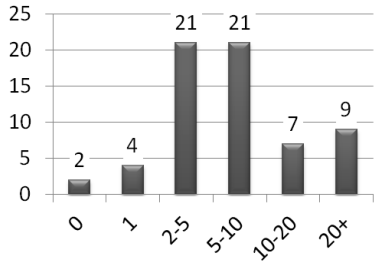
- (1) Long term ongoing re-search often shows signs of both diversification seeking and repetition seeking behaviors within the same trail. Half of all frequently re-searched queries lead to both new and repeat URL clicks. These new-findings can occur at any time. If we were to believe the user's sole intention was to re-find a particularly relevant URL, we would expect the user to "settle" on a URL, but this is rarely the case.
- (2) While seeking new content, users are selective in their new-findings, even when frequently re-finding. Often users do not appear interested in all types of subtopics equally. These new findings may reveal more about the user's underlying interest than the re-findings.
- (3) While repetition seeking is common, we find evidence that in some cases, repetition seeking may be the first step to novelty seeking. New-findings are more likely to happen at the end of the trail than the beginning.

4.1 Experimental Methodology

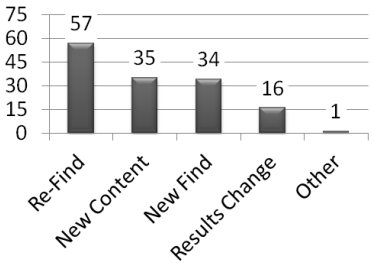
In this dissertation we seek to better understand ongoing search. As such, we only consider long multi-session re-search trails; trails that span at least two sessions. Unless otherwise noted, we only consider re-search trails with at least four query instances.

Since the focus in this chapter is on the re-search trail, we use the terms new-findings and re-findings to mean a finding of a new URL or repeat URL relative to the re-search trail, and not the user's full search history. Thus a re-finding in this context is when a user clicks a URL from a search result from query instance q_i, as well as query instance q_j in the re-search trail, q_j. A new-finding is the first time a URL is clicked in the re-search trail, and necessary the first time the URL has been found via a search. This definition differs slightly from the definition of re-finding and new-finding user in the rest of this dissertation.

In this investigation we also want to explore the notion diversification and repetition seeking by the users. In our user survey described in Section 3.3.3 we described four possible intentions behind re-searching: re-finding, new-finding, new content seeking, and comparing result lists. Re-finding can be viewed as repetition seeking as the user was first and foremost seeking the same URL, although the content on said URL could have changed. Participants in the user study who reported intentionally new-finding and or seeking new content regardless of the URL can be viewed as diversification seeking. When a user seeks to new-find, he or she may come across a URL that contains duplicate content as one previously visited. However, the user's primary stated purpose is to new-find, and thus find a new URL they had not previously selected. Another way to distinguish the re-finding (repetition seeking) from the other three intentions, new-finding, new content seeking, and comparing result lists for changes (diversification seeking) is that in the case of the intentional re-finding, the user may benefit from the search engine results remaining unchanged, while in the latter three cases the user may benefit from a change to the search results.



(a) The number of queries participants anticipate reissuing at some point in the future.



(b) User reported intentions behind queries they reissue. Multiple answers permitted.

Figure 4.1: Results of the user survey on re-searching. Participants stated (a) their typical intentions behind re-issuing queries and (b) and how many previously issued queries they intend to issue again in the future. Since users often had multiple re-search queries, they may have multiple typical intentions.

4.2 Characteristics of Re-Search

We begin our discussion by looking at re-searching in general, giving an overview of prevalence of re-searching. We explore the patterns of types of findings within the re-search trail, observe that users exhibit both diversification and repetition seeking behaviors, and often click on re-findings and new-findings within the trail. We find in some cases users may have multiple URLs they frequently re-find, and that users are selective in what new URLs they click.

4.2.1 Overview

Re-searching is observed to be common in the search logs and often reported to be intentional by participants of the user survey. All participants of the survey were able to recall an instance where they had issued the same query multiple times over multiple days, with 25

participants (39%) re-searching at least once a day. Almost all participants could recall at least one query they intended to re-search again in the future, while 37 participants (approximately 58%) had 5 or more queries they intended to re-search, and 9 participants (14%) had more than 20 (Figure 4.1 a). In the AOL search log, same query re-search accounted for 39.5% of all query instances over the three month interval. As suggested by the user survey, the log also shows many re-search queries are re-searched again. Trails of multiple re-searchings were common with 55.1% of same query re-searches belonging to a trail of length greater than 2.

Based on search log analysis and feedback from our survey participants, some common categories of re-search include: 1) names of individuals, including celebrities, sports players and TV shows. Participants in our survey also searched for themselves, and people they knew; 2) educational (i.e. "two year old milestones", "mother goose preschool lesson plans") as well as schools and universities; 3) activity, activity planning (i.e. "bridal shower ideas", "kids games", "paintball", "trip planning") and activities for kids; 4) news; 5) reviews; 6) inspirational, such as cooking ideas and topical blogs; 7) shopping for expensive items such as cars and home repairs; and 8) searching for coupons. These last two categories show a potential benefit of incorporating re-search analysis into improving other areas of search, such as advertising, and not just the user's information goals.

The intentions behind re-issuing the same query, as reported by participants of the user survey, are shown in Figure 4.1 (b). The most common intention was re-finding (repetition seeking), followed by searching for new content and new-finding (diversity seeking). More importantly participants also reported having multiple intentions for the same re-search query. A re-search query was used to both re-find a webpage and find new webpages (reported by 40 participants), re-finding and finding new content (38 participants) and finding new websites and

new content (29 participants). Thus there appears to be a dual need for repetition and diversity within the same re-search trail.

A combination of diversity seeking and repetition seeking behaviours within the same re-search trail can also be observed in the AOL dataset. Of same query re-search trails of length 5, 52.7% contain both repeat and new URL clicks in the re-search trail. Only 37.7% include exactly one clicked URL (i.e. are repetition seeking only) while 9.6% of the trails have only new URL clicks and no repeat clicks (i.e. are diversity seeking only).

4.2.2 Simultaneous Diversification and Repetition Seeking

In many instances the same URL is repeatedly clicked along side new findings throughout the re-search trail. Consider the example from the AOL search logs illustrated in Table 4.1. The user is interested in the British actor, Clive Owen. In general, an individual issuing this query my wish to watch clips from movies, read celebrity gossip, find movie reviews, purchase DVDs, etc. In this example the most frequently clicked URL for this trail, *http://imdb.com*, accounts for 63% of the clicks in the trail. The user also clicks a news site and two fan sites. While there is one URL the user frequently visits, he or she does not settle on it, clicking other URLs throughout the trail.

We refer to the URL that is clicked on more frequently than any other URL within the re-search trail as the *click based dominant URL*. By this definition, there can be only one dominant URL. Re-search trails that are solely new-finding, therefore, do not have dominant URLs, while for re-search trails that are solely re-finding, the re-found URL is the dominant URL. Therefore, in this section we only consider trails with at least one re-finding and one new-finding within the trail from trails with at least four clicked queries.

		Query: *Clive Owen*
Q_1	$C_{1,1}$	http://www.imdb.com
Q_2	$C_{2,1}$	http://www.imdb.com
Q_3	$C_{3,1}$	http://news.scotsman.com
Q_4	$C_{4,1}$	http://www.imdb.com
Q_5	$C_{5,1}$	http://www.imdb.com
Q_6	$C_{6,1}$	http://secondsight.sirfrancis.org
	$C_{6,2}$	http://cliveowen.net
	$C_{6,3}$	http://www.harsh-light.com
Q_7		-
Q_8	$C_{8,1}$	http://www.imdb.com
Q_9	$C_{9,1}$	http://www.cliveowen.net

Table 4.1: An example of a re-search trail in the logs for the query "Clive Owen" issued by a single user over 32 day period.

While we only have domain information for the AOL search log, analysis on the Sogou log leads to similar findings when conducted on the full URL or just the domain. In the Sogou log, 80.0% of trails have a dominant URL. For 94.0% of the trails with dominant URLs, the dominant URL is the only URL clicked on in the trail for the corresponding domain. Additionally, for 98.5% of re-search trails where a single domain is clicked on more frequently than any other domain in the trail, the trail also has a dominant URL, and that dominant URL is from said domain. Therefore, when a domain is observed to be more frequently clicked than any other domain in a re-search trail in the AOL search log, it is likely the trail may also also have a dominant URL and that clicks on the dominant domain likely correspond with clicks on the dominant URL. Regardless, a dominant domain is still worth noting, as it may also hold a strong level of user interest.

Dominant URLs are common among re-search trails with at least one re-finding and one new-finding, accounting for the majority of repeat URL clicks within the trail as shown in Table 4.2. Overall dominant URL clicks are common across all re-searching. In the AOL log, 75.9% of same query re-search trails, 83.9% of minimal change re-search trails, and 78.9% of term overlap trails have dominant URLs. For same query re-search, dominant URLs account for 53.1% of all URL clicks, and 84.6% of all repeat URL clicks.

Not only are dominant URLs common across all trails, but they are also common in trails that have fewer repeat URL clicks (i.e. are more diversification seeking than repetition seeking). Figure 4.2 plots the frequency of trails with dominant URLs (A) and the percentage of dominant URL clicks (B) as a function of the percentage of repeat URL clicks, while Table 4.2 shows the percentage of trails with dominant URLs and the percentage of clicks across all trails. Figure 4.2 (A) shows that regardless of the percentage of repeat URL clicks in the trails, dominant URLs are common in trails for all similarity metrics and data sets. In other words,

Type of Re-Search	# of Trails	% of URL Clicks	% of Repeat URL Clicks
Same Query, $\sim_{=}$	76.9%	53.1%	84.6%
Minimal Change, $\sim_{M}$	83.9%	52.2%	71.1%
Term Overlap, $\sim_{T}$	78.9%	43.9%	67.2%

(a) In the AOL Dataset

Type of Search Results	# of Trails	% of URL Clicks	% of Repeat URL Clicks
Full URL	80.0%	45.3%	75.7%
Domain of URL	80.1%	46.1%	75.3%

(b) In the Sogou Dataset

Table 4.2: The overall prevalence of dominant URLs, and clicks on dominant URLs in terms of the number of trails that have dominant URLs, the percentage of URL clicks that are on the dominant URL and the percentage of repeat URL clicks that are on the dominant URL.

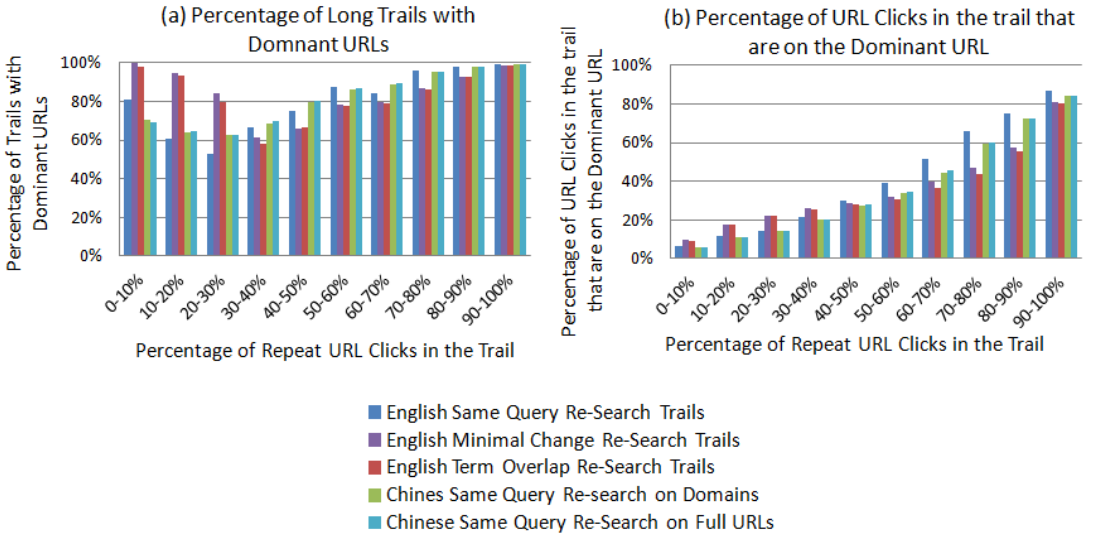


Figure 4.2: Frequency of dominant URLs in trails with at least one re-finding and one new-finding within the re-search trail, as a function of the amount of repeat URL clicks in the trail. Above: percentage of trails that have dominant URLs. Below: percentage of trail clicks that are clicks on the dominant URL.

even trails with many new-findings typically still have dominant URLs. These dominant URLs often accounts for at least half of all repeat URL clicks, as shown in Figure 4.2 (B).

Trails defined by the weaker similarity metric behind minimal change queries and term overlap query re-searching also exhibit a similar pattern. This is interesting as the definition of term overlap similarity was designed to incorporate the way a query may be modified, and accommodate a potential drifting information need. Yet, from Table 4.2 and Figure 4.2 we see that a dominant URL often exists and accounts for more than half of all repeat URL clicks. On the other hand, both minimal change and term overlap trails contain more non-dominant re-findings than same query re-searching. For same query re-search trails, 23.4% contain only

one re-found URL. In comparison, 13.5% of minimal change re-search and 15.9% of term overlap re-search respectively contain only one re-found URL.

Both Re-Finding and New-Finding Appear Intentional

Given prior emphasis on re-finding in the literature, one might expect that the intention of the user is to re-find the dominant URL, and thus expect the user to settle on the dominant URL. If this were true, most of the new URL clicks would appear at the beginning of the trail. Contrary to the notion of settling, it is rarer for new-findings to occur only in the first fourth of the re-search trail (10.3%) than the last fourth of the trail (27.8%)[1]. Of same query re-search trails, 39.0% have new findings throughout the trail. Therefore most re-search trails cannot be explained by the notion that the user is seeking a URL to settle on.

Another way to gain insight into whether an action is intentional is to explore the user's follow on actions. If the action was unintentional, the user may seek to correct it. The user may re-issue the query and click a different URL, the one the user initially intended, after only a brief period of time. For same query re-search, a re-search resulting in a new URL click is more likely to be re-searched again within the same session (39.1%) than a re-searching resulting in a re-finding (14.0%). However, a same session re-search query where the previous query resulted in a new-finding click is 1.67 times more likely to also result in an additional new-finding rather than a re-finding. Additionally, rarely is the query re-issued in under a minute (only 4.9%). Figure 4.3 shows the time before a query is re-searched again, given the type of finding it resulted in. Figure 4.3 reveals a bimodal distribution for time delays following a new-finding, one centered around 5-15 minutes and the other at 1 to 3 days, which could indicate two different behaviors.

[1]Just as users are not settling on URLs while re-searching, there are many instances where users do not settling on queries. We find 80.7% of term overlap re-search trails have a non-same query re-searching in the last half of the trail. In contrast, 66.1% of term overlap re-search trails have a non-same query re-searching in the first half of the trail. This indicates the information need may be continually evolving in some instances.

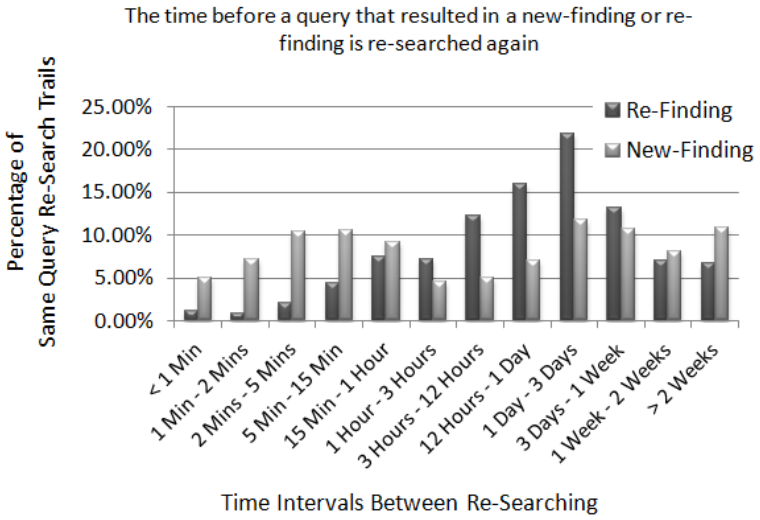


Figure 4.3: The distribution of time intervals before a same query re-searching that resulted in either a single click re-finding or single click new-finding is re-searched again.

One possible explanation as to why the user re-searches and new-finds within the same session is that he or she is actively seeking new information, and may not be satisfied with just one new URL. On the other hand, re-searching spread out over multiple days could be a sign of looking to see how information has changed (e.g. sports scores, weather or stock performances). We also find that, on average, same query re-searching resulting in a new-finding are re-searched again after a longer time period (27 hours longer), greater number of sessions (0.52) and greater number of queries (21.5) than re-searchings resulting in re-finding. Users do not appear to be 'correcting' a new-finding by re-finding. Thus, while it is possible that in some cases the new-finding may have been a mistake, it is not likely to be true for all re-searching resulting in new-findings.

Both re-finding and new-finding appear important to satisfying the user's ongoing information need. Since diversity seeking and repetition seeking behaviors often occur together,

it may be best to accommodate both behaviours simultaneously to address the user's evolving needs.

Individualization Occurs in all Aspects within the Re-Search Trail

A query that is a re-search query for one user, is often not a re-search query for another user. From the AOL log, we see that Re-search queries tend to differ for different users. Even considering highly navigational queries that are typically very common, only 4.4% of the user base has a four query re-search trail with the most common re-searched query. However, when a user re-uses a query, he or she may continue to re-use it frequently. 20% of the users had at least one same-query re-search trail of 7 or more queries. The mean average same query re-search trail length was 6.23. In our user study, participants acknowledged their re-search queries were often personal, including topics like hobbies and health. In a few instances participants declined to give specifics about their re-search queries, citing the personal nature of the query. Therefore, while the re-search query may not be issued frequently by other users, it is still important to the individual user.

Along similar lines, two users that issue the same re-search query could have different dominant URLs. For example, the re-search query trails for "jobs", "prom dresses", "airline tickets", and "Sudoku" have different dominant URLs for different users. In cases where the same query is issued by multiple users, 96.1% of the trails have at least two different dominant URLs for different users. Even queries normally considered to be navigational can also have a personalized dominant URL. In the case of the query string "google", some users prefer the news portal to the general search page while others prefer image search or the scholar tool.

The rate at which a dominant URL is clicked on is also user and query specific. The percentage of new URL clicks in each user's re-search trail can indicate a degree of diversification

Query: *Prom Dresses*		
Q_1		-
Q_2	$C_2, 1$	http://www.promgirl.net
	$C_2, 2$	http://www.prom-dresses.com
Q_3	$C_3, 1$	http://www.prom-dresses.com
Q_4	$C_4, 1$	http://www.cybernetplaza.com
	$C_4, 2$	http://www.dressesonline.com
	$C_4, 3$	http://www.yourprom.com
Q_5	$C_5, 1$	http://www.metrofashion.com
	$C_5, 2$	http://www.bargainweddinggowns.com
Q_6	$C_6, 1$	http://www.shopshop.com
	$C_6, 2$	http://www.promdressshop.com
	$C_6, 3$	http://www.prom-dresses.com
	$C_6, 4$	http://www.promgirl.net

Table 4.3: An example of a re-search trail in the logs for the query "prom dresses" issued over 33 days.

that the user may be seeking. For example, the dominant URL for the re-search trail "prom dresses" in Table 4.3 tends to correspond with different stores. Across users with this query, the average percentage of new findings per trail across users is 73.4%. While the user may prefer one store, he or she is willing to consider others. The queries "airline tickets" and "jobs" have a lower average percentage of new findings, 48.2% and 43.6% respectively. A user may be more settled on a preferred place to purchase tickets or peruse job openings in this instance.

4.2.3 Multiple Dominant URLs per Trail

For long re-search trails a user may frequently revisit a set of URLs. While we consider only multi-session trails in chapter a multi-session trail can contain instances of same-session re-searching. A re-finding that occurs multiple times in-session may show a different need than one that occurs in multiple sessions, as cross session re-findings may reveal a persistent interest in the URL by the user. We will explore this idea more fully in Section 5.2.5.

We consider a URL to be a *session based dominant URL* if it is clicked for re-search queries at least half of those sessions. Thus while clicked based dominant URLs can encapsulate the notion of most relevant URLs, session based dominant URLs may signal on-going interest over time. By this definition, a re-search trail can have more than one session based dominant URL, a URL can be a session based dominant URL and not a click based dominant URL and vice versa.

While it's more common for trails to have on session based dominant URL, we found many examples of trails with multiple session based URLs. Of trails with at least 4 queries, 16.7% of same query re-search trails have more than one session based dominant URL as shown in Table 4.4. The mean average number of session based dominant URLs for all trails was 2.3 for same query, 1.79 for minimal change, and 2.05 for term overlap re-search trails. Examples from

	% of Multi-Session Trails with	
Type of Re-Search	1 Dominant URL	≥ 2 Dominant URLs
Same Query, $\sim_=$	92.3%	16.7%
Minimal Change, $\sim_M$	88.6%	18.1%
Term Overlap, $\sim_T$	79.2%	29.8%

Table 4.4: The percentage of multi-session trails with session based dominant URLs.

same query trails include searching for news (i.e. "las vegas news"), shopping (i.e. "2007 honda crv", "discount bridesmaids dresses"), sports and entertainment (i.e. "green bay packers") and pornography. As the similarity metric is relaxed, the percentage of trails with one session based URL decrease, but the percentage of trails with more than one session based dominant URLs increase. This may be because the weaker similarity metrics are capturing two similar and overlapping, but not quite identical, information needs. Multiple session based dominant URLs were more common among term overlap queries (29.8%), mostly shopping and services (i.e. "gwinnett county public schools"). These multiple session based dominant URLs is that they might complement each other, such as a local school newspaper and a township paper, or offer contrary perspectives, such as sites that discuss different investment options.

4.2.4 Intentional Diversification Seeking

Users who are interested in diversity may be seeking new URLs, new content or checking to see whether and how the result list changed.

One way we can explore this idea is to examine the click entropy in the re-search trail. Entropy is a measure of uncertainty defined by equation 4.1, where $p(x_i)$ is the percentage of

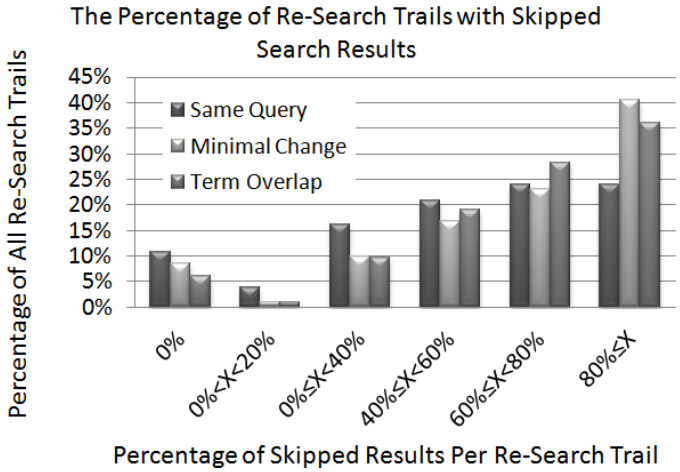


Figure 4.4: The distribution of re-search trails according to the percentage of skipped search results over the course of the entire trail in the AOL search log.

clicks on url x_i. Larger click entropies indicate more uncertainty, where smaller values indicate more determinism. As may be expected, non-dominant URLs in the trail typically account for more of the click entropy of the search trail than dominant URLs (an average of 1.53 vs 0.73). Therefore, there is more uncertainty in the non-dominant URLs. By studying a user's diversification seeking actions we can gain a better understanding of a user's breadth and depth of interest in a given topic covered by a query.

$$H(X) = -\Sigma_i p(x_i) log(p(x_i)) \tag{4.1}$$

Selective Nature in Diversification seeking

To get an idea of how selective users are within a re-search trail, we look at the ratio of non-clicked, or skipped results, to clicked search results over the course of the full re-search trail. In order to filter out navigational queries and ensure a non-trivial number of new-findings,

we consider only re-search trails with at least three different search result URLs clicked. We estimate the number of skipped results by counting the non-clicked ranks across the query instances for a given user's re-search query. We only consider skipped results that rank higher (and thus closer to the top of the result page) than the deepest clicked result, as the user may not have fully explored the result list for each issue of the query.

