Augmenting Interactive Information Seeking with System-Level Assistance

by

Ryan E. Burton

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Information)
in the University of Michigan
2024

Doctoral Committee:

Professor Kevyn Collins-Thompson, Chair
Professor Eytan Adar
Professor Jaime Arguello
Professor Xu Wang

Ryan E. Burton
ryb@umich.edu
ORCID iD: 0009-0006-6492-1361

ACKNOWLEDGEMENTS

I'd like to begin by thanking my advisor, Kevyn Collins-Thompson, for his direction and guidance throughout my time in the PhD program as well as his patience and understanding, even as I developed health issues during the PhD that severely impeded my progress. I would also like to thank Eytan Adar, Jaime Arguello, and Xu Wang for serving on my committee and providing helpful feedback to my dissertation proposal as well as this dissertation.

I would additionally like to extend my appreciation to the faculty at the School of Information for their time and feedback over the years. In particular, I would like to thank Yan Chen, Libby Hemphill, and David Jurgens.

My SI colleagues have been some of the best, including (but not limited to) Tawfiq Ammari, Sungjin Nam, Lia Bozarth, Rohail Syed, Allan Martel, Teng Ye, Yue Wang, Tzu-Yu Wu, Melody Ku, Lindsay Blackwell, Penny Trieu, Padma Chirumamilla, Jackie Cohen, Carol Moser, Cindy Lin, Hari Subramonyam, Sangseok You, and Warren Li. Thanks for your friendship.

Special shout-outs go to Sam Carton, Chanda Phelan, and Daphne Chang for being the best drinking buddies. Happy hours were good times.

I'd like to thank my mother, who nutured my curiosity. I wouldn't have gotten this far without you.

Finally, I extend a special thanks to Heeryung Choi. Thank you for your endless support.

TABLE OF CONTENTS

LIST OF FIGURES

FIGURE

LIST OF TABLES

TABLE

LIST OF ALGORITHMS

ALGORITHM

LIST OF APPENDICES

LIST OF ACRONYMS

AI Artificial Intelligence

CDF Cumulative Density Function

CTR Click-Through Rate

HCI Human-Computer Interaction

IR Information Retrieval

LLM Large Language Model

RQ Research Question

SERP Search Engine Result Page

SET Search Economic Theory

STEM Science, Technology, Engineering, and Mathematics

ABSTRACT

Within the past two decades there has been an increased amount of interest in human-centered information retrieval research beyond the traditional system-focused view embodied by Cranfield- and TREC-style evaluation. From blending search and recommender systems to work on search as learning, the limits of conventional search engines' focus on query relevance to deliver ten blue links at millisecond speeds has become evident as use cases become more varied. My work focuses on the types of scenarios that typical systems neglect in searching and browsing, particularly tasks involving multi-attribute queries and information seeking as learning, and explores ways to guide users towards optimal behavior through system design. Enabling this across the three studies comprising this dissertation is a sidebar affordance, serving as a means for enabling complementary information seeking interactions. The contributions of this work will have implications on the effective design and implementation of new types of user-centered interactive IR systems.

We begin with an investigation of time-quality tradeoffs with *slow search*. Taking inspiration from movements such as slow food, slow travel, and slow technology, slow search serves as an acknowledgement of the fact that there are tasks for which users have indicated a willingness to wait for the perfect set of results. By implementing a user study where searchers were exposed to a system that embodied characteristics of slow search, where speed could be traded for an better results, we analyzed user behavior as they performed tasks which typically required multiple queries with a baseline Web search engine and saw how their effectiveness in using the system improved as they used this novel interface.

Next, we performed a simulation study to explore the implications of changing attributes of our slow search system on the behavioral outcomes of synthetic users modelled based on human interaction log data. By incorporating the users' cost model, we were able to identify fruitful directions for further interactive search experiments. In this way, we showed the potential for guiding low-performing users towards higher performance.

We finally focus on search as learning, using a large language model (LLM) as an enabler of slow search. Here, our study tests a contextual chatbot assistant that aids in users' reading and searching in a specialized domain – data science. The chatbot can provide responses to questions about documents and domain-specific vocabulary. Using mixed methods, we

identify patterns of use and investigate learning and interaction behaviors. Results show learning gains reveals that trust is a prominent factor in users' perceptions of usefulness. We furthermore propose an extension to develop a retrieval framework that can be used to directly optimise the set of interactions that a user may take in order to extract the maximum utility of a document. Using vocabulary learning and searching-to-learn as a foundation, we propose both an algorithm and user study to evaluate effectiveness in jointly considering relevance and familiarity with technical terms to learn to ensure users get the most out of the documents they search for.

The theme linking these studies together is a focus on improving user behavior to reduce effort or time-on-task, and increase value over time during interactive search. This dissertation serves as a basis for future system design and experiments that preserve interactivity, encourage effective mental models, and reduce user effort while increasing the value users receive during the search process.

CHAPTER 1

Introduction

The goal of this dissertation is to understand user behavior as they use novel interactive information system interfaces and explore the mechanisms we have as system designers to enable users to achieve optimal behavior. Such improvements may be driven by interface elements to expose operational transparency in the way the system works, system properties such as ranking algorithms tuned for precision or recall, and fundamental changes to the mode of search as seen in the recent exploration of chat for information seeking to allow for interaction more akin to question-answering. In service of this, the work presented here represents analysis and evaluation of task-centric search behavior with unconventional systems in contrast to Web search, and an exploration of a simulation framework that models multi-round interactions towards a complex goal satisfying multiple attributes in order to predict the necessary changes to system and user behavior in order to reduce time on task, increase relevance over time, and improve knowledge gain. In all, this dissertation offers an exploration of the use of systems augmented with both atypical interface elements and flexible system-level tradeoffs such as tradeoffs between retrieval time and result quality, as well as approaches for optimizing combinations of dynamic system and user behavior. As such, this work *in toto* probes an unexplored space in the literature.

Information retrieval (IR) has undergone a series of shifts in the evolution of the field. Beginning with boolean search allowing access to literature databases, the dominant paradigm shifted towards ranked retrieval, as typified by Web search engines such as Google, Bing, and Yahoo Search. With the prominence of Web search, users have come to expect a model of search that mirrors this paradigm.

We know however that this form of search is not sufficient for all forms of information seeking. Marchionini [132] outlined a trichotomy of search activities: lookup, learn, and investigate. Of the three activities, only lookup search is well served by our current search systems. Marchionini made the case that learning and investigation could be addressed by systems that encouraged and assisted in exploratory search. In cases such as these, we must

be sure that we are not only providing value to users, but that we are conveying this value in the most effective way so that users understand and trust what the system is doing.

In addition to exploratory search systems, there has been work on providing new ways of ranking [97, 70, 96] and presenting results [85], as well as providing new ways of querying [142, 216]. Systems are in development to trade quality off for time [36], and systems are being developed to, for instance, rank for diversity rather than merely topic relevance [161]. We must ask ourselves as system designers and implementers whether users can develop an understanding of the capabilities of these new types of systems: systems that many users, despite their familiarity with search engines, may not have ever used before.

Marti Hearst has presented examples of uncommon search user interfaces [85] and design principles for effective human-computer in information retrieval. As Hearst notes, information seeking as a process is imprecise. As such, the interface of an IR system should do as much as it can to guide the user in fulfilling his or her information needs. It should be noted that representing an information need—"the current cognitive state of an information seeker, [which is] fluid and constantly changing" [84]—may very well be impossible philosophically [172]. Because relevance is truly dynamic, where what a user needs is affected by intermediate results throughout the process, systems should ideally take this into consideration. This is usually considered in the realm of exploratory search [132], and has received attention in the personalization of search systems [23, 210, 217].

There has been a wealth of research in human-computer interaction on how people build mental models of interactive systems [220, 164, 148, 41]. For the concept of mental models, we may consider the definition by Jonassen and Henning: "...the internal, conceptual, and operational representations that humans develop while interacting with complex systems" [98]. We must understand what in particular about information retrieval systems is distinctive in this respect. There have been numerous publications on mental models in information retrieval [27, 28, 29, 56, 60, 86, 115], but none account for the dynamic aspect of the search process, that is, how a user's mental model changes *as* they learn to use an IR system. Additionally, we must understand how these systems encourage different methods of information seeking. In particular, we are interested in systems that guide users toward an efficient mental model.

Additionally, human-computer interaction researchers have developed a base of literature on the topic of interface design, interaction, and the psychological factors present in the prior two. This is a good starting point for determining how search interfaces may be designed, not necessarily to simply be more transparent, but to effectively build understanding and trust.

Economic models have made a bit of an impact in information retrieval, giving us insight

into optimal interaction and search strategies, as well as decision making and cost evaluation [9, 12, 10]. Such costs may take the form of scrolling [13], clicking on a document [10], examining the document, and/or issuing a query – either new or reformulated [12] We believe that exploring the relationships between economic theory and information retrieval will prove fruitful in this endeavour. Particularly, we are interested in a more efficient and effective learning experience when using search for learning purposes *as well as* reduced interaction costs, which leads to what we will describe as optimal behavior. In situations such as these, although improved learning outcomes my be a desirable objective, approaches that reduce time or interaction costs can allow for users to turn their attention to other learning goals using the time and effort saved. As an example, [186] shows a study in which participants can learn the same amount in half the time, and we may also see other changes such as increased curiosity in a given topic.

This dissertation embraces the interactive nature of search in which a user iteratively negotiates with a system to satisfy their information need explores the resultant behaviors by both system and user in this interchange. As such, it explores the types of mechanisms we have as system designers in positions ranging from interaction designers to engineers to enable users to achieve optimal behavior through interface design to improve users' mental models of the system, system properties through, for example, algorithmic changes, and fundamental changes to the mode of search in a shift from relying solely on a query-response paradigm to, as we introduce in Chapter 3 for example, a dynamic ranking that complements a user's search interaction as they complete their task. This particular chapter revolves around a study published at SIGIR 2016 based on the idea of "slow search" – the notion that a system may be able to "take its time" to process results for better results in certain circumstances. I presented the first working prototype of such a system by introducing a sidebar for asynchronous results in addition to a typical Web search interface, and showed that users' behaviors adapted to the system from session-to-session as they learned the capabilities of the system. We further saw that this incarnation of slow search provided benefits such as reduced session time and better worse-case performance.

Following from this, Chapter 4 seeks to explore the question, "what user outcomes can we expect as we explore various novel interface elements?" To that end, I designed and evaluated a simulation framework to explore fruitful directions for further interactive search user experiments involving such novel interface elements with flexible time and risk tradeoffs. This is based on the simple idea that, in interactive search tasks, users are involved in a cycle of performing actions and receiving feedback in the form of information presented to them by the system. I show, by investigating changes to user and system behavior, characteristics of high and low performance, as well as ways to guide low-performing users toward higher

performance. This work is directly inspired by economic models of search, the notion of "optimal behavior", and the evolution of a user's mental model of system performance.

In Chapter 5, I turn to a specific scenario—vocabulary learning during technical reading—and applied large language models as an analogue for slow search to provide assistance to users in the process. Through building a system that supported contextual question-answering with a sidebar-positioned chatbot affordance, I was able to provide a way for users to ask about the content of articles they were reading, and for definitions of keywords. Through log data analysis and interviews, we could see improvements to prior knowledge and vocabulary knowledge before the study. Taking this direction further, I propose the case for an algorithm that jointly optimises the relevance of documents that a learner searches for, and the particular words that the learner should become familiar with in order to get the most out of each document. This aims to balance a learner's prior knowledge with the educational relevance of a resource and provides a pathway for them to maximise their expected future gain.

The considerations of the tradeoffs of time and risk with quality combined with the employment of novel interface elements should serve as a demonstration to system designers as to the value of exploring these tradeoffs and the circumstances in which provide such value. More to the point, there is still a lot of value that we can provide to users during search tasks, and this dissertation explores interventions to lead users toward such value.

CHAPTER 2

Background

This review of past literature will serve to expand on the major themes of *mental models* in the human-computer interaction (HCI) and information retrieval (IR) fields, how *individual differences* and *patterns of interaction* factor into user behaviour to guide personalisation, and option pricing as a gateway to measuring the value of an interface element – and hence giving us an objective to maximise when leading users towards optimal behaviour.

2.1 Mental Models

Mental models in human-computer interaction generally refer to what a user knows of how he or she can interact with a complex system. Precise definitions however, are numerous. Makri et al. [128] outlined examples, one of which comes from Carroll et al. [41], who define mental models as "knowledge of how the system works, what its components are, how they are related, what the internal processes are, and how they affect components." This forms a somewhat procedural view of these models. Schumacher and Czerwinski [174] categorised these various definitions into at least three classes: as collections of knowledge structures, metaphors and analogies, and as process descriptions [128].

Norman [148] was an early proponent of considering the mental models that designers lead user to construct through their artefacts. He distinguishes a *user's* mental model from the design model of the *system*. The user's model is how a user believes that a system works. This model is dynamic – being constructed through use of a system. The *system image* is what Norman considers the implementation of the designer's conceptual model of the system. The system image forms all the parts of the system that a user perceives, and with which a user is able to interact. This could therefore include documentation, training courses and materials, and error messages. The *design model* is what the designer has in his or her head as they design the system; primarily how the use of a system is perceived by the user. This is where metaphors and analogies can come in, such as the desktop metaphor of

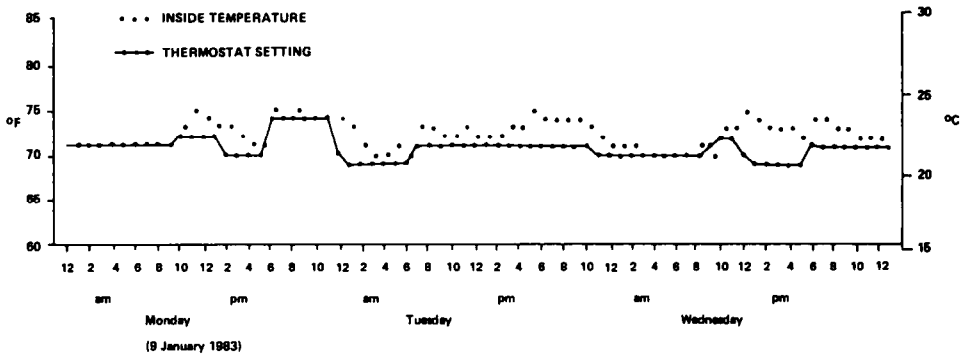


Figure 2.1: Thermostat adjustment patterns consistent with the feedback theory [114].

graphical user interfaces.

There have been numerous studies performed that address mental models directly. It should be noted that mental models are not exclusive to human-*computer* interaction. The system in question may be a thermostat, a calculator, or a sewing machine. As an example, in a influential cognitive science study, Kempton [114] interviewed participants about their home heating and how they set their thermostats, and connected them their conceptions about the mechanisms behind the heating system. The author was interested in the idea of *folk theories*, which he considered to be socially-shared beliefs—acquired through interaction with the world or social interaction, both of which form everyday experiences. He notes that being a "theory" implies some degree of abstraction; knowledge that can be applied in situations that are similar and analogous, and which can predict the result of an action and therefore guide behaviour. The concept of a folk theory stems from cognitive anthropology and thus is more general than a mental model; it encompasses it. Other, related constructs include "naive theory", or "naive problem representation".

In his study, Kempton [114] interviewed users of residential thermostats and collected behavioural data. His interviews included exploratory dialogues with residents of 30 houses in Michigan about energy management and a set of 8 interviews from 12 Michigan residents about thermostat control. He additionally collected automatic thermostat recordings from 26 New Jersey homes. This data was used to compare behavioural patterns between uses with different folk theories of thermostat operation, of which he descovered two folk theories through the course of the study: the feedback theory (patterns of which are shown in Figure 2.1), and the valve theory (shown in Figure 2.2). In the feedback theory, the thermostat turns on and off in response to the temperature of the room. By contrast, the valve theory sees the dial of the thermostat as controlling the rate of heat flow.

Although one, the feedback theory, is closer to how the system does in fact work, Kempton

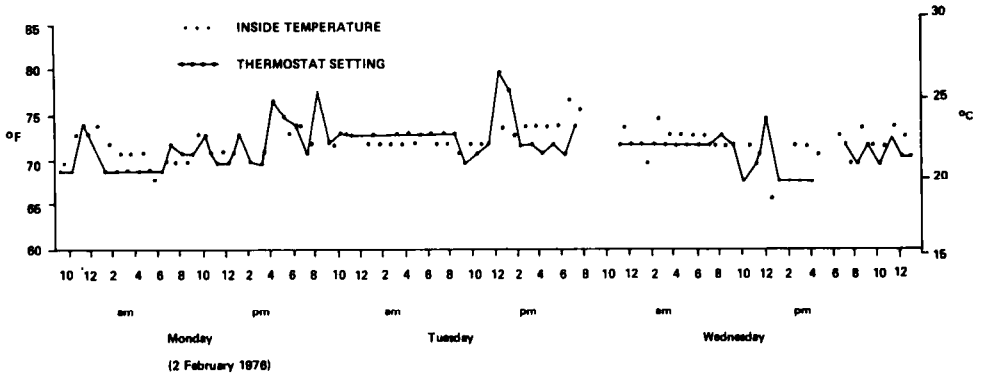


Figure 2.2: Thermostat adjustment patterns consistent with the valve theory [114].

notes that these models are still, for the most part, incomplete [114]. Regardless, the author holds that in examining the thermostat data, the "incorrect" valve theory is still functional; users still make reasonable predictions and decisions of how to operate their thermostats.

In more recent study, Wash [205] used a similar method as Kempton did to assess the common folk theories that non-expert computer users commonly hold about viruses and hackers. Wash [205] here conducted a qualitative study using an iterative process of multiple interviews. Although the study is not generalizable due to a biased, non-random sampling method—which the author notes was intentional for a wider breadth of experiences and demographics—the paper serves to outline the variability of folk models regarding information security. In this case, these models of viruses, hackers, and botnets determine how people will react to, e.g., a security threat, and what they will do to prevent it. For example, someone who has the idea that hackers only target "big fish" might not be concerned, while another who views them more as burglars or criminals believe they could potentially be a target.

A common theme in these two papers is their respective outlooks on models and design. Wash [205] says it best in the following: "[W]hether the folk models are correct or not, technology should be designed to work well with the folk models actually employed by users.". Wash was specifically interested in models of security and individuals' understanding of threats; poor understanding may have serious consequences. Both Kempton and Wash would agree that a user's mental model need not be accurate to be useful or effective. Wash in particular believes that the responsibility falls into the hands of designers as to how a user constructs a mental model through their interactions with the system. Although users can be re-educated to have more correct models, "it is more difficult to re-educate a society than it is to design better technologies" [205]. Designers may try to tell users explicitly what to do, but without an understanding of the rationale, users may ignore the advice if they are

Abstract Machine Arena		Task Arena		Action Arena
"execution cycle" of machine	⟷	basic calculation	⟷	⟨terminated-calculation⟩
function register	⟷	operator	⟷	⟨operation⟩
1st argument	⟷	1st operand	⟷	⟨1st-number⟩
2nd argument	⟷	2nd operand	⟷	⟨2nd-number⟩
execution of computational unit	⟷	evaluation	⟷	"="
result	⟷	value	⟷	answer on display

Figure 2.3: Young's task/action/abstract machine mapping of his basic calculation task for a four-function calculator. Later, he elides the abstract machine domain to focus on tasks and actions [220].

not convinced that it will help. Shaping the construction of mental models without explicit instruction is a topic that we will return to when we discuss persuasion.

In more familiar territory of HCI, Young [220] investigated in a the use of pocket calculators and the *conceptual* models of how they operate. He compared the functionality of three types of calculators (Reverse Polish Notation or RPN, four-function, and algebraic), and attempted to construct conceptual models of all three. The endeavor provided a few key insights, one of which is that we gain confidence in a model when it predicts behaviors not specifically accounted for in the model's construction. These models can be relatively abstract, but must be simple enough to be understood; the assumption that each calculator has a *single* conceptual model that covers the entire range of its usage and the interactions between features is neither realistic nor attainable.

Young was able to construct a relatively simple and elegant model of RPN calculators, but found it difficult to do so for four-function and algebraic calculators with what may seem to be a straightforward framework ("implied register models") due to its unwieldiness in describing their complex interactions [220], and employed Moran's task/action mappings [143] as an alternative. Task/action mappings enables one to analyze the relationship between the tasks carried out by the machine in response to actions performed by the user. It uses the Command Language Grammar, a representational framework that abstracts a system into its Conceptual Component (tasks and semantics), its Communication Component (syntax and interaction), and its Physical Component (spatial layout and device construction).

The author found that by looking at connecting task-level operations to conceptual actions taken by the calculator as well as what the work that the machine actually performs

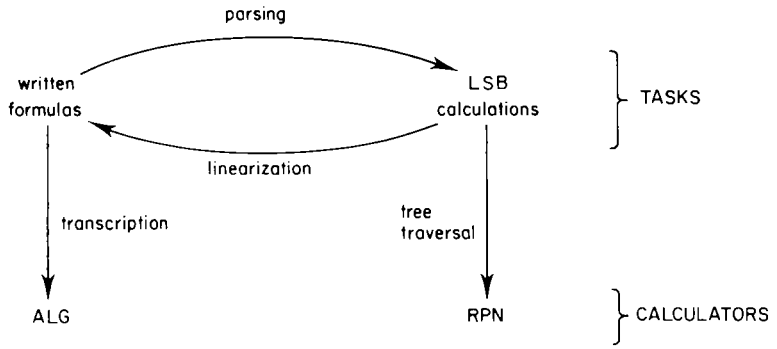


Figure 2.4: Complementary domains of algebraic and RPN calculators [220].

(Figure 2.3), task/action mappings helped to overcome many of the prior difficulties and allowed him to compare and contrast different kinds of tasks and the correspondence between them (see Figure 2.4), but notes that these mappings often represent an idealized, or optimal set of behaviours, and therefore does not capture the full scope of use. Task/action mappings proves useful regardless: we can analyze the complexity of mappings and a comprehensive model is, in any case, typically infeasible. As a final note, task/action mappings here are distinct from other common conceptual mappings; whereas other conceptual mappings predict the machine's process of execution in response to inputs, task/action mappings are intended to capture the behaviours that the user needs to perform in order to get the expected output.