Since we are aggregating clicks across multiple search instances, it is important to note whether or not the search result lists are likely changing. In the AOL log 92.7% of same query re-search trails do not involve two different clicked domains at the same rank for the same user over two different instances of the same re-search query. The result lists from different instances of the same query, however, can still change in ways undetectable in the log, such as at non clicked ranks. Thus our estimation of the number of possibly relevant skipped results is likely to be conservative.

The results are shown in Figure 4.4. For 10.8% of same query re-search trails, users are not selective. That is over the course of the re-search trail they eventually clicked on all search results up to a certain depth of the search result list and no further. It may be the case that the most relevant URLs may already be at the top of the result list, or it may be the case that the user may be more affected by the click position bias for the given query. On the other hand, in 48.2% of all same query re-search trails users skipped at least 60%, or 3 in 5 search results and in 24.1% users skipped at least 80%, or 4 in 5, search results. These users appear to be selective in what they are searching for. The distribution of skipped URLs was similar in the Sogou dataset with 22.4% of trails having 60-80% skipped search results, while 19.7% have 80% or more skipped results.

Personalized Diversification Seeking

The most common difficulty reported by participants in the user survey associated with re-searching was irrelevant results (those they felt did not match the query), followed by not useful results (those they felt did not match their need) and not enough results. Yet the user survey and log analysis both showed a strong need for diversification in that users were often seeking new content and new URLs. Clearly not just any diversification is sufficient; results need to be diversified according to a user's interest.

We find many examples where users do not click on search results evenly across subtopics related to a given query in the AOL log. For example, one user same query re-search trail for the query "*32 Weeks Pregnant*" is shown in Table 4.5. While the user appears to have a very deep interest in woman's health, baby health and general pregnancy information, the user also appears to have a broad and shallow interest in photos, presumably maternity photos, and personal accounts (blogs) of pregnant women. Thus each subtopic is not represented equally in the user's result clicks.

Query: *32 Weeks Pregnant*

Q_1	$C_1, 1$	http://www.pregnancyguideonline.com	Information Site
	$C_1, 2$	http://pregnancy.about.com	Information Site
	$C_1, 3$	*anonymized* URL_1	Personal Blog
	$C_1, 4$	http://www.faqfarm.com	Information Site
	$C_1, 5$	http://parenting.ivillage.com	Information/Community Site
Q_2	$C_2, 1$	http://www.babycentre.co.uk	Information/Community Site
	$C_2, 2$	http://www.amazingpregnancy-pictures.com	Photography Site

Continued on next page

Table 4.5 – continued from previous page

	$C_2, 3$	http://www.pbase.com	Photography Site
	$C_2, 4$	*anonymized* URL_2	Personal Blog
Q_3	$C_3, 1$	http://www.faqfarm.com	Information Site
	$C_3, 2$	*anonymized* URL_1	Personal Blog
	$C_3, 3$	http://www.babycentre.co.uk	Information/Community Site
	$C_3, 3$	http://www.flickr.com	Photography Site
	$C_3, 4$	http://parenting.ivillage.com	Information/Community Site
Q_4	$C_4, 1$	http://parenting.ivillage.com	Information/Community Site
Q_5	$C_5, 1$	http://www.womenshealthcaretopics.com	Information Site
Q_6	$C_6, 1$	http://pregnancy.about.com	Information Site
	$C_6, 2$	http://parenting.ivillage.com	Information/Community Site
Q_7	$C_7, 1$	http://www.flickr.com	Photography Site
	$C_7, 2$	*anonymized* URL_3	N/A, Expired and Not Archived
	$C_7, 3$	http://pregnancy.about.com	Information Site
	$C_7, 4$	http://www.babycentre.co.uk	Information/Community Site

Table 4.5: An example of a re-search trail in the logs for the query "32 weeks pregnant" issued over 33 days.

In order to explore how similar users are in terms of their non-dominant URL clicks, we compare the conditional probabilities of a non-dominant URL clicks given the same query, the click based dominant URL as well as at least one non-dominant URL in common. By

requiring the dominant URLs to be the same, and not just the re-search query, we can lessen the effect of inherently ambiguous queries. For example, the query "jaguar" could refer to the car manufacturer or the animal. If two users with this re-search query share the dominant URL jaguarusa.com, they are more likely to have the same topic interest in the query: the sports car. The results are shown in Table 4.6. We find that the conditional probability of two users having a non-dominant URL in common given they have the same re-search query and dominant URL (9.6%) is only slightly higher then the conditional probability give only that they have the re-search query in common (9.1%)[2]. Recall earlier in Section 4.2.2 we found that 96.1% of queries had at least 2 different dominant URLs for different users. This shows that the dominant URL may not be an indicator as to which subtopics a user is interested in. If two users who have the same re-search query have a non-dominant URL in common, however they are 17.8% likely to have another non-dominant URL in common. Thus, surprisingly, the non-dominant URL provides more predictive capability than the dominant URL. If the two users have both a dominant URL and a non-dominant URL in common, the conditional probability increases to 22.7%. As shown earlier, the click entropy in the dominant URLs is lower than the click entropy in the non-dominant URLs. In many cases two users with slightly different interests in the same query may still have the same dominant URL. Knowing at least one non-dominant URL the users share in common may help reduce this uncertainty. This also supports the notion that users may only be interested in specific subsets of the non-dominant URLs.

The results of our log analysis show that many users engaged in diversification seeking while re-searching, do not click on all results, and do not click on non-dominant URLs randomly. This finding, combined with the feedback from the user study about irrelevant results, show that in many cases a user may not be interested in diversification with all possible subtopics.

[2]Having more than two non-dominant URLs does not increase the conditional probability by a statistical significant percentage.

Given	Conditional Probability of Non-Dominant URL Click
Query	9.1%
D-URL & Query	9.6%
ND-URL & Query	17.8%
D-URL, ND-URL & Query	22.7%

Table 4.6: Conditional probability of two users both having a non-dominant URL click in common given the re-search query is the same (Query), the click based dominant URL is the same (D-URL), and they have another non-dominant URL in common (ND-URL).

Connection with Abandoned Queries

Re-searching typically lead to more clicks than non re-searching and thus may be viewed as more fruitful searches. In the AOL dataset, same query re-search queries lead to a mean average of 0.77 clicks, compared to 0.63 clicks for the average query. Yet there is also a large percentage of abandoned queries while re-searching. An abandoned query is a query issued by the user on which the user clicks zero results. The percentage of trails with abandoned queries is given in Table 4.7. Of same query re-search trails, 17.6% contain only abandoned queries, i.e. are "abandoned trails". Same query abandoned trails account for 24.1% of abandoned queries in re-search, and 13.9% of all abandoned queries throughout the entire log. On average, a re-search trail with no clicks contains 2.72 queries, and spans 15.1 days. The top 25% longest abandoned re-search trails, however, averaged 8.8 query instances for same and minimal change, and 13.4 query instances for term overlap multi-session re-search trails respectively.

	Similarity Metric		
Trails	$\sim_=$	$\sim_M$	$\sim_T$
Contain No Abandoned Queries	20.3%	12.8%	18.5%
No More than Half Abandoned	29.8%	27.2%	39.6%
More than Half Abandoned	32.3%	40.8%	33.4%
All Queries Abandoned	17.6%	19.2%	8.5%

Table 4.7: Percentage of re-search trails that contain Abandoned Queries. The probability of any query in general being an abandoned query is 42.3%

While many early works generally consider query abandonment as a sign of failed search, it is unlikely that a rational user would continue to issue the same (or similar) queries that do not yield fruitful results. If the user is not finding relevant links to click, and the URL list is not changing, he or she would likely cease issuing the queries. One explanation is that the user may simply be satisfied with the snippet returned as suggested by [90]. Another possibility is that the user may be looking to determine whether the result list has changed. In some examples of re-searching, such as in the politician and cancer patient examples, the searcher may be primarily checking to see if new information is available. In these instances the absence of a new scandal for the politician or drug treatment for the cancer patient also provides useful information to the searcher. This user intention behind re-search was reported by 28% of our participants in our user survey, and was the leading intention associated with the difficulty of navigating the results with an abundance of irrelevant results.

4.3 Summary And Contributions

In this chapter we studied on-going re-search, which occurs when a user continues to issue the same query or similar query over time. As observed in the logs and confirmed by the survey (1) long re-searching trails are prevalent; (2) re-search consists of both repetition and diversification seeking behaviors and that these behaviors appear to be intentional; (3) users do not appear to be equally interested in all subtopics, even when re-finding the same URL and (4) re-finding may be the first step to new-finding. These findings have several implications for the design of our relationship based framework, as well as search engine design beyond the scope of this dissertation.

4.3.1 Implications for ARTEMIS

The evidence suggests novelty seeking may be along personal interests. New-findings do not appear to be random as evidence by the percentage of skipped results, and by the conditional probabilities of other new findings. Additionally, users appear to be intentionally new-finding while re-finding. This was a common situation as reported by our user survey and thus shows a need to support both actions simultaneously.

We saw in our exploration of session based dominant URLs that in as much as a third of term overlap re-search trails users had multiple session based dominant URLs. We theorized the user may be comparing or contrasting their content on his or her own. The search engine could present these results together in a way that aids in the user's understanding of the different URLs' content. For example, it could attempt to highlight key phrases or sentences that distinguish between the documents in the result snippets. Since the user has already identified the most interesting documents through his or her revisitations, the search engine would only need to compare a handful of sites on average. Even if the user was not attempting to interpret

the relationship of information on these session based dominant URLs, presenting them together may still be desirable, as it may help the user find these dominant URLs faster.

4.3.2 Contributions Beyond ARTEMIS

Since users do not appear to be interested in all search results, it may be desirable to filter out search results that the user is not interested in. In their study of re-ranking re-findings Shokouhi et al found that skipped results were not likely to be clicked during subsequent searches, but that clicked results might continue to be clicked [128]. We observed that the non-dominant URL clicks can help us predict which URLs a user will likely click in the future. The search engine could promote other non-dominant URLs clicked by similar users. It might also be advantageous to group non-dominant URLs into categories based on how they relate to the dominant URLs. For example, personal blogs and photography in our "32 Weeks pregnant" example. The search engine could then return more results like these to the user.

Another way a search engine could take advantage of these findings is to diversify the search results for abandoned queries. Half of all abandoned re-searches are re-searched again within the same session. While it is possible the searcher was satisfied by the snippet, he or she is likely not using all snippets from all search results. 28% of our participants in our user survey reported re-searching to see how the result sets have changed. Several methods have been studied to help understand the difference between good and bad abandonment, as well as detecting if the snippet was useful [66, 36]. Such methods could be used to detect valuable snippets. URLs corresponding to less valuable snippets could then be replaced in an effort to diversify results. Such an approach would likely not negativity impact clicked based methods of evaluation, such as normalized discounted cumulative gain (NDCG), since a searcher who abandoned queries in a re-search trail before, is likely to abandon it again. If the first two

queries in the re-search trail have been abandoned, the probability of future re-search queries being abandoned is 92.9%.

Chapter 5

Observations of Re-Finding

In Chapter 4 we explored ongoing re-search, the phenomenon where users continue to re-issue the same query over time. We found that users typically show both diversification and repetition seeking behaviours in terms of clicking on new search results (new-finding) and previously clicked search results (re-finding). We also explored in depth the personal nature of the diversification seeking. In this chapter we seek to explore the repetition seeking behaviour (re-finding) more closely.

In addition to studying the aspects of re-finding that a search engine typically encounters, we also study aspects of the re-found result to better understand why the searcher might have been looking for that particular page and what they wanted to do once there. To this end, we supplement the proprietary Microsoft query log analysis with analysis of the page's content, crawled daily, and with Web browser logs. This additional data enables us to study things like how the page content changes between visits and the consistency of the trails [165] people follow from the re-found pages. These datasets are describe in detail in Section 3.3.2

The key findings of this chapter include:

- (1) The query used to re-find a result is typically better than the query used to initially find it. Re-finding queries are shorter than the first observed query associated with a given URL click (the previous query), and rank the re-found URL higher. When re-finding occurs across multiple sessions, the re-finding query is also more common than the previous query.
- (2) Re-finding queries tend to converge. When a person repeatedly uses a search engine to find the same result, the query used may differ some initially, but typically becomes consistent over time.
- (3) The need associated with a URL appears to be consistent when it is found by the same individual. A user who clicks a previously clicked result is more likely to follow the same path than other users clicking the same URL.
- (4) Session-level and cross-session re-finding are very different. Cross-session queries change more substantially and in different ways than intra-session queries do. Cross-session re-finding may involve picking up a previous task. The queries at the beginning of a session are particularly likely to involve re-finding results found at the end of a previous session.

5.1 Experimental Methodology

In this chapter we focus on the re-finding behavior. While we include discussion about re-search, our primary focus is on re-finding. As such, we no longer constrain ourselves to re-searches nor re-findings in the re-search trail. Thus in this chapter we refer to re-finding and new-finding relative to the users entire search history, not just similar queries.

Since we are no longer focusing on queries in this chapter, we no longer differentiate between small query differences like those between same query and minimal change query

similarities. Instead we consider non-substantial and substantial change between the previous instance query and the re-finding instance query. Such a change may indicate a substantial change in the way the user expresses the re-finding URL.

5.2 Characteristics of Re-Finding

We begin our discussion by looking at re-finding in general, giving an overview of how prevalent re-finding is and what basic re-finding queries look like. We then explore how re-finding queries change, and show that when there are changes the re-finding query appears to be a better query than the previous query. We find that for multiple instances of re-finding of same URL by the same user, the series of queries issued by the user tends to converge to a single high quality query. We observe that people follow consistent trails from re-found results. We then investigate what may motivate the observed session-level differences, and show that re-finding may sometimes be a means of carrying tasks across sessions.

5.2.1 Overview

In general, 21.9% of all of the queries we studied were observed instances of re-finding. This is somewhat lower than the 38.8% reported by Teevan et al. [139]. The difference almost certainly reflects the shorter time period studied in our analysis (there is less opportunity to re-find with only one month of history versus a year) and the fact that we did not filter users to ensure a baseline amount of activity with the search engine per user (users who only appear in the logs for one query cannot re-find). Thus our value is a lower bound on the true incidence of re-finding during this time period.

Searchers appear to be targeting a particular URL more often during re-finding than new-finding. Participants in the critical incident user study described in Section 3.3.2 reported

intentionally seeking the clicked URL 48% of the time during re-finding and 30% of the time during new-finding. One participant was an outlier, and reported intentionally searching for the URL only 5% of the time. Excluding this participant, the difference is even more striking, with 72% of re-finding instances, and still only 30% of new-finding instances being intentional.

Single-click queries are particularly likely to involve re-finding; 29.6% of all single-click queries are re-finding queries. In contrast, the probability that a click during a multi-click query involves re-finding is only 5.3%. Although the first click following a query is always more likely to involve a previously found result than subsequent clicks, no click position has higher than a 7.2% probability of re-finding, regardless of click count for multiple click queries.

URLs that are re-found once are likely to be re-found again. On average, 66.1% of re-finding queries are also previous queries for a later re-finding. And if a re-finding query is not substantially different from the previous query, the result is even more likely to be found again (69.2%). Re-finding trails are discussed further in Section 5.2.3.

About half (48%) of all re-finding instances occur within a single session; the rest occur across sessions. The number of sessions between a re-finding query pair follows a long tail distribution, and averages 3.51. Re-finding is bursty, with re-finding queries appearing in groups. In a session, the query immediately after a re-finding query involves re-finding 59.3% of the time. Over half (51.1%) of the subsequent three queries are likely to be re-finding, as are 46.6% of all remaining queries in the session, all much higher than the probability of a random query involving re-finding (21.9%). Thus if a search engine observes a single instance of re-finding, it is likely to observe many more.

Most (79.2%) of the time when a result is re-found, the query used to re-find is exactly the same as the previous query, and an additional 11.4% involve only non substantial changes. These findings are consistent with previous work [139]. The remaining 9.4% of re-finding queries

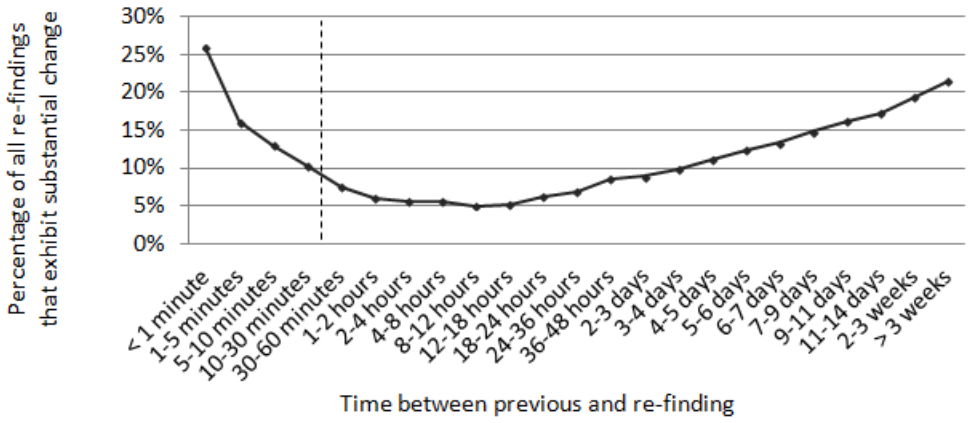


Figure 5.1: The percentage of re-finding queries that are substantially different from the associated previous query, as a function of the interval between the two queries. Dashed line indicates the cut-off for the session boundary. Queries are more likely to differ substantially when there is a very short or very long time interval between the re-finding and previous queries.

are ones that undergo substantial changes between the previous query and the re-finding query. The data collected via the user study suggests substantial changes are more likely to occur when the re-finding query was not specifically intended to lead to a particular URL. When participants reported that their query was intended to find the re-found URL, the query changed substantially 25% of the time; in contrast, when the URL was not being sought in particular, it changed substantially 48% of the time. Since a substantial change can indicate that the searcher has a new way of expressing their information need based on previous information interactions, we look more closely at this subset of re-finding queries in Sections 5.2.2 and 5.2.3.

The percentage of re-finding queries that are substantially different from the associated previous query are shown in Figure 5.1 as a function of the interval between the two queries. The dashed line indicates the earliest possible session boundary. Recall a session was defined as 30 minutes of inactivity from the search engine. Re-findings that occur before 30 minutes are by

definition same-session re-findings. Over half the sessions in our AOL dataset consisted of 2 or fewer queries, so re-findings over long periods of time are likely to be cross session re-findings.

From Figure 5.1 we see re-finding queries are least likely to change at intervals of about a day. Revisitation of popular pages commonly follows a cyclical daily pattern [4], and this behavior may reflect re-finding using oft repeated, well learned query bookmarks. Substantial query changes happen more often after short (less than an hour) or long revisitation intervals (a day or greater). These differences may reflect a qualitative difference in re-finding within a session as compared to across multiple sessions. How people use the search results they re-find is explored in greater detail in Section 5.2.4, and the differences between session-level re-finding and cross-session re-finding are discussed in Section 5.2.5.

5.2.2 Re-Finding Queries Are Better Queries

In this section we dive deeper into substantially changed re-finding queries. These queries provide a picture of how users modify their queries when the way they refer to their intended URL changes. The evidence suggests that searchers sometimes learn information about what they are looking for after the previous query that allows them to better express what they are looking for in the re-finding query. Our analysis shows that re-finding queries tend to be better queries than their corresponding previous queries; the queries become shorter, more common, rank the re-found result higher, and relate more directly to the text of the result.

Re-Finding Queries Shorter

Queries associated with re-finding are substantially shorter than queries not associated with re-finding. On average, a re-finding query is 12.1 characters long, and its associated previous

query is 11.7 characters long. In contrast, queries used to find new results are 18.9 characters long.

Re-finding queries that change substantially from the previous query are much more likely to be longer queries. They have an average length of 18.6 characters, similar to that of new-finding queries. This may be a reflection of intent. As discussed earlier, our user study suggests substantial changes tend to occur when the searcher is not seeking a specific URL. In contrast non substantial change queries have an average length of 11.4 characters.

The way length changes between queries varies as a function of the time interval between queries, as can be seen in Figure 5.2. We observe that queries get longer within a session, and shorter across sessions. When a re-finding query occurs within an hour of the previous query, it is 173% more likely that a word will be added to the query rather than a word being removed from the query. After an hour has elapsed, it is 106% more likely a word will be removed than added.

We hypothesize that the change in length reflects a fundamental difference between intra- and cross-session re-finding queries. For within session re-finding, people sometimes continue searching after a previous visit to the URL because the result does not initially appear to meet their need. When the same result is later returned for a longer, more targeted query, that can prompt a revisit to re-access the result's potential relevance. In contrast, across sessions users may be more likely to want to re-find a specific URL. In these cases, the shorter query reflects the user's ability to better express the target result based on information learned during previous interactions. In Section 5.2.5 we discuss these hypotheses and the evidence for them in greater detail.

Regardless of whether the re-finding query is longer or shorter than the previous query, it is very likely to substantially overlap with the previous query. In 52.8% of all re-finding

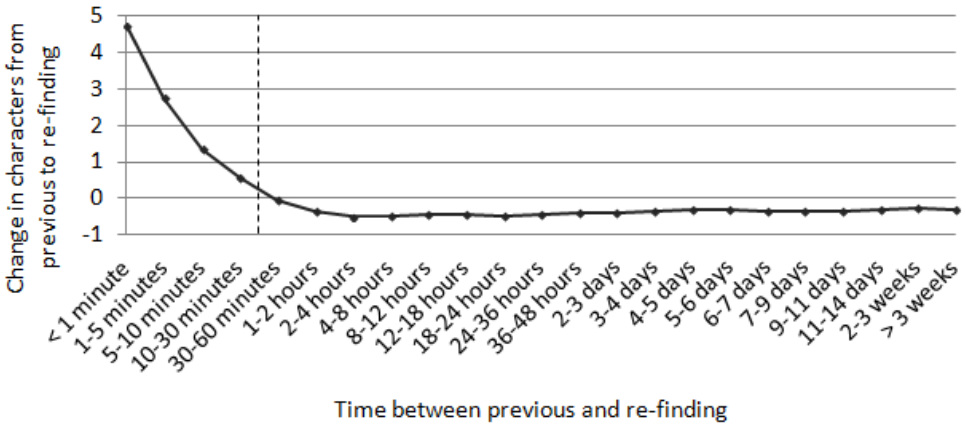


Figure 5.2: The change in query character length and query specificity for substantially changed queries, as a function of the time between the previous and re-finding queries. Within a session, re-finding queries are typically longer than their previous query counterpart, whereas across session they are typically shorter.

instances with substantial change, either the previous query is a proper subset of the re-finding query, or vice versa.

Re-Finding Queries More Common

Next we explore how common the query used to re-find a result is. A URL may be returned in the search results for multiple queries. Some of those queries may be more typically used when finding the URL, while others may be more atypical. For example, the queries *free music* and *pandora* both return the result http://www.pandora.com, but people more commonly search for the site using the latter query. To measure how common a query is for a URL, we look at the set of all queries which result in a click by any user on the URL, and measure the percentage of time we observe the query in question in that set. The measure is query and result

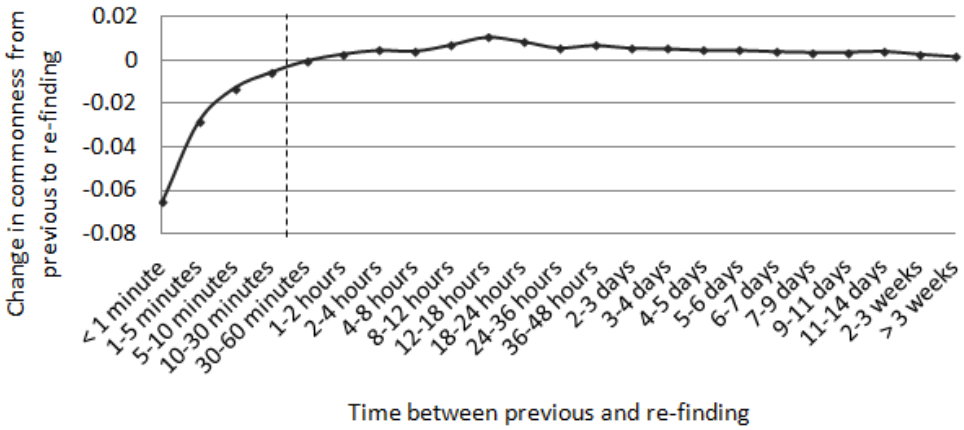


Figure 5.3: The change in how common the query is for substantially changed queries, as a function of the time between the previous and re-finding queries. Within a session, re-finding queries are typically less common than their previous query counterpart, whereas across session they are typically more common.

specific, but not user specific; a user may always click the result in question following the query, but if others search for the result using a different query, the query is not very common.