The task/action mappings employed by Young [220] constitute a formal framework for analysing usage, but it is unlikely that users construct such types of models explicitly in their minds. Mental models could be represented in a variety of forms, each of which is difficult to elicit from a user due to the challenge for individuals to articulate the complete picture. A set of studies by Zhang [225, 226, 227, 228] investigated mental models in the realm of information retrieval, in which students were asked in semi-structured interviews about their usage of particular systems, such as a digital library or a Web search engine. As his method of eliciting the users' mental models, he asks his participants to draw a diagram of how the system worked, as well as conducting semi-structured interviews and collecting drawing descriptions [226]. Finally, he combined these into a collective mental model of the Web (Figure 2.5). Through this study, he found that drawing was a sufficiently useful method for identifying the more concrete aspects of a user's model, such as the elements of systems and the relationships between them. However, to determine the abstract features such as matching and ranking mechanisms, it falls short. Conducting interviews is better at this, which makes the two methods complement each other.

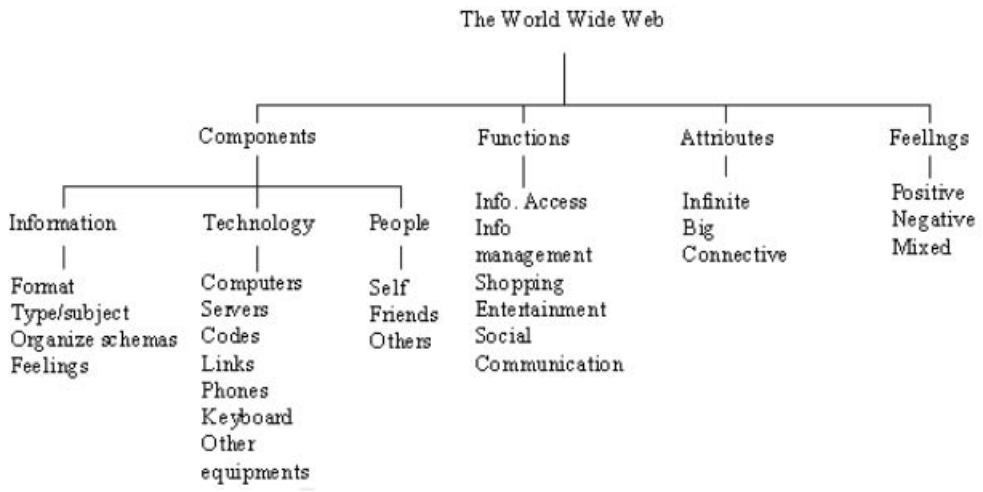


Figure 2.5: Undergraduates' collective mental model of the Web [226].

2.1.1 Mental Models of Search

Zhang (2008) also investigated the relationship between mental models and search behaviour [225]. He elicited mental models through drawings (an example of which is shown in Figure 2.6), which he then categorised as follows:

1. Technical view, where the Web is a composition of computers, servers, modems, and CPUs

2. Functional view, where the Web is viewed as a place for shopping, communication, entertainment, looking for information, or doing research

3. Process view, or search engine-centred view, where the Web is seen as a set of informational branches from search engines

4. Connection view, where the Web is considered the connection between information, people, and devices, and a means for them to communicate

.

By this process, the author was able to analyze mental models without the need to relate them to a complete reference model of the system. He connected these different types of models to user behaviours in the types of movements (that is, visiting a page via a hyperlink or URL in the address bar), backtrackings (use of the "Back" button), and query reformulations. In this study however, there were no significant differences in interaction or query formulation, feelings of difficulty or satisfaction with their performance, and no

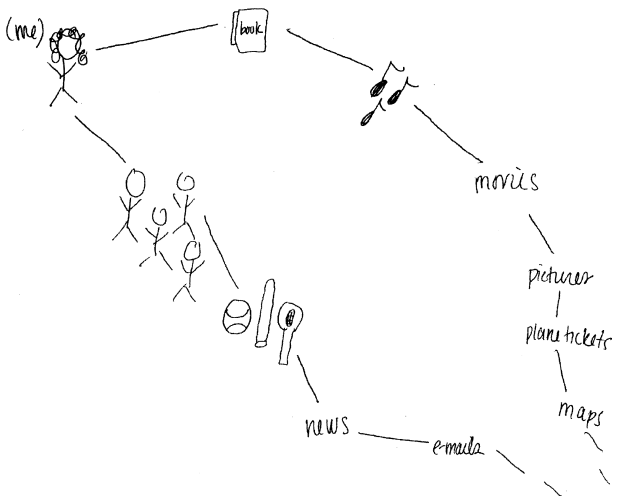
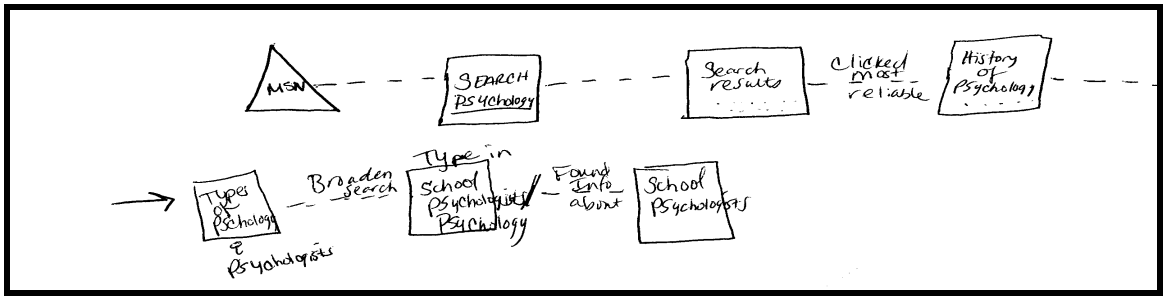


Figure 2.6: One user's functional view of the Web [225].

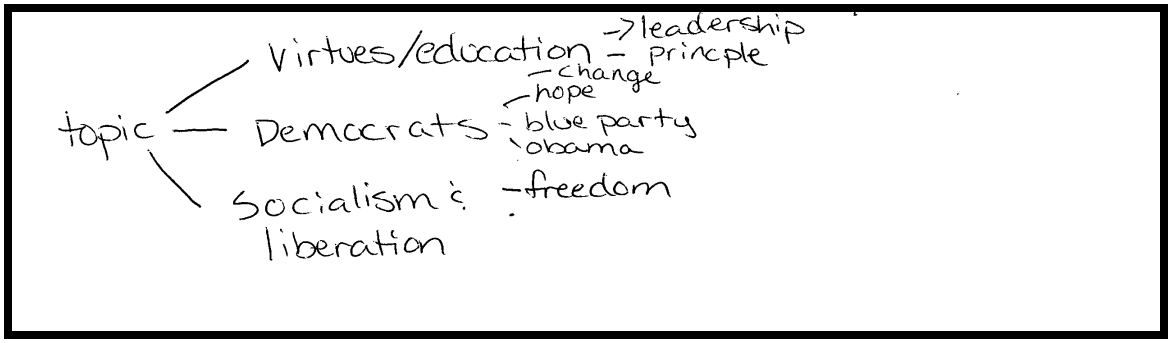
consistently repeated search patterns were identified. He however did see a slight difference in the number of movements—users with connection views, where the Web is seen as a highly connected network of entities, make the most movements.

A comparable study to Zhang's [225] is one by Holman [86], in which the author used a combination of contextual inquiry and concept mapping to analyze millenial students' understanding of how search tools generated a list of results. Millenials have lived their entire lives in the information age, but it is unclear whether this corresponds to the "Google generation" having good mental models of search, despite the fact that they may conduct Web searches on a regular basis. The author also asked the participants to draw a diagram as Zhang [225] did, but of the relationship between the search tools and results, and interviewed each participant. Three categories emerged from this method: a *process view*, a *hierarchical view*, and a *network view* (Figure 2.7). The process view is seen as the simplest, which characterises the search engine as a black box. There was no significance testing done for the analysis, but the author notes that those with more sophisticated models (such as the network view) performed the most searches and used advanced search features. In comparison, users with a process view used only the basic search features.

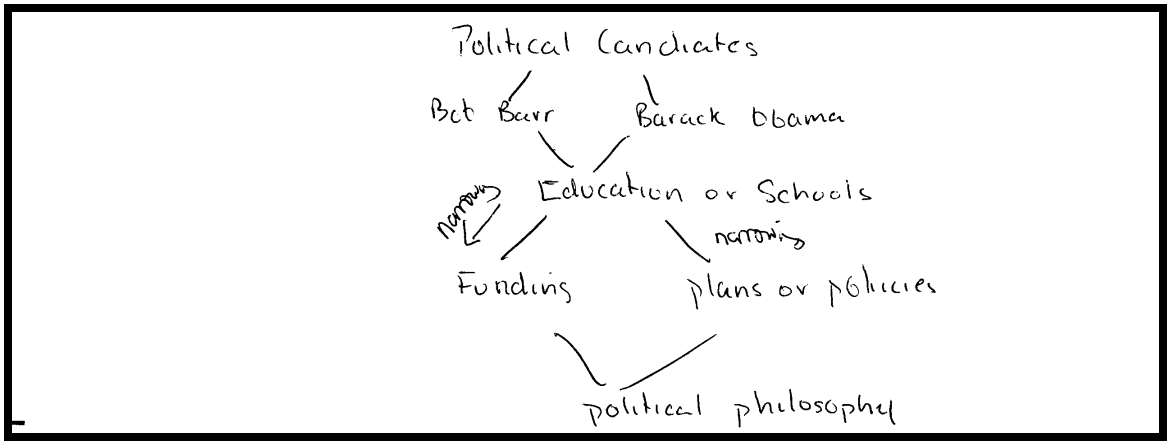
Borgman (1985), in an early study of mental models of an information retrieval system [27], asked participants to perform both simple and complex tasks, as determined by the use of indexes and boolean operators, after "training" them with conceptual models. Borgman here defines the conceptual model as what is portrayed by the designer through the design of the system. This is distinct from a user's mental model, which lives in his or her head; a mental model is built from the conceptual model, but may have differences. In her study,



(a)



(b)



(c)

Figure 2.7: Examples of a process view (2.7a), a hierarchical view (2.7b), and a network view (2.7c) [86].

the "model group" received an introductory narrative in the form of an analogy to traditional card catalogues. The "procedural group" got a narrative of background operational information in the form of the IR system's manual. The author also elicited models from the user, and found that these models were typically abstract, with no resemblance to a card catalogue. Previous studies of human-computer interaction [42, 81, 69] argued that model-based instruction was more effective than procedural training, but Borgman's results were inconclusive; she observed a difference as predicted in the conditions that were suggestive, but not statistically significant.

Through her study, Borgman (1985) uncovered a few issues—one of which concerned the fidelity of the models given by users. Because they must articulate these models, the question is left of whether differences in the models were due to how well the participants articulated them. The results also indicated that subjects constructed two types of models: one for the task, and one for the system, where the task model is built first and the system model is built subsequently through usage and practice.

Borgman (1985) also raised an issue of the effects of individual differences. The results showed that there were differences between majors in who passed the benchmark tests—namely between those in STEM majors – who fared better – and those studying the Humanities. She posited that these groups have different cognitive styles, and continued to investigate these individual differences in a followup study [28], which will be elaborated on in Section 2.2.

As Borgman [28] did, Zhang and Chignell [224] also performed a study at the intersection between mental models and individual differences, in this case, examining their interaction. To elicit users' mental models, the authors used a psychological instrument called the Repertory Grid Technique, or RGT [71]. RGT is based on the theory that people understand the world through personal constructs—their set of representations/model of the world formed through one's social experience [113]. The process involves the generation of concepts and the various attributes about them. Principal component analysis can then be employed to compare the constituent factors of these constructs. This comparability is one advantage of the method; another is that it does not rely on the participant describing their model verbally, which is typically a challenge for subjects.

Zhang and Chignell's aforementioned study [224] study involved 64 subjects across spectrums of academic levels (high school, college, graduate, professional), first language, discipline (STEM or Humanities) and computer experience. The analysis found differences between groups depending on education level, academic discipline, and computer experience. Individually, each factor had a suggestive or statistically significant effect, but there were no significant effects from interactions *between* these variables. The results focused on

four factors: 1. the purposefulness of querying (targeted/untargeted), 2. the applicability of data organization to other information systems outside of information retrieval, 3. the function of querying as a form or process, and 4. the function of the structure of data. This supports the idea of individual differences affecting the generation of mental models, though the authors were careful to point out the novelty of the method and the exploratory nature of the study.

2.1.2 Cognition and Personality

Turner and Sobolewska [195] published a related study that examined both mental models and individual differences. It took a more emotional approach to mental models using EQ (empathy quotient), which they contrasted with the more common approach of investigating from a purely cognitive perspective. The work of Reeves and Nass [162] highlighted the degree to which users anthropomorphise technology, and this determines how we view, use, and react to these artefacts. Referencing work in Psychology by Baron-Cohen [15, 16, 17, 18], the authors looked at the dichotomy between *systemizing* and *empathizing* cognition. The prior is associated with the common behaviour of analyzing and exploring sytems, whereas the latter leverages theory of mind [156]; this may be because computers seem autonomous, complex in purpose, adaptable and unpredictable, which are all human-like qualities. The gist indicates that people tend to use some amount of both cognitive styles, and the degree to which they do can be measured. In this interpretivist study, Turner and Sobolewska found that this individual difference in SQ (systematizing quotient) and EQ related to the extent to which users gave longer, more technical answers (high SQ), or the degree to which users ascribed agency to their technology (high EQ) [195]. This may indicate that high SQ subjects had a higher depth of model fidelity than high EQ subjects.

On this issue of mental model fidelity, Khoo and Hall [115] performed a study that considered the ways in which individuals thought about digital libraries through use. The authors analyzed the issues that participants reported having with regards to the interface, navigation, and the ease of use of searching and browsing. They found that subjects held Web search engines as a baseline, and therefore, when compared to a digital library (which, as seen in [128], users tend to compare with search engines rather than traditional libraries), the digital library was lacking. The paper also presented an interesting phenomenon: participants seemed to attribute features that they would like to be added to another system such as a particular search engine, despite the fact that this search engine itself may not the features in question. This strengthens our belief that users are likely to have incomplete models of the systems they use.

Another, previously mentioned paper by Makri et al. [128] explored mental models of digital libraries. The method of data collection here comprised interviews and observations with eight participants to elicit mental models of the usage of both traditional and digital libraries. The authors asked participants—masters students—to find relevant documents using a traditional library, digital library, or another library of their choice. They found that users' perceptions shaped how they approached each system in terms of what a digital library is in relation to what they believe a traditional library or search engine to "be". The authors perhaps expected users to exploit their knowledge of traditional libraries, but instead found that users made analogies to common Web search engines like Google. Although students were able to articulate the differences in the ways in which each system may be used from a procedural or task-oriented standpoint, few drew explicit comparisons of the various systems on their own. This led to users being unable to take advantage of all the features or capabilities of the digital libraries, in the study, such as browsing (which participants ascribed only to traditional libraries).

Taking a more mechanistic approach by contrast, which in this case can be seen as also assisting users in mental model construction, Muramatsu and Pratt [144] conducted a user study in which they provided users with feedback within the search interface to explain the query transformations used for a given request, such as stemming, term closeness scoring, or automatic boolean operations. Users were often unaware that these transformations were being done, or knew why they were being done. This feedback gave users more transparency into the workings of the system. The study participants indicated that suggestions for how to reformulate queries would have been helpful—applying the model is still an important step that may be difficult for users to do. The authors did not investigate the effect of model accuracy on retrieval performance, but other studies in the literature explore the model/behaviour relationship [29, 225, 27].

Commonly, the majority of these papers address the construction of mental models from a cognitive perspective. However, a study by Turner and Sobolewska [195] approached the topic from an emotional angle. Affective computing is a burgeoning field, but for our purposes we are more concerned with cognition. However, these studies were focussed on the initial construction of a mental model; we are at present interested not only in the initial construction, but also subsequent alterations and reconstructions that come through repeated use. Learning is continuous and dynamic, and it seems reasonable to expect systems to regard it as such. Systems may guide users towards the optimal path of usage (discussed further in Section 2.3). Many of these model elicitation methods are rather invasive: they involve asking users to articulate their models, draw their models, or taking questionnaires and participating in interviews. More appropriately, a system should be less invasive in

learning what kinds, classes, or levels of fidelity of models exist in the minds of users. The question of how to represent these models appropriately for machines to use is also important to answer. Considering the economics of search may give clues as to how this could be done. This will be covered in Section 2.3.

We are interested in methods of helping users form more useful mental models. We can approach this as a reversal of the *machine teaching* problem, where the goal is to find an optimal training set to provide to a machine learning algorithm in order to learn a particular target model [230]. This is necessary if we cannot simply make use of the model, which is the case for a human learner. Machine teaching assumes a cognitive model of the learner which, given the optimal lesson, can be adjusted toward the desired cognitive state.

Socially Guided Machine Learning [190] is an approach that addresses looks at machine learning in a similar manner as machine teaching. In other words, it takes an interactive perspective for machine learning, where the task is designed as a partnership between the human teacher and machine learner. Thomaz and Breazeal [191] performed a study with a simulated environment they dubbed "Sophie's World". They modelled a problem in this world–to bake a cake with provided ingredients– as a Markov Decision Process to be trained. In a reinforcement learning paradigm, individuals provided feedback to the machine based on the actions it takes.

The issue raised by Borgman [28] – that mental models and individual differences are closely entwined – is an important one. Learning style, academic interests, technical aptitude, and personality traits are all potential factors that may influence the development of mental models. We will therefore turn our attention to individual differences, and their connections to mental models and behaviour.

2.2 Individual Differences

The study by Borgman [27], covered in Section 2.1, had results that indicated an effect stemming from individual differences. She noted that if there are differences in performance due to individual differences, then these systems being designed and used might not be *equitable*. Hence, in a followup study [28], Borgman further investigated these effects by examining the link between academic interest, information seeking style; and personality traits, technical aptitude and reasoning ability. The author thus hypothesized that technical aptitude and personality traits determined academic orientation, which then intermediates the aptitude for performing information retrieval tasks (Figure 2.8).

The author measured the results of a number of tests, including the SAT, symbolic reasoning test or SRT [173, 168], the number of math, science, and programming courses taken,

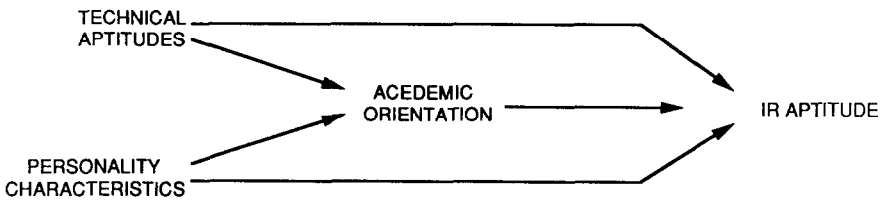


Figure 2.8: Hypothetical model of information retrieval aptitude proposed by Borgman [28].

and personality indicators such as the Myers-Briggs Type Indicator or MBTI [145], and the Kolb learning style inventory [119]. She found that choice of major could be explained by personality tests alone, and could be predicted by the number of prior courses taken along with aptitude test scores (SRT, SAT). There was thus some clustering by technical aptitude and personality characteristics, but these two were independent of each other. She also found that the effect was stronger for those who stuck with one major than for those who switched.

Ford et al. [68] performed a related study to investigate the relationship between cognitive style and information seeking behaviour in particular among one hundred and eleven subjects from various disciplines. This study considered a variety of factors, including global/analytic style, the problem stage, uncertainty, and task complexity, among many others. The subjects were asked to complete literature search tasks, while the battery of aforementioned factors were measured. The analysis focused on a holist/serialist (or alternatively by correlation, global/local) learning style. This is referred to as *field dependence*. Field dependent individuals prefer to take more of a "spectator" approach to learning, where the process is structured and analyzed for them. By contrast, field independent people are more effective at structuring their analytic activities on their own. Field independent individuals are also less socially-oriented than those who are field dependent. The analysis found that field-independent individuals are indeed more focused and analytic as well as more active in their behaviour. It also found that holists were more exploratory, taking a more comprehensive approach to search, supporting the idea that individual differences—here, of cognitive style—can be a determinant of search behaviour. Field dependence is also considered in numerous other studies in IR and HCI, amongst other fields such as psychology in investigating differences in gender [40] and culture [133, 147].

Field dependence is a major aspect of cognitive style, and may also be viewed as global/analytical style. The authors of a paper on the impact of thinking style on Web search strategies [105] asserted that incorporating the consideration of thinking style into the design of IR systems would help to predict user intention and give individuals a better basis for comprehending search results. In this context, thinking style refers to one's *personal*

preferences in how he or she solves problems. This is distinct from personal *abilities*; two individuals with similar abilities may have very different preferences in when and how to use them. This paper thus includes a study to determine whether a specific thinking style emerges over time when searching, as it does for other daily tasks.

Their study consisted of three hundred and fifty five Taiwanese 5th grade students. All the participants had two years of computer training, and their thinking styles (categorized as *global* or *local*) were determined by questionnaire. For their search tasks, they were asked to write down their target query terms and revise them as their intentions changed. Participants also had their interactions recorded and formatted into navigation flow maps to show the relationships between queries, documents, and tasks. This allowed the authors to examine the number of keywords, visited pages, depth of exploration, revisited pages, and frequency of query refinement. Their analysis found differences in interaction: more local thinking styles correlated with more in-depth understanding and answer refinement. In comparison, more global styles correlated to more high-level exploration.

From these results, it seems reasonable to believe that we may observe distinctions in interaction behaviors that would reflect individual differences. In a large-scale study of individual behavioural differences from search logs [38], Buscher et al. analyzed the mouse movements that users performed on the results page of a large commercial search engine for 1.8 million queries. They extracted clusters of behaviour and related these clusters to the results of other smaller-scale studies. They also considered the effect of the task (navigational vs. non-navigational) on these clusters. Analysing the clusters, they found three clusters of search behaviour for non-navigational tasks: 1. Economic (fast, focused movements, little time on SERP, few clicks) 2. Exhaustive-Active (detailed examination, many clicks, infrequently abandon searches), and 3. Exhaustive-Passive (similar to Exhaustive-Active, but spends more time on SERP, and idle and abandon more searches).

The extent to which demographic factors as the individual differences in question affects search behaviour has also been studied. Weber and Castillo [206] specifically investigated the effect of age, race, gender, and income on predicting which URLs would be clicked on for a given query. They also had other objectives and applications in mind, such as targeted advertising, improved query suggestion, and more relevant news article recommendations. The authors performed an analysis of 28 million users with Yahoo! accounts that had demographic information filled in, with (Query, URL) pairs, (URL, Query) pairs, and (Query Term 1, Query Term 2) (i.e., bigram) pairs. They found differences in behaviour regarding the more distinctive queries that the different groups issued, and differences in behaviour such as the relationship between education level and query length, click entropy, and URL depth.

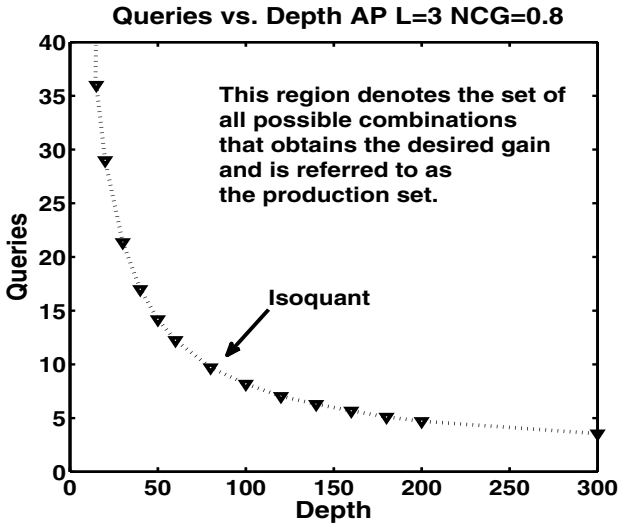


Figure 2.9: An example plot of queries vs. depth on a particular document collection and ranking algorithm. The labelled isoquant is the minimum amount of the inputs that produce the given level of gain [9].

From our exploration, we can see that the construction of mental models is closely related to an individual's cognitive style. Recent work in information retrieval has looked at cognition and decision-making from an economic perspective, which we shall explore next in Section 2.3.

2.3 Economic Models of Interaction

In 2011, an influential paper by Azzopardi [9] established Search Economic Theory (SET) to explain interaction with information retrieval systems. With analogies to Production Theory, which models a firm's outputs from its inputs, Azzopardi modelled search strategies as a combination of inputs (Q, D) for a query of length L, where Q is the number of queries to be issued, and D is depth, or the number of documents to examine. Azzopardi then formulated a production function to determine the maximum cumulative gain possible from employing this strategy. This forms the basis of our investigation into what we consider to be optimal behaviour.

Influenced by SET and information foraging theory [154], Maxwell and Azzopardi [136] performed an experiment to determine the effect of delays on behaviour in interactive information retrieval. Driven by this overarching research question, they tested hypotheses that document download delays and query response delays would increase the time spent on the documents and search engine result pages respectively. They also examined the effect that

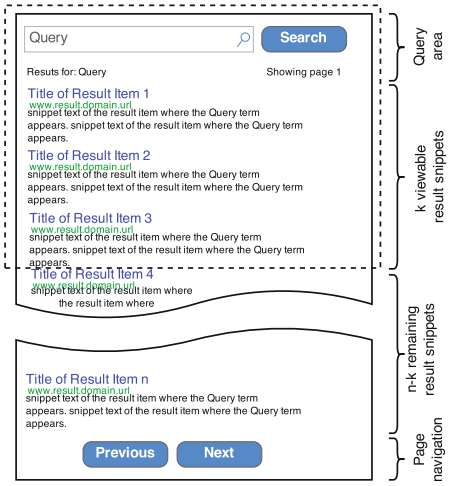


Figure 2.10: Cost-based choices in search engine result page layout portrayed visually. The dotted area is the visible portion of the page on load [13].

the increased cost of document access (i.e., download delays) had on the number of queries issued and number of documents examined. In the study, the authors asked university students to assume the role of a journalist, and to find as many documents relevant to a given topic as possible within a 20 minute timespan. They found that subjects spent longer on documents and on result pages due to an increase in the sum of document, query delays, and query formulation times. Additionally, with increased document access times, subjects performed fewer queries and examined more documents per query. However, the time of the delays was not a variable that was considered; they conjectured that 2–4 seconds was the 'tipping point' of where behaviour changes, but this claim was not investigated. Additionally, the subjects were given information needs explicitly; the effect of delays on users' own information needs was not explored. This type of cost-based analysis can help us to determine optimal choices more broadly in information seeking and retrieval.

In [13], the authors established a cost model of browsing search engine result pages based on estimates of the time required by both the system and user for clicking, scrolling, and inspecting snippets, while taking into account the size of the page, and the size of the screen used. The motivating scenario involves a user being presented with a search engine result page immediately after issuing a query. Optimally, for a given device, we may consider whether the interface should optimally show as many results as can fit above the fold with pagination, or should it allow for some scrolling before going to the next page. We can see the scenario more clearly in Figure 2.10. This paper follows in a line of other endeavours to model *cost*, as opposed to the more common avenue of optimising *gain*. Kashyap et

al. [106] formulated a cost model of faceted navigation for a system called FACeTOR that accounts for the time to examine results, the cost of choosing a facet and hence refining the list of results, or expands an attribute, revealing more facets of the particular attribute. Through simulated navigation and a user study via Amazon Mechanical Turk, they tested the predictions of their model and found that their cost model was realistic. The cost model of Russell et al. [170] served to inform the activity of sensemaking – a task more general than information retrieval. The authors decomposed the process into different types of subclasses, and characterized their costs. By doing so, they made the point that by trading off costs in one task, we can take advantage of the reduced costs in other aspects. As such, sensemaking becomes an anytime algorithm [6]. As an example, by saving time expenses from automated clustering methods, the designers of an educational course were able to extend the comprehensiveness of their search.

For examples of work which examine *gain* in information retrieval, we may turn towards analyses conducted by Smucker and Clarke [181]. Following on their proposition of time-biased gain, where the gain from a relevant document is equal to the probability of viewing it subjected to a time decay [181], Smucker and Clarke [180] explored a simulation-based approach to approximate this gain as an alternative to estimation using their closed-form solution. This allows for more flexibility in analysis – one can model a distribution of gain while changing other variable with less effort for easier "what-if" experiments. Taking this approach also potentially allows one to model a sample from a population of users.

A number of papers have also taken a behavioural economics approach towards search. In examining search persistence and failure, Mansourian and Ford [129] analyzed interviews of academic members of staff to ascertain their perceptions about missing information – drawing a direct connection to Simon's concept of bounded rationality [130]. March and Simon [130] outlined the problems with "classical" rationality, and instead reformulated rationality as being relative to a frame of reference. This subjective viewpoint is more realistic than the classical, objective viewpoint in that it does not presuppose the attainment of complete information with regards to alternatives, consequences, and utilities that are needed for decision making. Mansourian and Ford [129] coded their interview transcripts with a coding scheme that accounts for the various aspects of both bounded rationality and satisficing [130], which we have diagrammed in Figure 2.11.

From their data, it was evident that many participants often stopped their searches before they felt truly satisfied that they had found all of the information relevant to their needs. As far as perceptions go, the authors proposed the following impressions:

- The "Inconsequential" Zone: the best case scenario (Low volume; low importance)

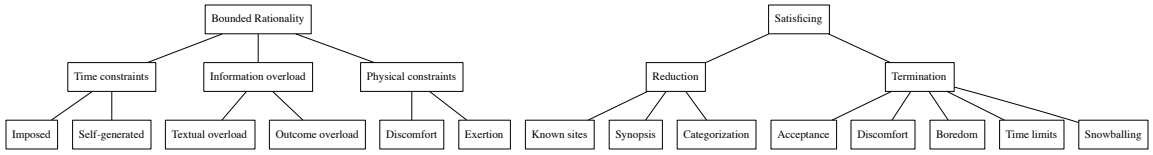


Figure 2.11: A visual representation of Mansourian and Ford's (2007) bounded rationality and satificing coding scheme [129].

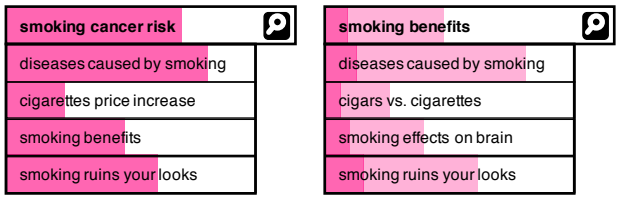
- The "Tolerable" Zone: when missing the relevant information does not lead to search failure (High volume; low importance)
- The "Damaging" Zone: for high-recall situations where users are more concerned about the relevance of the information than the quantity (Low volume; high importance)
- The "Disastrous" Zone: the worst case scenario, when users might miss a large amount of important information, leading to search failure (High volume; high importance)

The paper also points out a number of satisficing search strategies that were observed in the data, which fall along the continuums of effort and the perceived extent of missed information. As per [4], these strategies were categorized as follows:

- Reducing the search task
- Categorization of types of searches and resources
- Formulating synopses, rather than comprehensive consumption
- Termination of search, from acceptance, discomfort time limits, boredom, or snowballing (where a users reach a fixed point in the information they find)

This study by Mansourian and Ford [129] did not address the ways in which interviewees arrived at their perceptions of the quantity of missed information. To address the question of how users might arrive at this estimate of missed information, Umemoto et al. [197] developed an interface called ScentBar, that visualizes the intrinsically diverse aspects of a query, and shows how much information in each aspect a user could potentially miss at any given time, were they to stop their search (Figure 2.12).

Their system adds an interface element that contains a horizontal bar chart, where each bar represents an aspect, and the amount filled in with two different shades represents the total amount of information in each aspect, and the amount of *unexplored* information in those



(a) At beginning of search task (b) At end of search task

Figure 2.12: The design of ScentBar [197].

aspects. The authors were primarily focused on stopping behaviour, and how introducing this kind of feedback would affect it. They formulated a gain metric that satisfies the need for importance, relevance, and novelty, and define missed information as the additional gain from results that are unexamined. In their experimental study, participants were asked to perform an exhaustive search on various aspects of a topic, with no time limit. They found that participants missed significantly less information before stopping when using ScentBar and formulated more expansive query strategies that allowed them to find more information than they would have otherwise missed without feedback.

What *governs* stopping behaviour is another subject of study. Zach [222] conducted a study in which twelve arts administrators were interviewed, and a model of information acquistion using the collected data was constructed. These administrators viewed information seeking as non-explicit; it was a means to making a decision and as such was thought of as an auxiliary process. With the process not being explicitly considered, this may give support to the idea of search as not necessarily constituting a set of rational choices. In their interviews, the administrators indicated that their stopping criteria were never determined beforehand, but was the result of feeling comfortable with the results they gathered. When time and comfort were at odds, participants often satisficed. The primary factor that determined stopping behaviour therefore was their confidence in being able to address their primary task.

A factor related to satisficing that affects decision making relates to the paradox of choice [175]. This paradox in essence states that if one is provided with more choices where each option is highly relevant and the individual perceives success from making the correct choice, it is likely that the individual will make a poorer decision and will have reduced satisfaction. In [150], the authors demonstrated in an experiment that when given a list of search results, those who were given a larger result set were in fact less satisfied and less confident in their choices. This is a finding that is consistent with prospect theory—introduced by Kahneman and Tversky [102].

Prospect theory was proposed as a critique of the prevalent conceptualization of decisions

being motivated by expected utility. By the expected utility hypothesis, risks are modelled as a multinomial distribution ("lotteries"), where each alternative is one in a set of discrete outcomes. Here, the preferred alternative is the one with a higher expected utility value than another. Expected utility theory is also subject to two axioms: namely that of independence and continuity, which establish that preference orderings are not affected by either a convex combination with an additional lottery, or by small changes in probabilities. Similar to lotteries, Kahneman and Tversky [102] introduced "prospects", expanded in [196] as having their expected value being rescaled by functions $\pi(p)$ and $v(x)$. These two functions reflect a subjective appreciation and subjective value to the outcomes. Guided by psychology, prospect theory considers phenomena such as loss aversion, where the subjective pain of a loss is greater than the pleasure of a corresponding gain in value. In the [196] expansion, the authors also considered cumulative decision weights as well as prospects with any possible number of outcomes.

Prospect theory allows us to analyze and predict decision-making under risk. The nature of the information at our disposal can influence the actions of users, and the lack of information about uncertainty can be misleading. In prior work, Joslyn and LeClerc [99] argued that despite the fact that it is often difficult for non-experts to make use of uncertainty estimates in the decision-making process, the presence of specific numerical uncertainty estimates lead to more optimal decisions that point estimates. In [100], Joslyn and LeClerc demonstrated that when students were given the scenario of making daily decisions to salt roads during winter months, their decisions were closer to the optimum than participants when given the probability of freezing in addition to a single-value forecast. It should be noted however that the authors also found that the decisions made were not necessarily rational; all participants tended to be more risk averse in comparison to an optimal strategy.

With the benefit of information on uncertainty in mind, Kay et al. [109] proposed a space-efficient visualisation for arrival times in mobile transit apps called a quantile dotplot. This tactic discretises a probability distribution, which they find is easier to reason about, by drawing randomly from the inverse cumulative density function (CDF). The result is less noisy than discretising the density by taking random draws. In an experiment comparing different types of visualisations, they demonstrated that less granular discrete plots performed better than finer-grained discrete plots, which in turn performed better than a continuous density plot. An earlier study by Kay et al. also demonstrated in an earlier work that individuals trust measurements more when they are also provided with information about uncertainty [110]. Giving a point estimate can give the impression of being more precise than it actually is.

Further investigations into the methods and limitations of uncertainty visualization in-

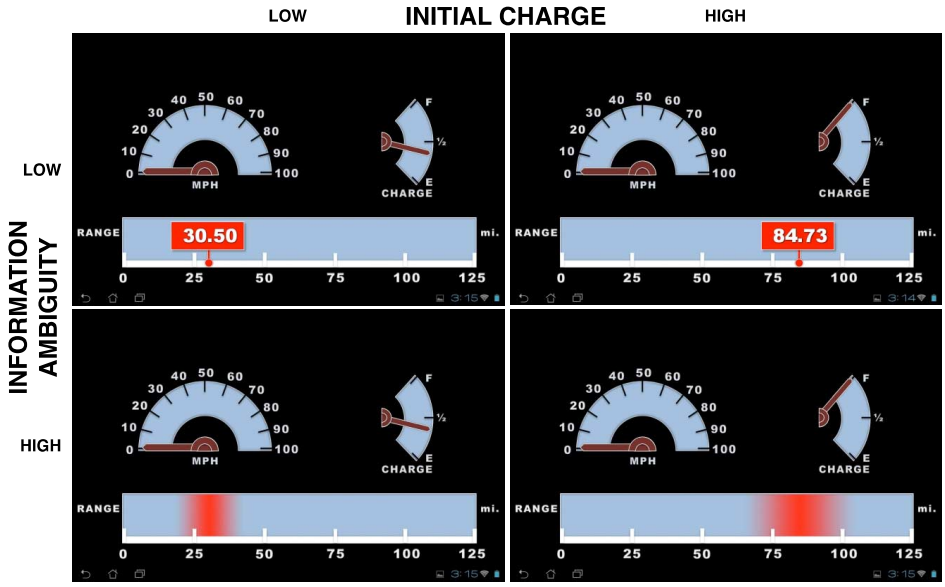


Figure 2.13: The instruments of investigation by Jung et al. [101]

clude [75], [91], [101], [51], [79], [92], and [19]. Greis et al. [75] distributed an online questionnaire where user preferences towards visual representations are elicited, and an experiment involving a turn-based farming game, where a weather forcast is given for the next three days using these visual representations. After the experiment, the players' decisions were compared, and participants viewing a probability density function made the most optimal decisions. This representation also received the most selections in terms of preference.

An experiment by Jung et al. [101] investigated the role of precision in reflecting the state of the current charge of an electric vehicle. The authors asked participants to drive in a car in which the instruments were covered with a tablet showing gauges designed by the experimenters to present different levels of information ambiguity through the fuzziness of the colour band indicating the remaining range during the experiment in either a condition of high or low initial charge. Qualitatively, the authors measured participants' driving experience and trust towards the vehicle, as well as the quantitative efficiency of their driving. They found that although the less ambiguous display led users to perceive higher information accuracy, there was evidence that there was lower range anxiety and more trust towards vehicles with more ambiguous displays.

The limitations of including uncertainty in visualizations must also be noted. Greis et al. [75] notes that the usual method of conveying quantitative information through probabilities can lead even well-educated adults toward difficulty in problem-solving. Alternatively, qualitative information, or even the framing of the uncertainty (negative vs. positive) can be misleading or bias decision-making. Hullman [91] goes further, identifying problems that

come with evaluating different representations. The complexity of the psychology of uncertainty leads to noise and bias in data collection without enough care. The work by Hullman [91] gives suggestions for mitigation.

Within other realms of decision making, Kelly and Azzopardi [112] examined the behavioural effects of varying the number of results per page, and also measured users' experiences using a modified Search Self-Efficacy scale. In terms of interaction, the results showed that those exposed to fewer search results per page viewed more search result pages, but viewed significantly fewer documents than those exposed to more results per page. There were no significant differences between users' perceptions of difficulty or success, but those in the condition with three results per page reported their tasks as less difficult, and felt they were more successful. This is in concordance with the results from Oulasvirta et al. [150] which investigated the paradox of choice.

In a paper by Yilmaz et al. [219], the authors posited that the traditional relevance judgements used in information retrieval evaluation do not fully capture the *utility* that a document has to a real user. The authors argue, rather, that extracting this utility may take a great deal of effort on the part of a user, whereas judges of relevance for collection-based evaluation is not only trained to perform the judgement, but is also more patient in carrying out this evaluation. Consequently, when users cannot see the immediate relevance of the document, or that it requires too much effort to evaluate, these users often give up to move to another document. In some sense, this could potentially constitute missed information. The authors proposed a two-stage model of user behaviour in this respect: first, users perform a quick judgement to determine if there's any value present and how much time and effort it takes to make use of it. Secondly, users who expect the get use out of the document will then commit that effort to do so.

Also of interest is *switching behaviour*, which can occur due to a number of reasons including dissatisfaction with search results, the need for broader coverage, or for verification. We can view switching behaviour as related to stopping behaviour, but rather than being the result of satisfaction, it is the result of dissatisfaction. Guo et al. [80] collected users' reasons for switching *in situ* with a browser extension called *SwitchWatch*. On occasions when users switch between one of Google, Yahoo!, or Bing, the add-on presents a questionnaire that asks whether users are interested in finding the same information on both search engines, and their reason for the switch. Dissatisfaction constituted the majority of the switching occasions (57%), whether due to expecting better results, frustration, or simply not being satisfied with the results in general. Next most common was for additional coverage and verification (26%), and search engine preferences (12%).

However, we can imagine a situation in which the primary cost is time, rather than effort.

We can see this in question-and-answer scenarios. Aperjis et al. [5] performed an analysis on postings on Yahoo! Answers and found that users are willing to wait a longer period of time after initially receiving a few answers to their question. In estimating the probability that a user closes his or her question given the number of answers and time between answers, the probability of closing the question increases with these variables. There is a tradeoff at play: the user wishes to get a high-quality answer, but does not want to wait too long. Using the data, they estimate the parameters of a concave utility function.

An economic model is not an end in and of itself. It should be informative and predictive. In the case of information retrieval, what we aim for here is insight into user behaviour; in our case to modify it towards higher value usage. One option towards modification is *nudging*, which I took as the primary mechanism to be employed in Chapter 4 and now discuss below.

2.3.1 Nudging

With an interactive system such as a search engine, we can adapt the system's behaviour in response to a user's behaviour. A system can therefore present information or change its interface in response to a user's actions or inferred goals. A popular example of this principle in production software was seen in Microsoft Office, which has experimented with adaptive menus to cater to different levels of user expertise [140], and its Office Assistant – character-based intelligent agent that offers tips and help based on a user's background, actions, and queries [87].

The Lumière Project [88] served as the basis for Microsoft's Office Assistant. Among its contributions, it applied Bayesian networks to construct user profiles of behaviour and expertise, and reasoned about a user's goals and needs based on their actions within the system. Goals here, are target tasks or subtasks that form the basis of a user's attention. In contrast, needs are the information or steps that are required to achieve the user's goals. Describing Lumière's integration into Microsoft Excel, the authors demonstrated the system's ability to reason temporally about a user's actions, and goals, and likelihood that the user requires assistance, and makes recommendations to help in performing actions *in situ* and in more broadly topics which would assist for longer-term needs and goals. Although character-based assistive agents have fallen out of favour, the Bayesian User Model formulation and its use in reasoning and decision making about user needs is useful for decision theory and preference elicitation [176, 223].

We have also seen active assistance at work in studies on the benefits of ambient information interfaces and augmented reality. In one such study [104], the authors present a system

under the assumption that it is better to provide just enough information in the right way to facilitate good choices than it is to simply provide *more* information. An activity that is commonly subject to information overload is that of shopping at the supermarket in aisles of similar products. The authors consider the information visualisation and design aspects of representing multiple dimensions of products in a way that is easy to grasp at a glance, enabling quick comparisons and easy inference [193]. As an example, they present a design for an in-store trolley interface that visualises a shopper's current progress towards their goal of, for instance, buying healthier foods with lower fat or buying more locally-grown produce. The trolley here is made to project a running overall score onto its handle to signify the shopper's adherence to their goal. In a user study, they asked 18 participants to shop with either an ordinary shopping cart or with the same cart with a clip-on apparatus that lights up a series of LEDs that indicates food mileage – how "local" the item was, as well as whether the item was organic or conventially grown. They found that when using the augmented shopping cart, 72% of the products selected had lower mean food mileage than those selected by participants using the ordinary shopping cart. The authors make the case for salience rather than recommendation in order to not disrupt a shopper's experience.

In [157], the authors characterise interaction as a collection of input methods and system responses that form an interactive prospect, subject to loss aversion. Interactions in the service of a task under a goal are assigned some utility by the user, which is based on psychological perception. Following this, one can assume that one interaction will be preferred to another if it has higher utility. The system's response to a user's action making up the interaction can be considered a multi-dimensional consumption bundle, where dimensions might include the visual feedback of a character being typed on screen, an audible "click" sound, and a vibration being produced. Loss aversion comes in to play as an extension of prospect theory, however, whereas economic prospects such as monetary investments are easy to manipulate on the experimenter side and reason about on the subject side, interactions as prospects are more difficult to handle. The costs and benefits of interactions are multifaceted, and may involve dimensions such as time, physical effort, and cognitive overhead. This paper analyses text selection methods as prospects, comparing a letter-by-letter technique, and a word-by-word technique where the selection is snapped to word boundaries. Subjects' preferences were observed and given an option to disable snapping behaviour, where the mechanism results in a loss of character progress. The authors found that there was an asymmetry between the subjective losses and gains in the interaction, with losses being more pronounced than the gains.