As with our earlier analysis, we compare the commonness of the query used to re-find a result to the previous query over different time intervals. The difference can be seen in Figure 5.3, as a function of the time interval between the two queries. For intra-session re-findings, the re-finding query is 2% less common than the previous query, whereas for cross-session re-finding it is 0.5% more common.

Re-Found Results Rank Higher

When a re-finding query differs substantially from the corresponding previous query, we find the rank of the re-found result also differs. On average, the result is initially found via the previous query at rank 1.65 (i.e., it is the 1.65th result from the top of the list). When it

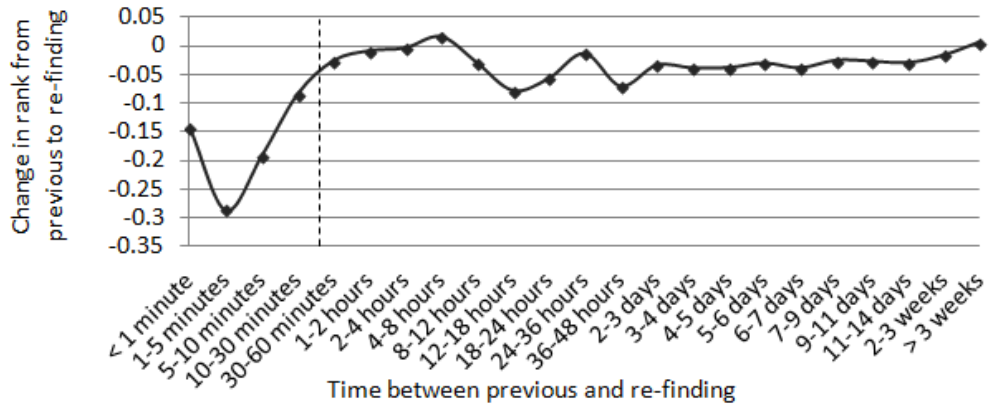


Figure 5.4: The change in position of the re-found result from the previous query to the re-finding query for substantially changed queries. The result almost always moves up in the result list.

is later re-found via a different re-finding query, it is ranked 1.57, or closer to the top of the list. The change in position of the result between the previous query and the re-finding query as a function of the time interval can be seen in Figure 5.4. We observe that in 21 of our 23 time buckets (consisting of 95.8% of all instances of re-finding) the average position of the result during the previous query is further from the top of the result list than the position of the re-finding query. Again, we observe somewhat different behavior when re-finding happens within a session versus across sessions. For intra-session re-findings, the position is decreasing by 0.19 ranks, whereas cross-session re-finding is only decreasing by 0.03 ranks. It may be that the significant change in rank of a previously found result within a session inspires the searcher to return to the result to see if what they are looking for can indeed be found there.

Query Type	Percentage of time slices query in page content				
	100%	75-99%	50-76%	1-49%	0%
Re-finding	85.8%	3.1%	0.5%	0.7%	9.9%
New-finding	78.3%	2.5%	1.0%	1.1%	17.0%

Table 5.1: The percentage of time the query terms are present on the page for each time slice. Re-finding queries occur more often in the static content of the page.

Re-Finding Queries More Related to Page Text

We also looked at how closely the query used to re-find a page matched the text content of the page, in order to understand how well the query reflected a consistent picture of the page. Our hypothesis was that while queries used to find content initially might reflect transient content on a Web page, query terms used for re-finding would reflect the static page content. Such queries would be more likely to consistently return the page in the result list, even as the page content is re-crawled by the search engine.

We measured how often re-finding and new-finding queries pointed to the static portion of the found result page using the percentage of time slices in the Web crawl which contained the given query words. For example, the query *times* might be in 100% of the crawled versions of the New York Times homepage, where as the query *obama* might be in only 80%, and the query *banana* in less than 1%.

We found, as expected, that re-finding queries were more likely than new-finding queries to refer to content that was consistently present in the page, and that new-finding queries were more likely to never actually appear in the page (see Table 5.1).

In this section, we have seen that when queries change substantially, they become shorter, more common, more consistently tied to the page content, and rank the re-found result higher. Further, the queries exhibit different patterns of behavior depending on whether the re-finding query occurs in the same session as the previous query or in a different session. These differences are discussed in greater detail in Section 5.2.5.

5.2.3 Re-finding Converges

In addition to observing that re-finding queries tend to improve over their previous query pair, we also find that for commonly re-found results, searchers tend to converge quickly on a good query to use for re-finding and stick with that good query over time. In this section we discuss re-finding trails, or instances of multiple re-findings of the same URL by a given user.

The queries used in re-finding trails appear to settle quickly. The conditional probability of the next instance in the re-finding trail involving a non substantial change between the previous and re-finding queries given that the current re-finding instance involves a non substantial change is 98.2%, whereas the probability of a re-finding instance involving a non substantial change is, in general, only 90.6%. The conditional probability that the next re-finding instance in a trail is a substantial change given that the current re-finding instance is a substantial change is only 18.9%. This shows that re-finding queries are unlikely to transition from involving non substantial change to involving substantial change. Further, 17.5% of the trails start with a substantial change, which is greater than the probability of a re-finding being a substantial change in general (9.4%). Substantial change re-finding instances are more likely to occur at the beginning of chains, and chains are more likely to end with non substantial change re-findings.

5.2.4 Need Consistent across Queries

In addition to looking at how queries change and evolve, we explored how the searcher used the re-found Web page. We did this by looking at the click trails of a user following a click from a re-found search results page. Recall that a click trail is a series of clicks following links on pages after the initial search result click. If the user were attempting to re-find previously viewed information reached by the re-found page, the click trail from the same search result to other pages outside the search engine is likely to also be the same, while if the user wanted to find new information on the re-found result, the re-finding trail may be different than the previous trail.

To explore the overlap in re-finding trails, we measured the percentage of time a given hop was the same across re-finding query trails with the same initial result click for a given user in the Microsoft search logs. For comparison, we computed a comparable value for the same URL using data collected from people who found it using a new-finding query. We looked at a number different types of hops, including: the second hop from the result page (the first hop being the click through to the result page), third hop from the result page, the first hop the user dwelled on for more than 30 seconds, and the final hop in the trail that started at the query result page.

As can be seen in Table 5.2, we find that trails are specific to users; the overlap between trails taken when a result is re-found is much higher than the overlap between trails taken from the same result by different users. When a person re-finds a result, they do the same thing more often than might be expected. Further, when the re-finding query is not substantially changed from the previous query, we find users are even more likely to follow the same path. The user tasks in these cases may be highly repetitive.

Change to re-finding query	Hop			
	Second	Third	Dwell	Final
Substantial	26.30%	21.27%	13.00%	18.44%
Non Substantial	43.93%	30.54%	21.07%	26.96%
All Re-finding	38.80%	26.87%	19.35%	19.67%

(a) Trails following re-finding clicks

Hop			
Second	Third	Dwell	Final
10.43%	3.97%	5.41%	5.01%

(b) Trails following new-finding clicks

Table 5.2: Percent of hops that are repeat hops in search trails following (5.2a) re-finding for a given user, and (5.2b) new-finding across users, given the first hop is the same.

Session	Hop			
	Second	Third	Dwell	Final
Same	39.96%	33.43%	27.27%	24.95%
Different	41.76%	27.87%	17.14%	25.90%

Table 5.3: The percentage of hops that are the same following a re-finding query as they were following the associated previous query, broken down by whether the two queries occurred in the same session or not.

The trails people follow after a re-finding result click varies as a function of whether the re-finding occurs within the same session or within a different session. As shown in Table 5.3, users are at least as likely to follow a consistent path from a re-found result when it is re-found in the same session as when it is re-found in a different session. One reason for this could be that if a user intentionally wants to re-find to retrace a given trail, it is easier for the user to retrace previous steps within the same session. But all of our data taken together suggests the picture may be richer.

5.2.5 Session-Level Differs from Cross-Session

The analysis we have presented thus far suggests re-finding is very different when it occurs at the session-level as compared to across sessions. In Section 5.2.2 we saw that when re-finding occurred within a session, the re-finding query was more likely to be longer, less common, and rank the result much higher, where as when the re-finding occurred across session, the query was more likely to be shorter and more common. In Section 5.2.4 we saw that people were more likely to follow the same path when re-finding within a session than across a session.

In this section we look at what all of these findings together tell us about how re-finding is being used at the cross- and intra-session level.

Intra-Session Re-Finding is Reevaluating

We hypothesize that some instances of session-level re-finding may involve the user returning to a previously found result that the user initially believed did not satisfy the user's information need, but that the user was willing to revisit to see if it now satisfies that user's need. The re-finding query within a session is typically longer, and the re-found result is typically ranked closer to the top of the list.

In contrast, we hypothesize that cross-session re-finding involves people trying to intentionally re-find the same result they have seen before as easily and directly as possible. The queries across session are less likely to change, and are likely to be short and common when they do, and rank the result somewhat higher. If the re-finding interval is at least a day, the amount of change to a cross-session re-finding query generally increases. We suspect this may reflect the user forgetting the previously used query terms, as well as changes to the search results and page content.

However, although we suspect results re-found across session are more likely to be actively sought out than results re-found within a session, we also suspect they are more likely to be visited to find new information. We saw in Table 5.3 that intra- and cross-session re-finding had almost the same percentage of second hop overlap; however same session re-findings was much more likely to have the same third hop and to dwell on the same hop. In different sessions, the users may be looking for new content; for example, checking a news website and navigating to the sports page. Such repeat trails would likely have periodic but cross-session patterns. In

these cases, it is also likely that the user would choose a new article to read after repeating the first step.

To better understand the validity of these hypotheses, we look to our user study. We observe that only 41.4% of the re-finding that occurs within the same session was labelled as intentional. This number is closer to what is typical of new-finding queries than is typical of re-finding queries. There are a number of examples of unintentional findings within a session. In some cases this happens after the user substantially changes their query (e.g., from assembly programming to assembly tutorial). In these cases, the user may have felt on first glance that that result did not adequately meet their need, but was more confident in the result when it appeared again for a different query. As we saw in our analysis of the log data (Section 5.2.2), in the user study the average change in result rank was the greatest in the first half hour. The large move up the result list may have influenced our participants' beliefs that the result would satisfy their information need. We also observe that there are multiple URL clicks to other pages in-between the intra-session re-finding instances. It may be that those other pages do not satisfy the information need, so the user chooses to return to one that might.

Cross-Session Re-Finding is Picking up a Task

When looking more closely at cross-session re-finding, we see some evidence to suggest users may sometimes be picking up a task they left off when re-finding across sessions. Figure 5.5 (a) shows the probability that a query will be a previous query while Figure 5.5 (b) shows the probability that a query will be a re-finding query, both as a function of its position in the session. The fact that the last query in a session is much more likely to be re-found in the future could indicate sessions often represent tasks that are not yet complete. Similarly, the initial query in a session is more likely to be a re-finding query than other queries in the session, and

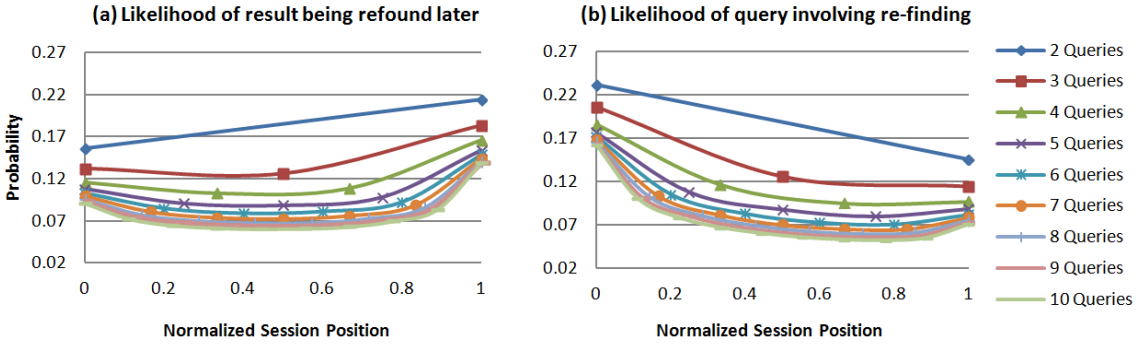


Figure 5.5: The probability that a query at the position in the session is (a) the previous query in a re-finding instance, or (b) the re-finding query in a re-finding instance. Each line shown represents cross-session re-finding probabilities for sessions with a given number of queries. Results found at the end of a previous session are more likely to be re-found, while results found at the beginning of a re-finding session are more likely have been seen before.

could indicate searchers may be picking up a task that was previously abandoned. When there are at least five queries in a session, the first query is over twice as likely to be a re-finding query than the last query in the session. Similarly, the last query in five-query sessions is 1.4 times more likely to be a previous query associated with future re-finding.

In the Figure 5.5 (a) we also see a slight increase in probability that the last query in the session is also a re-finding query, as well as an increase in the probability that the first query is the initial query in Figure 5.5 (b). This is because, as mentioned in the discussion of query chains, queries that are re-finding queries are also more likely to be previous queries for a later instance of re-finding than queries in general are.

5.3 Summary And Contributions

In this chapter we studied the re-finding, which occurs when a user uses a search engine to return to a document he or she has previously found via search. Our analysis yielded four key findings: (1) the query used to re-find a result is typically better than the query used to initially find it; (2) re-finding queries tend to converge; (3) The need associated with a URL appears to be consistent when it is found by the same individual; and (4) session-level and cross-session re-finding are very different. These findings have implications both for our framework design, described in Chapter 6, and for search engine design outside the scope of this dissertation.

5.3.1 Implications for ARTEMIS

The most straight forward way to improve search based on this study of re-finding would be to provide commonly found URLs to the user. Many re-findings are re-found again, so simply keeping track of previous re-findings may be the easiest way to see an immediate benefit. In some cases, even when the result is not going to be returned, a search engine may be able to identify that it should. If, for example, the user's current query is a substring of a previous query, the search engine may want to suggest the results from the history that were clicked for the longer query. In contrast, queries that overlap with but are longer than previous queries may be intended to find new results more than previously viewed results.

Intra-session and cross-session re-findings may be two different behaviors and thus we might want to support them differently. Users typically follow the same path and report the re-finding as intentional more often when re-finding cross session than when re-finding intra session. At the beginning of a session, when people are more likely to be picking up a previous task, a search engine should provide access into history. In the middle of the session, it makes

sense to focus on providing access to new information or new ways to explore previously viewed results.

5.3.2 Contributions Beyond ARTEMIS

The findings in this chapter also have implications for search engine design beyond this dissertation. Re-finding queries may make useful labels. The query used to re-find a URL is often better than the query used to initially find it. They are shorter, more commonly used when finding the URL, and more often contains words commonly found in the contents of the webpage. We believe the re-finding query may express how the person has come to understand this result. Such labels may be useful in clustering, either as the full label or as an annotation. In a group of clustered documents and re-found URLs, the re-found URL may be the most interesting URL in the cluster to the user. By using the query as the cluster label, the user may be more easily able to re-find the URL. According to the cluster hypothesis, clustered documents are likely similar, so the label will likely be somewhat applicable to the rest of the documents.

The search engine may be able take advantage of the repeatable nature of re-findings. For example, we observed that when a person issued the same query twice and clicked on the same result each time, a future identical search was highly predictive of a repeat click. In these cases, the search engine can treat the result specially and, for example, taking additional screen real estate to try to meet the user's information need with that result. The search engine may be able to predict the users final destination, given the consistency in the click trails following a re-finding. Possible destinations could be provided as either deep links within the snippet, or as an additional result to help the user reach their intended destination faster.

Thirdly, given that re-finding queries may represent a new, and by many definitions better, way of thinking about the re-found URL, it may be possible to use re-findings to help

novice searchers. For example, if a novice searcher uses a common previous query, the search engine may suggest the common re-finding query.

Part III

ARTEMIS Design and Implementation

Chapter 6

The ARTEMIS Framework

In this chapter we describe the ARTEMIS (Assisted Relationship Tracing for Exploratory Multi-Session Informational Search) framework based on the findings in Chapters 4 and 5. ARTEMIS is designed to aid ongoing information searches that cannot be satisfied within a single query or session, by simultaneously accommodating both the desire to return to past sources (repetition seeking behaviors), and aiding in the discovery of new documents (diversification seeking behaviors). Prior work has focused on task resumption or the discovery of new information separately. Our unified approach goes a step further by showing how new documents may be relevant given a subset of past documents.

ARTEMIS employs a similar concept to that of justification in recommender systems, described in detail in Section 2.3.3. In recommender systems a justification is sometimes given to illustrate to the user why an item might be relevant to him or her. This justification is often separate from the underlying recommendation algorithm, and based on something the user might understand such as "your friends liked this item" or "this movie has the same actor as others you have liked." In a similar vein ARTEMIS shows relationships between currently

unexplored documents in the search results and past high valued results that might "justify" the result to the user. In the recommendation system analogy the search engine is the recommender, and the search result is an item being recommended. Results clicked on by the user may be viewed as items purchased. The past high valued search results that are highly relevant to the user's underlying need are those items with positive ratings. The relationship of an unexplored document in search results to highly rated documents is then a form of justification of the result to the user.

Using this justification concept, ARTEMIS makes the assumption that when a user is shown how new documents that might be useful to his or her underlying task, the user can better identify which documents are relevant. Additionally the user can concentrate his or her attention on that aspect of the new document. For example, a user engaged in a literature review may not understand why a new paper is relevant. If a new paper cites the same probabilistic modeling papers that the user has previously indicated a strong interest in, the user could infer that the new paper may use a similar approach. Thus the user is given an indication as to which part of the new paper (the modeling section) may be relevant, and can focus his or her attentions on that part of the document.

In addition to helping users return to past information, ARTEMIS is designed to support:

- *Identification of the relevant documents* - Relationships may help users identify which documents may be more (or less) relevant to their underlying task.
- *Targeting of relevant information within documents* - Relationships can help users determine how a document is relevant, which then may allow users to target the relevant portions of a document.

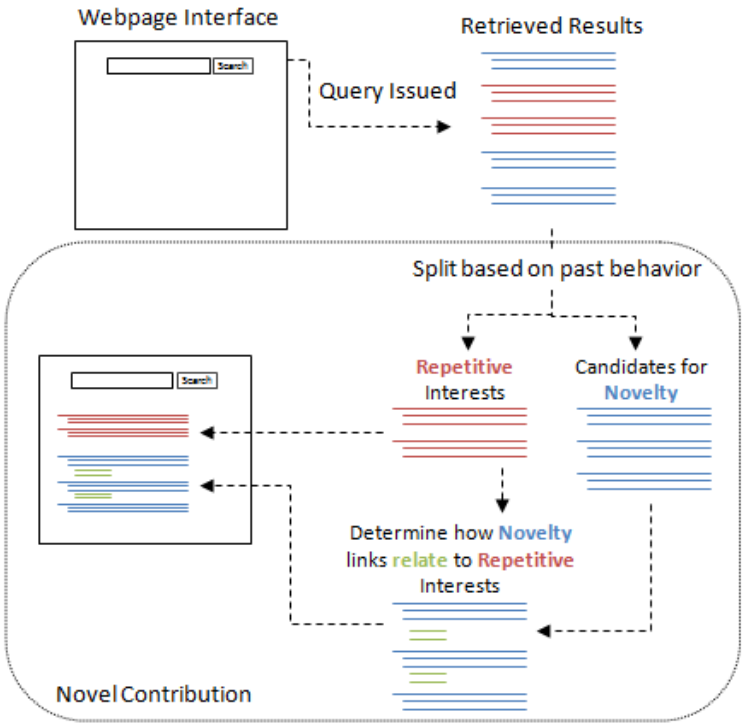


Figure 6.1: Workflow of our proposed system.

6.1 System Design

In this section we describe ARTEMIS. We begin by discussing the overall workflow, and then delving into each of the components in depth.

First, let us describe some notation. Let D be the past search results the user has seen. D can be partitioned into two distinct sets, $D = D_{\bar{\star}} \cup D_{\star}$ where $D_{\bar{\star}} \cap D_{\star} = \emptyset$. $D_{\star}$ represents the set of starred results and $D_{\bar{\star}}$ the unstarred results. We will discuss the concepts of stars in Section 6.1.2. Further, let D^n be the search results for the nth, and current, query. Then $D^n \subseteq D$. We can also partition D^n similarly, $D^n = D^n_{\bar{\star}} \cup D^n_{\star}$. Let d be a specific document $d \in D$.

6.1.1 The ARTEMIS Workflow

Figure 6.1 shows our proposed workflow. We begin with a black box search engine. Search results returned by the black box search engine are split into two categories: those that are *high valued repetitive* results ($D^n_\star$), referred to as starred documents, and those that are candidates for *diversification and novelty*, ($D^n_\star$), referred to as unstarred documents.

The repetitive results can be loosely interpreted as the user's core interests in, or possibly primary understanding of, a topic. We assume the user is both familiar enough with these documents that no further information about them is needed to assist the user in his or her understanding of these documents and that they continue to hold some level of users interest for future queries due to the repetitive nature of the user's interest in them. The candidates for novelty may also have been previously visited by the user, or may be new, never before seen URLs. The key distinction between starred and unstarred documents is that the user does not appear to be overly familiar with unstarred documents and that new information about these unstarred documents might help augment the user's understanding of either these documents or their underlying task.

To make our approach more transparent to the user, we signal the importance of documents to the user through the use of stars. The three stars in ARTEMIS, ordered by importance, are "User Starred" a gold star, (), "System Starred" a blue star, (), and "Unstarred/No Star" an outline ()[1]. The first, "User starred" are stars that the user indicates are important, while "System Starred" results are results the system deems important to the task and have no positive or negative user feedback and "Unstarred/No Star" is given to a search

[1]Stars have been used in previous online systems to flag important items. Most notably, Gmail (https://mail.google.com/) allows users to "star" important messages. However, Gmail differentiates between what the system deems important (by a flag icon) and what the user deems important (by a star icon). Thus the user feedback does not override the initial system classification.

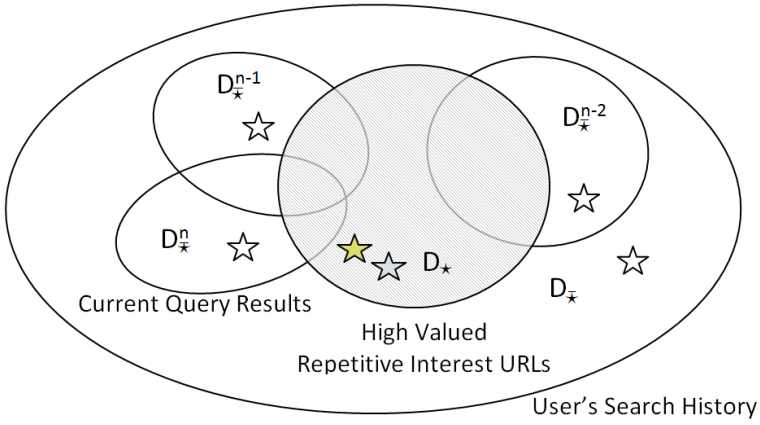


Figure 6.2: Digram of search space for a user. The shaded region indicates starred documents, both user starred and system starred. We predict the search results for each query related to the task will likely consist of some starred and some unstarred documents.

result that the user has either explicitly said is not important, or has not provided any feedback and the system does not predict it to be important.

Figure 6.2 shows the document space of search results presented to the user. The search results may or may not have been clicked. As shown in Figure 6.2 since starred documents are likened a user's core interest, it is likely some subset of them is often present in the search results, $D_{\star}^{n} \neq \emptyset$. We will explore this hypothesis in the next chapter.