We posit that this economic interaction framework can be applied to search systems, where a user's needs and goals are information based, but must also account for interaction

efficiency. We may apply nudging as a technique to encourage mental model development based on the state of the system and a user's goals in order to maximise the efficacy and increase the overall perceived utility of the system. Chapter 4 was designed under this assumption.

2.4 Patterns of Use in Interactive Information Systems

In evaluating search systems, researchers have a variety of tools and techniques for answering their research questions. Log analysis allows us to study click interaction behaviour. Qualitative techniques let us know users' subjective opinions and perceptions of a system. In our case, we are interested in creating a correspondence between a user's interaction patterns and what they expect from their interactions, that is, an idea of their mental models.

Buscher et al. [37] used eye-tracking techniques to determine how people's perceptions of ad quality affect their perceptions of search result quality. In their study, they were sure to cache the results for consistency between queries and gave different classes of tasks to participants (i.e., informational, navigational, transactional). In order to generate bad ads, the authors used a non-descriptive subset of queries. This still allowed for some keywords to be highlighted. They measured areas of interest, fixation impact, clicks, and time on result pages, and found that high quality ads led to more attention paid to ads and comparatively less attention pair to organic search results.

In the realm of online advertising, it is imperative to have good estimates of an ad's click-through rate (CTR). Wang et al. [204] proposed a formulation to smoothen the estimates of CTR under the assumption that the behaviour on similar pages will itself be similar; that is, more similar than it would be on another random page. Estimating CTR is often troublesome for rare events, because of data sparsity. Using pre-existing subject hierarchies or automatic cluster discovery, the researchers enrich rare events with information from other "close" events to make reasonable inferences about behaviour.

In experimental settings, researchers take advantage of the control that performing studies in a laboratory allows. Käki and Aula [103] made a few recommendations for maximizing the validity of a study involving a new search interface. This serves to reduce the variability in things like query formulation skills, query refinement style, and the thoroughness that one exercises in reading and evaluating results. In particular, some tactics may include using balanced task sets in which tasks are similar in topic and difficulty between conditions, but also different in a relatively minor aspect, such as changing one query term to something different yet equivalent for the purpose of evaluation. This is particularly useful in within-subject experiments, where it is also recommended that counterbalancing is employed as a

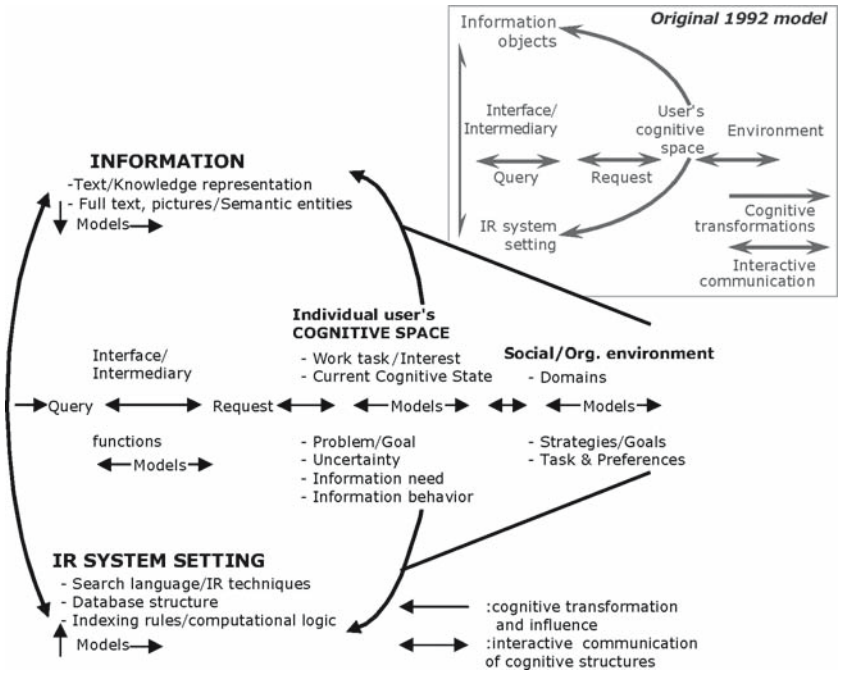


Figure 2.14: Ingwersen's Cognitive Model of IR interaction [118].

technique to alleviate learning effects. The authors also outline metrics that are meant to be useful for interactive IR, such as "search speed" (answers per minute), "qualified search speed" (answer per minute given relevance), interactive precision and recall, and immediate accuracy. It should be noted that some of these recommendations are proposed in the context of simple fact-finding experiments where each task has one correct answer.

Knight and Spink [118] outlined a comprehensive survey of models of information behaviour, applicable to the Web. Of particular interest is a set of interactive information seeking retrieval models, which include Marchionini's Information Seeking in Electronic Environments Model, Bates's Berrypicking Model, and Ingwersen's Cognitive IR Interaction Model. For our purposes, we are interested in a cognitive model of information interaction, where we consider not only the interaction process between the user and the system, but also between the user and the documents, and with other information objects. With Ingwersen's model, we can not only account for the cognitive dimensions of system interaction, but also information interaction.

Chen and Macredie [44] considered the human factors involved in interaction patterns with a survey of prior research on behaviour and three factors: gender differences, prior knowledge, and cognitive styles. A comprehensive study design should be made with these factors in mind, and analyses done to account for these differences. It should be noted that not all findings are conclusive, but prior work shows differences in navigation patterns,

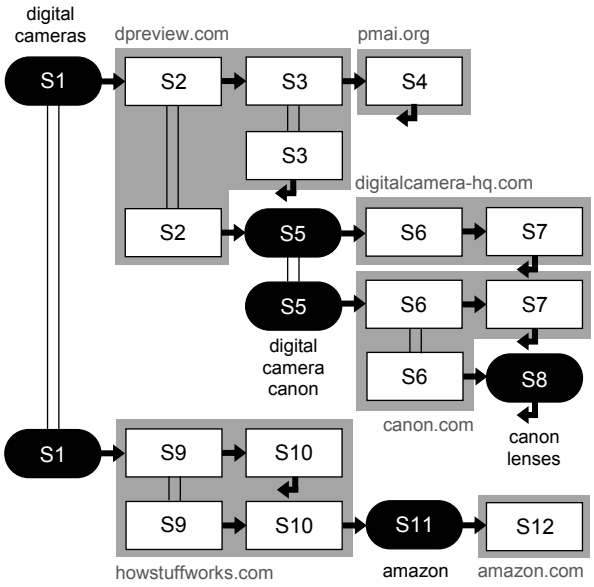


Figure 2.15: Example of a search trail extracted by White and Drucker [211].

attitudes and perceptions, adjustment to information structure, and querying behaviour, depending on these factors.

To better understand the behavioural variability seen in Web search interaction, White and Drucker [211] analyzed log data collected over a period of five months. The authors extracted search trails, which they were able to treat as a string and compute differences in features such as time, number of queries, number of steps, number of revisits, number of branches, and average branch length. Through factor analysis, they found that forward and backward motion accounted for most of the variance between users, followed by branchiness and time. The authors also noticed differences in query behaviours, with particular terms being associated with the type of query—whether navigational or informational.

In another set of studies, researchers use these aforementioned techniques to explore intentionality in behaviour and cognitive search strategies. Work by Marchionini [131] introduced a framework for considering search strategies, which he defined as general approaches to information seeking problems, such as using a particular search engine for particular kinds of information. He categorised by their level of goal-directedness, planning, and formality, which led to two categories: analytical strategies and browsing strategies. Analytical strategies are the more systematic, goal-driven and planned of the two, whereas browsing strategies are more extemporaneous, informal, and data-driven. Marchionini [131] was clear to discuss strategies in relation to "moves" and "tactics", which are seen at different levels of abstraction. Moves are individual actions, such as clicking a link or entering a query. Tactics

are groups of behaviours, such as issuing a query and performing subsequent refinements. Tactics are at the basis of a strategy, which is considered at the level of problem-solving for particular needs, and "patterns", at the highest level, are interactions used for *all* information seeking tasks, such as a particular search style.

Using this conceptualisation as a basis, a study by Thatcher [189] explored the manners in which different tactics constituted search strategies through a bottom-up approach. With a combination of interviews, giving search tasks to participants, gathering log data, and collecting retrospective verbal protocols with the help of screen recording, the author identified seventy-eight tactics. With this set of tactics, the author used participants' intentions at key decision-making points to group tactics together into strategies. He found twelve distinct strategies, and that participants often changed their strategy multiple times during a single task. This might occur if one strategy was unsuccessful or they wished to use different strategies in separate browser windows.

An earlier, yet related study by Stelmaszewska [182] aimed to better understand the different information interaction patterns that emerge as users developed search strategies for using digital libraries. These authors used a recorded observational study with a think aloud protocol among seven participants, who were instructed to use several digital libraries to achieve their own personal objectives. In congruence with Marchionini [131], the authors saw two distinct types of strategies: searching and browsing. The authors argued that the choice of strategy was related to the user's prior knowledge of the domain and familiarity with the collection. Much of the focus of this study is on result examination, which seems to be the pivot for the next search strategy to be applied. Patterns emerged for which strategies would be applied for different scenarios: for no matches (where users abandon the library, reformulate their query, or change their preference settings), for too many results (where users reformulate, change preferences, and change query terms in that order), and for a manageable number of results (where users scanned results and examine documents in detail). When faced with an "OK" number of results, users generally use this opportunity to determine relevance of the results; their assessment determines whether to stop or continue. It should be noted that result evaluation is only one step of the process; in work by Sutcliffe [184], the author identified four main activities: problem identification, need articulation, query formulation, and results evaluation. Sutcliffe also developed a comprehensive predictive model of information retrieval, considering these four activities and synthesizing cognitive theories from IR and experimental results of information seeking [184].

2.5 Option Pricing

In capital investment, real options represent an opportunity to undertake an initiative. We can view it as a staged investment or decision. Traditionally, one may use options valuation in the service of operational production flexibility in one of four central ways [22]: reducing set-up time at installed equipment, multipurpose stations, parallel assembly lines, and/or a flexible work force. We can reframe these approaches for information retrieval, such that reducing set-up time can be seen as the cost of switching search engines or rankings, multipurpose stations (or a flexible manufacturing system) can be seen as IR system integration, parallel stations could be framed as parallel searches, and a flexible work force can be viewed in terms of ranking and matching algorithms. Seen this way, real options seem to have particular applicability to the optimisation and time-quality cost tradeoffs that we see in slow search, both from a user's perspective and for an IR system implementation.

To see this more clearly, let us consider the type of system implemented in the study described in Chapter 3 – a Chrome extension that added a sidebar and a "Work Harder" button to Google Search. This particular system incorporated—in essence—two search engines with a low cost of switching between the two, tight integration between the two search engines, ranking running in parallel, and two different ranking algorithms. We can consider the costs and value that these capabilities present to the user as well as for the provider. For instance, will more sophisticated matching or ranking algorithms be worthwhile for a user performing the current task? Providing the sophisticated algorithm will come at a cost for the provider as well. Considering the option valuation will be useful in making the decision of whether or not to employ the more advanced ranking algorithm.

For companies performing options valuation, their main source of uncertainty is the demand of the goods and/or services they produce. For the user of a search engine, their source of uncertainty may be whether a feature will be useful for their task. The designer of a system has the responsibility to convey the value of the feature in a way that is easy to understand and supports their decision-making. When performing a search, a user has a particular mental model that guides their expectations of their use of a feature. "What kinds of results can I expect from the sidebar when I click the 'Work Harder' button on this search? Is this slow search working, or should I stop and abandon it?" Very simply, we can distill this further to, "How likely am I to get better results using this feature, and how much better are these results?" The uncertainty here maps cleanly to the typical characterisation of relevance.

To influence a user's conception of the usefulness of the feature, we may convey that using the "Work Harder" button will be helpful. Furthermore, when using it, we may try

to convey the improvement in quality in a tangible way, or one that allows easy comparison to not using the feature. A hypothetical extension to the Work Harder button could say that "most users got better results using the Work Harder button for this query", or even more simply that most users *used* the button for the query. The magnitude of the win is more difficult to convey, but one method that is easily measurable is the odds of success using the button versus not using the button. Other methods could try to directly convey the improvement in document matching and ranking with visualisation techniques, or the savings in time from using the feature.

Real options valuation provides a framework to think about slow search in a quantifiable way that considers uncertainty, value, and cost. More sophisticated options valuation models can help both users and search engine providers make informed decisions about using or employing features in their interfaces. I use real options in Chapter 4 as a way to quantify the benefits of system-level interventions to explore flexible time and risk tradeoffs.

2.6 Search-as-Learning

Search-as-learning comprises as substantial aspect of Chapter 5, which outlines the evaluation of a system that assists in vocabulary learning using a conversational large language model (LLM). This chapter includes a review of the literature in Section 5.2.1.

We now turn our attention toward a set of studies that encapsulate these principles and themes in Chapters 3 to 5.

CHAPTER 3

Exploring Time-Quality Tradeoffs through Slow Search

3.1 Introduction

Current search systems are heavily optimized for speed: commercial search engines often conspicuously display the fraction of a second that it takes to return the list of results to a query. Traditional systems take numerous shortcuts for efficiency, such as making simplifying linguistic assumptions for query processing, document matching and ranking [188, 135]. As a result, much semantic richness is discarded in the process of retrieval, and much of the potential in terms of relevance quality may not be realized. The implicit time budget to which system developers must adhere also limits the scope and effectiveness of creative and useful extensions that may be considered for search processing and interfaces, such as enhanced personalization or novel ways of diversifying or summarizing results [117].

Slow search – the notion that a system may be able to "take its time" to process results for increased effectiveness – has been proposed, but only at the level of advancing the concept and exploring user attitudes to waiting for queries [188, 187, 58]. In this paper, we present a study that investigates the effect that an actual slow search system that supports asynchronous (background) query processing has on user behavior.

Search that focuses on speed, sometimes at the expense of quality, may be underserving users with particular needs or devices. For instance, the growth of mobile phone usage is outpacing that of desktop PCs—especially in developing countries—but there is a capability gap not only between phones and PCs, but between different phones as well. This may lead to lower levels of information seeking and engagement [146]. This study would therefore be useful to search engine implementers and interface designers targeting developing regions. Exploiting intrinsic diversity in search queries aids in exploratory search—useful in education, for students learning about new topics [213]. Demonstrating the feasibility of this new slow search paradigm would also encourage implementers of conventional search engines to further

explore the importance of the time–quality tradeoff, potentially leading to more systems that can automatically adjust their performance along a scale that effectively trades off urgency and quality.

The contributions we present in this paper include an extensive analysis of a search system that embodies characteristics of slow search. We are primarily interested in the practical value of trading speed for quality. To that end, we developed a novel system which improves the topic relevance of a query asynchronously over time while the user continues to work. This allows us to investigate the types of tasks for which users are willing to tolerate a delay in processing for more relevant search results. Using log data, we show how users behave when given asynchronous slow search capabilities and compare it to a baseline without these features. We also trained a logistic regression classifier to predict task success depending on the capabilities given to the user and interaction features. We also provided an anonymized data set[1] to allow for analysis by other researchers.

As the primary purpose of this work is concerned with understanding patterns of interaction behavior when users have the ability to run a slow search in the background, we consider the following research questions:

RQ1: What are the types of queries for which users initially report they would have a willingness to wait?

RQ2: How much time will users typically wait for results from a slow query?

RQ3: In terms of search activity, how do users spend their time while waiting for a slow query to finish?

RQ4: How does typical user behavior change when provided with the ability to run a slow query?

RQ5: Do users perform search tasks more effectively with slow search?

We address RQ1 in Section 3.3.2, and RQ2–RQ5 in Section 3.4.

3.2 Related Work

The concept of slow search was introduced to the literature by Dörk et al. [58] and Teevan et al. [188]. With inspiration from other "slow" movements, including slow food, slow travel, and slow technology, the authors posit the changes in how individuals and groups approach the process of search if a system emphasized slowness over speed. The authors propose that users will be encouraged to be more mindful and reflective, allowed to revisit previous journeys in their search tasks, and invited to explore the inner workings of a system as well

[1] http://umich.edu/~ryb/slow

as the relationships between items in search results. Poirier and Robinson [155] described a model of how slow principles may be applicable to information behavior. These initial papers provide insight in how proposed slow systems might be built, or survey-based results on how users might be willing to use such systems.

Previous work has shown that users would be willing to engage in slow search for certain kinds of queries and tasks. A study by Teevan et al. [188] based on user surveys and empirical analysis of search query logs found that, while increased load times for search results led to increased abandonment for typical queries, for tasks in which result quality was poor, users were willing to wait for better results or try alternative methods of finding information. However, none of this previous work built or studied a working slow search system with real users: to our knowledge, our paper represents the first study of how users interact with an actual Web search scenario providing slow search features.

Asynchronous search has been studied previously, but primarily in the context of bandwidth limitations and without recognition to the notion of improving search results [43]. Prior work has examined the relationship between time delays and user behavior [33, 188, 136], but these were in the context of conventional search systems, where users have the expectation of rapid responses, and with no benefit to waiting.

Slow search also has parallels with question answering systems. Aperjis et al. [5] found that users wait longer to get an additional answer after receiving a small number of responses on Yahoo! Answers. By analogy, a user performing a search may be willing to repeatedly check in on the results of a slow search as it builds a final result set, and decide whether to stop the search or continue waiting. However, the authors do not show how the number of answers for a question relate to their quality or the times in which they arrive. Liu et al. [126] demonstrated in a field experiment that the frequency, quality and time of solutions to tasks on the crowdsourcing site Taskcn reflect strategic decisions depending on the reward level, the existence of a reserve (i.e., a prior high-quality solution), and expertise of crowdworkers. The authors however do not investigate the waiting behavior of the requester in light of the solutions. In other words, for slow search, we are interested in a mix of the two—examining when a user believes the information received is "good enough" to stop waiting for additional information.

Büttcher et al. [39] compared the effectiveness of different systems while accounting for the CPU time involved in query processing. Generally, the systems that used more CPU time showed better results in effectiveness. In the efficiency task, comparing each system's best run to its fastest run, the differences in ms/query can be quite appreciable. This shows that there is often a benefit to extra processing time, and a system that takes advantage of this time when appropriate could satisfy users better, provided they are willing to wait.

There is also increasing recognition of time as an important factor in the evaluation of search systems. Clarke and Smucker [46] proposed a metric of time-based gain to measure an information retrieval system's effectiveness to reflect the value that a user gains over time in interacting with the system. For slow search, this metric is applicable to the value gained from waiting as the system works to provide better results. A recent user study by Crescenzi et al. looked at a design somewhat contrary to ours, namely, the effect on search behavior when users were given *less* time to search [53].

To build a result set, we explore not only relevance, but also comprehensiveness of subtopic coverage in the form of intrinsic diversity. Radlinski et al. [158] define intrinsic diversity (as opposed to extrinsic diversity in the form of ambiguity about an information need) as being the various aspects that are by their nature part of the information need. As an example, given a query of "jaguar", the extrinsic diversity of the information need lies in disambiguating whether it is in reference to the automobile maker, the animal, or the codename of a version of the Mac OS X operating system. In contrast, the intrinsic diversity of the information need lies in its subtopics, provided it is not ambiguous. That is, if the user is indeed searching for the animal, then relevant subtopics may include the jaguar's habitat, its diet, and its physical characteristics. Radlinski et al. [158] outline five different scenarios where intrinsic diversity is required. These include cases where there is no single answer, where the user would like different viewpoints on an issue (political; product reviews), where the user would like a selection of options to choose from, when the user would like an overview of a topic, and where a task requires gathering disparate evidence to build confidence in an answer's correctness. There has been much work done on query ambiguity and extrinsic diversity; intrinsic diversity has received less treatment. Raman et al. [161] demonstrated that it is possible to identify queries that signal the beginning of intrinsically diverse tasks and re-rank results by their various aspects. Azari et al. [8] and Crabtree et al. [52] prior to this applied an approach of exploiting different and diverse aspects and reformulations of queries for the purpose of query expansion. These query expansion approaches increase the set diversity of the candidates of the search results, but do not organize the results by aspect. The portfolio theory, applied by Wang and Zhu [202] to balance relevance and risk for ranked lists, has been used to optimize for diversity in search results [160]. Rafiei et al. [160] used correlations between pages based on heuristics such as entity mentions, numbers, site names, and query extensions to measure diversity. It may be possible to combine these approaches to increase or reduce overall diversity as well as diversity within subsets or subtopics of results. It should be noted that diversity has long been a topic of research in recommender systems [199, 73], where there are many different perspectives of diversity, including a relationship to novelty. All of these approaches ignore the dynamics of time: how the intrinsic diversity

of a set of results changes with time.

Compared to the existing literature, this work presents a working system that embodies the principles of slow search and directly improves the relevance of search results, while investigating the relationship between types of tasks, user impatience, and quality improvement over time.

3.3 Method

To measure user behavior characteristics, we designed an extension for the Chrome Web browser that works in conjunction with Web search engines to capture the current query and send it to a server for extended processing when the user clicks a "Work Harder" button to the right of the main search editbox on the search engine page, as shown in Figure 3.1. Doing this adds the query to a sidebar (a) on the search engine result page, which shows a progress bar (b) as well as the top three results at any given time (c). We call this extended-time background query a 'slow' query. The user may click on the "(more results)" link (d) at the bottom of the sidebar to view the full list of re-ranked results, as they are improved and updated asynchronously by potentially adding new documents to the list and re-ranking them. This page also displays a progress bar, and may be left open while the user continues to search on the main search page.

This 'slow' query processing occurs as a background process, during which users are free to continue performing their own searching and query reformulations in the main interface while the 'slow' query completes. In this study, we allow at most one slow query at a time, which may be cancelled before its progress is complete and removed from the sidebar. The extension also serves to log interaction through queries, clicks, and mouse movements.

3.3.1 Study Participants

Our study consisted of 44 participants (18 Male, 26 Female; mean age = 23.5 SD = 5.9), recruited through the University of Michigan School of Information. Most were undergraduates (n = 18) or holders of an undergraduate degree (n = 12). The majority reported being very experienced with search engines: we asked about their familiarity on a scale from one to five; the mean response was 4.6 (SD = 0.6). Additionally, 38 reported using search engines more than once per day, while 5 reported using them more than once per week. The remaining participant reported using them more than once per month. We also asked participants to report their confidence in their abilities to find the information they need while searching on a scale from one to five; the mean response was 4.36 (SD = 0.65).