The unstarred documents in the current search results ($D_{\bar{\star}}^{n}$) are compared against the complete history of high valued repetitive URLs ($D_{\star}$), not just those in the current result set returned by the search engine for the current query, to uncover relationships between the two sets of documents. We then present the relationship information to the user by displaying the accompanying relationship under the unstarred result. Thus the snippet of a search result shows how the document relates to the query, and the relationships show how the document might relate to the underlying task.

While it is possible for this extra information to distract the user by taking up screen real estate from the search page, prior work via an eye tracking study by Curtrell and Guan has shown that increasing the length of the contextual snippet improves user performance for informational tasks [40]. Surprisingly, Cutrell and Guan showed that users tend to look at more results when shown larger snippets. Thus we might expect the additional relationship information, especially those supported by text from the document, to be similarly beneficial to the user in this ongoing information seeking case.

6.1.2 Repetitive Interest URLs

In this section we describe the repetitive interest URLs, $D_\star$. The motivation behind starred documents concepts comes from the click based and session based dominant URLs encountered in Chapter 4. We theorized that these dominant URLs may be the user's primary interest in the search query as the user continues to return to these documents when re-issuing the query. The idea behind starred documents in a task is similar, that they may represent the user's core interest in the topic and likely have a connection to re-findings. As we have seen in Chapter 5, URLs that have been re-found before are likely to be re-found again, thus by keeping track of highly re-found URLs, we may help the user re-find again.

Over time the list of highly important repetitive URLs may change as the user's underlying interest may change. While user stars are permanent, and persist over the course of the task until explicitly removed by the user, system stars are ephemeral. Depending on the star strategy, the system may update it's prediction of document importance over time. A document that was deemed important and thus given a system star for one query, may not be viewed as important and starred for another. Thus system stars may only exist for the duration of a single query.

	Group	Description
Control Groups	*NoStars*	Both star system & relationships disabled.
	NoRelations	Users allowed to star documents, but relationships disabled.
Simplistic strategies	*PastClicks*	All documents visited in the past are starred.
	NoSystemStars	No documents are starred by systems. User must star all documents
	RandPastResults	Past search results randomly starred according to how frequently they appear in the search results.

Table 6.1: Strategies for identifying repetitive interest URLs.

Automatic Idenitification of Repetitive Interest Support URLs

We propose several simple strategies for identifying repetitive interest URLs in Table 6.1. For our control groups we consider a group with no stars or relationships (group *NoStars*) and one with stars and no relationships (group *NoRelationships*). Book marking and storing past visit information is a popular technique in the literature for remembering, returning to and re-finding past documents, as discussed in detail in Section 2.2. Thus we might expect one potential use of the star system is to star documents the user deems possibly relevant to be read later (i.e. bookmark the document). The purpose of control group *NoRelationships* is to identify what benefit comes from relationships beyond simple bookmarking, thus this group also does not have system stars. Users can still star documents they wish to return to, but no support relationships are offered. Thus stars in this context are effectively a form of bookmarks.

We include 3 simple strategies for starring documents. The first is to star all past clicked documents and therefore potential re-findings (Group *PastClicks*) which will include both session based and click based dominant URLs as discussed in Chapter 4. Next we propose a system strategy that leaves all documents unstarred, thus the user must explicitly star all important documents (Group *NoSystemStars*). Finally we propose randomly selected documents that have previously appeared in the search results according to how frequently they have appeared in the search results (Group *RandPastResults*). While it may seem that random past results would make poor candidates for starred documents, we note the fact that a document persists in the search results for multiple queries in the same task is itself a signal of possible relevance to the underlying task. Along similar lines, we previously theorized in Section 5.2.1 that the multiple occurrence of a particular URL in different result sets for different queries may affect the user's perception of relevance in our exploration of unintentional re-finding.

Analyzing Star Quality

There are two goals behind our simple star strategies. The first goal is to evaluate how effective simple strategies can be. The second purpose is to test the influence of system stars on the perceived relevance of the documents and relationship accuracy. Prior research has shown that users often do not scrutinize recommendations [42]. Similarly it has also been shown that users can be biased by system ratings and tend to give ratings similar to ratings that were assigned by the system, regardless of the accuracy of those system ratings [39]. Whether or not the system stars a document might sway the user's opinion on the importance of that document. These multiple naive strategies allow us to explore to what extent the system stars influence the users' perception of importance. If the system does not influence the user much, and the user is actively engaged in starring and unstarring results, we would expect the final list of starred

documents, with either a system or user star, to be similar across the groups. We explore this notion more fully in Section 7.3

The star system allows the user to explicitly state what is important and relevant to their task. This not only improves the quality of the starred documents and, by extension, the relationships of unstarred documents for the user, but it also provides a means of explicit feedback as to what should and should not be starred. Since overall performance in the task is affected by both the quality of the stars and the quality of the relationships, this allows us to analyze system starring strategies separately. However, we note that there may be a discrepancy between what the user wishes to see starred, and which starred documents would yield the highest task accuracy.

6.1.3 Relating New Results to Past Results

In this section we discuss possible relationships between documents. While a relationship based search interface has similarities with faceted search interfaces, we note the primary goals of faceted search and relationships are not the same. As such, the characteristics of good facets are not characteristics of good relationships. In faceted search the goal is to provide the user with a meaningful perspective of the search space as a whole, to allow the user to more easily navigate the search space. Thus facets that provide full coverage over the result set and partition the search space roughly evenly are preferred as discussed in Section 2.3.1. Relationships, on the other hand, are designed to show how a specific document relates to the underlying task, and as such should be as unique and specific as possible. A relationship that highlights the fact that two papers were published at the same conference is not as informative as one that says they were published in the same track in the same conference, or presented in the same

	Legal Decisions	Scholarly Articles	Web URLs
Entities	E.g. Litigants, Circuit System, Judges	E.g. Authors, Institutions, Venues	E.g. People, Locations, Organizations
Link Structure	Decision A cites B C cites both A & B A & B both cite C	Paper A cites B C cites both A & B A & B both cite C	Page A links to B C links to both A & B A & B both link to C
Vocabulary		E.g. General Terms	E.g. Keywords

(a) Structure Based Relationships

	Legal Decisions	Scholarly Articles	Web URLs
Agreement	Upholds/Overrules	Supports/Contradicts	
Depth		Shallower, Deeper	
Coverage		Broader, Narrower	

(b) Semantics Based Relationships

Table 6.2: Types of possible relationships between two documents (A & B), for different domains. Some relationships may involve a third document (C) that may not have been previously visited, or even appeared in the search results.

session as sessions tend to group like papers. Thus it may be beneficial to have many highly specialized and sparsely used relations that only apply to a small sub domain.

Table 6.2 lists possible relationships between two documents that a user may find useful in their information journey. We divide the relationship types into two categories: structure based and semantic based. By structure based we mean relationships that typically can be found within the structure of the document in at least some domains, for example key words and general terms in ACM style papers. Semantic based relationships are more challenging, requiring an understanding of meaning in the document.

Structure Based Relationships

We include in our description 3 types of structure based relationships: links between the two, entities in common and keywords in common. We note in some domains these relationships might require more semantic understanding of the documents. Litigants in legal documents tend to be identifiable via structure, as do authors in scholarly articles. Additionally, scholarly articles may have special formats to indicate keywords (such as the keywords and general terms sections of ACM style papers.)

Researchers have also explored identifying entities in free text. The task of identifying entities is usually referred to as named entity recognition. Approaches include statistical models [98, 58], machine learning [178] and others [104].

One method for distinguishing key phrases in documents is to use vector space models such as tf-idf [133, 124], however researchers have also explored point-wise KL divergence [1, 146] for identifying key terms in market intelligence [55] and political blogs [56] as well as graph based methods [57].

Relationships involving citations or links between documents is perhaps the most intuitive relationship for our goal. On the world wide web links are expressed via explicit structure through document meta data. In the legal and scholarly domains, the link structure needs to be inferred through the citations. While the legal domain has specific formatting for citing cases, cited documents in scholarly citations may be abbreviated to conserve space. Thus one area of research is on entity resolution [53, 18] which seeks to identify when two references refer to the same entity. For example, author Sarah K Tyler has been cited as "S K Tyler" and "S Tyler". A citation graph may erroneously create three entities for all three variants of her name. One challenge is that multiple nodes may have the same textual representation. There are currently two published authors named and cited as "Sarah K Tyler" currently shown in Google Scholar.

Semantic Based Relationships

Semantic based relationships are relationships that cover the meaning in the documents. We list three type relationships: supports/contracts, shallower/deeper and broader/narrower. The supports/contradicts relationship could be supported using sentiment analysis [105, 112, 113], which is usually done to understand the favorability of items in reviews. The favorability score could be used in connection with a citation to get a sense of whether the citation was favorable, or not. Topic modeling could be used to determine how closely the topics of two documents relate, and whether one subsumes the other.

While in general these relationships must be inferred, the legal domain documents follow a defined structure. Phrases such as "Plaintiff-Appellee" in the preamble not only signal who the plaintiff is, but who won the previous judgement and who is appealing - in this case the defendant in appealing the previous decision. Additionally, in the end of the document are phrases like "AFFIRM", "AFFIRM in part", "REVERSE in part", "VACATE" (cased

Figure 6.3: Layout of the ARTEMIS search page. Search results appear to the right in red (Starred Results) and blue (Unstarred Results). Below the blue search results are the relationships shown in green. The left hand side has two menus containing meta information

accordingly) make automatic parsing of the documents easier. Scholarly articles and the open web may provide more of a challenge. However, because these items are free text items, the entity resolution problem exists.

6.2 Interface

In this section we discuss the interface design of ARTEMIS. We begin with the typical search engine design common to Google, Bing and Yahoo with search results presented on the right of the screen, and a left side panel reserved for meta information & search tools as shown in Figure 6.3.

6.2.1 Primary Search Interface

The search panel shows the starred results under the heading “Starred Documents”. We opt not to refer to it as “Repetitive Interest URLs” to avoid user confusion. Unstarred

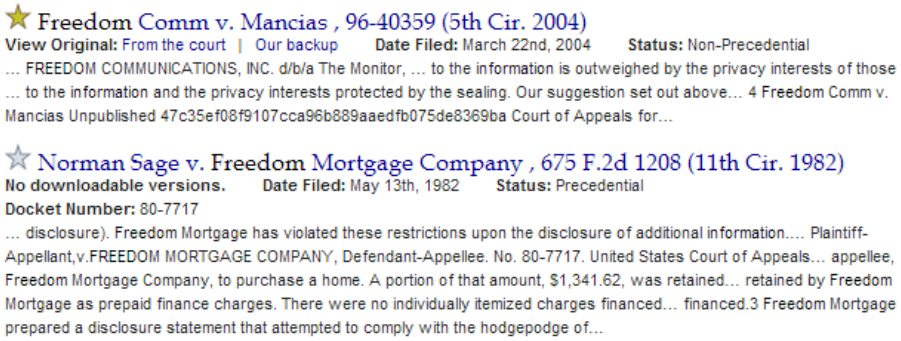


Figure 6.4: A screen shot of starred documents. The yellow star indicates that the document was starred by the user while the blue star indicates the document was starred by the system.

documents are displayed just below starred documents, under the heading "Other Documents". We choose this heading rather than "unstarred results" as we did not wish to imply to the user that these results are not relevant to their task.

Starred Documents

Figure 6.4 shows a screen shot of two starred results from our proof of concept system described in Chapter 7. Our black box search engine provides meta data about the document, primarily the date filed and the status (precedential or non-precedential) of the document and possibly the docket number. We left the snippet unaltered with one exception: the presence of a star to the left of each search result title. In the case of Figure 6.4, *Freedom Comm v. Mancias , 96-40359 (5th Cir. 2004)* was starred by the user and *Norman Sage v. Freedom Mortgage Company , 675 F.2d 1208 (11th Cir. 1982)* was starred by the system.

Documents with user stars are displayed above documents with system stars as shown in Figure 6.4. We do this for two reasons. First, we hypothesize that documents with user stars are more valuable to the user than documents with system stars. Again, we test this hypothesis

Figure 6.5: A screen shot of an unstarred document with relationships to starred documents.

with our proof of concept system in Section 7.3. Second, the ordering may aid in the case of intentional re-finding, where a user may be explicitly searching for a specific document. We hypothesize that re-findings are more likely to be starred documents. While the user might remember whether he starred the document, system stars are not permanent. There is no guarantee that a document with a system star for one search query will remain starred for another. Therefore, in the case where the user recalls the user starred status of a document he or she intends to re-find, he or she would only need to scan a subset of results looking for that document. This is the only re-ranking preformed with our system.

Unstarred Documents

Figure 6.5 shows the anatomy of an unstarred result. The snippet is similar to that of the starred result, however below the snippet is a list of relationships. In this case, only one relationship type is present for this unstarred document, that of a co-citation, and there are four

instances of this relationship. The search result *In Re the Reporters Committee for Freedom of the Press , 773 F.2d 1325 (D.C. Cir. 1985)* is co-cited with a high valued repetitive interest documents 4 times. Relationships displayed under the search also results serve a similar goal as the backwards highlighting approach in mSpace [169]. As the user peruses the search results set, he or she can see which facets belong to each document.

We display one relationship block per relationship type to avoid cluttering the interface. When there are multiple starred documents for the relationship type, we select the example relationship where the starred document has the highest perceived value to the underlying task based on the starred strategy. Thus we give preference to user stars first, then system stars. If, for example, we were using the strategy *PastClicks* where any past clicked document was starred, we would select the relationship with the starred document that has been clicked the most. For strategy *RandPastResults*, we choose the relationship with the starred document that occurred the most frequently in the prior search results. If the starred documents in multiple relationships are equally preferable, we give preference to the one that has been encountered most recently. A link below the relationship type shows how many more relationships of that type the search result has with starred documents. Clicking the link will reveal all relationships of that type for that result.

Each relationship block has up to three components: the relationship description, the supporting snippet if possible, and the number of additional relationship blocks available for that type. For structure based relationships, the snippet can be easily pulled from the document. In this case we highlight the supporting text around the citation. We provide multiple visual cues to distinguish this relationship types from supporting snippets, including color coding, indentation and bulleting. We color code the relationship description purple as it is complementary to green (the typical color of a URL) and noticeable against the black text. If multiple relationship types

were present, the user could scan the number of purple blocks of text to get a quick count of the number of relationship types. This could provide a visual indication of the relevance of a document, the implication being that the more relationship types and the more relationships present, the more likely that the document is relevant to the underlying task.

All titles of documents in the relationship description are linkable to the actual document. In addition, if the relationship description includes a starred document, a star is shown next to the starred document's title, which shows the pedigree of the importance of the starred relationship. In the case of the relationship shown in Figure 6.5, the starred document *Donald Wayne Engelking v. Drug Enforcement Administration* was starred by the user. The star icon is also useful in relationships like this one where the relationship includes a third document *Arizona Christian School Tuition Organization v. Winn*, which was not previously starred and may or may not be present in the search results for the current query.

6.2.2 Side Panels

Figure 6.6 shows the two components of the left hand side menu. While the panels appear next to each other in the above figure, in the actual system they are stacked with the Recent Stars panel listed first as indicated by Figure 6.3.

The Recent Stars panel is designed to provide additional re-finding support. The panel shows the stars the user has most recently encountered. We previously found cross session re-finding to be indicative of a user picking up a task as seen in Section 5.2.5. Additionally, we found results clicked at the end of the session were more likely to be re-found, and queries at the beginning of a session were more likely to lead to a re-finding. Given these facts, it makes sense to present the user with the URLs likely to be re-found when he or she initiates a new search session, perhaps even before a query is ever issued. This list is always populated, even when the

(a) The Recent Stars Menu

(b) The Relationships Overview Menu

Figure 6.6: A screen shot of the side menus

user is initiating a new search session after an absence, or viewing a document as re-finding is more likely to occur in short time intervals.

The Relationship Overview menu provides at a glance what relationships have been found between the search results and the starred documents. This panel effectively provides a grouped (clustered or faceted view where the relationships types represent the meta data the results are grouped according to) view of the documents. The user can select a relationship type and only those documents with that relationship are displayed. Thus the user can choose to narrow his focus to relationships that support his interests (and possible bias) or he can scan the result lists looking for relationships that may support an alternate view point as discussed in Section 4.2.4. Key choices include: Number of facets, as well as which and what order of facets. We collected click data from our pilot studies to order the relationship types by descending importance and then fixed the order.

While it is possible to order the relationships by clicks from each individual user, we opt not to keep the relationship order fixed. It has been found that individuals sometimes exhibit *Change Blindness* [129]. Change blindness is the term given when a change in visual stimuli goes unnoticed to the user. While studying change blindness in the context of interface designs, researches have found that users must focus their attention in order to see change and that users can only pay attention to 4 to 5 items on a page at a given time [120, 119]. Thus change in text headings may be too small to be perceived as there are many elements on the page, including multiple search results and other menus. It may take extra cognitive effort to detect and perceive changes in the order of relationships. Along similar lines, studies have also shown that re-ordering search results can negatively impact performance when re-finding [138]. Thus, without visual cues that the relationship ordering is changing it is possible for such a change to go either unnoticed by the user, or require extra time on the part of the user to process the change [119].

6.2.3 Additional Views

In addition to our main search page, we provide two additional views: the complete stars view and the in-depth relationship view. In our main search page we made several ad hoc decisions about how many items to display in the left side menus, and how many of each relationship type to show per unstarred result. The side menus provide only a limited view of stars and relationships. The primary function of these two views is to present the user with a complete picture of stars and relationships. We also collect click data on these views, both opening the page and clicking on a link contained within. These views therefore serve a secondary purpose of providing a means for us to explore the validity of these design decisions.

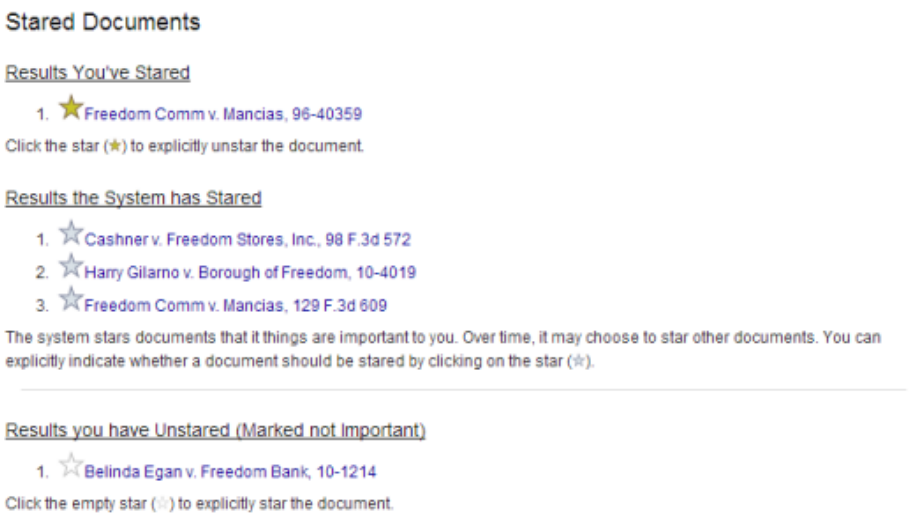


Figure 6.7: The Complete Stars View.

Complete Stars View

The complete star view shows the full list of starred documents, grouped first by type (User Starred, System Starred and Unstarred) and then ordered by the time of action as shown in Figure 6.7. The purpose of listing unstarred documents in addition to starred documents is to allow the user the opportunity to correct a mistake; to either unstar a user starred document or star a result the user previously unstarred.

In-Depth Relationship View

The In-Depth Relationship View shows all relationships for a given unstarred result, grouped first by relationship type and then ordered by the relevance of the starred document. The display is similar to Figure 6.4, but without the cap on the number of relationships.

6.3 Summary and Discussion

In this chapter we presented the ARTEMIS framework. ARTEMIS is designed to split search results into two categories: highly repetitive and important documents (i.e. "starred" documents) and candidates for novelty and diversification (i.e. "unstarred" documents.) These unstarred documents are compared against a past history of starred documents in order to uncover relationships. We also described three simple strategies for starring documents as well as providing an overview of relationship types common across multiple different domains.

6.3.1 Implications for non Mulit-Session Information Searches

While initial estimates were that 50%-60% of queries were informational [24, 122], more recent studies predict the number may be closer to 80% [71]. In this section we consider the implications of ARTEMIS on non-informational queries, specifically transactional (to complete a transaction or preform an action, such as buy a product, download multimedia, or launch a webapp) and navigational (to locate a particular page, such as that belonging to a specific person or organization).

While Cutrell and Guan found that increasing snippet length aided information queries, it inversely effected navigational queries [40]. Nevertheless, we still believe navigational queries will not be negatively effected by the ARTEMIS framework. First, relationships are only displayed if the search result is not starred and has relationships to past starred documents. We find that navigational queries[2] are often re-searched (70.8% and 86.2% in the AOL and Sogou logs respectively) and tend to lead to re-findings (70.5% and 85.0%.) As we will see in Chapter

[2]To find calculate these numbers, we used Liu et al's method for detecting navigational queries in large scale log data [95]. This approach is based on two assumptions, the *less effort assumption* which states users are less likely to click on URLs that are not the intended target, and the *cover page assumption*, which states users are likely to issue queries where the target is returned high in the results list. This method has been shown to be one of the more effective methods of automatically detecting navigational queries [23]. Other methods include characterizing the skew of the click distribution [87], click entropy [37] and exploring how well the anchor text matches the query [76, 87].

7 there is a strong correlation between re-finding, re-searching and user stars. In the starred strategies outlined above, it is likely that the navigational result would have a system star and thus would not have a lengthened snippet. Additionally, Cutrell and Guan's eye tracking study involved varying snippet lengths for snippets comprised solely of blocks of text. While we have effectively lengthened the snippet of a document, we provide visual cues so that the user can disassemble the snippet to the relevant components.

While we designed ARTEMIS specifically for informational queries, it may also have benefit for transactional queries as well. Recall from Chapter 4, the example of the re-search trail with query *prom dresses*. In this example the dominate URL corresponded to a given store, where non-dominant URLs corresponded to other stores. Comparisons across products are typical in most shopping situations and being able to indicate how stores relate (e.g. same distributer, same brands) may help users identify which stores are likely to have the products they seek. As described in our discussion of faceted search engines in Section 2.3.1, many of the web's largest e-Commerce websites employ faceted search.

6.3.2 From Re-Searches to Information Seeking Tasks

The framework for ARTEMIS was built after studing the simple case of exploratory multi-session information searches, namely that of re-search. We saw in our study of re-search in Chapter 4 that the number of session based dominant URLs increased as we relaxed our definition of similarity. Term overlap trails had nearly double the rate of trails with more than one session based dominant URL (29.8%) than were observed with same query trails (16.7%). We also note term overlap re-searchings are a subset of substantial change re-searching. Researchers have shown searches related to the same task tend to occur in clusters [72, 157]. Therefore, it's possible the term overlap re-searches occur in many of the sessions relating

to the same task. Thus the session based dominant URLs for the term-overlap re-searchings may also be session based dominant URLs for queries relating to the same information seeking task. The phenomenon of dominant URLs may continue to exist as we group queries by less stringent similarity metrics. Queries for multi-session information searchers are similar in that they relate to the same information need, even if they might be exploring different aspects of that information need.

Chapter 7

Proof of Concept Implementation

In this chapter we explore the validity of the ARTEMIS framework with a proof of concept implementation based on a search engine of legal documents. We note that the task in this study is a skilled task, that often requires expertise, but that our study is conducted over unskilled workers. As such the results may not necessary extend to all of web search. Nevertheless, this category of task, where a relative lay person tries to understand a complex problem, is an important subset of the types of task-oriented information searchers generally present on the web.

Completing a skilled task by an unskilled or under skilled worker is a challenging, and not all together uncommon, search tasks for an unskilled person to do while searching. Legal, Medical, Scholarly, topics that we have discussed throughout this dissertation, are examples of domains in which lay people may want to search. A searcher facing a serious illness may find themselves needing to search and understand the medical domain, while someone facing a civil suit may find themselves similarity needing to search and understand the nuances of his or her particular circumstances. A researcher may need to understand several complex topics to

adequately position her paper. A cub reporter, or opt-ed contribute may find themselves writing on topics they are non-experts and may need to search to fill in the holes of their knowledge. Even grade school students often turn to the web when needing to research a topic for a report.

In Chapter 6 we previously hypothesized that starred documents may represent the user's primary interest or core knowledge and that those starred documents are more likely to be the most revisited websites, and the most relevant documents. We explore this hypothesis and others in this chapter.

Key findings in this chapter include

- Users of our implementation of ARTEMIS are able to identify and understand relevant documents (identification). Users with stars had higher rates of precision and recall, and summaries from users with relationship were deemed to be of higher quality.
- Users of our implementation of ARTEMIS are better able to identify the relevant pieces within a relevant document (targeting). Users reported using the text for relationship support snippet to search for the relevant section within the document.
- Starring extraneous documents that may not be as relevant to the underlying task have negligible effects on user performance.

7.1 Proof of Concept System

Legal documents are notoriously difficult for lay people to understand as they are often long, terse, and filled with specialized jargon. Legal documents also tend to be written for a more educated audience with a higher reading level than typical of most web documents. Therefore, a user engaged in a search of legal documents may benefit from a relationship based framework such as ARTEMIS. Thus we conduct our proof of concept legal study in this domain.

7.1.1 Underlying Search Engine

We begin with "black box" search engine from CourtListener.com. CourtListener is an archival site and search engine for all precedential opinions for the 13th federal circuit courts and the Supreme Court of the United States as well as the non-precedential opinions from the circuit courts excluding the D.C. circuit. To date, it has indexed 767,498 documents. While we refer to CourtListener as a black box, it is an open source system[1] utilizing Solr[2], which is built on top of Apache's Lucene. The search engine uses a vector based similarity metric with porter steamer to identify relevant documents. One of the key advantages of Solr is it's customizability. CourtListener has gone through several iterations with different underlying engines and has been optimized for CourtListener's collection of documents and domain. The change most relevant to this dissertation is the doubling the default snippet length.

Solr allows for both fielded (faceted) queries and clustering of documents. Currently CourtListener utilizes the facets capability. The facets CourtListener use are court id (including the 11 circuit courts, the District of Columbia Circuit, the Court of Appeals for the Federal Circuit, and the Supreme Court of the United States), status (precedential, non-precedential, published, unpublished, relating-to, in-chambers), case name, judge, docket number, citation and case number. While we disabled the faceted view for our study, the option to do a fielded query based on the facets was still available to all users in our user study, including the control groups. Additionally, an tutorial to explain advanced query options was available via a link under the search box.

Of these fielded query options, Citation and Court are most similar to our relationships, however, the user experience between using fielded queries and the ARTEMIS framework is very

[1]The source of CourtListener can be found at https://bitbucket.org/mlissner/search-and-awareness-platform-courtlistener/src

[2]http://lucene.apache.org/solr/

different. In order to use the cited relationship in a fielded query, effort is required on the part of the user. A user must first realize that the citation relationship may be beneficial in their context. The user must next identify for which document the set of citations will be useful. Thirdly the user must correctly specify the document in the query. The format for specifying a citation relationship is *citation:'{VOL#} {ReporterName} {Page#}'*. This is not intuitively obvious to a lay person. ARTEMIS, on the other hand, automatically looks for and displays these relationships without user involvement, however, ARTEMIS only looks for citation relationships for a subset of documents, those relating to starred documents.

7.1.2 Limitations with CourtListener

The documents archived by CourtListener are Optical Character Recognition (OCR) text versions of the original court PDFs. When OCR errors are discovered, documents are re-scanned creating slightly different versions of the same document. While CourtListener makes an effort to de-duplicate these multiple versions from the search results, both versions are maintained in the database with different keys. These keys are embedded in the URL of the document. Thus it is possible for the user to click on two different URLs that correspond to two different versions of the same document, or for two different users click on two different versions of the same document in their search history. Therefore, we consider a document to be unique based on it's title key (referred to as the slug by CourtListener) as opposed to the URL. It is possible, though rare for the OCR error occur in the title, resulting in two different slugs for the same document. Thus we compare the gold citations by hand to ensure proper de-duplication of slugs ids.

Name	Description
CoCited	The search result is co-cited with starred documents
CoCites	The search result and a starred document cite an overlaping set of documents
Cited	The search result is cited by a starred document
Cites	The search result cites a starred document
Court	The search result is in the same court system as a starred document.

Table 7.1: Description of relationships identified in our proof of concept system.

7.1.3 Extracting Relationships

The relationships used in our proof of concept system are described in Table 7.1.

CourtListener extracts the *cites* relationship for use in their fielded query, by relying on the predictable structure of a citation in the OCRed documents [123]. The *cites* fielded query returns documents that cite the one specified in the query. In addition to this relationship, we also included (1) *cited*, (2) *co-cited*, and (3) *co-cites*.

We opted to skip most entity based and semantic relationships for our study. Those surveyed while designing our proof of concept system did not feel that people or organizations referenced in the document would be useful in the typical case nor whether a case received negative treatment (e.g. was vacated or overruled) as the original opinion may have addressed multiple issues, some of which may still be considered good law. Additionally, these relationship types are not straight forward to identify. Thus we did not explore these relationship types.

Since CourtListener already identifies the court system for each document, we opted to use this relationship, even though those surveyed did not think it would be of much use. The Judge metadata information was made available after our proof of concept system was

implemented and before the beginning of our study, so we allowed it to remain an option for advanced queries, but did not turn this field into a relationship.

The next step was to identify support snippets for relationships. The *Court* relationship does not lend itself to support snippets, however *Cited* and *Cites* clearly do. For both these relationships types the support snippet is the text where a document is being cited. Since the citations within legal documents follow a specific format, we were able to easily identify the location of each reference and the surrounding blocks of text in the document. Most cited decisions were referenced only once per document. Of a random sample of roughly 21,000 documents, each cited decision was referenced 1.11 times in the text of the document and 91% of the cited decisions are referenced only once. Thus for the majority of citations there is only one possible support snippet. When multiple references to the same decision are present in a single document, CourtListener creates a hyperlink from the first reference to the document of the decision. We follow suit, using the block of text around the first reference as the support snippet in cases where the citation is referenced more than once. For the *co-cited* and *co-cites* relationship where two documents are either being cited by, or are citing the same third document, we extract two support snippets; one for each citation.

7.1.4 Modifying the User Interface

Our interface is similar to the one outlined in Chapter 6.2 with a few minor modifications to support the study.

First, we provide a system tutorial that explains the star system and relationships to the participants. Users are automatically shown this tutorial upon registering with our system, and can revisit the tutorial at any time. Only the relevant parts of the tutorial were shown to

the user. For example, users in our *NoRelations* group did not see the parts of the tutorial that relate to system stars or relationships.

Second, we wrap the ARTEMIS search page into a two frame display. The left frame is the ARTEMIS system as previously described in Chapter 6. The right frame provided a textarea for participants to complete their summaries while engaging our system.

Third, we provide a system time clock. Each session with our system required users to spend a minimal amount of time searching. Once they spent sufficient time, the system would allow them to take the end of session survey. Time spent reading the tutorial and task description does not count as time spent on their task. Therefore we provided a system clock to help participants keep track of remaining time required for credit in our study.

7.2 User Study Design and Setup

The purpose of this study is to determine how the ARTEMIS framework can help a layperson pick up a difficult task, such as understanding legal documents. We turned to Mechanical Turk (mTurk)[3] to conduct this study as it has a large and diverse pool of potential study participants from across the US. We have previously noted several limitations with Mechanical Turk in Section 3.3.3. First and foremost is the satisfying problem [82], whereby a user selects the first satisfactory response on a survey question rather than the best response. A closely related problem is that users may have only minimal engagement in the task to get the payout. We will address these issues below.

[3]https://www.mturk.com/mturk/welcome

7.2.1 Study Methodology

Our goal is to aid in on-going search broken up over time, thus we conduct our study over 3 thirty minute sessions per participant. We acknowledge that this is an artificial session boundary, but serves our purpose of focusing users to abandon and pick up their task at a later date.

Each participant was assigned one of the six tasks described in Table 7.2. While researching their task they were instructed to provide a list of up to the 24 most important decisions. For each document they cited, they were asked to write a 1 to 3 sentence summary as to the importance of the case or why they were citing it. The purpose of these summaries is to explore how well the participants understood the documents they were citing. Thus we evaluate our system using 2 of the 5 criteria outlined by White and Roth, (1) task success and (2) learning and cognition [167]. At the conclusion of each session they filled out a brief survey about their task.

As we will discuss, our study had a large drop out rate. Therefore, our experiment was divided up into two Mechanical Turk "Hits" or jobs. The smaller, less intimidating initial phase instructed the participants to complete one session. The second phase, offered only after the first one was completed, included two more sessions. This is known in social psychology as the foot-in-the-door technique [51]. Participants who agree to do something small (i.e. complete a single session), are more inclined to follow through with further requests (i.e. complete two more sessions). Additionally, this gives us a chance to weed out participants who are minimally engaged in their task and our study.

	Brief Summary of Tasks
1	Identify the most relevant copyright and trademark litigation decisions since the date of the Supreme Court's *Bilski v. Kappos* decision
2	Identify the most important asylum decisions in the Ninth Circuit since June 1, 2010.
3	Identify the most important FOIA (Freedom of Information Act) decisions in the Ninth Circuit since June 1, 2010.
4	Identify the most relevant appellate decisions for secondary liability of an online service provider for trademark infringement since June 1, 2010
5	Identify the most relevant appellatelevel precedence for a case involving allegations of price-fixing, bidrigging, and geographic allocation of the market.
6	Identify the most important appellate-level developments in copyright law.

Table 7.2: A brief summary of tasks given to participants. Participants were given scenarios in which they were acting at the behest of a senior lawyer. The senior lawyer needs to review relevant case law for a given topic. He requests the participant provide the 24 most important decisions and to explain the relevance of each case or its key holdings in at most three sentences to aid in his review. Full descriptions can be read in Appendix Section B.

Task and Group Assignment

Time of day may be correlated with worker ability. A number of factors including type of employment and social responsibilities could effect available times for different demographics. Retirees may prefer to participate in the middle of the day, those in college may prefer the times after classes and those with young families may find time in the evening after children have gone to bed. In order to ensure both sufficient randomness and roughly equal participant group sizes, we assigned each participant a group and task pseudo randomly. As each participant signs up, the system assigns him to a random participant group with the smallest size. If, for example, group *NoStars* and *NoRelations* have 3 members each, but *PastClicks*, *NoSystemStars* and *RandPastResults* only have 2, the system will randomly assign the next person to be in either the *PastClicks*, *NoSystemStars* or *RandPastResults* group. Once a person is given a group they are randomly assigned a task outlined in Table 7.2, utilizing the same strategy in order to keep the number of participants assigned to each task within each group roughly equal.

One challenge encountered during our Mechanical Turk study was the large dropout rate. Typically mTurk users would create an account with CourtListener first, before accepting the mTurk hit. Of those that abandoned the study, 53% never made it past the tutorial. The other 47%, however, issued queries and conducted some searches before abandoning the study. These abandoned accounts needed to be removed from consideration in order to continue to ensure roughly equal task and group sizes. The system, however, has no way of predicting which one of these accounts would be abandoned when 2 or more participants sign up concurrently. Lack of user activity could indicate the user had abandoned the study, or the user was busy reading a document and not interacting with the system. To give us time to remove abandoned accounts we offered our Task in small 10 batch hits. When the batch was full (after an average of 11 hours, 55 minutes), and the abandoned accounts had been abandoned for sufficiently long, the

abandoned accounts were removed from the databased and the next request for participation issued. We note there is still possible to have unequal group/task sizes as some participants continue to drop out over the course of the study.

Initial Phase of the Study (Session 1)

For the first hit, participants were asked to complete one session with our system. We offered users $2 for completing the 30 minutes with the search engine, submitting a list of citations with short summaries and taking the survey with a possible $3 bonus based on the quality of work. The goal of the $3 bonus was to incentivize the Mechanical Turk Participants (colloquially referred to as 'turkers') to work hard, thus helping alleviate the satisficing problem. The turkers were unaware of possible follow on sessions at this initial stage.

Upon signup the turkers were first shown a tutorial of the system. The tutorial described CourtListener, then the star system and relationship system if applicable. Next the tutorial showed how to submit a summary and complete the end of session survey. The tutorial gave no instructions on how to read or interoperate the legal documents, however we did give examples of good summaries for hypothetical documents to help guide the user toward the types of summaries we were after.

Turkers were asked to supply one to three sentence summary for each citations as to the relevance of the document to their task. Additionally turkers were also asked to cite a document by it's full name or URL. In a few instances turkers failed to follow these directions. Therefore, at the end of the initial phase we conducted a quick blind evaluation of each turker's summary. In order to pass this evaluation the turker must cite at least one document with a summary.[4] The evaluator was only aware of the turker's system ID, not the task or group the turker was assigned. We did not evaluate the correctness of the citation nor the summary, just

[4]As a point of reference, the average citation count for the first session was 5, with a standard deviation of 3.

that the turker was following directions. This evaluation is a form of instructional manipulation check [109], which is a task designed to weed out participants who are not fully engaged enough in the study to follow directions. If a participant is not paying enough attention to follow study directions, he or she may also be minimally engaged in the task and his or her results may be noisy and incomplete. Six participants were removed from our study after this evaluation, 4 for failing to adequately cite documents and 2 for failing to provide summaries. Turkers who passed this evaluation were awarded the $3 bonus, and invited to the second phase.

Second Phase and Completion of the Study (Sessions 2 and 3)

For the second phase of our study, turkers were instructed to complete two more sessions over the course of a week. As before, each session included 30 minutes of searching, a writeup with summaries and small end of session survey. During this phase the system would prevent a user's return for at least 12 hours from the time of the last completed session, ensuring a minimal time interval between tasks. The stated payout for this second phase was $5, with a stated potential bonus of $7 based on work quality.

The participants were not required to view the tutorial prior to the second or third sessions, however it was available should a participant wish to view it.

Creating a Gold Set of Citations for Evaluation

In order to test the effectiveness of our system we needed to derive a gold standard of citations. To do this we asked a group of experts to identify the relevant decisions. The experts were law students at Berkeley Law School, and ranged in experience from first year law students to those who had just completed their degree. The experts were first asked to conduct their

own searches to find said relevant documents. They then convened in small groups to decided on gold standard set to be used.

It is important to note that the inter-annotator agreement amongst the law students was very low. For two of the tasks, tasks 5 and 6, we had two separate groups of annotators for each task. The inner annotator agreement for task 5 was 1 cited document in common out of 19 (for an overlap of just 5%) and for task 6 it was 3 documents in common out of 14 in total (21% overlap.) The number of documents cited per group were roughly equal, but the groups typically cited different documents for different tasks even after three hours of searching. The point of discrepancy appears to be the phrase 'most important'. It was not uncommon for two students to agree that a document had some relevance but disagree on the significant points of a document or the degree of importance. Therefore while we use the union of the law student's findings to represent our gold standard for standard evaluation measures of precision and recall, it is not the only evaluation criteria we use.

7.2.2 Participants

As mentioned earlier there was a high self selection drop out rate in the initial phase to the user. Of the 112 participants who signed up for an account with CourtListener, only 36 completed a session for a retention rate of only 32%. Of those 36, 28 went on to complete the second phase for a total retention rate of 25%.

The participants that completed all three sessions ranged in age from 21 to 61 years old, with a mean average age of 36 and a standard deviation of 11. There was a roughly equal gender split with 12 participants self identified as male, 16 female.

Since the retention rate of our study was so low, and the participants were self selecting in their willingness to accept the initial phase as well as and follow through with the second

phase, it is worth noting their level of expertise and interest in the legal domain. Therefore, during the second stage of our task we asked users about their background in the legal domain. We order the responses in our survey from greater amounts of experience to less, thus any bias introduced by the satisficing problem will lead to an over estimation of the amount of experience rather than an underestimation in the user's level of law experience. Multiple responses were permitted. The results are shown in Figure 7.1. Five Participants had some legal training but no degree (for example, may have worked in ligation support gathering documents), 10 read legal commentary (e.g. Harvard Law Review), only 7 have read precedential or non-precedential opinions before, 10 read or watched non-fiction legal TV (e.g. court TV), 22 read or watch fictional legal dramas (e.g. Law and Order) although we note such shows quite popular, and 4 reported no prior interest in the legal domain. No participant had a law degree or was licensed to practice law. Thus while our participants are self selecting, and appear to have some interest in the law they still are not legal experts.

7.3 Results

In this section we discuss our findings of the ARTEMIS system in our user study. We find participants used relationships frequently, both to predict document relevance and to find the relevant information within documents. Participants demonstrated a better understanding of the documents they cited, and were better able to identify the relevant documents.

Due to the sample sample size of this user study, we use ($p < 0.01$), using a two sided tests, as our threshold for statistical significance in this chapter. When computing the statistical significance using the t-test (described in detail below) we need to also specify the degrees of freedom, df. The degrees of freedom is the number of variables in the experiment that are free to vary. In general, the more degrees of freedom present in the system, the less likely any difference

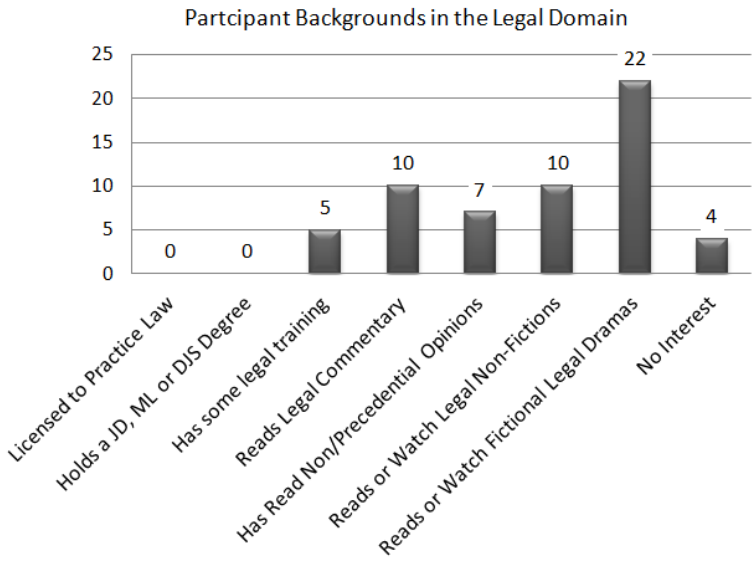


Figure 7.1: The self reported level of legal experience of the participants. Participants were instructed to select legal training only if they had training in how to read legal documents but no degree. Reading legal commentary includes sites like Harvard Law Review

between two sample means is due to random chance. A typical value for the *df* is the number of events observed, but this assumes each event is independent of each other. While users are likely independent of each other, the queries issued by a single user in a task are likely somewhat related. Similarly the result clicks of a single user are not independent. Thus we use number of users in the current analysis when determining the degrees of freedom.

In this analysis we sometimes observe differences between two variables that are not statistically significant. We note lack of statistically significant difference does not imply the two variables are the same, just that our data sample may be too small to detect the difference between them, and that any difference is likely not large.

Unless otherwise noted, all differences are significant.

	Citation Rate	Number of Examples
Clicked Documents	55%	364
Re-Found Documents	61%	53
User Starred Documents	77%	178
PastClicks System Starred Documents	70%	25
RandPastResults System Starred Documents	21%	71

Table 7.3: The probability of a each document type being cited. Note, the difference between User Starred Documents, and *PastStars* system starred documents are not statistically significant.

7.3.1 Evaluating our Hypothesizes and Design Decisions

In this section we explore the validity of our hypothesizes and design decisions from earlier chapters.

Hypothesis: Both system and user starred documents are typically the more relevant than non-starred documents.

We previously hypothesized that starred documents are typically the most relevant documents. In order to evaluate this hypothesis we treat the citation in the user summary as a indicator of relevance. The citation rates are given in Table 7.3. Documents with a click were 55% likely to be cited, documents with more than one click (and thus re-findings) were 61% likely to be cited, and documents starred by the user were 77% likely to be cited. This difference is statistically significant ($p < 0.01$) according to the z-test for population proportions. The z-test for population proportions, defined by equation 7.1, calculates the probability that two

	Ave Time (s)	Stdev Time (s)	Number of Documents
User Starred Documents	244	298	111
System Starred Documents	218	281	103
Not Starred Documents	99	112	25

Table 7.4: The average time spent reading the clicked decisions. Differences between time spent reading user starred and system starred documents are not significant.

populations differ significantly in one aspect (i.e. the citation rate). In this equation p_i is the sample population proportion, n_i is the sample size for sample i, and p is the weighted average between two populations, $p = (p_1 * n_1 + p_2 * n_2)/(n_1 + n_2)$. There was no statistical significance between the citation rate of user starred documents in each of the experiment groups, or the *NoExplanations* control group.

$$z - score = \frac{(p_1 - p_2) - 0}{\sqrt{p * (1 - p) * [(1/n_1) + (1/n_2)]}} \tag{7.1}$$

Since documents with user stars are more likely to be cited than documents that are clicked, it stands to reason that starred documents might be more relevant than clicked documents to the underlying task.

Hypothesis: Users may be more familiar with starred documents

ARTEMIS is built around the notion that there exists a subset of documents in the user's search history that the user may be more familiar with, and that by providing relationships between new documents and these familiar documents, a user may be able to better understand

the relevance of these new documents. To evaluate the first hypothesis, we explore how long users interact with each document. Our assumption is that the more time a user spends reading the document, the more familiar he or she is with it. We calculated the total time on the page for all documents the user clicks on, thus if the user spends 30 seconds on the page, then spends another minute on the page after re-finding, the total time on page is 90 seconds. Results are shown in Table 7.4.

Users spend an average of 4:04 minutes reading user starred documents, and 1:38 minutes on non starred documents. The difference in time spent on page between starred documents and unstarred documents are significant using the student's t-test for significance (defined by equation 7.2) with a $p < 0.01$, and 13 degrees of freedom ($df = 13$). This finding supports our hypothesis that users may be more familiar with starred documents than unstarred ones.

$$t = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{s_1^2}{n_1} + s_2^2 n_2}} \tag{7.2}$$

We have previously shown starred documents are highly correlated with cited documents, and thus we might expect users to spend more time with cited documents than non cited. However, we also find that users spend almost as long on system starred documents as starred documents, even under the star strategy *RandPastClicks*, which has a much lower citation rate. The difference between time spent on page of system starred and user starred documents is not significant according to the t-test, even under system star strategy *RandPastClicks*.

Adding further support to this hypothesis, we find unstarred documents are much more rare in the search results multiple queries issued by the user (occurring in the result set for an average of 3 queries) than documents with system stars under *PastClicks* (29 queries) or

user stars (47 queries queries). Thus there is more opportunity to become familiar with these documents.

Hypothesis: The list of starred documents will converge.

We hypothesized that if system stars had little influence over users' perception of relevance in the document, and the user is actively engaged in the star system, the final list of starred documents will be similar across the different experiment groups. In actuality users rarely unstar a system starred document. On average users starred 12.2 documents and unstarred 1.9. Users who were shown system stars starred an average of 3.1 out of 19.8 system starred documents.

Interestingly, there is a large overlap between the *PastClicks* star strategy which includes all past findings receiving system stars and the *RandPastResults* star strategy which includes random past search results receiving system stars. There were 25 unique documents starred by the system under strategy *PastClicks*, and 71 unique documents under strategy *RandPastResults* with an overlap of 19 documents. This may be because documents starred by the system in *PastClicks* occur more frequently in the search results than unstarred results (for an average of 29 queries vs 3). Therefore, the random strategy based on frequency in the search results is more likely to assign a system star to a document that would have been starred under *PastClicks*, then ones that would not.

Hypothesis: User starred documents are more important to the task than system starred documents.

Next we hypothesized documents with user stars are more important to the user's underlying task than documents with system stars. As before, we use citation as a proxy for importance. The validity of this assumption depends on the system star strategy.

As we see from Table 7.3, documents starred under *PastClicks*, the star strategy where any document previously clicked is starred by the system, had a 70% citation rate. This is not statistically significant from the citation rate of user starred documents[5], 77%, and shows this simple strategy may be reasonable. On the other hand, documents starred under *RandomPastResults*, the star strategy were random prior search results were starred, had a 21% citation rate of system starred documents. As a point of comparison, the overall citation rate was just 8%. Thus documents starred by *RandomPastResults* were statically significantly more likely to be cited than unstarred documents, but not as likely to be cited as user stars. We note it is possible that the presence of the system star is biasing the user's perception of relevance, and the user is more incline to cite the document simply because it has a system star. In our surveys, 4 of the 13 members in the experiment group said they used system stars to decide which documents to read.