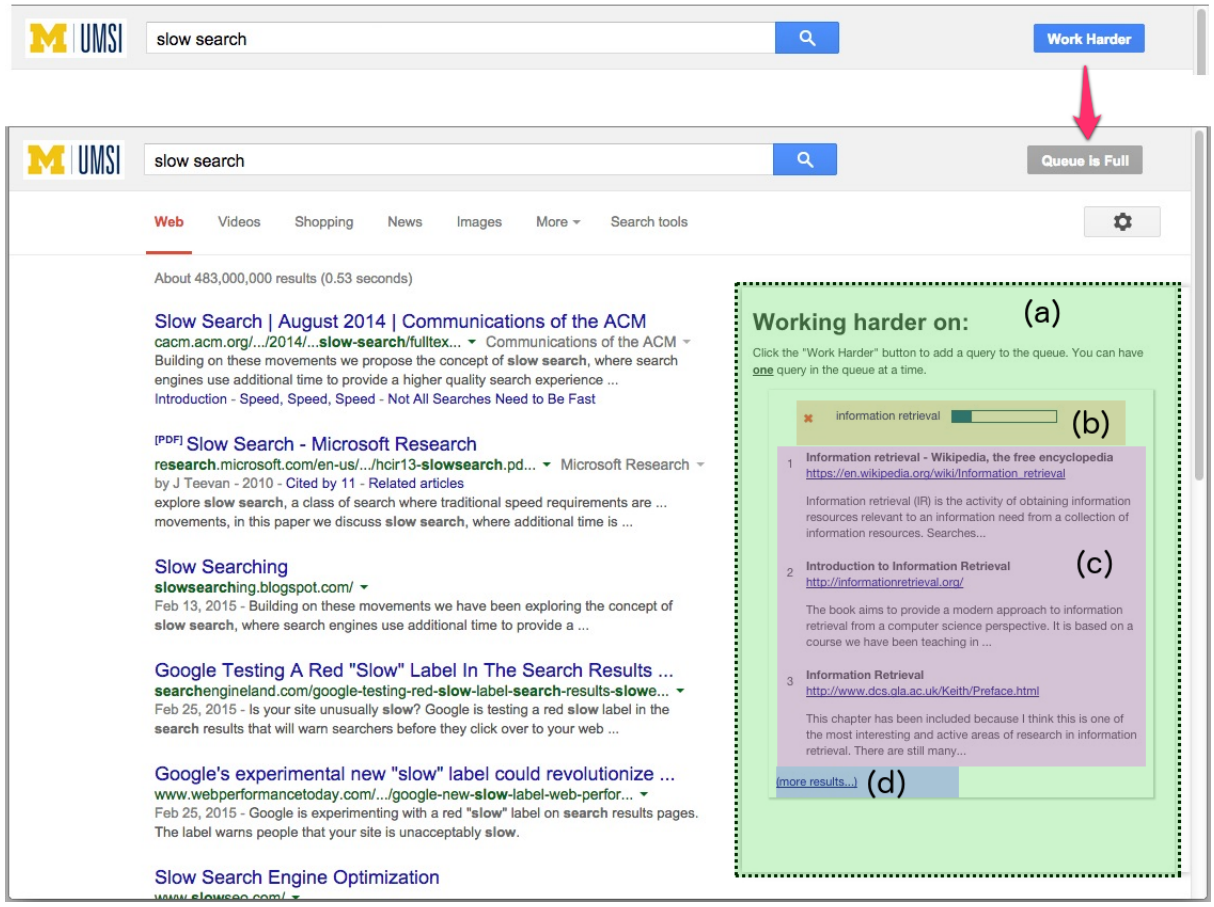


Figure 3.1: Interface with "Work Harder" button and sidebar (a). Colors added for illustration. Clicking the "Work Harder" button in the upper right adds the current query to the queue (b). The top three results at any moment are presented below (c), and a full list of re-ranked results is available by clicking on (d). These interface additions are always present.

3.3.2 Background Survey

To better understand tasks for which people might be willing to wait for a better answer (RQ1), we asked participants to provide a description of the last search task they performed in which they failed in satisfying their information need. We report these tasks as well as their anticipated willingness to wait for the perfect results below.

3.3.2.1 Prior tasks users reported as difficult

We first asked participants the following question:

Think back to the last time you had trouble finding information with a Web search engine. What was the information that you were trying to find? Please be as specific as you can, as best as you can remember.

We coded responses by topic, summarized in Table 3.1. As most participants were students, the majority had issues finding information for classes or assignments, as seen in the response from an ecology student who tried to find information for a course on birds: "Our team was trying really hard to find the specific information needed to support our study. For instance, we hope to find if the tree bark thickness affect the foraging preference of the bird. Most of the study we found were of the bird but not related to our topic."

The common trends for difficult education-related needs involve finding new and novel information (e.g., finding articles on a topic that has not been seen before), finding reliable scholarly articles on a topic, and expressing the problem in the right way for the search engine to yield useful results ("It was difficult to search for because I wasn't sure what I was searching for.").

For many other topics, the problem involved finding a specific item, such as a person, product, or song. This was most common among the Career, Entertainment, and Shopping topics. The main issue in these cases involved expressing the right criteria to find these items. For instance, one subject tried to find a song by its lyrics, but the lyrics alone were not specific enough. She knew that they were from a pop song from an indie artist, but could not express this to the search engine. Instead, the results were dominated by a popular Jay-Z song. Similarly, another subject tried to find a particular drawer slide, but was not able to use the right search terms. Instead, he had to iteratively search related topics in order to pick up more useful search terms.

For other topics, users had difficulty finding a specific item, such as a person, product, or song, most commonly among the Career, Entertainment, and Shopping topics. The main

issue in these cases involved expressing the right criteria to find these items. For instance, one subject tried to find a particular drawer slide, but was not able to use the right search terms. Instead, he had to iteratively search related topics in order to pick up more useful search terms.

Topic	Count
Education	16
Shopping	6
Entertainment	5
Health	5
Career	3
Technology/Troubleshooting	3
Food	2
Sports	2

Table 3.1: Topics of tasks reported as difficult.

We also categorized participants' reported tasks according to the nature of information they were seeking. Overall, 16/44 (36%) of difficult/unsatisfied needs involved searching for specific items or facts that satisfied multiple attributes; 10/44 (22%) were questions seeking a specific factual answer; 4/44 (9%) needs were for the latest version of information; 4/44 (9%) involved searching for a person. The remaining needs involved more vaguely-defined needs, more exploratory research needs, or procedural information on how to solve a problem. This predominance of multi-attribute search needs, the nature of which we can find examined in [117], motivated our design of tasks for slow search as described in Sec. 3.3.4.

3.3.2.2 User willingness to wait

As part of studying the time-quality tradeoffs that users might find acceptable in a search engine (RQ2), we asked participants:

Given your experience, if a search system was able to provide the perfect results, how long would you be willing to wait for the search engine to process your query while you did other tasks and you were notified when it found these results?

Figure 3.2 shows users' self-reported willingness to wait (the 'impatience curve') as a function of waiting time. The y-axis shows the proportion of users, as estimated from survey responses, who would be willing to wait at least t minutes for search results using the slow search system.

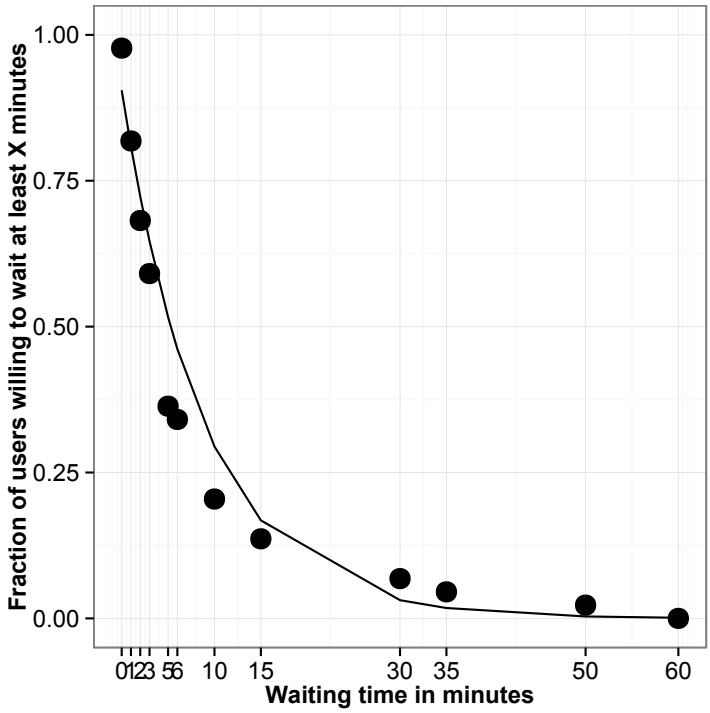


Figure 3.2: Users' self-reported willingness to wait decays exponentially as a function of waiting time.

We are interested in what users self-report as acceptable, not only to calibrate our experiment, but also to see if their actual behavior matches the expected behavior (which we compare in Sec. 3.4.1). Users reported a willingness to wait 9.5 minutes on average (SD = 13.2). An exponential decay in acceptable waiting time is evident from our analysis: with the survey response data, we fitted an exponential decay model $w = \exp(-at)$ to estimate the empirical probability w that a typical user would be willing to wait at least t minutes. The fitted exponential parameter was $a = 0.11$, meaning that for every additional minute of waiting time, about 10% of remaining users were not willing to continue waiting. We note that this rate of decay 'impatience factor' is in accord with that reported by Teevan et al. [188] that asked a similar question about willingness to wait for perfect results.

3.3.3 Experiment Design

For the purpose of this study, we focused on investigating the effects of an improvement in relevance for multi-attribute tasks. To that end, we implemented a server that communicates with the Chrome extension to simulate an improvement in relevance over the course of five minutes for each slow query submitted. For each task that a user may choose to tackle, we manually selected five to ten high-quality documents and snippets that, collectively, allow a participant to correctly solve the problem posed by the task. When the "Work Harder" button is used, the server selects documents from the pool to insert into the ranking every twenty seconds. Similarly, another process on the server periodically moves high-quality documents closer to the top of the ranking over the course of the five minute period, until these documents reach the top of the ranking.

We randomly assigned participants to one of three conditions. In the baseline condition ($n = 16$), the interface resembles a conventional Web search engine, with no "Work Harder" button or sidebar. In the "*static gain*" condition ($n = 15$), the interface adds a persistent "Work Harder" button and sidebar to the conventional interface. Furthermore, the system inserts highly-relevant documents in the middle of the ranking "below the fold" of the re-ranked results page and the rank position of these documents stays the same over the course of the five minutes. Finally, in the "*dynamic gain*" condition ($n = 13$), the interface is the same as in the "*static gain*" condition, but the system inserts documents at the last position of the re-ranked list and then continuously increases the position of documents at 20 second intervals, over the five minute time window, until they finish at the top of the ranking. In this study, we used a dynamic gain that was linear with respect to time. With this design, we introduce the two new capabilities of an improved result list and a dynamic ranking. We chose to contrast the "*static gain*" condition with the "*dynamic gain*" condition to determine

whether users actually perceived the improved relevance as well as to study the effect of the dynamic ranking.

3.3.4 Description of Search Tasks

Participants were presented with a list of four topics, with each topic having three tasks within it. Each participant was required to select one task from two separate topics. We allowed participants to choose tasks and topics of interest to them with the goal of increasing their intrinsic motivation to complete each task.

We prepared the total set of twelve tasks such that each task was presented in the form of a question to be answered, and each task called for the participant to find five items that satisfy multiple attributes specified within the problem. We did this to control the cognitive effort required for each task – users were expected to find a set of candidate answers and verify that each of them satisfied all constraints in order to receive the full reward. For any particular item that the user submitted in their answer, we considered it "correct" if and only if it satisfied all of the required constraints. We believed that having a slow search system which reduced this high expected effort would encourage use of that capability when available.

Local Businesses (32 tasks completed)
Name five I.T. companies in Ann Arbor with at least 50 employees.
Entertainment (28 tasks completed)
Name five video games in which Pharrell Williams's music has been featured.
Education (21 tasks completed)
Who are the five most influential professors in the United States in the field of sociology?
Shopping (7 tasks completed)
What are five smartphones that are thinner than a standard No. 2 pencil and usable on AT&T?

Table 3.2: Examples of search tasks and their topics.

In Table 3.2 we present examples of search tasks for each of the four topics, along with the number of task completions by topic. Local Businesses had the most interest, with its tasks being chosen 32 times in total. Conversely, Shopping received the least attention, which users choosing these tasks only 7 times.

3.3.5 Study Procedure

The user study took place in a laboratory setting at the University of Michigan School of Information. Participants volunteered to attend one of eight study sessions, with each session lasting a maximum of 90 minutes. Each participant was placed at a computer set up with the Chrome extension, which in turn was randomly associated with one of the three study conditions (Baseline, *static gain*, and *dynamic gain*). Participants completed two search tasks, with each task lasting a maximum of thirty minutes.

To introduce participants to the capabilities of the system before they began the first task, users were asked to perform an exploratory search task—in this case to explore the topic 'snow leopards' —as a warmup for five to ten minutes.

As motivation to finish the task within the allotted thirty minutes, we compensated users based on their performance in answering each question, which called for an answer that addressed each attribute of the problem, as well as three relevant documents that they found useful in solving the problem. This gave us a way of verifying whether participants found the documents inserted into the ranking, and explicit relevance feedback of these documents. An answer that perfectly met the criteria of the task led to a bonus of $2 with partial credit being possible, and giving relevant documents led to a bonus of $1 per URL – this served as motivation to give explicit relevance feedback.

3.3.6 Data Preparation

Missing Data. For each request made by the extension to log interaction data, the system associated a session ID with a particular interaction event, with this session ID being linked to the user's ID, which is randomly generated when a user begins the study. After the data collection was complete, there were 34 out of 1149 clicks without session IDs in our log database, and hence, we were unable to associate these clicks with a task, user ID, or condition. We therefore manually inspected the click data in an attempt to re-associate each click with a session ID. We were able to re-associate all clicks but one due to ambiguity in candidate tasks: 13 of 34 clicks were recoverable from session IDs included in page URLs, and 20 of 34 clicks could be manually recovered based on analyzing clicks with session IDs from closely associated contemporaneous queries. We outline our process below.

One set of clicks without session IDs originated on the slow search results page. In order to show the correct user the correct page, we included a session ID and query ID in the URL in the page, which was incidentally logged on each event. Therefore, we were able to pull the session IDs directly out of the URLs. This enabled us to recover 13 of the 34 clicks.

The other set of sessionless clicks originated from the regular search results page. Of

these, almost all of the remaining clicks seemed to originate from unambiguous tasks: that is, only one user at a particular time was solving one of the candidate tasks at any given time. Thus, to re-associate the clicks, we inspected the queries from which each click originated (queries were logged with clicks, and all clicks had queries associated with them in the data) and also the queries of contemporaneous clicks with session IDs. We were able to determine the session that each click belonged to based on the subject of the query. For instance, if a user searched for coffee shops, clicked on results which were logged without session IDs and then searched for specific coffee shops at which point session IDs were logged, we were reasonably confident that the two clicks belonged to the same session, as these sessions had one user performing the task at once.

There was once click that we were unable to associate with a session ID unambiguously, because at that moment, multiple users were working on the same task. We therefore do not include that click in our results.

Relevance Judgements. As a part of completing the task, we asked users to provide three relevant documents that helped in answering the task's question. We made final judgements on these documents in the process of calculating bonuses – if the document indeed provides information relevant to answering the question correctly, then the document was deemed relevant for the bonus. Otherwise, we considered the document not relevant.[2]

3.4 Experiment Results

In this section we conduct an analysis of user activity, addressing the remaining research questions (RQ2–RQ5) and examining users' behavior in more detail.

3.4.1 How Long Participants Waited For Results

Our second research question (RQ2) concerns the amount of time users typically wait for results. In our background questionnaire described in Section 3.3.2.2, our participants expressed a mean willingness to wait 9.5 minutes for the perfect results for their difficult information needs. Our system imposes a maximum wait time of five minutes, which was not explicitly communicated to our participants. Five minutes was selected in order to give users more time to submit multiple slow queries within a single task session. 2/15 participants in the *static gain* condition and 2/13 participants in the *dynamic gain* condition explicitly

[2]If the document only provides information that would result in answering the question incorrectly, as, for example, only providing a Web page for a coffee shop outside of the location we ask, the page was considered not relevant.

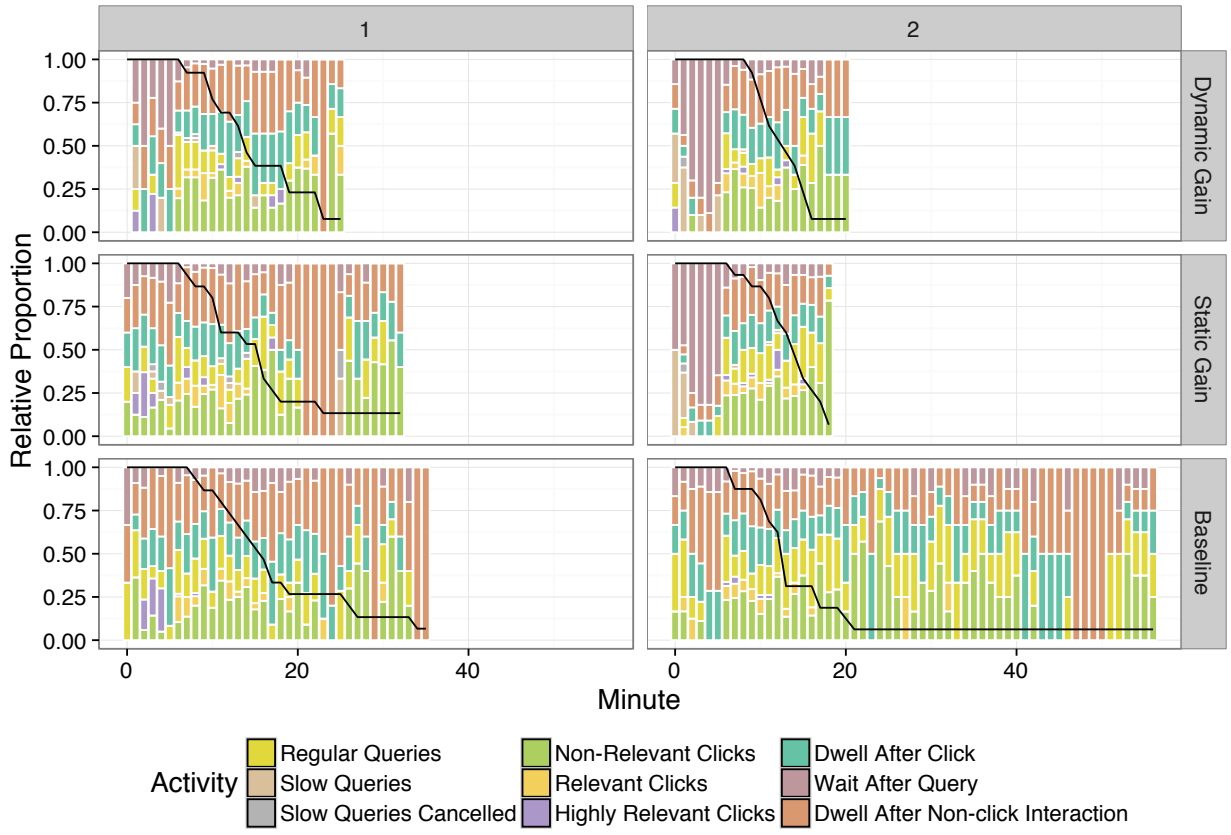


Figure 3.3: The actions that users perform over the course of the two tasks by condition. The black lines show the proportion of remaining participants in the session.

mentioned without prompting that the waiting time was too long when asked about their impressions of using the system in post-task questionnaires.

Analyzing the behavioral data, we find that in both the *dynamic gain* and *static gain* conditions, approximately twice as many participants used slow search in the second task ($n = 11$) compared to the first task ($n = 5$). Table 3.3 presents expected wait times for each condition and task, representing the time for which each slow query is processed until either completion, or cancellation by a user. Users in the *dynamic gain* condition waited an average of 227 seconds in both tasks 1 and 2. By comparison, for users in the *static gain* condition, their wait times increased from an expected 136.1 seconds to 266.6 seconds: participants cancelled more queries in the first task than in the *dynamic gain* condition, but by the second task, these users cancelled fewer than in the *dynamic gain* condition, which likely contributed to the increase in mean slow processing time observed for the *static gain* condition. This difference was likely due to the continuous improvements in the dynamic gain condition making the utility of the sidebar more apparent sooner to these users in comparison

Condition	Task	Query Processing Time (sec)	SD	Users
dynamic gain	1	227.3	109.8	4/13
dynamic gain	2	227.2	82.4	3/13
static gain	1	136.1	145.6	5/15
static gain	2	266.6	85.0	1/15

Table 3.3: Mean slow query processing times by task, with standard deviation (SD) and fraction of users who cancelled their slow query.

to the single improvement seen in the *static gain* condition. An independent two-group t-test shows that the difference between the mean wait time in the first and second tasks of the *static gain* condition is statistically significant ($t(9.7949) = -2.318$, $p < 0.05$).

3.4.2 How Participants Spent Their Time

To obtain an overview of user activity as participants progressed through each task, we aggregated the actions that users performed and averaged across users for each minute of activity. The resulting plot of these aggregated actions is shown in Figure 3.3. Generally, participants took slightly longer to complete their first task than their second (944 seconds vs. 804 seconds on average). For the two tasks, the users in the dynamic gain condition had the shortest completion times (879 seconds for the first task, and 735 seconds for the second task). These differences were not statistically significant.

Differences between first and second session. In general, there appeared to be a period of slight acclimatisation as users in the slow search conditions made and cancelled slow queries throughout the session. By comparison, in the second task, users started by making slow queries and committed more to this decision rather than cancelling and restarting. More precisely, in the *static gain* condition, users made an average of 0.53 slow queries and cancelled 0.33 of them in the first task. By the second task, they made 0.93 slow queries and cancelled 0.07 of them. Similarly, in the *dynamic gain* condition, they made an average of 0.62 slow queries in their first task and cancelled 0.23 of them; by their second task, they made 1 slow query and cancelled 0.08 of them. Potentially, this small number of slow queries could reflect an optimal interaction strategy, which we will discuss further in Section 3.4.8. We present the results for comparison in Table 3.4.