Hypothesis: Most queries will have both starred documents and unstarred documents.

If starred documents represent a core set of documents relating to the user's underlying information need, it stands to reason that some subset of them will appear in the search

[5]Since system stars appear below user stars in the result list, it is possible their click rate, and therefore their citation rate is affected by the click position bias. The click position bias is a bias towards favoring results closer to the top of the search results. We find, however, that there is still no statistical significance between the citation rate of user starred and system starred by the *PastClicks* star strategy even at the same ranks.

	Probability of a Search Result Set Having		
	Unstarred Documents	Starred Documents	Both
1st Session	99%	62%	61%
2nd Session	100%	80%	80%
3rd Session	98%	88%	88%
Overall	96.7%	72.1%	71.2%

(a) Considering either user starred documents or system starred documents as 'starred' documents

	Probability of a Search Result Set Having		
	Unstarred Documents	Starred Documents	Both
1st Session	99%	46%	46%
2nd Session	100%	68%	68%
3rd Session	98%	72%	71%
Overall	96.7%	59.8%	59.6%

(b) Considering only user starred documents as 'starred' documents. (System starred documents ignored.)

Table 7.5: The probability of the search result set containing documents of a given type. Differences between the percentages of search results that contain starred results, and those that contain starred and unstarred results are not significant.

results as the user explores different aspects of his or her information need. As noted above, unstarred documents tend to occur in fewer query result sets (3) than documents with system stars under *PastClicks* (29) or user stars (47). Table 7.5 shows the probability of a search result set containing both starred and unstarred documents. On average we expect 71% of search results sets to have both an unstarred document and either a system starred or a user starred document. Additionally, 60% of all search result sets have an unstarred document and at least one user starred document. As expected, the more time a user has with the system, the greater the likelihood of both document types in the search results. This result is significant according to the z-test, ($p < 0.01$).

Hypothesis: There exists a correlations between re-finding, re-search and user stars.

We find for users in our experiment groups, re-finding and re-searching are more likely to lead to user starred documents. The search results for non-substantial change re-search queries typically have more stars than search results for non re-search queries (2.7 vs 0.9). Additionally re-findings are more likely to be starred than new-findings (90% vs 60%).

Exploring our design decisions.

In Chapter 6 we described our implementation with two side panels (Recent Stars, and Relationships Overview) and two views, the complete stars list and the complete relationship list. The decision decisions in terms of the number of elements to display in the Recent Stars and Relationships Overview panels were ad hoc, motivated by our pilot studies. In this section we explore the effectiveness of these two panels.

We found that users used both the Recent Stars panel and the Stars View. Participants viewed the full list of stars in the Stars View an average of 5.2 times. They re-found documents

from that page an average of 1.9 times, or roughly 40% of the time they viewed the stars page. On the other hand, the re-found from the recent stars menu 2.4 times. Re-finding was the most common action when viewing the Stars View. Participants rarely unstarred or starred documents from this view.

The cut-off of 3 starred results in the recent stars menu seems reasonable. The average rank of the re-findings via the full stars list view was 9.3, with 44% of re-findings occurring at or below the tenth rank. Of the 25 re-findings via the full stars list view, 6 (24%) were in the top 3 and thus also in the recent stars menu.

Participants were less inclined to use the full relationship view, however, they interacted with the relationships overview menu. Participants clicked on documents in the relationships overview menu 2.3 times. 4 users starred an average of 2.2 documents in the Relationships Overview menu. This is interesting, as the Relationship Overview menu displays only the title of the document. A user would have to decide that a document is worth starring based on the type of relationship, number of relationships, or a combination there of. We find that users rarely expand the relationships into the full relationship view (an average of 1.3 times per user.)

7.3.2 Increased Understanding of Relevant Documents

One goal of ARTEMIS is to aid in understanding of the documents. There are multiple ways a relationship based system could help aid the user on their information journey. Most intuitively, the relationship can explain why a document is relevant. Even if the relationship itself is not sufficient for the user to understand a document's relevance, it may provide a starting point for the user to explore the matter more fully, for example by indicating what passage in the document may be relevant, or giving a user background to understand the potential relevance.

Sample User Summaries for Task 1
The Supreme court ruled that an invention that departs from the prior art only in its use of a mathematical algorithm is patent-eligible only if the implementation is novel and non-obvious. The algorithm itself must be considered as if it were part of the prior art. This case is relevant because it is the second member of a trilogy of Supreme Court decisions on the patent-eligibility of computer software related inventions.
This case is Precedential and cites Bilski. It also has relationships to a bunch of other cases I have starred
"Notably, the district court, following our precedent in In re Bilski 545 F .3d 943 (Fed. Cir. 2008), relied solely on the machine-or-transformation test, although not the exclusive test for patentability, is 'a useful and important clue.' Bilski v. Kappos, 130 S.Ct. 3218, 3227 (2010)"
This case is very relevant because this case is one of the judicially refining software patent eligibility cases that has come after the Bilski decision. After deciding the Bilski case, the Federal Circuit Court has not created a new test, or expanded the machine-or-transformation test. This case has shed some light on patent-eligibility; specifically related to the broad treatment of the Internet as a machine.

Table 7.6: An few summaries collected from users for various cited documents for task 1. The first example shows a clear understanding of the decision and it's importance, whereas the second example is clearly not useful, nor does it show any understanding of the document. The third example is a direct quote from document, a quote that was also present in the support snippet, but was not part of the original snippet.

To explore this goal of understanding we turn to the citation summaries. Table 7.6 shows four example summaries from our participants for task 1. The first summary shows a clear understanding of the document and it's importance. The second example illustrates an over reliance on the relationship system; the participant cited the document because it had relationships with previously starred documents. In our user post-session surveys, 4 of the 13 participants in the experiment group reported relying on the presence and/or number of relationships to gain a sense of whether the document was relevance, which matches what we observed with the user click behavior in Section 7.3.1. Perhaps most interestingly is the third example where the user's summary of the document is a direct quote. The document, *Fort Properties, Inc. v. American Master Lease LLC, 2009-1242* has cited the document *In Re Bilski, 2007-1130*, a document central to the underlying task and previously starred by the user. The quote the user selected is embedded in the support snippet shown for the citation relationship.

To get a better feel for the users' ability to interpret the legal documents, we examined the summaries according to three subjective criteria:

- *Useful*: Would the summary make it easier or faster for the senior partner to find the relevant information, or allow the senior partner to predict the relevance of the document prior to reading it.

- *Comprehension*: Does the participant appeared to comprehend the significance of the document, and it's key findings.

- *Sufficient*: Would the summary stand on it's own, or would the senior partner need to read the document for clarification. In other words, does this summary satisfy the partner's request.

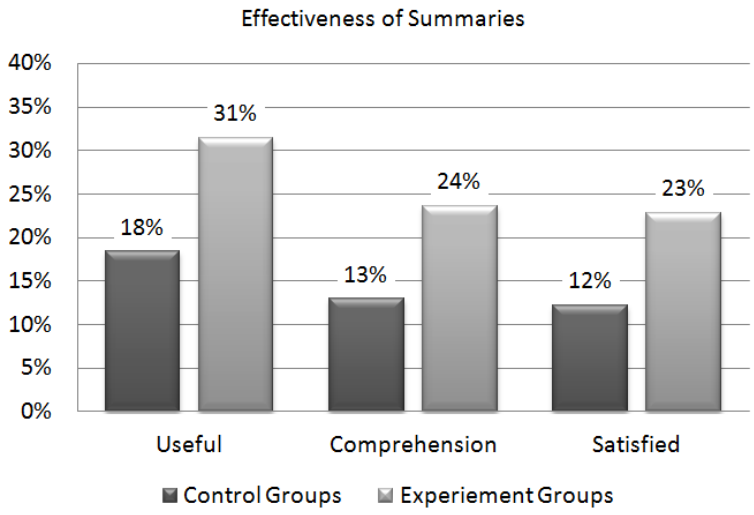


Figure 7.2: The percentage of times our evaluator agreed that the summary was (1) useful, (2) showed comprehension or (3) was sufficient.

These three criteria are similar, but not identical. For example, the third summary in Table 7.6 is a direct quote from the document. The partner might gain enough information from a quote to comprehended the significance of the document, but the quote does not necessary show that the participant comprehended the document. On the other hand, a summary might show comprehension, but for a document that is not relevant for the task and therefore not useful. A summary that explains the subject matter, without going into the significance of the court's ruling or how it affects the task, like the forth example, might show comprehension and be useful, but would not be sufficient.

We asked a legal expert to evaluate the summaries based on the three criteria above. The lawyer was asked to envision herself as the senior partner in our sceneries and was randomly presented with each summary and gave a yes/no response for each criteria. The proportions of 'yes' responses for the summaries are shown in Figure 7.2.

We observe that participants in the three experimental groups with star system strategies show greater understanding than those in the control groups. Summaries were deemed more useful, showed better comprehension and were sufficient more often than those in the control group a statistically significant percentage of the time according to the z-statistic, ($p < 0.01$). Performance in each of the experiment groups, however, did not differ in a statistically significant manner. This suggests that while some star strategies might be better at predicting which documents will be cited, the specific star strategy has less impact on the overall understanding of the document.

7.3.3 Improved Targeting of Relevant Information

Next we evaluated performance with recall (equation 7.3), precision (equation 7.4) and f-measure (equation 7.5) to see how well participants are able to target relevant documents. Simply put, recall is a metric of how many of the relevant documents the user was able to find, and precision is the metric of how often the documents that the user found were relevant. F-measure is the harmonic mean of recall and precision and is a balanced measure of these two metrics which keeps one from dominating the other. We show our results in Table 7.7.

$$r = \frac{tp}{tp + fn} \qquad (7.3) \qquad p = \frac{tp}{tp + fp} \qquad (7.4) \qquad f = \frac{2(p * r)}{p + r} \qquad (7.5)$$

The ability to star documents appears to help users identify relevant documents. From Table 7.7 we see that group *NoStars* had a statistically significantly worse f-measure than all other groups. On the other hand, control group *NoRelations*, which had stars and no relationships had similar performance as the three experimental groups that was not statistically significant. The ability to see relationships does not appear to significantly influence precision

Participant Group	Precision	Recall	F-Measure
NoStars	0.27	0.15	0.19
NoRelations	0.69	0.25	0.35
PastClicks	0.56	0.22	0.31
NoSystemStars	0.77	0.25	0.36
RandPastResults	0.62	0.25	0.34

Table 7.7: The Precision, recall and f-measure scores for the given participation groups. The *NoStars* participation group preformed statistically significantly worse than all other groups. The differences in performance for the control group *NoStars* was statistically significantly worse than the other groups. There was no statistically significant difference between the other four groups, ($p > 0.01$).

and recall, however, have other benefits as we have discussed. Interestingly, group *NoRelations* had slightly worse useful, understanding and sufficient scores than group *NoStars*.

There is also evidence to suggest that participants using ARTEMIS are able to target the relevant information within documents more effectively. When asked how they used our system, four participants indicated they used the relationships to gain an overall feel for the documents. One said he used relationships "to preview case content." Another participant reported searching for the text from the support snippet inside the document to locate the citation faster. While the participant who gave the summary that matched the support snippet in Table 7.6, we find 5 examples where a document was cited without the user having ever clicked on the document.

7.3.4 Increased Ability to Re-Find

One of our stated goals of ARTEMIS was to support re-finding. We theorized that re-finding would be supported by the star system alone. Surprisingly, we find that the addition of the relationships along with stars boosts the re-finding rate, whereas the stars alone do not. Users in our three experiment groups had an average number of 11 re-findings, nearly double the rate of both the control groups. Users without relationships had 5, and those without stars or relationships had 6, however this difference is not statistically significant, ($p > 0.01$).

Perhaps counter intuitively given the nature of our task, we also find that users continue to have interest in documents even after they had been cited. Most documents were starred in the same session or the session before they were cited, however 7% of starred documents were starred in a session following the session they were cited. One possible explanation as to why a user would star a document after it has been cited is to increase the number of explanations.

7.3.5 Effectiveness and Robustness of Relationships

Unstarred results had an average 22 relationships (standard deviation of 33) with starred documents. The large number of relationships is due in part to the duplicate versions of the same documents, which increased the number of *cocites* and *cocited*. Of the unstarred results, 76.2% had relationships. 52.3% of Unstarred documents had less than 10 relationships and 34.8% had less than 5. The most common relationship type was *court* (34%) followed by *cocites* (30%), *cocited* (24%), *cited* (6%) and *cites* (6%).

The probability of an unstarred search result with explanations being clicked (3.3%) was not statistically significant from the probability of an unstarred result being clicked (3.2%), however *CoCited* and *Cited* had a statistically significantly lower probability of being clicked (2.5%, and 1.5% respectively.) This may be due to the construction of our task where partic-

ipants were asked to find recent information. The *Cited* relationship appears when a search result has been cited by a starred document. In order for this to happen, the search result must be older than the starred document.

As might be expected, users reported preferring the citation relationships to the court relationship. Users overwhelming reported wanting to see more of the *CoCites* relationship, very few wanted to see more of *Court* relationship, however, users also did not request less of that relationship. Documents with the *Court* relationship also had an increased click rate, however this could also be an artificial of the task design as some tasks required users to cite documents in certain circuits. Interestingly, even though *CoCited* and *Cited* were negatively correlated with citing a document, 60% of the participants reported wanting more of this type of relationship.

We find that while users tend not unstar system starred documents, erroneous system starred documents do not appear to negatively impact the user's ability to understand a document. Documents cited by users in the experiment groups *RandomPastResults* and *NoSystemStars* had comparable Understanding, Useful and Sufficient rates as those in the experimental group *PastClicks*. These scores are not statistically different, even though documents starred by the system in the former group was far less likely to be cited. All three experiment groups had statistically significant improved Understanding, Useful and Sufficient scores then the control groups as shown in Section 7.3.2. A likely explanation is that non-cited starred documents have fewer relationships with new documents (55) than cited starred documents (134), thus they do not have a large effect future search results.

7.4 Summary and Contribution

In this chapter we created a proof of concept implementation of our ARTEMIS framework for a legal document search system built on top of the existing CourtListener System.

Our proof of concept implementation was designed to test our hypothesis that there are important documents central to the user's underlying tasks (i.e. starred documents), and that finding relationships between these documents and other documents (i.e. unstarred documents) might aid both in understanding and the finding of relevant documents. Key findings in this chapter include (1) users with ARTEMIS show a better understanding of the documents they cite; (2) Users with the star system are better able to identify relevant documents (diversification seeking); (3) ARTEMIS aids in re-finding (repetition seeking); and (4) the effectiveness of relationships in ARTEMIS may be somewhat robust to erroneously starred documents.

In this chapter we also validated the hypotheses behind ARTEMIS. We had previously theorized that there exists a subset of documents most relevant to the underlying task and that the user would be more familiar with in the user's search history. We found users had a higher citation rate of starred documents than clicked documents, and tended to spend more time reading clicked starred documents than clicked unstarred documents. We theorized that a subset of starred documents would likely continue to be returned with each additional query, and found this to be true for roughly 70% of queries in the follow on sessions. While it is likely that more complex starring strategies will further improve understanding, we showed simplistic star strategies can be an effective at uncovering relationships that aid in understanding and that our approach is valid.

7.4.1 Implications for System Star Strategies

The results of or user study suggest that users place some level of trust in system starred documents. While they didn't cite system starred documents with *RandomPastResults* at the same rate as user starred documents, or system starred documents with *PastClicks*, they spent a comparable amount of time reading them. They also rarely unstar system documents.

Our study also suggests simple strategies for starring documents might be sufficient. We observed that system stars under the *PastClicks* strategy have a similar citation rate as the user stars, suggesting simple strategies in predicting which documents to star may be sufficient. We found that 71% of user starred documents would be starred under the past clicks strategy and only 4% of system starred documents are unstarred. This strategy may be effective even for users who do not actively star documents themselves.

7.4.2 Implications for Relationships

While users were willing to explore the full stars list displayed in the full star list view, participants very rarely explored the full relationship list. They appeared to prefer to interact with the relationship overview menu, and the relationships below the documents themselves. Thus the choice of which relationships to display, and how many may have a significant impact on performance.

We also saw some evidence that users may be considering the presence of relationships, and not the details of the relationships when judging documents. In our end of session surveys almost 40% of the participants reported wanting more *Cited* or *CoCited* relationships, even though these were negativity correlated with click rates. It might be that users have difficulty discerning the difference between the *Cited* and the *Citing* relationship. Therefore we may want to limit relationships displayed to be the highest valued ones.

Chapter 8

Conclusion & Future Work

In this chapter we summarize the contributions of our dissertation, as well as give future directions for this work.

8.1 Contributions of this Dissertation

In this dissertation we addressed the dual need to return to valuable content as well as discover new content. We first performed extensive analysis of prior search logs, and then proposed a new framework for explanations of search results akin to justifications in recommender system. We then proved the validity of our framework with a proof of concept implementation and user study.

In Chapter 4 we studied repetitive queries called re-searches. We discovered a dual need to return to valuable content and discover new information when issuing the same query, even for queries typically thought of as navigational. Within the re-search query trails there are often URLs (i.e. *dominant URLs*) that are re-found frequently, many accounting for more than

half the repeat clicks in the trail. Even in the presence of these frequent re-finding, however, users continue to seek out new information and new URLs. These new URLs, in particular, seem to be more useful in predicting what a user is likely to click, than the dominant URL, and may give greater insight into the user's underlying interest in the query. We also found evidence that re-finding may often lead to new-finding.

In Chapter 5 we explored repeat findings, or re-findings. We discovered that users are more likely to re-find at the beginning and end of a session. Cross session re-findings typically exhibit similar click trails indicating they may be picking up the same task, where intra session re-findings could indicate a renewed interest in the re-found URL. We find that users often settle on a query when re-finding and that this query tends to be 'better' in terms of being shorter, and the rank of the re-finding result tends to be higher. Thus re-finding and re-searching are tightly coupled behaviors.

In Chapter 6 we defined the ARTEMIS framework which can help users both return to valuable content (re-find) while simultaneously aid in the discovery of new information. ARTEMIS works by identifying repetitive interest URLs, which are typically the most relevant to the underlying task, called repetitive interest URLs. ARTEMIS then identifies relationships between these repetitive interests URLs and new URLs, presenting them to the user akin to a justification approach in recommendation systems. These relationships can help a user understand why a document that does not appear relevant may actually be so.

In Chapter 7 we evaluated a proof of concept implementation with a user study on a task in the legal domain. We implemented the ARTEMIS framework as tied to a legal search engine and conducted a user study with lay people. We found even naive system star strategies were effective in aiding users to find relevant documents, and better able to understand found documents. Further, we found the relationships with erroneously (random) starred documents

had little to no negative impact on understanding. Participants used relationships both to identify relevant parts within the document, and to identify which documents may be more relevant. While we don't expect the results to extend to all of web search, the results are promising for this important and common subset of tasks, where non-experts preform complex information search tasks such as in the legal, medical and scholarly domains.

8.2 Possible Extensions of This Work

Our user study showed ARTEMIS was effective in the legal domain for a single structured task. In this section we discuss several avenues for expanding ARTEMIS.

8.2.1 The Role of Serendipity and Task

In our user study the participants were engaged in a single task. In reality users are likely to have multiple tasks, multiple topics of interest, and often even engage in them simultaneously within the same session [72]. Researchers have studied methods for identifying queries relating to cross session tasks [72, 81, 5, 157]. Partitioning queries by task is a straight forward way for implementing multi-task support for ARTEMIS.

The results from Chapter 7 suggest ARTEMIS may be somewhat robust of errors relating in task partitioning. Let's suppose the user is engaging in two tasks simultaneously, T_A and T_B. There are two ways that task T_A may interfere with the effectiveness of ARTEMIS on task T_B : effecting which documents are label repetitive interests (i.e. starred documents) and creating non-useful relationships.

In the case of two tasks that are completely disjoint and in different domains, it is unlikely there will be much overlap in the search results for each task. Thus a user starred document for task T_A is unlikely to appear in the search results for queries related to task T_B.

Thus the user starred documents for one task would likely not appear starred in another. If following a simple star strategy described above, either *PastClicks* or *RandPast*, the same principle would hold, and that documents for task T_B would be just as likely to be starred by the system regardless of the fact that the user was also engaged in T_A.

Erroneous, or less useful, relationships may occur when starred documents for task T_A have relationships with the documents in the search Results in task T_B. As we observed in Chapter 7, however, less relevant (not cited) documents tended to generate fewer relationships than relevant (cited) ones. Thus this would suggest there would be few relationships between task T_A and T_B. On the other hand, relationships that bridge tasks may be worth exploring. Discovery of new relationships between two, seemingly disjoint interests may lead to the discover of unexpected information and serendipity. The goal behind a faceted search, or clustering search results is to facilitate browsing of information. Unexpected relationships may be those most interesting to the user in an open task setting.

Further study is needed to test these hypotheses.

8.2.2 Effects on User Bias

Bias on the web has previously been observed by researchers. Ieong et al showed users exhibit a domain bias [68]. By changing the domain associated with the snippet, user's perception of relevance in the document was changed. White showed people seek evidence that confirms their beliefs, and as such show bias in search with regard to health [164]. White's study went a step further and suggested bias was present in the search engine results as well. Yue at al explored presentation bias by exploring result snippets and title attractiveness [172].

Relationships in ARTEMIS can be used to find documents that conform to the user's underlying bias, or may help a user over come it. Users may use relationships as a form of

faceted search and drill down into the clusters that interest them, or they may use relationships to reveal results they did not believe were relevant in fact actually are. Further study is needed to understand the effect of bias.

8.2.3 Account for Negative, Dissimilar and Compound Relationships

In this dissertation we focused on positive relationships, and the similar characteristics between two documents. Although in Table 6.2 we did include the possibility of a *contradicts* relation, the rest of the relationship types involve the commonalities between two documents. In some cases, however, it may be the differences that are important. Our cancer patient, for example, may be looking to consider a new treatment that he or she hasn't seen before.

While we generally expect users to use relationships as a form of support for documents, they can also be used to weed ones possibly irrelevant search results. We saw in Chapter 7 that some relationships are negatively correlated with citation rate. This may be because the relationships tend to exist with documents that are not relevant, or it may be that the relationship signals to the user that the document is not relevant to the user. In either case, filtering out documents with these relationships may be beneficial to the user. Many users in our re-searching user survey discussed in Chapter 4 reported looking to see how the result list had changed. Two mentioned specific instances of searching for their own name to find others with the same name. In these cases, filtering out known results (or results connected to the same person) might bring to light more results otherwise obscured.

Yet another possible relationship type is compound relationships. There may also be instances where the relationship itself is only interesting in the context of other relationships. Let us return to the literature example. Our user may be focusing on a narrow aspect of her literature review. Literature reviews typically include different documents in different domains

that may be relevant to different parts of a researcher's central thesis. The fact that a document has the same author as one set of cited documents, or that a paper was published in the same venue as another set might both indicate some base level of relevance. If the same document has such relationships to two subsets of cited documents within the researcher's literature review, however, it may signal the paper is bridging the two domains and is highly relevant to her task. Thus the presence of both relationships may indicate to the user that the document is worth reading. One benefit a faceted search approach has above ARTEMIS is the ease at which users can explore multiple facets simultaneously.

8.2.4 Drifting of Query Intent

We've previously noted that as the user explores the search space, his or her need may evolve. Therefore it is desirable to have a starring strategy that can accommodate the changing need.

Just as ARTEMIS appears somewhat robust to extraneous starred documents, it likely is somewhat flexible to changing needs. When the information need changes, documents that were previously high valued may no longer be so. As the information need changes, the user will likely issue queries related to the new need. Since the intent behind these queries is different, the result lists returned will likely differ. Thus previously starred documents may not appear in these results for new queries. Additionally, when designing ARTEMIS we use ordering relationships partial based on recent activity. Relationships under starred documents were ordered by the importance of the starred document according to the star strategy, and then by recency. Therefore, as the information need drifts, and new new documents are starred, relationships will include these new documents first.