Relevance gains over time. In Figure 3.4, we plot the median relevance of the documents clicked by users for the two tasks. We use the median to reduce the effect of outliers. We will discuss this further in Section 3.4.3.

Here, we include the preselected highly relevant documents as well as the documents that

Condition	Task	Submitted	Cancelled
dynamic gain	1	0.62	0.23
dynamic gain	2	1.00	0.08
static gain	1	0.53	0.33
static gain	2	0.93	0.07

Table 3.4: Slow queries submitted/cancelled by task.

users considered relevant for solving the tasks. The highly relevant documents are worth twice as much as the user-selected documents. This gives us a profile of how users manage to make use of the system to find the documents used to solve the given tasks. As a means of comparison, we also present the mean cumulative clicks performed during each task in Figure 3.5.

In the first task of Figure 3.4, we see that in the *static gain* and baseline conditions, users perform similarly, eventually leading to a cumulative relevance score of 4 at 1000 seconds. However, in the *dynamic gain* condition, users do not perform as well: for these users it takes approximately 1500 seconds to reach their cumulative relevance score of 2.

As we will see in Table 3.6, where we calculate session-level features for each condition, users in this condition click on fewer documents on average than in other conditions. However, the trajectories are in fact similar (see Figure 3.5). The difference is in relevance. This implies that users in *dynamic gain* do examine documents, but not the most relevant documents. The relevance trajectories in the beginning are more similar in the second task than in the first, but by the tenth minute they begin to diverge. The baseline condition ends up with a median score of 2.5 at approximately 2000 seconds, while the *dynamic gain* and *static gain* conditions gain a score of 3 by approximately 1000 seconds.

We suspect that the *static gain* condition performs as well as it does because all the relevant documents are injected into the ranking when it is first available to be examined; by contrast, in the *dynamic gain* condition, these documents are injected into the ranking at the bottom of the list and their ranks increase over time. It may have taken the users getting accustomed to the system in the first task before they were able to build a mental model of how the system worked and employ that model in the second task.

3.4.3 How users progressed towards a goal over time

Similar to Figure 3.4, in Figure 3.6, we plot the average relevance of the documents clicked by users, faceted by topic. The two most popular topics, Entertainment and Local Businesses, were the only two to have representation from all conditions in both sessions (with 27

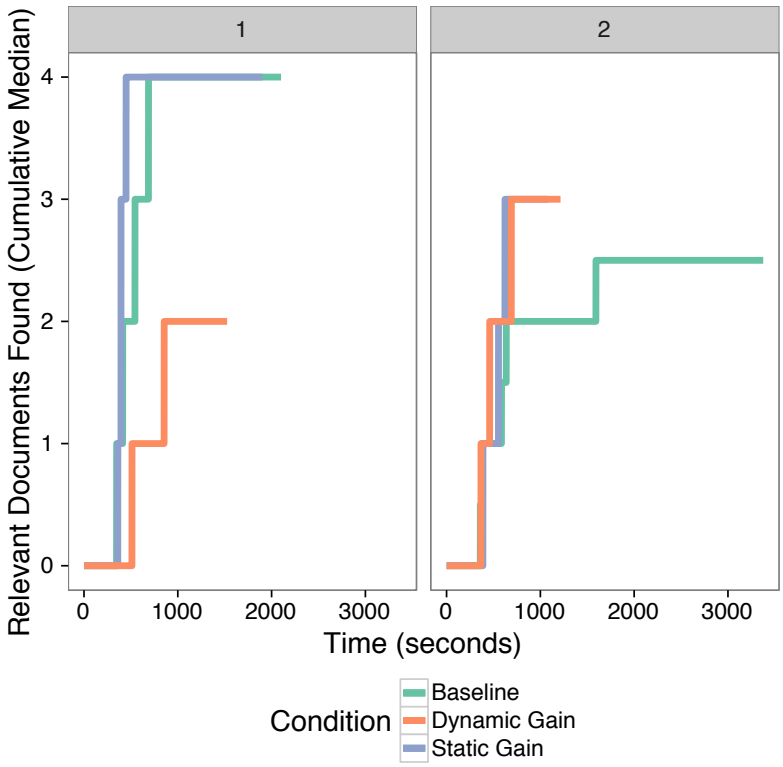


Figure 3.4: Median time-relevance curves by task.

Entertainment sessions and 31 Local Businesses sessions in total). As such, we will compare the characteristics of the sessions for each session as they correspond to the conditions within these topics. To do this, we performed a Mann-Whitney U test on each of the conditions and topics to compare the characteristics on the first task to those on the second task.

We found that for the Entertainment topic, the users in the *dynamic gain* condition made more slow queries ($p < 0.05$) during the second task ($CntQ_S = 2$) versus the first task ($CntQ_S = 1$). In comparison, in the Local Businesses topic, we see that for the No Button condition, users examined significantly fewer documents per query in the baseline condition ($p < 0.05$) in the second session ($CPQ = 1$) than in the first session ($CPQ = 2.07$). For the Local Businesses topic as well, in the *dynamic gain* condition, users also began making use of slow queries ($p < 0.05$) for the second task ($CntQ_S = 1$) compared to the first ($CntQ_S = 0$).

We also found interesting differences in the way users progressed in these tasks, though the differences were not statistically significant. Entertainment sessions typically became shorter in duration in the second session across all conditions (994 seconds to 702.4 seconds on average), and Local Businesses sessions became longer (658.8 seconds to 893.3 seconds on average). In aggregate, the number of queries performed (8.3 to 5.4 for queries in session

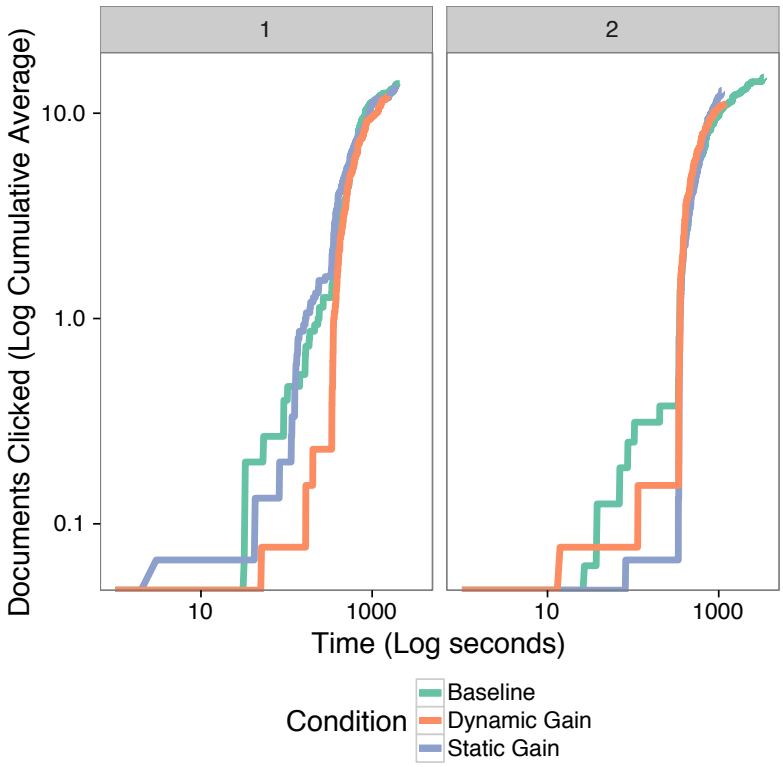


Figure 3.5: Average time-click curves by task. This includes non-relevant clicks.

for Entertainment and 5.2 to 9.7 for Local Businesses) changed accordingly. However, users examined more documents per query in Entertainment (2.6 to 2.8) and fewer in Local Businesses (3 to 2.5). Furthermore, the number of slow queries issued per session is concomitant with the number of regular queries by task and condition, however, we found that the number of highly relevant documents clicked increased in the second task compared to the first for both topics in the *dynamic gain* condition, which was unusual in other conditions. Thus, while users adjusted their behavior differently depending on the topic as they progressed through the experiment, users in the *dynamic gain* condition were able to consistently find the highly relevant documents regardless of other changes in interaction.

It should be noted that two time-relevance curves stand out in Figure 3.6: that of the baseline condition for the first task of Education, and that of the *static gain* condition for the second task of Shopping. This is due to two users who found a substantial number of relevant documents in comparison to the other conditions in the same topic and task. These curves in fact affect the aggregate profile of Figure 3.4 to the extent that the baseline condition in the first task in this figure and the *static gain* condition ended with the highest cumulative relevance score.

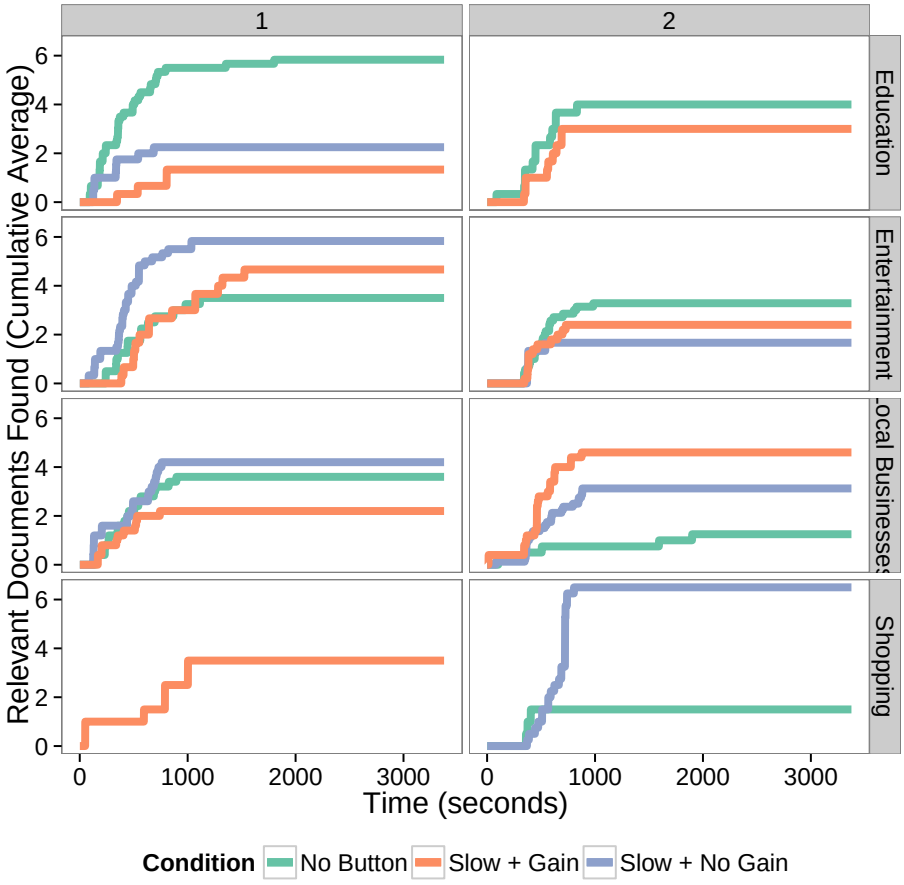


Figure 3.6: Average time-relevance curves by topic.

3.4.4 Behavior While Waiting

Our third research question (RQ3) pertains to activity while waiting for slow search results to finish processing. To answer this question, we looked at how users continued to interact with the system after submitting a slow query. We note that for the conditions with the slow search button, many users spent their time waiting after submitting a slow query. This was especially pronounced in their first five minutes of each task, where more users waited on average after making a query than at subsequent time periods, Comparing the two slow search conditions in the first task, we see in Figure 3.3 that more users spend time waiting after querying in the first five minutes of *dynamic gain* than in *static gain*. However, the profiles are more similar by the second task: in both groups, users made heavy use of the "Work Harder" button initially, waiting before eventually clicking on results.

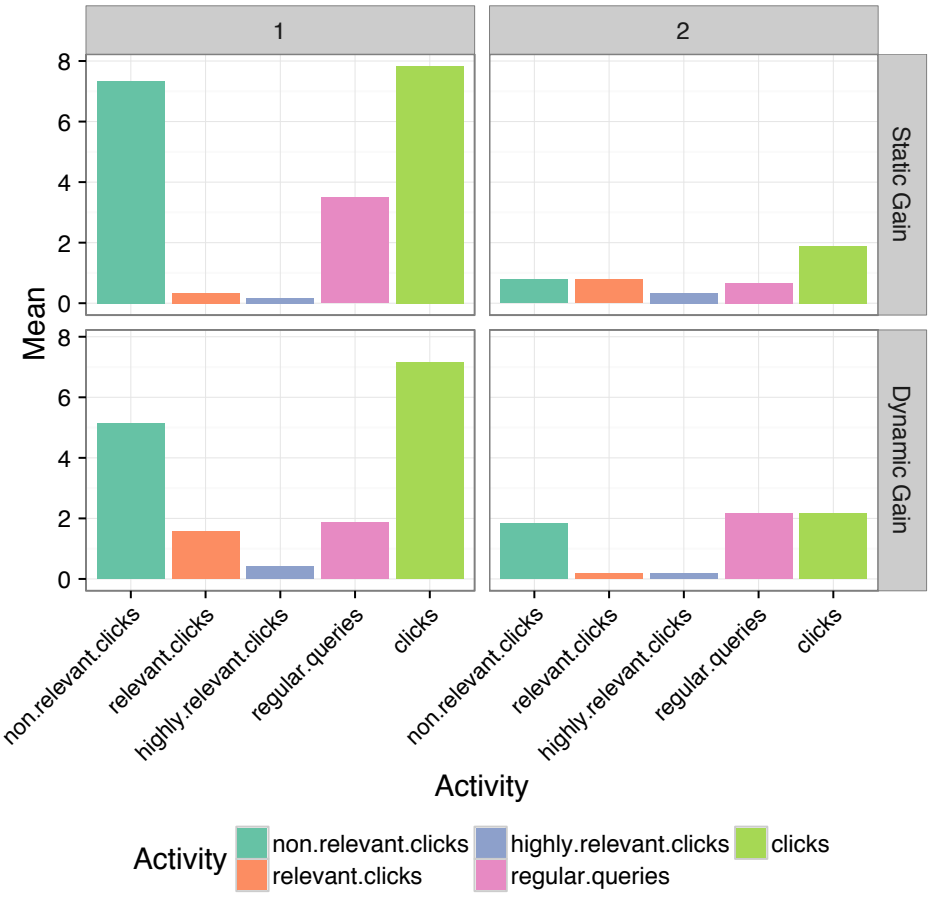


Figure 3.7: How users spend their time while waiting for slow queries to finish.

Figure 3.7 shows the activities that users performed while slow queries were processing for the *dynamic gain* and *static gain* conditions, i.e. after a slow query was submitted, and

before the query either finished processing or was cancelled. Most activity in this interval is focused on examining documents: the number of clicks is relatively high in the first task (6.6 for *static gain* and 7.6 for *dynamic gain*). In comparison, the number of queries is relatively low (2.2 for *static gain* and 3.8 for *dynamic gain*). By the second task, users do less in this interval, perhaps relying on the system more than examining documents and conducting additional queries themselves. In both conditions, the average number of queries and clicks both decrease (*static gain*: Queries = 1.55, Clicks = 3; *dynamic gain*: Queries = 0.91, Clicks = 3.09). Of additional note is that for *static gain*, the number of relevant clicks dropped from 1.2 to 0.18, while the number of highly relevant clicks stayed relatively consistent (0.6 in task 1 and 0.55 in task 2). This may indicate that users' time was better spent in the second task with regards to finding the most relevant documents to solving the given task. We can compare this to *dynamic gain*, where the number of relevant clicks increased (0.4 to 0.73), and also the number of highly relevant clicks increased from 0 to 0.36. Thus, while the number of non-relevant clicks and queries decreased, users made better use of their time in finding helpful documents.

3.4.5 Feature Analysis of Search Behavior

To investigate research questions RQ4 and RQ5, we computed a list of features characterizing search behavior, as inspired by previous studies such as [3]. The features we calculated are outlined in Table 3.6 and Table 3.7. The features we calculated are outlined in Table 3.5.

Dwell time (C_{IT}) was determined by calculating the time between a click and any subsequent interaction with a search page (mouse movements, scrolling, keyboard events, queries, or clicks). As we ask users to provide five correct items that satisfy multiple attributes for each task, we calculate Precision as the proportion of items included in their answer for a task that satisfy all attributes.

3.4.6 Behavioral Analysis of Searchers by Condition

For our fourth research question (RQ4), we investigate the types of changes seen in users' behavior when given asynchronous slow search capabilities. Having randomly assigned users into a condition either with or without such capabilities, we compare the session-level features for each condition. We present the values of these features in Table 3.6.

Compared to the two slow search conditions, users in the baseline condition on average were the slowest in completing a session ($\Sigma\Delta t = 961$ seconds), issued the highest number of queries ($CntQ_R = 8.81$) and the longest queries ($QWL = 5.72$; $QCL = 34.88$), and had the longest dwell time ($C_{IT} = 358.86$). These differences were not statistically significant,

Baseline Features	
$\Sigma\Delta t$	length of session
$CntQ_R$	count of regular queries in the session
$CntQ_S$	count of slow queries in the session
Q_RPS	Regular queries per second ($\frac{CntQ_R}{\Sigma\Delta t}$)
Slow Features	
Q_SPS	Slow queries per second ($\frac{CntQ_S}{\Sigma\Delta t}$)
$CntQ_SC$	Slow queries cancelled
Q_SCPS	Slow queries cancelled per second ($\frac{CntQ_SC}{\Sigma\Delta t}$)
Query Features	
QWL	Query word length
QCL	Query character length
Click Features	
$CntR$	Number of pages in the session
CPQ	Result clicks per query ($\frac{CntR}{CntQ}$)
Q_{DT}	Query deliberation time (time to first click for a query)
C_{IT}	Dwell Time (inactive time after click)
P	Precision

Table 3.5: Representation of session features.

but they may reflect a greater degree of effort for users in this condition, as users take more time to examine and possibly evaluate documents, and conduct more queries to address the various facets of the problem. We also found that users did indeed make use of slow search when given the option: features that quantify the use of slow search such as $CntQ_S$, Q_SPS, $CntQ_SC$, and Q_SCPS were significantly greater than zero in the *dynamic gain* and *static gain* conditions ($p < 0.05$). For most of these features (that is, $CntQ_S$, Q_SPS, and $CntQ_SC$), the values were highest in *dynamic gain*, though not significantly more so than in *static gain*. In contrast, Q_SCPS was highest in the *static gain* condition, though this was not statistically significant. No other differences were significant (with all tests here based on paired Mann-Whitney U tests with Bonferroni correction). Cognitive factors such as the evolving degree of user trust in result quality for the slow search conditions may contribute to these cross-condition differences and exploring these is a topic for future work.

3.4.7 Behavioral Analysis of Successful Searchers

Our fifth and final research question (RQ5) investigates whether users perform tasks more effectively with the help of slow search. As Table 3.6 shows, users in the *dynamic gain* condition received the smallest reward, and had the lowest precision. This raised the question of what factors played a part in increased performance. We compared user features for successfully completed tasks to those of the remaining tasks. We define success as a precision of 1 for a particular task, such that the answers given satisfied all the criteria set by the task's question. We present the features computed based on success in Table 3.7. We performed Mann-Whitney U tests to determine whether there were significant differences by success.

We found significant differences in session length ($\Sigma\Delta t$; $p < 0.05$) and time to first click ($p < 0.05$). The average session length for successful tasks (891.10 seconds) was significantly higher than that for less successful tasks (847.69 seconds). Despite this, the number of queries issued is not significantly different, though the average time to first click is significantly higher for the successful (343 seconds) than for the less successful (260.44 seconds). This indicates that the time spent examining the search results was a major factor in success, as we also notice that the dwell times were not significantly different. We also investigated click relevance, as described in Section 3.4.2. Users who were successful had an expected click relevance of 3.69, compared to a click relevance of 3.42 for the rest of the users. However, this difference was not statistically significant.

We also performed logistic regression to predict user success using the above feature set. We found that the intercept ($\beta = -6.524$) and the clicks per query ($\beta = 0.887$) were significant predictors of success ($p < 0.05$). For the intercept, a participant is not likely to be successful, with all other predictors held constant. An additional click per query increases the odds of success by 143%. Other predictors that were marginally significant ($p < 0.1$) include the rate of regular queries (Q_RPS, $\beta = -0.009$, $p < 0.1$) and the effect of session length when the condition is *static gain* ($\beta = -0.009$, $p = 0.1$). In the case of the query rate, an increase in this rate predicts an increase in the odds of success, whereas an increase in the session length in the *static gain* condition predicts a decrease in the odds of success. No other terms were significant predictors.

Overall, the logistic regression analysis shows that making good use of one's time is the main factor in success. That is, searching and examining documents in a short period of time usually means that the user will be successful. The interaction between the session length and using the *static gain* system also suggests that, as a longer session length implies difficulty in satisfying an information need, not being able to take adequate advantage of the system's assistance decreases the likelihood of success. We present the values of the model coefficients in Table 3.8.

3.4.8 Analysis of Interaction Strategies

Because the ability to perform a slow search was a new feature for participants – the training period built into the start of the study not withstanding – we examined how participants' choice of search strategies changed across sessions as users became more familiar with the feature.

In particular, we were interested in how users in the two slow conditions adapted their decision-making and use of the feature in relation to more optimal strategies. Each of the slow conditions could be considered to have an optimal strategy in terms of the number of regular queries issued, the number of snippets examined, the time taken to invoke the "Work Harder" button for the first time, and the waiting time for slow processing.

For the *dynamic gain* condition, we consider one optimal strategy to be the following: 1. Issue a query; 2. Click "Work Harder"; 3. Wait for 5 minutes as the results automatically improve to maximum effectiveness; 4. Examine the first 10 slow results[3]. In comparison, the *static gain* condition has a very different strategy: 1. Issue a query; 2. Click "Work Harder" 3. Examine the first 30 slow results immediately. The differences stem from the fact that the *static gain* condition happens to improve relevance immediately, but to a much lesser degree than in the *dynamic gain* condition at 100% completion. Thus, for the *static gain* condition, it is in the user's best interest not to wait, but this is not evident from the interface.

To examine how behavior changed relative to these strategies, we analyzed whether these strategic components shifted toward optimality from the first task to the second task, in each condition. To do this, we estimated the number of snippets examined by using time on page from our baseline condition to determine the time to examine one snippet (8s), and used the time on page from the improved results page with slow results with the assumption that the times to examine a snippet are comparable. This value is capped at the number of results on the page (50). We employed a bootstrap hypothesis testing procedure [74], and present our findings in Tables 3.9 and 3.10.