A better approach might be to detect which starred documents are less relevant and down weight their importance. Our current approach selects only one relationship of each type for each unstarred document to display. We currently estimate the best relationship example to be the one with the most important starred document. Yet our strategy for deciding starred document's importance, and thus the relationship ordering, was somewhat ad hoc. The original intent behind starred documents was to identify a subset of documents from the user's history which are the most relevant and which the user is most familiar with. Topic modeling and topic drift techniques may be able to determine when and which previously starred documents are no longer as similar to the user's underlying interest, and therefore which relationships to down weight.

Appendix A

User Surveys

A.1 Re-Searching Intentions Survey

- What is your gender?
  - Male
  - Female
- What is your age?
- How often do you use a search engine?
  - Multiple times a day
  - Once a day
  - A couple of times a week
  - Once a week
  - Less frequently than once a week

- Which search engines do you use? Check all that apply.
  - Google
  - Bing
  - Yahoo
  - AOL
  - Other
- Think about the types of queries you use. How often do you use the same query you have used before?
  - Multiple times a day
  - Once a day
  - A couple of times a week
  - Once a week
  - Less frequently than once a week
- For a typical query that you have used before, what is your intention when issuing it again? Check all that apply
  - To find a website you have seen before. For example, my query is apple pie and I have found a recipe that I liked before which I want to reuse.
  - To find new websites you do not know about. For example, I didn't like the recipe I tried last time, and I want to find a new one.
  - To find new information, regardless of whether it is on a page you have seen before or not. An example might be to find out the score of a sports game.

  - To see how/if the search results have changed. For example, to see if any new websites are discussing an issue I care about.
  - Other (please explain in the comments below)

- Think about the most common queries you use. Do you ever have multiple intentions for the same query? If so, what are they?

  For example, I use the query Apple Pie on Monday and Tuesday. On Monday I want to find the same apple pie recipe I was using before (and revisit that webpage) and on Tuesday I want to see if there are other recipes I could try (and want to visit new pages as well.) Or perhaps on Monday I want to find the same apple pie recipe I was using before (and revisit that webpage) on Tuesday I want to find out how to make a pie crust, and dont care where I find that information (To find new information, regardless of whether it is on a page you have seen before or not) Check all that apply

  - To find a website you have seen before AND To find new websites you do not know about
  - To find a website you have seen before AND To find new information, regardless of whether it is on a page you have seen before or not
  - To find a website you have seen before AND To see how/if the search results have changed
  - To find new websites you do not know about AND To find new information, regardless of whether it is on a page you have seen before or not
  - To find new websites you do not know about AND To see how/if the search results have changed

  - To find new information, regardless of whether it is on a page you have seen before or not AND To see how/if the search results have changed
  - To find a website you have seen before AND To find new websites you do not know about AND To find new information, regardless of whether it is on a page you have seen before or not
  - To find a website you have seen before AND To find new websites you do not know about AND To see how/if the search results have changed
  - To find new websites you do not know about AND To find new information, regardless of whether it is on a page you have seen before or not AND To see how/if the search results have changed
  - To find a website you have seen before AND To find new websites you do not know about AND To find new information, regardless of whether it is on a page you have seen before or not AND To see how/if the search results have changed
  - None of These
  - Other (please explain in the comments below)

- When you cant find what you are looking for, what are the reasons? Check all that apply.
  - Too many search results irrelevant to your query
  - Too many search results that are relevant, just not useful for your goal
  - Takes too long to navigate through the search results
  - Not enough search results
  - Results are too basic for my expertise level. For example, I am an expert chef and the results all assume I can't boil water.

  - Results are too advanced for my expertise level. For example, I can't boil water and all the results assume I'm an expert chef.
  - Other (please specify in the comments below).

- Of the queries that you have used in the past, how many do you plan to use again in the future?
  - 0
  - 1
  - 2-5
  - 5-10
  - 10-20
  - 20+

- What queries do you typically reuse?
- Please provide any comments you may have

A.2 Proof of Concept End of Session Feedback Survey

Search Engine Quality

- How good were the search results, were they relevant to the query?

	1	2	3	4	5	
Few Results Were Relevant	○	○	○	○	○	Most Results Were Relevant

- Were you happy with the number results per query?

	1	2	3	4	5	
Too Few	○	○	○	○	○	Too Many

(a value of 3 indicates the number of results was just right).

- Was our system fast enough for you?

	1	2	3	4	5	
Fast/Usable	○	○	○	○	○	Too Slow/Not Usable

Accomplishing Your Task

- How confident were you in your summary?

  I.E. How ACCURATE do your think your summarization of the court cases you liste d were?

	1	2	3	4	5	
Not Confident	○	○	○	○	○	Very Confident

- How Confident are you that the decisions you cited were relevant?

  In other words, how confident are you that the partner in our hypothetical scenario would agree that your summary contained only relevant cases? I.E. How CORRECT do you think the citations in your summary were?

	1	2	3	4	5	
Not Confident	○	○	○	○	○	Very Confident

- How confident were you that you found all of the relevant decisions?

  In other words, how confident are you that someone knowledgeable with the subject matter would not feel you are missing important cases in your summary? I.E. How COMPLETE do you think your summary was?

	1	2	3	4	5	
Not Confident	○	○	○	○	○	Very Confident

Stars

- How often did you use the star system?

  Use cases for stars may include clicking a document because it was starred, or looking for a document by searching the star list for it.

	1	2	3	4	5	
Rarely	○	○	○	○	○	Often

- How did you use the star system?
- Were the decisions starred by the system (indicated with the blue star) interesting/worthy of being starred?

	1	2	3	4	5	
Most Starred Documents Were Undeserving of Stars	○	○	○	○	○	Most Starred Documents Were Deserving of Stars

- How happy were you with the number of decisions the system starred (indicated with the blue star)?

	1	2	3	4	5	
Too Few	○	○	○	○	○	Too Many

(a value of 3 indicates the number of starred decisions was just right).

- Did you star anything yourself?
  - Yes
  - No

  Why/Why not?

- Did you find our star system useful?

	1	2	3	4	5	
Not at all	○	○	○	○	○	Yes, very

Relationship

- How often did you use the relationships?

  Did you click on a search result because of an relationship? Did you avoid clicking on a search result because of the relationship (either because the relationship made it clear it wasn't useful, or because the relationship provided all the information you needed?)

	1	2	3	4	5	
Rarely	○	○	○	○	○	often

- How did you use the relationships?

- How often were the relationships given by the system useful?

	1	2	3	4	5	
Rarely	○	○	○	○	○	Often

- Which relationships would you like to have seen MORE of?
  - When a search result and a starred document are often cited together
  - When a search result and a starred document cite the same decisions
  - When a search result is cited by a starred document
  - When a search result cites a starred document
  - When decision in the search results was in the same circuit (court system) a starred decision

- Which relationships would you like to have seen LESS of?

  - When a search result and a starred document are often cited together
  - When a search result and a starred document cite the same decisions
  - When a search result is cited by a starred document
  - When a search result cites a starred document
  - When decision in the search results was in the same circuit (court system) a starred decision

- (Optional) Any additional feedback?
- How can we further improve our system?

A.3 Legal Knowledge Background Survey

- Gender:
  - Male
  - Female
- Age:
- Before helping us evaluate our search engine, how familiar were you with the US Legal System and precedential and non-precedential opinions (like the documents you read while using CourtListener)?
  - I am licensed to practice law
  - I hold a Juris Doctor, a Master of Laws, or a Doctor of Juridical Science degree
  - I have had some legal training or schooling, but do not have a law degree

- I have read precedential and non-precedential opinions before
- I read legal commentary (e.g. The Harvard Law Review)
- I read or watch non-fiction legal thrillers or legal dramas (e.g. Court TV)
- I read or watch fictional legal thrillers or legal dramas (e.g. Law and Order)
- I have no interest in the legal domain

Appendix B

Task Descriptions for User Study

B.1 Task 1

Scenario: You are a junior associate at a large national law firm. You have been assigned to assist a partner with an upcoming appeal to the Federal Circuit in a patent case. He meets with you to explain that he has been focused on copyright and trademark litigation over the last two and 1/2 years and has not paid close attention to developments in patent law over that time period. He is familiar with the Supreme Court's holding in Bilski v. Kappos, but has not kept up with developments at the Federal Circuit or Supreme Court since the date of that ruling. The appeal you will be working on involves questions of patentability under Section 101 of the Patent Act, specifically whether the computer-implemented invention in the case is an unpatentable abstract idea. He asks you to identify the relevant precedents issued by the Federal Circuit or Supreme Court since the date of the Bilski ruling. He indicates that he will be reading the cases himself, and so asks for only the most concise summary of why you believe the case is relevant, at most three sentences, and asks that you focus more on identifying the

most important cases. He also asks that you limit your survey to the two dozen cases you deem most relevant.

Your Task: Produce a list of the most relevant Federal Circuit or Supreme Court decisions since the date of the Supreme Court's Bilski decision, limiting yourself to at most 24 cases. Explain the relevance of each case or its key holdings in at most three sentences.

B.2 Task 2

Scenario: You are a junior associate at a small firm in San Francisco that specializes in asylum appeals before the Ninth Circuit. Having impressed the managing partner while working on a recent appeal, she decides that you should update a training document for the firm's newly hired associates. The firm often hires associates without experience in asylum appeals and has created a short overview of the key appellate issues. Unfortunately, the associate that used to keep this training document up-to-date left the firm at the end of May 2010 and the document has not been updated since. New associates are expected to familiarize themselves thoroughly with the most important precedents, so this document focuses on identifying the most important decisions and provides only the briefest summary of the case's holdings, less than three sentences. Recognizing that the training document has fallen significantly behind, but that you have other responsibilities, the managing partner asks that you bring it up-to-date by identifying only the 24 most important asylum decisions since it was last updated.

Your Task: Identify the 24 most important asylum decisions in the Ninth Circuit since June 1, 2010. Explain the relevance of each case or its key holdings in at most three sentences.

B.3 Task 3

Scenario:

You are a junior associate at a San Francisco-based non-profit that provides legal assistance to newspapers and other journalists, particularly with respect to Freedom of Information Act (FOIA) litigation. Your organization regularly creates a pamphlet to distribute to clients that summarizes the most important FOIA decisions issued within the last two years. The last pamphlet covered up to May 31, 2010, and the new one is now overdue. Your supervising attorney has informed you that you have been drafted to create this pamphlet in time for it to go out with the Christmas letter to clients. He explains that the size of the pamphlet will require you to limit yourself to the twenty-four most important FOIA decisions in the time period, and that you must summarize each case concisely, in at most three sentences. This particular pamphlet will be sent to clients within the Ninth Circuit, so he also asks that you focus your work on precedents relevant there.

Your Task: Identify the 24 most important FOIA decisions in the Ninth Circuit since June 1, 2010. Explain the relevance of each case or its key holdings in at most three sentences.

B.4 Task 4

Scenario: You are a junior associate at a large national law firm. You have been assigned to assist a partner with an upcoming federal appeal in a trademark case. He meets with you to explain that he has been focused on copyright and patent litigation over the last two and 1/2 years and has not paid close attention to developments in trademark law over that time period. He is familiar with the state of the law up through May 31, 2010, but has not followed trademark law closely since that time. The appeal you will be working on involves questions of

the possible secondary liability of an online service provider for trademark infringement where the online service provider's users engage in direct trademark infringement. He asks you to identify the relevant appellate-level precedents issued since June 1, 2010. He indicates that he will be reading the cases himself, and so asks for only the most concise summary of why you believe the case is relevant, at most three sentences, and asks that you focus more on identifying the most important cases. He also asks that you limit your survey to the two dozen cases you deem most relevant.

Your Task: Produce a list of the most relevant appellate decisions since June 1, 2010, limiting yourself to at most 24 cases. Explain the relevance of each case or its key holdings in at most three sentences.

B.5 Task 5

Scenario: You are a junior associate at a large national law firm. You have been assigned to assist a partner with an upcoming federal appeal to the United States Court of Appeals for the Ninth Circuit in a antitrust case. He meets with you to explain that he has been focused on intellectual property litigation over the last two and 1/2 years and has not paid close attention to developments in antitrust law over that time period. He is familiar with the state of the law up through May 31, 2010, but has not followed antitrust law closely since that time. The appeal you will be working on involves allegations of price-fixing, bidrigging, and geographic allocation of the market. He asks you to identify the relevant appellate-level precedents issued since June 1, 2010. He indicates that he will be reading the cases himself, and so asks for only the most concise summary of why you believe the case is relevant, at most three sentences, and asks that you focus more on identifying the most important cases. He also asks that you limit your survey to the two dozen cases you deem most relevant.

Your Task: Produce a list of the most relevant appellate decisions since June 1, 2010, limiting yourself to at most 24 cases. Explain the relevance of each case or its key holdings in at most three sentences.

B.6 Task 6

Scenario: You are a summer associate at a Silicon Valley-based law firm hoping to receive an offer of a fulltime job. One of the partners you are most hoping to impress writes his own survey or treatise on appellate-level developments in copyright law. He informs you that his treatise was last updated to cover decisions through May 31, 2010 and that it is past time to update it with more recent decisions. He asks you to identify the relevant appellate-level precedents issued since June 1, 2010. He indicates that he will be reading the cases himself, and so asks for only the most concise summary of why you believe the case is relevant, at most three sentences, and asks that you focus more on identifying the most important cases. He also asks that you limit your survey to the two dozen cases you deem most relevant.

Your Task: Produce a list of the most relevant appellate decisions since June 1, 2010, limiting yourself to at most 24 cases. Explain the relevance of each case or its key holding

Bibliography

[1] Jose Abracos and Gabriel Pereira Lopes. Statistical methods for retrieving most significant paragraphs in newspaper articles. In *ACL/EACL Workshop on Intelligent Scalable Text Summarization*, pages 51–57, 1997.

[2] David Abrams, Ron Baecker, and Mark Chignell. Information archiving with bookmarks: personal web space construction and organization. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, CHI '98, pages 41–48, New York, NY, USA, 1998. ACM Press/Addison-Wesley Publishing Co.

[3] Anne Adams and Ann Blandford. Digital libraries' support for the user's' information journey'. In *Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries*, pages 160–169. ACM, 2005.

[4] Eytan Adar, Jaime Teevan, and Susan T. Dumais. Large scale analysis of web revisitation patterns. In *Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, CHI '08, pages 1197–1206, New York, NY, USA, 2008. ACM.

[5] Eugene Agichtein, Ryen W. White, Susan T. Dumais, and Paul N. Bennet. Search, interrupted: understanding and predicting search task continuation. In *Proceedings of the*

35th international ACM SIGIR conference on Research and development in information retrieval, SIGIR '12, pages 315–324, New York, NY, USA, 2012. ACM.

[6] Omar Alonso, Michael Gertz, and Ricardo Baeza-Yates. Clustering and exploring search results using timeline constructions. In *Proceedings of the 18th ACM conference on Information and knowledge management*, CIKM '09, pages 97–106, New York, NY, USA, 2009. ACM.

[7] Liliana Ardissono, Anna Goy, Giovanna Petrone, Marino Segnan, and Pietro Torasso. Intrigue: personalized recommendation of tourist attractions for desktop and hand held devices. *Applied Artificial Intelligence*, 17(8-9):687–714, 2003.

[8] Anne Aula, Natalie Jhaveri, and Mika Käki. Information search and re-access strategies of experienced web users. In *Proceedings of the 14th international conference on World Wide Web*, WWW '05, pages 583–592, New York, NY, USA, 2005. ACM.

[9] Anne Aula, Rehan M. Khan, and Zhiwei Guan. How does search behavior change as search becomes more difficult? In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 35–44, New York, NY, USA, 2010. ACM.

[10] Ricardo Baeza-Yates, Carlos Hurtado, and Marcelo Mendoza. Query recommendation using query logs in search engines. In *Current Trends in Database Technology-EDBT 2004 Workshops*, pages 588–596. Springer, 2005.

[11] Evelyn Balfe and Barry Smyth. An analysis of query similarity in collaborative web search. In *Advances in Information Retrieval*, pages 330–344. Springer, 2005.

[12] Krisztian Balog, Pavel Serdyukov, and Arjen P de Vries. Overview of the trec 2010 entity track. Technical report, DTIC Document, 2010.

[13] Krisztian Balog, Pavel Serdyukov, and Arjen P de Vries. Overview of the trec 2011 entity track. Technical report, DTIC Document, 2011.

[14] Marcia J Bates. The design of browsing and berrypicking techniques for the online search interface. *Online Information Review*, 13(5):407–424, 1989.

[15] 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*, SIGIR '04, pages 321–328. ACM, 2004.

[16] Adam L. Berger and Vibhu O. Mittal. Ocelot: a system for summarizing web pages. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '00, pages 144–151, New York, NY, USA, 2000. ACM.

[17] Marianne Bertrand and Sendhil Mullainathan. Do people mean what they say? implications for subjective survey data. *The American Economic Review*, 91(2):67–72, 2001.

[18] Indrajit Bhattacharya and Lise Getoor. Collective entity resolution in relational data. *ACM Transactions on Knowledge Discovery from Data*, 1(1), March 2007.

[19] Mustafa Bilgic and Raymond J Mooney. Explaining recommendations: Satisfaction vs. promotion. In *Beyond Personalization Workshop, IUI*, volume 5, 2005.

[20] Daniel Billsus and Michael J Pazzani. A personal news agent that talks, learns and explains. In *Proceedings of the third annual conference on Autonomous Agents*, pages 268–275. ACM, 1999.

[21] Roi Blanco, Diego Ceccarelli, Claudio Lucchese, Raffaele Perego, and Fabrizio Silvestri. You should read this! let me explain you why: explaining news recommendations to users. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, CIKM '12, pages 1995–1999, New York, NY, USA, 2012. ACM.

[22] Roi Blanco and Hugo Zaragoza. Finding support sentences for entities. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '10, pages 339–346, New York, NY, USA, 2010. ACM.

[23] David J. Brenes, Daniel Gayo-Avello, and Kilian Pérez-González. Survey and evaluation of query intent detection methods. In *Proceedings of the 2009 workshop on Web Search Click Data*, WSCD '09, pages 1–7, New York, NY, USA, 2009. ACM.

[24] Andrei Broder. A taxonomy of web search. *SIGIR Forum*, 36(2):3–10, September 2002.

[25] Marc Bron, Jasmijn van Gorp, Frank Nack, Lotte Belice Baltussen, and Maarten de Rijke. Aggregated search interface preferences in multi-session search tasks. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '13, New York, NY, USA, 2013. ACM.

[26] Harry Bruce, William Jones, and Susan Dumais. Keeping and re-finding information on the web: What do people do and what do they need? *Proceedings of the American Society for Information Science and Technology*, 41(1):129–137, 2004.

[27] Bruce G Buchanan and Edward H Shortliffe. *Rule Based Expert Systems: The Mycin Experiments of the Stanford Heuristic Programming Project (The Addison-Wesley series in artificial intelligence)*. Addison-Wesley Longman Publishing Co., Inc., 1984.

[28] 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, 2011.

[29] Katriina Byström and Kalervo Järvelin. Task complexity affects information seeking and use. *Information Processing & Management*, 31(2):191–213, 1995.

[30] Robert Capra and Manuel A. Pérez-Quiñones. Re-finding found things: An exploratory study of how users re-find information. *CoRR*, cs.HC/0310011, 2003.

[31] Robert G. Capra and Gary Marchionini. The relation browser tool for faceted exploratory search. In *Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries*, JCDL '08, pages 420–420, New York, NY, USA, 2008. ACM.

[32] Claudio Carpineto, Stanislaw Osiński, Giovanni Romano, and Dawid Weiss. A survey of web clustering engines. *ACM Computing Surveys*, 41(3):17:1–17:38, July 2009.

[33] Claudio Carpineto and Giovanni Romano. Exploiting the potential of concept lattices for information retrieval with credo. *Journal of Universal Computer Science*, 10(8):985–1013, 2004.

[34] Lara D Catledge and James E Pitkow. Characterizing browsing strategies in the world-wide web. *Computer Networks and ISDN systems*, 27(6):1065–1073, 1995.

[35] Ed H. Chi, Peter Pirolli, Kim Chen, and James Pitkow. Using information scent to model user information needs and actions and the web. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '01, pages 490–497, New York, NY, USA, 2001. ACM.

[36] Lydia B Chilton and Jaime Teevan. Addressing people's information needs directly in a web search result page. In *Proceedings of the 20th international conference on World wide web*, WWW '11, pages 27–36. ACM, 2011.

[37] Paul Clough, Mark Sanderson, Murad Abouammoh, Sergio Navarro, and Monica Paramita. Multiple approaches to analysing query diversity. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '09, pages 734–735, New York, NY, USA, 2009. ACM.

[38] Andy Cockburn, Saul Greenberg, Steve Jones, Bruce Mckenzie, and Michael Moyle. Improving web page revisitation: analysis, design and evaluation. *IT & Society*, 1:159–183, 2003.

[39] Dan Cosley, Shyong K. Lam, Istvan Albert, Joseph A. Konstan, and John Riedl. Is seeing believing?: how recommender system interfaces affect users' opinions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '03, pages 585–592, New York, NY, USA, 2003. ACM.

[40] Edward Cutrell and Zhiwei Guan. What are you looking for?: an eye-tracking study of information usage in web search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '07, pages 407–416, New York, NY, USA, 2007. ACM.

[41] 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 *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '92, pages 318–329, New York, NY, USA, 1992. ACM.

[42] Marek Czarkowski. *A Scrutable Adaptive Hypertext*. PhD thesis, The University of Sydney, 2006.

[43] Marek Czarkowski and Judy Kay. A scrutable adaptive hypertext. In Paul Bra, Peter Brusilovsky, and Ricardo Conejo, editors, *Adaptive Hypermedia and Adaptive Web-Based Systems*, volume 2347 of *Lecture Notes in Computer Science*, pages 384–387. Springer Berlin Heidelberg, 2002.

[44] Wisam Dakka, Luis Gravano, and Panagiotis G Ipeirotis. Answering general time-sensitive queries. *Knowledge and Data Engineering, IEEE Transactions on*, 24(2):220–235, 2012.

[45] Arjen P De Vries, Anne-Marie Vercoustre, James A Thom, Nick Craswell, and Mounia Lalmas. Overview of the inex 2007 entity ranking track. In *Focused Access to XML Documents*, pages 245–251. Springer, 2008.

[46] Gianluca Demartini, Tereza Iofciu, and Arjen P De Vries. Overview of the inex 2009 entity ranking track. In *Focused Retrieval and Evaluation*, pages 254–264. Springer, 2010.

[47] Gianluca Demartini, ArjenP. Vries, Tereza Iofciu, and Jianhan Zhu. Overview of the inex 2008 entity ranking track. In Shlomo Geva, Jaap Kamps, and Andrew Trotman, editors, *Advances in Focused Retrieval*, volume 5631 of *Lecture Notes in Computer Science*, pages 243–252. Springer Berlin Heidelberg, 2009.

[48] Marian Dork, Sheelagh Carpendale, Christopher Collins, and Carey Williamson. Visgets: Coordinated visualizations for web-based information exploration and discovery. *Visualization and Computer Graphics, IEEE Transactions on*, 14(6):1205–1212, 2008.

[49] Marian Dörk, Sheelagh Carpendale, and Carey Williamson. The information flaneur: a fresh look at information seeking. In *Proceedings of the 2011 annual conference on Human factors in computing systems*, pages 1215–1224. ACM, 2011.

[50] P. Ferragina and A. Gulli. A personalized search engine based on web-snippet hierarchical clustering. *Software: Practice and Experience*, 38(2):189–225, 2008.

[51] Jonathan L Freedman and Scott C Fraser. Compliance without pressure: the foot-in-the-door technique. *Journal of personality and social psychology*, 4(2):195, 1966.

[52] Benjamin C.M. Fung, Ke Wang, and Martin Ester. Hierarchical document clustering using frequent itemsets. In *Proceedings of SIAM international conference on data mining*, pages 59–70, 2003.

[53] Lise Getoor and Christopher P. Diehl. Link mining: a survey. *SIGKDD Explor. Newsl.*, 7(2):3–12, December 2005.