Table 3.9 shows for that for *static gain* users, the time it took for users to first use slow search significantly decreased from the first to second task: from 386 s to 130 s ($p < 0.01$). Wait time increased significantly from 136.1 s to 277.3 s ($p < 0.01$). We see users moving closer to the optimal strategy for when to invoke slow search, but not for wait time. The number of queries did not change significantly, but increased slightly from 6.4 to 6.6. This could suggest that users did not know to take advantage of the fact that the preselected

[3]The times taken for each step would be as short as possible, and a user might elect to do other things, including regular searches, during the waiting interval.

[4]Insufficient data due to lack of use of this feature during the session.

documents were always in the middle of the ranking, and continued to search on their own even as they waited more for an effect.

Table 3.10 shows that in the *dynamic gain* condition, users invoked slow search much sooner for the second task (190.8 s to 135.9 s; $p < 0.05$), and significantly increased their waiting time (266.6 s to 277.1 s; $p < 0.05$). Additionally, although not statistically significant, we notice a decrease in the number of queries issued (5.38 to 3.68). This seems to suggest that these users had begun adjusting their behaviors toward the optimal strategy, as they developed a better mental model of how the system responded to their use of the "Work Harder" button.

3.4.9 Post-task Survey Results

After each task, we asked users about their experience using the system. We also asked participants in conditions having the "Work Harder" button to give their impressions on whether the button made the task easier, whether they noticed an improvement in the quality of results, as well as to write about their thoughts on the usefulness and ease of use of the system.

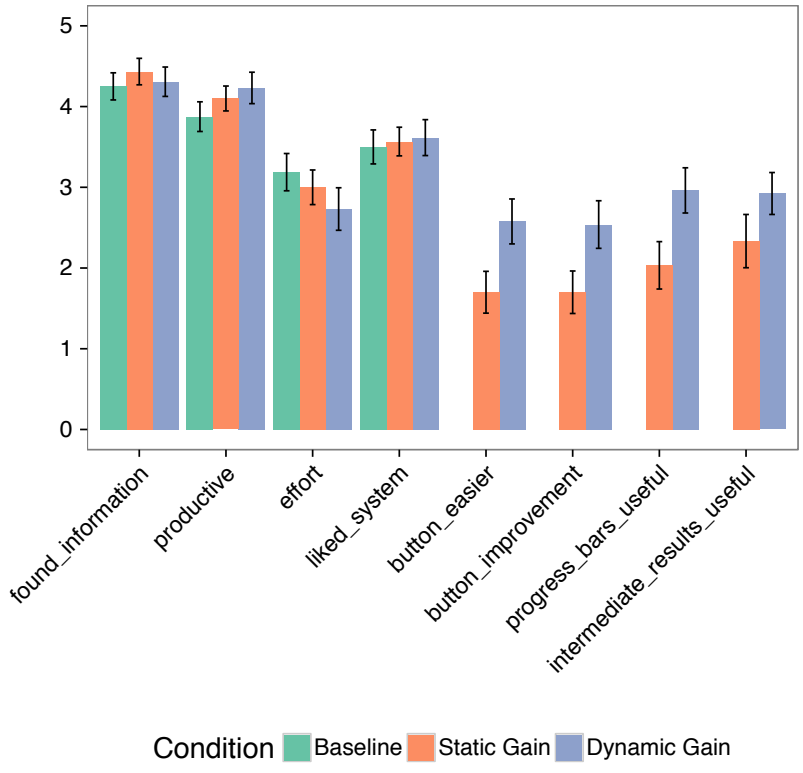


Figure 3.8: Post-task survey scores by condition.

Figure 3.8 shows the mean ratings on a five point Likert scale of participants' experiences of using the system by condition. The error bars represent the standard errors of the means. We performed ANOVAs on these results to see if exposure to the different conditions affected the ratings given to whether they were able to find the information they were looking for, their productivity, the effort extended, if they liked using the system, if the button made the task easier, if the button improved the quality of results, if the progress bars were useful, and if the ability to check the intermediate results was useful. Of these, we saw significant differences between conditions in response to the button making the task easier ($F(1, 54) = 5.324$, $p < 0.05$), the button improving the quality of the results ($F(1, 54) = 4.529$, $p < 0.05$), and the progress bars being useful ($F(1, 54) = 5.146$, $p < 0.05$).

In general, users' perceptions of the slow search features (the latter four in Figure 3.8) were higher in the *dynamic gain* condition than in the *static gain* condition. Among the more general experiential questions (the former four), we see that users in the *static gain* condition gave the highest rating for whether they found the information they were looking for ($M = 4.43$, $SD = 0.90$), while in the *dynamic gain* condition they gave the lowest rating ($M = 4$, $SD = 1.17$). For productivity, we see that having the button improved users' perceptions over the baseline ($M = 3.88$, $SD = 1.04$), with the *static gain* condition having the slight edge ($M = 4.1$, $SD = 0.84$) over the *dynamic gain* condition ($M = 4.07$, $SD = 1.01$). Users in the baseline condition also reported exerting the most effort ($M = 3.19$, $SD = 1.31$), which might have been reflected in their interactions, with users in this condition taking longer on average to complete tasks, perform more queries, and examine documents. Compared to the other conditions, users in the *static gain* condition reported liking the system the most ($M = 3.57$, $SD = 0.97$). Indeed, among these former four questions, the *static gain* condition has the highest ratings, though, once again, the differences were not statistically significant.

3.4.10 How users progressed towards a goal over time

In Figure 3.6, we plot the average relevance of the documents clicked by users, faceted by topic. The two most popular topics, *Entertainment* and *Local Businesses*, were the only two to have representation from all conditions in both sessions (with 27 *Entertainment* sessions and 31 *Local Businesses* sessions in total). We will first compare user behavior by topic, and then turn our attention to contrasting the characteristics of the sessions for each session as they correspond to the conditions within these topics.

Comparing behaviors by topic. In Table 3.11, we outline a list of features that we computed to characterize search behavior. We present these features by topic, as a central

question for the development of such a system is whether it will be used differently depending on the type of information need. We aggregated the needs by topic and compared the top two most addressed topics, *Entertainment* (28 completed tasks) and *Local Businesses* (32 completed tasks), against the combination of *Education* (21 completed tasks) and *Shopping* (7 completed tasks). We also performed pairwise Mann-Whitney U tests adjusted with Bonferroni correction to compare differences across topics.

As Table 3.11 shows, *Other* tasks had the longest mean session length at 1001 seconds. In comparison, *Local Businesses* had the shortest sessions on average at 793.4 seconds, while *Entertainment* had a mean session length of 841 seconds. *Local Businesses* also had the most regular queries in its sessions, with an average of 7.656. With this combination, *Local Businesses* also had the highest rate of regular queries in a session at 0.007928 queries per second, which is significantly higher than that of *Other* at 0.00497 ($p = 0.029$). *Local Businesses* additionally had an average dwell time of 247.2 seconds, which is significantly shorter than that of the *Other* category (392.8 seconds; $p = 0.023$).

Comparing rewards (and precision, which is related) in Table 3.12, *Entertainment* ends up having the worst performance outcome by users' answers at a $3.943 average reward and an average precision of 0.7286. This is significantly lower than the respective outcomes of the *Other* category, which has an average reward of $4.348 and an average precision of 0.8593.

Comparing task sessions. We computed the same behavioral features, separated by topic, to compare sessions being completed first versus sessions being completed second for the same topic. We additionally performed Mann-Whitney U tests to determine whether the differences observed between sessions were statistically significant. The full table was omitted for space. Most of the significant differences were seen in the *dynamic gain* condition.

For users in the *dynamic gain* condition, those who performed *Entertainment* queries cancelled fewer slow queries ($p = 0.05$) in the second session (2 slow queries cancelled to 0.2 cancelled on average). It should be noted that these users also performed fewer slow queries in the second session, but this difference was not statistically significant (a drop from 2 slow queries on average to 1; $p = 0.49$). This was the only condition and topic of the two for which there was a meaningful difference in cancellation behavior based on task order. For the same condition, users performing *Local Businesses* tasks did not cancel any queries in either session.

We also observed that *dynamic gain* users performing *Entertainment* tasks saw a decrease in the average relevance score of documents clicked in the session (from 4.67 to 2.4, $p = 0.03$). While the difference for *Local Businesses* was not significant, we note for comparison that the average relevance for *Local Businesses* increased from 2.2 to 4.6 between tasks ($p = 0.16$). Similarly, the relevance score for Education, which has data for this condition for both

task sessions, increased from 1.33 to 3 ($p = 0.35$). Users in this condition conducting *Entertainment* in fact viewed fewer (27.3 on average to 8; $p = 0.23$).

Dynamic gain users as well performing *Local Businesses* tasks switched from performing no slow queries in the first session to performing an average of one slow query in the second session. This increase was statistically significant ($p = 0.023$).

There were no statistically significant differences within the *static gain* condition for these topics, but we note that the number of slow queries dropped from 1.17 to 1 ($p = 1$) for *Entertainment* and increased from 0.2 to 0.625 ($p = 0.31$) for *Local Businesses*. This behavior is actually consistent within the topics: as previously noted, the slow queries performed in for *Entertainment* also dropped for *dynamic gain* from 2 to 1, and slow queries increased from zero to 1 for *Local Businesses* in the *dynamic gain* condition.

The only statistically significant difference between tasks for the baseline condition was seen in *Local Businesses* users, where their result clicks per query dropped from 2.07 to 0.996 ($p = 0.027$).

The differences within the *dynamic gain* condition and the lack of many significant differences in other conditions seems to point to a stronger effect of adjustment to the system for users exposed to both the asynchronous capabilities and time-biased gain [46].

We also found interesting differences in the way users progressed in these tasks, though the differences were not statistically significant. *Entertainment* sessions typically became shorter in duration in the second session across all conditions (994 seconds to 702.4 seconds on average), and *Local Businesses* sessions became longer (658.8 seconds to 893.3 seconds on average). In aggregate, the number of queries performed changed accordingly (8.3 to 5.4 for queries in session for *Entertainment* and 5.2 to 9.7 for *Local Businesses*). However, users examined more documents per query in *Entertainment* (2.6 to 2.8) and fewer in *Local Businesses* (3 to 2.5). Furthermore, the number of slow queries issued per session is concomitant with the number of regular queries by task and condition, however, we found that the number of highly relevant documents clicked increased during the second task compared to the first for both topics in the *dynamic gain* condition, which is the opposite of what was observed in other conditions. Thus, while users adjusted their behavior differently depending on the topic as they progressed through the experiment, users in the *dynamic gain* condition were able to consistently find the highly relevant documents regardless of other changes in interaction.

3.4.11 Performance Robustness

To measure performance, we used the reward earned per task by participants in each study condition. We present a histogram of these rewards in Figure 3.9 and summary statistics in Table 3.13. We performed Mann-Whitney U tests on these statistics by condition. However, none of these differences were statistically significant.

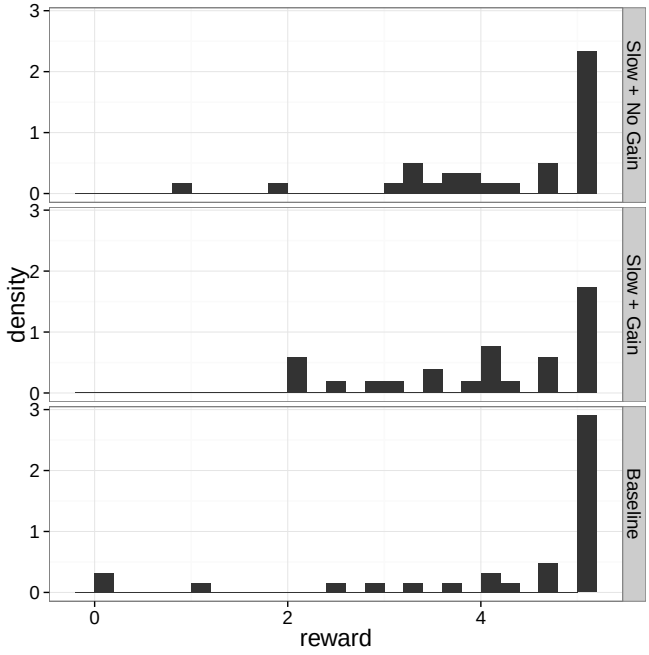


Figure 3.9: Distribution of rewards by study condition.

The *static gain* and *dynamic gain* conditions had the highest and lowest rewards respectively ($4.21 for the *static gain* and $4.04 for the *dynamic gain* condition). One possible explanation for this difference is that in the *static gain* condition, the system inserts all of the same documents that would have been inserted in the *dynamic gain* condition, but in the lower half of the ranking. As a result, *static gain* users gained access to these documents immediately, and seemed to be willing to look for them in the ranking. *Dynamic gain* users however had to wait to see these same results: in the *dynamic gain* condition, the system inserts documents into the ranking and improves the ranking of relevant documents steadily over the course of five minutes.

With that said, we also noticed that these two conditions had the smallest variances ($1.13 and $1.12 for *static gain* and *dynamic gain* respectively), compared to the baseline ($2.14). This gives the slow search conditions a greater degree of stability and predictability; for the baseline condition, especially in comparison to the *dynamic gain* condition, there is a greater risk, but also a potentially greater reward to using it for these multi-attribute

tasks. Therefore, we believe that there is usefulness in both types of search; in fact, having traditional search as an option to additionally having slow search may indeed be a useful approach.

3.5 Discussion and Implications

Our study provides insights about how users engaged with a slow search system that provided an asynchronous query capability with improvements in search result quality over time. We now discuss our main findings and implications.

Users are willing to wait for multi-attribute queries (RQ1). We found through our background survey that many of the tasks and queries that users typically have trouble with are multi-attribute queries in which various constraints of a query must be satisfied (Section 3.3.2.1). This justifies our use of such queries in our study, and the use of slow search after users gain familiarity with the system shows that multi-attribute queries are a good fit for a slow search system.

Users will typically wait for results (RQ2). Our background survey revealed that users reportedly are willing to wait for a mean of 9.5 minutes for "perfect" results (Section 3.3.2.2). Placing users under time pressure and imposing a maximum time of five minutes for query processing also led to users waiting, as was seen in Section 3.4.1. Interestingly, in both *dynamic gain* and *static gain* conditions, users typically submitted more slow queries, waited more, and cancelled fewer slow queries by the second task. A future study may manipulate the processing time for these slow queries to examine users' tolerance for waiting, and whether users will wait under tighter time constraints. Future studies may also look at impatience under greater uncertainty.

Users spent time looking for additional documents while waiting (RQ3). As illustrated in Section 3.4.2, users in both the *dynamic gain* and *static gain* conditions spent their time performing queries and clicking on documents in the interval while a slow query was processing. By the second task, these activities were reduced, but not in a statistically significant sense. We also showed that users in both of these conditions performed more slow queries in the second task and also waited more after performing these queries instead of clicking on documents or cancelling. This may indicate that users could still have been learning to use the system by the second task despite the training period and using the system for the first task. The reasons why users appeared to make more effective use of slow search by the second task require further study: the change could be due simply to their experience with the system in the first task, or it could be due to their increased awareness of the feature due to our explicitly asking users about their experience in using the "Work

Harder" button between tasks. A future study may extend the training period to ensure that users are not only familiar with the system, but that they are also confident in predicting what the system will do. We also plan to do a longer-term online study in which users interact with the system for an extended period of time, which will help us to determine how long it takes for user behavior to stabilize and what it looks like when it does. Such a study will also help to understand usage in different scenarios without artificial constraints.

User search behavior did not significantly change with additional slow search capabilities (RQ4). Our analysis in Section 3.4.6 showed that user behavior in terms of search interactions was similar across conditions, with users in the two slow conditions making significant use of the "Work Harder" button. We observed that users in the baseline condition took longer to complete sessions, conducted more and longer queries, and clicked on fewer documents per query (Table 3.6). Users, by the end of the study, may have still not yet fully understood the capabilities of the system. However, these results may also indicate that slow search systems should cater to similar types of queries as current search systems, and support the kinds of interactions that users have grown accustomed to. A future study may serve to tease out these differences by looking at users who have become familiar with the system and users without such a system.

Users did not achieve higher final effectiveness with slow search, but showed evidence of higher efficiency (RQ5). For the tasks we evaluated, users achieved comparable final rewards across the three conditions, with the *baseline condition* showing slightly higher average reward, but overall differences were not statistically significant. However, as we note above, users obtained these rewards in less overall time for both slow search conditions compared to the baseline condition, giving some evidence of higher efficiency. We also note that the reward variance in the baseline condition is higher than either slow condition, the reasons for which may be useful to explore in future work. In addition, users in the *dynamic gain* condition did indeed report that they noticed more of a difference in the improvement in search results than users in the *static gain* condition, and gave higher ratings for the usefulness of the progress bars. This may have been because it would have been clearer in the *dynamic gain* condition that the results were changing, and continued to change during the five minute duration. In contrast, users in the *static gain* condition may have not noticed the change between the unmodified and the modified results. Regardless, this shows that users are able to notice the difference when the results change, suggesting there is some utility in having future systems expose progressive improvements in ranking to users.

We note that users found slightly fewer relevant documents on average in the *dynamic gain* condition compared to the *static gain* condition. One explanation for this difference is that in the *static gain* condition, the system inserts all of the same documents that would

have been inserted in the *dynamic gain* condition, but in the lower quarter of the ranking. As a result, *static gain* users had the opportunity to gain access to these documents more quickly if they were willing to look for them in the ranking. In the *dynamic gain* condition, however, users had to wait longer to see the same highly relevant results, since the system begins with those documents at the bottom of the initial ranking and improves their position steadily over the course of five minutes.

While our dataset and corresponding analysis has allowed us to gain insight into the research questions we posed, we also recognize a number of limitations in our current study. Our findings, particularly that of RQ1, would be more robust with a larger sample of users. A future study in a more natural setting may also reduce experimental demand effects that might have influenced user behavior, and users' choices of tasks may have also affected their performance.

For future work, there are multiple possible avenues in exploring user interaction with slow search systems The 'Work Harder' button might be removed altogether and replaced with a background process that can automatically find and attempt to improve results for failed or abandoned search sessions. A user with low time pressure and a high degree of trust in a slow search system may submit a query to be processed in the background while performing non-search tasks, especially in the transition between devices, in which case supporting the examination of intermediate results or performing more queries in the interim would not be vital. In another instance, a user may use the system as as a supporting agent in a search task: the system would gather additional relevant documents and present them to the user as they continue to search. The participants' tendency in this study to continue searching shows that this is a useful capability to have. Along these lines, Microsoft implemented such a feature into their Bing Web search engine called "Deep Search" in December 2023 [1], which uses extra time – up to 30 seconds – to search for a more comprehensive set of results in the background. With Deep Search, a query is expanded using the large language model GPT-4 to a description of what the ideal results should look like to capture intent and expectations, and additional queries are used to search for a larger set of results which are then reranked.

The concept of "slowness" could additionally be applied to many different scenarios in which other aspects of retrieval 'quality' may be improved. This study focused on improving relevance for multi-attribute queries—a difficult class of queries for many existing systems, but in principle a system could also improve intrinsic diversity, employ crowdsourcing to augment algorithms, or summarize and organize results. Implementing slow search should take the costs of the design and any algorithms into account. Considering the goals of the users, simulation can be a low-cost tool for exploring design and algorithmic interventions, as

we highlight in Chapter 4. Additionally, on the system-side, we could reduce the amount of resources required for processing large volumes of requests by identifying tasks and queries for which slow search would be most helpful and offering the feature primarily in these circumstances.

The results of our study show that slow search is a robust alternative to Web search for multi-attribute tasks, minimizing the worst case performance that one experiences in using the system. We also compared how users behaved in satisfying a particular information need when a topic is approached first versus when the same topic is approached second, and show that there are many behavioral characteristics that change depending on when the topic is attacked when users are exposed to slow search with a gain in quality. We believe that this shows that there is a period of acclimatization in this case, where users take some time to adjust to the capabilities of the system.

One limitation of this study is the length of time for which users are exposed to the system. Our study sessions were designed to give users a ten minute exploratory "training" period to probe the capabilities of the system, after which they would have thirty minutes to solve each of two problems. This may have in fact not been enough time for training, and in a future study, we will adjust the training period to reduce novelty effects.

3.6 Conclusion

We reported on a user study that investigated five research questions about user interaction with a slow search system that offered users the option of running a 'slow' query in the background, showing progressive results in a sidebar. Using surveys and log data, we analyzed users who interacted with the system in one of three between-subjects conditions: a 'dynamic gain' condition that steadily improved search result quality of the optional slow query over the course of five minutes, a 'static gain' slow query that inserted relevant documents immediately with no additional ranking improvements over time, or a baseline condition giving conventional Web search results. Our findings suggest that users elected to perform slow search queries when given the opportunity. Additionally, we show that users are willing to wait for multi-attribute queries (RQ1), users will indeed wait for results when using slow search (RQ2), and users continued to search while waiting for results (RQ3). User behavior did not significantly change with additional slow search capabilities (RQ4), and users did not achieve higher final effectiveness with slow search, but did finish in less time (RQ5) on the tasks we evaluated. Followup studies may explore different mechanisms to improve quality for slow search and further investigate the nature of time-quality tradeoffs and user choice.

3.7 Author Contributions

This work was prepared by Ryan Burton and Kevyn Collins-Thompson. Ryan Burton was the main contributor to the work, having designed and conducted the experiment, as well as analyzing results and writing the manuscript. Kevyn Collins-Thompson contributed the curve-fitting procedure used in Figure 3.2 and revisions to the manuscript.