[54] Fosca Giannotti, Mirco Nanni, and Dino Pedreschi. Webcat: Automatic categorization of web search results. In *In Proceedings of Sistemi Evoluti per Basi di Dati, SEBD'03*, pages 507–518, New York, NY, USA, 2003. ACM.

[55] Natalie Glance, Matthew Hurst, Kamal Nigam, Matthew Siegler, Robert Stockton, and Takashi Tomokiyo. Deriving marketing intelligence from online discussion. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, KDD '05, pages 419–428, New York, NY, USA, 2005. ACM.

[56] Natalie S. Glance, Matthew Hurst, and Takashi Tomokiyo. Blogpulse: Automated trend discovery for weblogs. In *WWW 2004 workshop on the weblogging ecosystem: Aggregation, analysis and dynamics*. ACM, 2004.

[57] Maria Grineva, Maxim Grinev, and Dmitry Lizorkin. Extracting key terms from noisy and multitheme documents. In *Proceedings of the 18th international conference on World wide web*, WWW '09, pages 661–670, New York, NY, USA, 2009. ACM.

[58] Jiafeng Guo, Gu Xu, Xueqi Cheng, and Hang Li. Named entity recognition in query. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '09, pages 267–274, New York, NY, USA, 2009. ACM.

[59] Jiyin He, Marc Bron, and Arjen P. de Vries. Characterizing stages of a multi-session complex search task through direct and indirect query modifications. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '13, pages 897–900, New York, NY, USA, 2013. ACM.

[60] Marti Hearst. Design recommendations for hierarchical faceted search interfaces. In *ACM SIGIR workshop on faceted search*, pages 1–5, 2006.

[61] Marti Hearst, Ame Elliott, Jennifer English, Rashmi Sinha, Kirsten Swearingen, and Ka-Ping Yee. Finding the flow in web site search. *Communications of the ACM*, 45(9):42–49, September 2002.

[62] Marti A. Hearst. Clustering versus faceted categories for information exploration. *Communications of the ACM*, 49(4):59–61, April 2006.

[63] 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*, SIGIR '96, pages 76–84, New York, NY, USA, 1996. ACM.

[64] Jonathan L. Herlocker, Joseph A. Konstan, and John Riedl. Explaining collaborative filtering recommendations. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work*, CSCW '00, pages 241–250, New York, NY, USA, 2000. ACM.

[65] Jeff Huang and Efthimis N Efthimiadis. Analyzing and evaluating query reformulation strategies in web search logs. In *Proceedings of the 18th ACM conference on Information and knowledge management*, CIKM '09, pages 77–86. ACM, 2009.

[66] Jeff Huang, Ryen W. White, and Susan Dumais. No clicks, no problem: using cursor movements to understand and improve search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 1225–1234, New York, NY, USA, 2011. ACM.

[67] Robert N Hughes. Intrinsic exploration in animals: motives and measurement. *Behavioural Processes*, 41(3):213 – 226, 1997.

[68] Samuel Ieong, Nina Mishra, Eldar Sadikov, and Li Zhang. Domain bias in web search. In *Proceedings of the fifth ACM international conference on Web search and data mining*, WSDM '12, pages 413–422, New York, NY, USA, 2012. ACM.

[69] Peter Ingwersen. Polyrepresentation of information needs and semantic entities: elements of a cognitive theory for information retrieval interaction. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '94, pages 101–110, New York, NY, USA, 1994. Springer-Verlag New York, Inc.

[70] Peter Ingwersen. Cognitive perspectives of information retrieval interaction: elements of a cognitive ir theory. *Journal of Documentation*, 52:3–50, 1996.

[71] Bernard J. Jansen, Danielle L. Booth, and Amanda Spink. Determining the informational, navigational, and transactional intent of web queries. *Information Processing & Management*, 44(3):1251 – 1266, 2008.

[72] Rosie Jones and Kristina Lisa Klinkner. Beyond the session timeout: automatic hierarchical segmentation of search topics in query logs. In *Proceedings of the 17th ACM conference on Information and knowledge management*, CIKM '08, pages 699–708. ACM, 2008.

[73] Rosie Jones, Benjamin Rey, Omid Madani, and Wiley Greiner. Generating query substitutions. In *Proceedings of the 15th international conference on World Wide Web*, WWW '06, pages 387–396. ACM, 2006.

[74] William Jones, Harry Bruce, and Susan Dumais. Keeping found things found on the web. In *Proceedings of the tenth international conference on Information and knowledge management*, CIKM '01, pages 119–126, New York, NY, USA, 2001. ACM.

[75] Anupam Joshi and Zhihua Jiang. Retriever: Improving web search engine results using clustering. *Managing Business with Electronic Commerce: Issues and Trends*, page 59, 2002.

[76] In-Ho Kang and GilChang Kim. Query type classification for web document retrieval. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, SIGIR '03, pages 64–71, New York, NY, USA, 2003. ACM.

[77] Tapas Kanungo and David Orr. Predicting the readability of short web summaries. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, WSDM '09, pages 202–211, New York, NY, USA, 2009. ACM.

[78] Anita Komlodi, Dagobert Soergel, and Gary Marchionini. Search histories for user support in user interfaces. *Journal of the American Society for Information Science and Technology*, 57(6):803–807, 2006.

[79] Jonathan Koren, Andrew Leung, Yi Zhang, Carlos Maltzahn, Sasha Ames, and Ethan Miller. Searching and navigating petabyte-scale file systems based on facets. In *Proceedings of the 2nd international workshop on Petascale data storage: held in conjunction with Supercomputing '07*, PDSW '07, pages 21–25, New York, NY, USA, 2007. ACM.

[80] Jonathan Koren, Yi Zhang, and Xue Liu. Personalized interactive faceted search. In *Proceedings of the 17th international conference on World Wide Web*, WWW '08, pages 477–486, New York, NY, USA, 2008. ACM.

[81] Alexander Kotov, Paul N. Bennett, Ryen W. White, Susan T. Dumais, and Jaime Teevan. Modeling and analysis of cross-session search tasks. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, SIGIR '11, pages 5–14, New York, NY, USA, 2011. ACM.

[82] Jon A Krosnick. Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied cognitive psychology*, 5(3):213–236, 1991.

[83] Carol C Kuhlthau. Inside the search process: Information seeking from the user's perspective. *Journal of the American Society for Information Science and Technology*, 42(5):361–371, 1991.

[84] Bill Kules and Ben Shneiderman. Users can change their web search tactics: Design guidelines for categorized overviews. *Information Processing & Management*, 44(2):463 – 484, 2008.

[85] Tessa Lau and Eric Horvitz. Patterns of search: analyzing and modeling web query refinement. In *Proceedings of the seventh international conference on User modeling*, pages 119–128. Springer-Verlag New York, Inc., 1999.

[86] Dawn J. Lawrie and W. Bruce Croft. Generating hierarchical summaries for web searches. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, SIGIR '03, pages 457–458, New York, NY, USA, 2003. ACM.

[87] Uichin Lee, Zhenyu Liu, and Junghoo Cho. Automatic identification of user goals in web search. In *Proceedings of the 14th international conference on World Wide Web*, WWW '05, pages 391–400, New York, NY, USA, 2005. ACM.

[88] Anton V Leouski and W Bruce Croft. An evaluation of techniques for clustering search results. Technical report, DTIC Document, 2005.

[89] Anton Leuski and James Allan. Improving interactive retrieval by combining ranked list and clustering. In *RIAO*, pages 665–681, 2000.

[90] Jane Li, Scott Huffman, and Akihito Tokuda. Good abandonment in mobile and pc internet search. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '09, pages 43–50. ACM, 2009.

[91] Xiaoyan Li and W. Bruce Croft. Improving novelty detection for general topics using sentence level information patterns. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, CIKM '06, pages 238–247, New York, NY, USA, 2006. ACM.

[92] Shinjeng Lin and Iris Xie. Behavioral changes in transmuting multisession successive searches over the web. *Journal of the American Society for Information Science and Technology*, page 12591283, 2013.

[93] Jingjing Liu and Nicholas J Belkin. Personalizing information retrieval for multi-session tasks: The roles of task stage and task type. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '10, pages 26–33. ACM, 2010.

[94] Jingjing Liu, Nicholas J. Belkin, Xiangmin Zhang, and Xiaojun Yuan. Examining users knowledge change in the task completion process. *Information Processing and Management*, 49(5):1058 – 1074, 2013.

[95] Yiqun Liu, Min Zhang, Liyun Ru, and Shaoping Ma. Automatic query type identification based on click through information. In HweeTou Ng, Mun-Kew Leong, Min-Yen Kan, and Donghong Ji, editors, *Information Retrieval Technology*, volume 4182 of *Lecture Notes in Computer Science*, pages 593–600. Springer Berlin Heidelberg, 2006.

[96] Bonnie MacKay and Carolyn Watters. Building support for multi-session tasks. In *CHI '09 Extended Abstracts on Human Factors in Computing Systems*, CHI '09, pages 4273–4278. ACM, 2009.

[97] Gary Marchionini. Exploratory search: from finding to understanding. *Communications of the ACM*, 49(4):41–46, April 2006.

[98] Andrew McCallum and Wei Li. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 - Volume 4*, CONLL '03, pages 188–191, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[99] Kevin McCarthy, James Reilly, Lorraine McGinty, and Barry Smyth. Thinking positively-explanatory feedback for conversational recommender systems. In *Proceedings of the Eu-*

ropean Conference on Case-Based Reasoning (ECCBR-04) Explanation Workshop, pages 115–124, 2004.

[100] Lorraine McGinty and Barry Smyth. Extending comparison-based recommendation: A review. *Poster acceptance for the British Computer Societys Specialist Group on Artificial Intelligence (AI-03)*, 2003.

[101] David McSherry. Explanation in recommender systems. *Artificial Intelligence Review*, 24(2):179–197, 2005.

[102] Dan Morris, Meredith Ringel Morris, and Gina Venolia. Searchbar: a search-centric web history for task resumption and information re-finding. In *Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, CHI '08, pages 1207–1216, New York, NY, USA, 2008. ACM.

[103] Meredith Ringel Morris and Eric Horvitz. S3: Storable, shareable search. In *Human-Computer Interaction–INTERACT 2007*, pages 120–123. Springer, 2007.

[104] David Nadeau and Satoshi Sekine. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26, January 2007.

[105] Tetsuya Nasukawa and Jeonghee Yi. Sentiment analysis: capturing favorability using natural language processing. In *Proceedings of the 2nd international conference on Knowledge capture*, K-CAP '03, pages 70–77, New York, NY, USA, 2003. ACM.

[106] Raymond S Nickerson. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, 2(2):175, 1998.

[107] Jakob Nielsen and Rolf Molich. Heuristic evaluation of user interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '90, pages 249–256, New York, NY, USA, 1990. ACM.

[108] Hartmut Obendorf, Harald Weinreich, Eelco Herder, and Matthias Mayer. Web page revisitation revisited: implications of a long-term click-stream study of browser usage. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, CHI '07, pages 597–606, New York, NY, USA, 2007. ACM.

[109] Daniel M Oppenheimer, Tom Meyvis, and Nicolas Davidenko. Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, 45(4):867–872, 2009.

[110] Stanislaw Osinski and Dawid Weiss. Conceptual clustering using lingo algorithm: Evaluation on open directory project data. In MieczyslawA. Klopotek, SlawomirT. Wierzchon, and Krzysztof Trojanowski, editors, *Intelligent Information Processing and Web Mining*, volume 25 of *Advances in Soft Computing*, pages 369–377. Springer Berlin Heidelberg, 2004.

[111] Steven Pace. A grounded theory of the flow experiences of web users. *International journal of human-computer studies*, 60(3):327–363, 2004.

[112] Bo Pang and Lillian Lee. A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, ACL '04, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics.

[113] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135, January 2008.

[114] 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*, CHI '97, pages 3–10, New York, NY, USA, 1997. ACM.

[115] Peter Pirolli and Stuart Card. Information foraging. *Psychological review*, 106(4):643, 1999.

[116] Peter LT Pirolli. *Information foraging theory: Adaptive interaction with information*. Oxford University Press, USA, 1 edition, April 2007.

[117] Tom Pyszczynski, Jeff Greenberg, and John LaPrelle. Social comparison after success and failure: Biased search for information consistent with a self-serving conclusion. *Journal of Experimental Social Psychology*, 21(2), 1985.

[118] Vijay V. Raghavan and Hayri Sever. On the reuse of past optimal queries. In *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '95, pages 344–350, New York, NY, USA, 1995. ACM.

[119] Ronald A. Rensink. Internal vs. external information in visual perception. In *Proceedings of the 2nd international symposium on Smart graphics*, SMARTGRAPH '02, pages 63–70, New York, NY, USA, 2002. ACM.

[120] Ronald A Rensink, J Kevin O'Regan, and James J Clark. To see or not to see: The need for attention to perceive changes in scenes. *Psychological Science*, 8(5):368–373, 1997.

[121] Soo Young Rieh. On the web at home: Information seeking and web searching in the home environment. *Journal of the American Society for Information Science and Technology*, 55(8):743–753, 2004.

[122] Daniel E. Rose and Danny Levinson. Understanding user goals in web search. In *Proceedings of the 13th international conference on World Wide Web*, WWW '04, pages 13–19, New York, NY, USA, 2004. ACM.

[123] Karen Rustad and Rowyn McDonald. Building a free, open source legal citator. Technical report, UC Berkeley School of Information, 2012.

[124] Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513 – 523, 1988.

[125] Mark Sanderson and Susan Dumais. Examining repetition in user search behavior. In *Advances in Information Retrieval*, pages 597–604. Springer, 2007.

[126] m.c. schraefel, Max Wilson, Alistair Russell, and Daniel A. Smith. mSpace: improving information access to multimedia domains with multimodal exploratory search. *Communications of the ACM*, 49(4):47–49, April 2006.

[127] Erik Selberg and Oren Etzioni. On the instability of web search engines. In *Proceedings of RIAO*, pages 223–235. Citeseer, 2000.

[128] Milad Shokouhi, Ryen W White, Paul Bennett, and Filip Radlinski. Fighting search engine amnesia: Reranking repeated results. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '13, 2013.

[129] Daniel J. Simons and Daniel T. Levin. Change blindness. *Trends in Cognitive Sciences*, 1(7):261 – 267, 1997.

[130] Rashmi Sinha and Kirsten Swearingen. The role of transparency in recommender systems. In *CHI '02 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '02, pages 830–831, New York, NY, USA, 2002. ACM.

[131] Vineet Sinha and David R. Karger. Magnet: supporting navigation in semistructured data environments. In *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, SIGMOD '05, pages 97–106, New York, NY, USA, 2005. ACM.

[132] Barry Smyth, Evelyn Balfe, Jill Freyne, Peter Briggs, Maurice Coyle, and Oisin Boydell. Exploiting query repetition and regularity in an adaptive community-based web search engine. *User Modeling and User-Adapted Interaction*, 14(5):383–423, 2004.

[133] Karen Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. In Peter Willett, editor, *Document retrieval systems*, pages 132–142. Taylor Graham Publishing, London, UK, UK, 1988.

[134] DW Stephens, JR Krebs, et al. Foraging theory. *Foraging theory.*, 1986.

[135] Emilia Stoica and Marti A. Hearst. Nearly-automated metadata hierarchy creation. In *Proceedings of the eighth Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, HLT-NAACL-Short '04, pages 117–120, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics.

[136] Emilia Stoica, Marti A Hearst, and Megan Richardson. Automating creation of hierarchical faceted metadata structures. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 244–251, 2007.

[137] Linda Tauscher and Saul Greenberg. How people revisit web pages: empirical findings and implications for the design of history systems. *International Journal of Human Computer Studies*, 47:97–137, 1997.

[138] Jaime Teevan. *The re:search engine: simultaneous support for finding and re-finding.* PhD thesis, Massachusetts Institute of Technology, 2006.

[139] Jaime Teevan, Eytan Adar, Rosie Jones, and Michael A. S. Potts. Information re-retrieval: repeat queries in yahoo's logs. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '07, pages 151–158, New York, NY, USA, 2007. ACM.

[140] Jaime Teevan, Susan T. Dumais, and Eric Horvitz. Personalizing search via automated analysis of interests and activities. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '05, pages 449–456, New York, NY, USA, 2005. ACM.

[141] Jaime Teevan, Susan T. Dumais, and Daniel J. Liebling. To personalize or not to personalize: modeling queries with variation in user intent. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, pages 163–170, New York, NY, USA, 2008. ACM.

[142] Cynthia A Thompson, Mehmet H Göker, and Pat Langley. A personalized system for conversational recommendations. *Journal of Artificial Intelligence Research*, 21:393–428, 2004.

[143] Nava Tintarev and Judith Masthoff. A survey of explanations in recommender systems. In *Proceedings of the 2007 IEEE 23rd International Conference on Data Engineering Workshop*, ICDEW '07, pages 801–810, Washington, DC, USA, 2007. IEEE Computer Society.

[144] Hiroyuki Toda and Ryoji Kataoka. A search result clustering method using informatively named entities. In *Proceedings of the 7th annual ACM international workshop on Web*

information and data management, WIDM '05, pages 81–86, New York, NY, USA, 2005. ACM.

[145] Anastasios Tombros and Mark Sanderson. Advantages of query biased summaries in information retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '98, pages 2–10, New York, NY, USA, 1998. ACM.

[146] Takashi Tomokiyo and Matthew Hurst. A language model approach to keyphrase extraction. In *Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment - Volume 18*, MWE '03, pages 33–40, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[147] Daniel Tunkelang. Faceted search. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 1(1):1–80, 2009.

[148] Sarah K. Tyler and Jaime Teevan. Large scale query log analysis of re-finding. In *Proceedings of the third ACM international conference on Web search and data mining*, WSDM '10, pages 191–200, New York, NY, USA, 2010. ACM.

[149] Sarah K Tyler, Jian Wang, and Yi Zhang. Utilizing re-finding for personalized information retrieval. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, CIKM '10, pages 1469–1472. ACM, 2010.

[150] Sarah K Tyler and Yi Zhang. Multi-session re-search: in pursuit of repetition and diversification. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, CIKM '12, pages 2055–2059. ACM, 2012.

[151] Davor Čubranić. Polestar: Assisted navigation for exploring multi-dimensional information spaces. In *Proceedings of the Second Workshop on Human-Computer Interaction and Information Retrieval*, pages 9–12, 2008.

[152] Pertti Vakkari, Mikko Pennanen, and Sami Serola. Changes of search terms and tactics while writing a research proposal: A longitudinal case study. *Information Processing & Management*, 39(3):445 – 463, 2003.

[153] Cornelis J van Rijsbergen and K Sparck Jones. A test for the separation of relevant and non-relevant documents in experimental retrieval collections. *Journal of Documentation*, 29(3):251–257, 1973.

[154] Ramakrishna Varadarajan and Vagelis Hristidis. A system for query-specific document summarization. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, CIKM '06, pages 622–631, New York, NY, USA, 2006. ACM.

[155] Jesse Vig, Shilad Sen, and John Riedl. Tagsplanations: explaining recommendations using tags. In *Proceedings of the 14th international conference on Intelligent user interfaces*, pages 47–56. ACM, 2009.

[156] Michail Vlachos, Christopher Meek, Zografoula Vagena, and Dimitrios Gunopulos. Identifying similarities, periodicities and bursts for online search queries. In *Proceedings of the 2004 ACM SIGMOD international conference on Management of data*, pages 131–142. ACM, 2004.

[157] Hongning Wang, Yang Song, Ming-Wei Chang, Xiaodong He, Ryen W. White, and Wei Chu. Learning to extract cross-session search tasks. In *Proceedings of the 22nd international conference on World Wide Web*, WWW '13, pages 1353–1364, Republic and Canton of Geneva, Switzerland, 2013.

[158] Qing Wang and Huiyou Chang. Multitasking bar: Prototype and evaluation of introducing the task concept into a browser. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 103–112. ACM, 2010.

[159] Yitong Wang and Masaru Kitsuregawa. Link based clustering of web search results. In *Advances in Web-Age Information Management*, pages 225–236. Springer, 2001.

[160] Pontus Wärnestål. User evaluation of a conversational recommender system. In *Proceedings of the 4th Workshop on Knowledge and Reasoning in Practical Dialogue Systems*, 2005.

[161] Claire Warwick, Jon Rimmer, Ann Blandford, Jeremy Gow, and George Buchanan. Cognitive economy and satisficing in information seeking: A longitudinal study of undergraduate information behavior. *Journal of the American Society for Information Science and Technology*, 60(12):2402–2415, 2009.

[162] Ji-Rong Wen, Jian-Yun Nie, and Hong-Jiang Zhang. Query clustering using user logs. *ACM Transactions on Information Systems*, 20(1):59–81, 2002.

[163] Michael White, Tanya Korelsky, Claire Cardie, Vincent Ng, David Pierce, and Kiri Wagstaff. Multidocument summarization via information extraction. In *Proceedings of the first international conference on Human language technology research*, HLT '01, pages 1–7, Stroudsburg, PA, USA, 2001. Association for Computational Linguistics.

[164] Ryen White. Beliefs and biases in web search. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '13, pages 3–12, New York, NY, USA, 2013. ACM.

[165] Ryen W White, Mikhail Bilenko, and Silviu Cucerzan. Studying the use of popular destinations to enhance web search interaction. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '07, pages 159–166. ACM, 2007.

[166] Ryen W White, Bill Kules, Steven M Drucker, et al. Supporting exploratory search, introduction, special issue, communications of the acm. *Communications of the ACM*, 49(4):36–39, 2006.

[167] 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.

[168] Ryen W. White, Ian Ruthven, and Joemon M. Jose. Finding relevant documents using top ranking sentences: an evaluation of two alternative schemes. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '02, pages 57–64, New York, NY, USA, 2002. ACM.

[169] Max L. Wilson, Paul André, and mc schraefel. Backward highlighting: enhancing faceted search. In *Proceedings of the 21st annual ACM symposium on User interface software and technology*, UIST '08, pages 235–238, New York, NY, USA, 2008. ACM.

[170] Yi-Fang Wu and Xin Chen. Extracting features from web search returned hits for hierarchical classification. In Hamid R. Arabnia, editor, *IKE*, pages 103–108. CSREA Press, 2003.

[171] Ka-Ping Yee, Kirsten Swearingen, Kevin Li, and Marti Hearst. Faceted metadata for image search and browsing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '03, pages 401–408, New York, NY, USA, 2003. ACM.

[172] Yisong Yue, Rajan Patel, and Hein Roehrig. Beyond position bias: examining result attractiveness as a source of presentation bias in clickthrough data. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 1011–1018. ACM, 2010.

[173] Oren Zamir and Oren Etzioni. Web document clustering: a feasibility demonstration. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '98, pages 46–54, New York, NY, USA, 1998. ACM.

[174] Oren Zamir and Oren Etzioni. Grouper: a dynamic clustering interface to web search results. *Computer Networks*, 31(11):1361–1374, 1999.

[175] Klaus Zechner. Fast generation of abstracts from general domain text corpora by extracting relevant sentences. In *Proceedings of the 16th conference on Computational linguistics - Volume 2*, COLING '96, pages 986–989, Stroudsburg, PA, USA, 1996. Association for Computational Linguistics.

[176] Hua-Jun Zeng, Qi-Cai He, Zheng Chen, Wei-Ying Ma, and Jinwen Ma. Learning to cluster web search results. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '04, pages 210–217, New York, NY, USA, 2004. ACM.

[177] Dell Zhang and Yisheng Dong. Semantic, hierarchical, online clustering of web search results. In *Advanced Web Technologies and Applications*, pages 69–78. Springer, 2004.

[178] Li Zhang, Yue Pan, and Tong Zhang. Focused named entity recognition using machine learning. In *Proceedings of the 27th annual international ACM SIGIR conference on*

Research and development in information retrieval, SIGIR '04, pages 281–288, New York, NY, USA, 2004. ACM.

[179] Ruiqiang Zhang, Yi Chang, Zhaohui Zheng, Donald Metzler, and Jian-yun Nie. Search result re-ranking by feedback control adjustment for time-sensitive query. In *Proceedings of thirteenth Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pages 165–168. Association for Computational Linguistics, 2009.

[180] Zhiyong Zhang and Olfa Nasraoui. Mining search engine query logs for query recommendation. In *Proceedings of the 15th international conference on World Wide Web*, WWW '10, pages 1039–1040. ACM, 2006.