Feature	SG	DG	Base-line	U Test
Baseline features				
Session length ($\Sigma\Delta t^*$, sec.)	839.47	807.23	961.00	-
Regular queries ($CntQ_R$)	6.50	4.58	8.81	-
Regular queries/sec (Q_RPS)	0.01	0.01	0.01	-
Slow features				
Slow queries ($CntQ_S$)	0.73	0.81	0.00	$SG > B*$; $DG > B^*$
Slow queries per second (Q_SPS)	9.03×10^{-4}	9.99×10^{-4}	0.00	-
Slow queries cancelled ($CntQ_SC$)	0.33	0.35	0.00	$SG > B^*$; $DG > B*$
Slow queries cancelled per second (Q_SCPS)	3.86×10^{-4}	3.65×10^{-4}	0.00	$SG > B^*$; $DG > B^*$
Query features				
Query word length (QWL)	4.48	5.20	5.72	-
Query character length (QCL)	27.14	31.20	34.88	-
Click features				
Pages in session ($CntR$)	13.27	11.62	14.48	-
Clicks per query (CPQ)	3.14	3.23	2.37	-
Time to first click (sec.) (Q_{DT})	24.93	25.53	24.46	-
Dwell Time (sec.) (C_{IT})	234.36	276.28	358.86	-
Outcomes				
Reward ($)	4.21	4.04	4.16	-
Reward Variance ($)	1.13	1.12	2.14	-
Precision	0.85	0.72	0.83	-
Click Relevance	4.03	3.08	3.55	-

Table 3.6: Comparison of behavioral features across conditions. SG = *Static Gain*; DG = *Dynamic Gain*. $^*p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

Feature	NS	S	U Test
Baseline features			
Session length ($\Sigma\Delta t^*$, sec.)	847.69	891.10	$NS < S*$
Regular queries ($CntQ_R$)	7.53	6.20	-
Regular queries/sec (Q_RPS)	0.01	0.01	-
Slow query features			
Slow queries ($CntQ_S$)	0.39	0.57	-
Slow queries per second (Q_SPS)	5.22×10^{-4}	6.72×10^{-4}	-
Slow queries cancelled ($CntQ_SC$)	0.14	0.27	-
Slow queries cancelled/sec (Q_SCPS)	2.97×10^{-4}	1.65×10^{-4}	-
Query features			
Query word length (QWL)	5.01	5.23	-
Query character length (QCL)	28.90	32.67	-
Click features			
Pages in session ($CntR$)	12.44	13.75	-
Clicks per query (CPQ)	2.55	3.14	-
Dwell Time (sec.) (C_{IT})	284.20	296.36	-
Time to first click (sec.) (Q_{DT})	13.65	32.86	$NS < S^*$
Outcomes			
Reward ($)	3.23	4.74	-
Reward Variance ($)	1.70	0.36	-
Precision	0.52	1.00	$NS < S^{***}$
Click Relevance	3.42	3.69	-

Table 3.7: Comparison of behavioral features by success level. NS = Not Successful; S = Successful. $^*p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

	β	S.E.	p-value
Intercept	**-6.524**	**2.927**	**0.026**
Session length ($\Sigma\Delta t^*$)	0.003	0.002	0.136
Condition: *static gain*	4.837	3.304	0.143
Condition: *dynamic gain*	3.637	4.203	0.387
Regular queries ($CntQ_R$)	-0.122	0.147	0.404
Regular queries/sec (Q_RPS)	265.700	161.500	0.100
Slow queries ($CntQ_S$)	6.089	6.428	0.360
Slow queries per second (Q_SPS)	-4954.000	4807.000	0.303
Slow queries cancelled ($CntQ_SC$)	-0.989	6.428	0.878
Slow queries cancelled/sec (Q_SCPS)	1840.000	4695.000	0.695
Query word length (QWL)	-0.164	0.176	0.351
Pages in session ($CntR$)	-0.025	0.072	0.725
Clicks per query (CPQ)	**0.887**	**0.437**	**0.042**
Dwell time (C_{IT})	-0.001	0.003	0.697
Time to first click (sec.) (Q_{DT})	-0.029	0.027	0.267
$\Sigma\Delta t * \times$ Condition: *static gain*	-0.009	0.005	0.080
$\Sigma\Delta t * \times$ Condition: *dynamic gain*	-0.009	0.006	0.158

Table 3.8: Logistic Regression for predicting user success

Component	Task 1	Task 2	p-value
Queries	6.4	6.6	0.479
Time to "Work Harder" (sec.)	386	130.1	0.005 **
Wait time (sec.)	136.1	277.3	0.002 **
Snippets examined	NA[4]	22.5	0.014 *

Table 3.9: Changes in interaction in relation to optimal strategies for *static gain*. The number of snippets examined is estimated. *p¡0.05; **p¡0.01; ***p¡0.001.

Component	Task 1	Task 2	p-value
Queries	5.38	3.68	0.131
Time to "Work Harder" (sec.)	190.8	135.9	0.04 *
Wait time (sec.)	266.6	277.1	0.048 *
Snippets examined	21.2	20.1	0.5

Table 3.10: Changes in interaction in relation to optimal strategies for *dynamic gain*. The number of snippets examined is estimated. *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

	Entertainment (E)	**Local Businesses (L)**	**Other (O)**	**U Test**
Baseline features				
Session length (sec.)	841	793.4	1001	-
Regular queries	6.821	7.656	5.593	-
Regular queries/sec	0.006827	0.007928	0.00497	$L > O$
Slow query features				
Slow queries	0.75	0.3438	0.4074	-
Slow queries/sec	0.000913	0.000416	0.000526	-
Slow queries cancelled	0.4286	0.0625	0.1852	-
Slow queries cancelled/sec	0.000439	0.000079	0.000232	-
Query features				
Query word length	4.982	4.907	5.565	-
Query character length	28.7	29.8	35.17	-
Click features				
Pages in session	13.14	14.41	11.85	-
Clicks per query	2.684	2.707	3.335	-
highly.relevant.clicks	0.6786	0.4688	1.037	-
relevant.clicks	2.321	2.281	1.815	-
non.relevant.clicks	10.14	11.66	9	-
Time to first click (sec.)	15.1	23.85	36.06	-
Dwell time (sec.)	245.9	247.2	392.8	$L < O$

Table 3.11: Comparison of behavioral features by topic.

	Entertainment (E)	**Local Businesses (L)**	**Other (O)**	**U Test**
Outcomes				
Click relevance	3.679	3.219	3.889	-
Reward (dollars)	3.943	4.075	4.348	$E < O$
Reward Variance (dollars)	1.025	1.579	1.79	-
correct.answers	3.643	4.094	4.296	$E < O$
Precision	0.7286	0.8187	0.8593	$E < O$

Table 3.12: Comparison of outcomes by topic.

	Static Gain	***Dynamic Gain***	**Baseline**
Reward ($)	4.21	4.04	4.16
Reward Variance ($)	1.13	1.12	2.14

Table 3.13: Reward means and variances by study condition

CHAPTER 4

Simulation Towards Optimal Behaviour

4.1 Introduction

Information retrieval (IR), starting from its earliest studies, has been driven by measurement and performance metrics. Starting with the traditional Cranfield evaluation framework [47], it soon became clear that human factors were also an important aspect to be considered and measured [171]. Since then, work in IR has taken cues from other areas focused on user effort and satisfaction, including but not limited to human-computer interaction [85, 209], recommender systems, and economics [9, 10].

The process of evaluating these human factors typically takes the form of user studies [111], but performing these at scale is often prohibitively expensive. Other, cheaper methods involving real users may involve crowdsourcing studies or log data analysis of existing systems. Alternatively, stochastic simulation [180, 20, 137, 138] presents an extremely efficient method for running experiments at scale, albeit with synthetic users. Besides cost and versatility in mimicking a wide variety of users and systems at scale for information retrieval, simulation is also heavily used in other domains such as computational physics, materials science, and as we will explore briefly, financial investments. This suggests an avenue of integration, and we will introduce the use of a financial analysis technique known as *real options pricing* as a means for quantifying the benefits of a new interface feature or potential system-level intervention.

In economics and management, flexibility may be seen as a competitive advantage that can be analysed through the lens of real options [22, 45], which gives a way to measure the value of a potential decision that takes advantage of flexibility in light of uncertainty. Real options represent an opportunity to undertake an initiative. We can view it as a staged investment or decision. Bengtsson [22] highlighted four central ways to increase operational production flexibility: reducing set-up time at installed equipment, multipurpose stations, parallel assembly lines, and/or a flexible work force. We can reframe these approaches for

information retrieval, such that reducing set-up time can be seen as the cost of switching search engines or rankings, multipurpose stations (or a flexible manufacturing system) can be seen as IR system integration, parallel stations could be framed as parallel searches, and a flexible work force can be viewed in terms of ranking and matching algorithms. Seen this way, real options seem to have particular applicability to the optimisation and time-quality cost tradeoffs that we see in interactive information retrieval (IIR), both from a user's perspective and for an IR system implementation.

Using this economic IIR formulation, we may consider a system that presents users with the choice of an new interface element as an *option*. Such an element presents some presumable benefit to the user, but with associated costs such as using the element through effort, suffering from a loss in screen real estate, or spending extra attention on this new element. For companies performing options valuations, their main source of uncertainty is the demand of the goods and/or services they produce. For the user of a search engine, their source of uncertainty may be whether a feature will be useful for their task. The designer of a system has the responsibility to convey the value of the feature in a way that is easy to understand and supports their decision-making. When performing a search, a user has a particular mental model that guides their expectations of their use of a feature. Hence, this chapter centers around showing how real options valuation may prove to be a useful framework with applicability to IR system design through methods that give quantifiable results while considering uncertainty, value, and cost. We will in fact use this framework to investigate value and risk of choosing different rankings for our system, using simulation as an instrument. More specifically, we use a user simulation model to perform a Monte-Carlo integration of a utility function that quantifies the option value to a user of having the choice to access an extra interface feature. In our experiment, we give the user a *sidebar* that may provide better results asynchronously during a search task as an option – this is an element that may present some uncertainty through use, that is, the user may not be sure if it is worth their effort. A fundamental contribution of this work therefore is a method for quantifying the potential future value of an information access option for a user.

Our previous experiences with conducting user studies on novel search interfaces made us curious not only if we could systematically ensure that users were willing to engage with a new element of the system, but also how changing various aspects of the given element could lead to improved outcomes. With a simulation framework to explore these types of *what-if* questions, we identified the following research questions as grounds for our study:

RQ1: *How close is our simulation to the behaviour of actual users?* As we create our simulation framework using the behavioural characteristics of the users of an experimental pilot as a basis, we would like to determine whether our simulation can appropriately deliver

similar outcomes as our pilot users.

RQ2: *What differences exist between real versus perceived quality of search results?* Using simulation, we attempt to answer the question, "What if a user believes the sidebar is better than it actually is?" To do this, we will manipulate the decision probabilities of the simulation to reflect this change in belief.

RQ3: *How can we increase the value of the sidebar – a novel interface element?* To do this, we introduce the measure *option value*, commonly used in financial investment and risk management to estimate the expected future value of a given investment and apply this measure to find system parameters that increase the value of our aforementioned sidebar. This gets towards the problem of optimising system behaviour given a fixed user.

RQ4: *What state transitions of our simulation characterise high and low performance?* Our simulation will make stochastic decisions about how to behave as it proceeds through a search task. For this question, we aim to explore what types of user decisions would lead to the best outcomes in terms of cumulative relevance over the course of the task, and conversely which would lead to the worst. Contra to RQ3, this research question gets towards the problem of optimising user behaviour given a fixed system.

RQ5: *How might we guide a user from a state of low performance to high performance through system or affordance changes?* Because we are able to run many simulations resulting in a spectrum of performance outcomes, we can explore the space of parameters that led to these outcomes and determine which changes in parameters may lead to better outcomes within the space. We will do this with the *t-SNE* dimensionality reduction technique.

4.2 Related Work

Simulation and Time-biased Gain. Simulation as tool for investigating session behaviour in interactive information retrieval has been receiving increasing interest in recent years. As a framework, it gives the ability to extend beyond the Cranfield view of information retrieval to an interactive one while remaining fast and inexpensive to perform experiments.

Azzopardi [13] established a cost model of browsing search engine result pages based on estimates of the time required by both the system and user for clicking, scrolling, and inspecting snippets, while taking into account the size of the page, and the size of the screen used. The motivating scenario involves a user being presented with a search engine result page immediately after issuing a query. Optimally, for a given device, we may consider whether the interface should optimally show as many results as can fit above the fold with pagination, or should it allow for some scrolling before going to the next page. This paper follows in a line of other endeavours to model *cost*, as opposed to the more common avenue

of optimising *gain*. Kashyap et al. [106] formulated a cost model of faceted navigation for a system called FACeTOR that accounts for the time to examine results, the cost of choosing a facet and hence refining the list of results, or expands an attribute, revealing more facets of the particular attribute. Through simulated navigation and a user study via Amazon Mechanical Turk[1], they tested the predictions of their model and found that their cost model was realistic. The cost model of [170] served to inform the activity of sensemaking – a task more general than information retrieval. The authors decomposed the process into different types of subclasses, and characterized their costs. By doing so, they made the point that by trading off costs in one task, we can take advantage of the reduced costs in other aspects. As such, sensemaking becomes an anytime algorithm [6]. As an example, by saving time expenses from automated clustering methods, the designers of an educational course were able to extend the comprehensiveness of their search.

For examples of work which examine *gain* in information retrieval, we may turn towards analyses conducted by Smucker and Clarke [181]. Following on their proposition of time-biased gain, where the gain from a relevant document is equal to the probability of viewing it subjected to a time decay [181], Smucker and Clarke [180] further explored a simulation-based approach to approximate this gain as an alternative to estimation using their closed-form solution. This allows for more flexibility in analysis – one can model a distribution of gain while changing other variables with less effort for easier "what-if" experiments. Taking this approach also potentially allows one to model a sample from a population of users.

Other work on simulation has used the technique to explore the diversity of strategies during search sessions [20], the effectiveness of query personalisation [201], and the usefulness of various implicit relevance feedback models [214]. More recent studies have increased the sophistication of their models to improve the believability of their agent behaviours. In work by Maxwell and Azzopardi [137], the authors noted that typical models lacked a degree of realism that accounts for the agency of simulated users. Their main contribution was a model that incorporated cognitive state – that is, data about what the user knows, saw during tasks, and found relevant. Such a formulation was seen as useful for more realistic interactions such as following information scent and exhibiting stopping behaviour through abandonment [138].

Real Options. Simulation is also effective for the purpose of real option valuation in finance. In effect, the pricing of an option involves computing a (possibly complex) integral over time, and simulating stochastic paths in asset pricing space is often the only feasible way to compute an approximate option integral for more complex types of option scenarios that have no closed-form solution. There are parallels between the application of real options

[1] https://www.mturk.com/

in finance and manufacturing, and a potential application in information systems in terms of the value of increasing flexibility.

Exactly the same principle applies to our stochastic information retrieval scenario, except that asset prices are replaced by cumulative relevance over time. The Datar-Mathews method [134] is a Monte Carlo approximation of the Black-Scholes pricing options formula that, given various cash-flow scenarios, creates a probability distribution of expected net present value, that is, discounted cash flows. This "pay-off" distribution allows one to calculate real option value by finding the probability weighted mean while mapping negative values to zero[2][48]. Intuitively, the real option value using this method is calculated as follows:

Real Option Value = Risk Adjusted Success Probability × (Benefits - Costs).

When used in a Monte Carlo simulation, the calculation becomes

$$E[\max(e^{-\mu t}X_p - e^{-rt}X_c, 0)]$$

where X_p and X_c are random variables for operating profits and launch costs respectively, and μ and r are the risky-asset and risk-free discount rates respectively [134].

In finance, the Datar-Mathews method may be preferred to the Black-Scholes formula because of its transparency [55]: whereas the Black-Scholes closed-form formula is more of a black box with easily violated assumptions, the Datar-Mathews method is more flexible in the types of cash-flow distributions it can deal with (such as non-lognormal distributions) and uses the standard inputs typically used in net present value analysis. Furthermore, the resulting distribution can be traced by following the algorithm. Thus the Datar-Mathews method is more suitable to non-financial scenarios like our stochastic interaction scenario.

User Perceptions of Quality. In recommender systems and algorithmic sensemaking [218], the effects of operational transparency – revealing the work behind a service – has been explored as a means of improving perceptions of effort and trust [34]. Ethical principles of design exhort that deception is a negative practice, but design naturally invokes the use of deception through abstraction, metaphors, and affordances. Computing pioneer Alan Kay has in fact described the correspondence between what a user sees on the screen and what they think they manipulate as an illusion. In his words, from his experience at Xerox PARC, "There are clear connotations to the stage, theatrics and magic–all of which give much stronger hints as to the direction to be followed." [108]. Despite its prevalence, it is rarely described in principled terms, but has recently become to be known as *benevolent deception* [2]. One example given by Adar et al. is a robot therapist that under-reports the amount of force that a stroke patient exerts in order to help in overcoming their self-imposed

[2]Because an agent is not obligated to exercise an option, it is assumed that they will never exercise if it leads to a loss, that is, below zero.

limits [2]. Research on recommender systems has focused on various ways of manipulating beliefs and perceptions of various attributes of these systems, including the relationship between user effort and *perceived* user effort with a loading screen [194], persuasion [76, 59], or nudging [95]. Misbeliefs and biases have been studied in information retrieval, primarily for the purpose of mitigating or correcting them [208, 212, 11, 14]. Benevolent deception has not received as much interest thus far in the field.

4.3 Method

Using the experience and log data we gained from conducting a pilot study, we formulated a simulation that aimed to capture the behavioural aspects of users and explored the implications of changes to both the aspects of the system and aspects of user behaviour reflected in simulated data. Much of our motivation stemmed from the perspective of looking at the costs and benefits of presenting a user with the choice of using a new interface option such as an alternative ranking, and asking the question 'How can we measure the potential future value of this interface feature to this user?'. As we describe below, we use real options pricing strategies that arise naturally from stochastic simulation. We will now briefly describe the specifics of our search system that motivated this question, which has a *sidebar* of high-quality documents delivered slowly to reflect the system's ability to explore alternative time-quality tradeoffs in parallel with the usual immediate ranking.

4.3.1 Pilot Study System Description

We built a custom search interface that mimics that of a typical commercial search engine, but with an added sidebar at the right of the screen that presents additional relevant results incrementally over time. This sidebar is evocative of an algorithm that might be used in a "slow search" system [187] – trading result timeliness for improved result quality. The system uses a simulated algorithm – as opposed to a real and functional algorithm – to provide these better results using documents that were pre-labelled with relevance judgements depending on the task that the user is currently in the process of solving. The fact that the process is simulated rather than the result of an algorithm working to find better results is not made explicit to the user. The task and system setup are heavily inspired by those in Chapter 3 and such, a "high-quality" document is considered to be one that directly addresses the multi-aspect task given, and one that therefore saves additional search time that would otherwise be time looking through documents in the main ranking or issuing new queries. A screen shot of the interface for this experiment is shown in Figure 4.1.

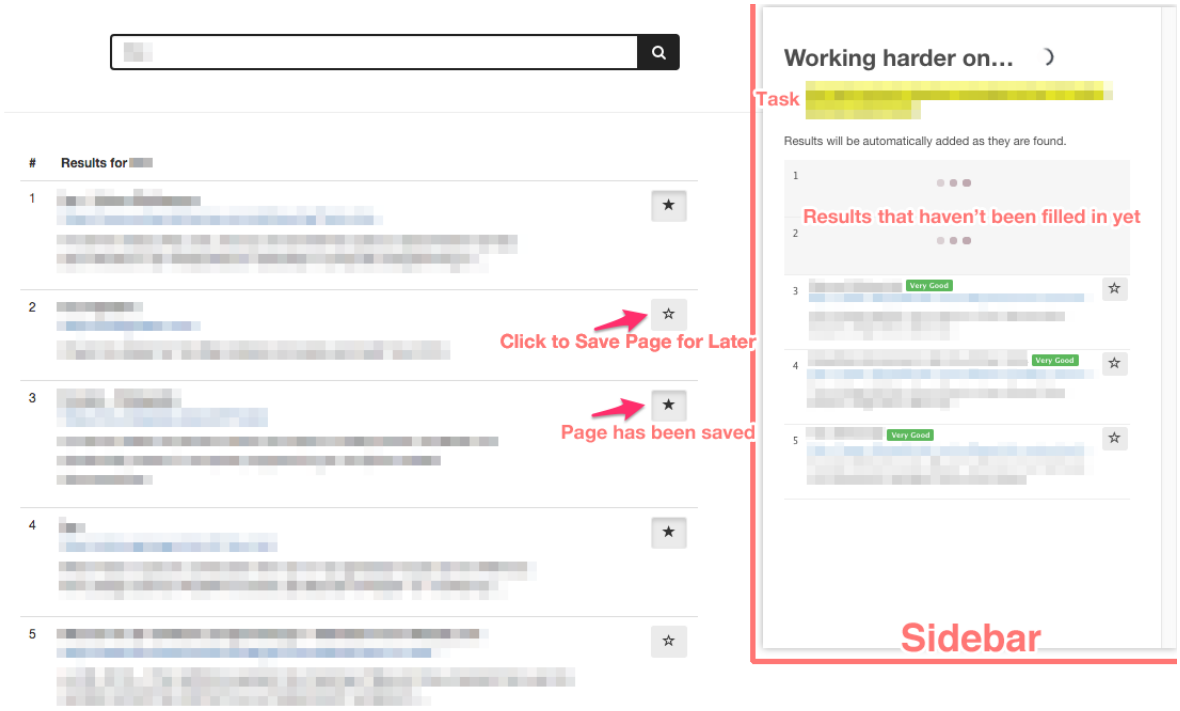


Figure 4.1: Screenshot of interface used in pilot study. The stars to the right of each result is a "save" button – the results saved can be called up at the end of the task for answering questions. Items in the sidebar are explicitly given labels denoting the "quality" of the results. In the sidebar, results arrive and are populated in reverse order starting with the lowest ranked. Sidebar animations in the form of a spinning indicator beside the title ("Working harder on...") and throbbers in the unfilled slots of the results provide a degree of operational transparency to show that more results are coming.

As opposed to interleaving higher quality search results in the pre-existing ranking or performing re-ranking, the use of the sidebar presents the user with a *clear option* of using another interface element and set of results. Besides quality, this allows us to examine the effects of other aspects such as wait time and the rate at which results are added to the sidebar on users' preference for this option for various tasks.

The tasks given fell into four categories: education, local businesses, shopping, and entertainment. These tasks were all multi-aspect with various constraints to be satisfied; an example task in *local businesses* was, "Find five local I.T. companies with at least 50 employees." Participants should satisfy all the constraints to receive the full reward for completion. For our pilot however, participants were uncompensated.

Our pilot consisted of 6 participants, each of whom completed 2–6 tasks. The study flow was completely automated after the calibration of an eye tracker to record gaze information, and began with a tutorial in which participants were given a primer on the system's functionality and usage. Each task that was performed by the user was followed by a ques-

State
Start (*start*)
Submit Query (*query*)
View Next Sidebar Result (*view_sidebar*)
Click Sidebar Result (*click_sidebar*)
Save Sidebar Result (*save_sidebar*)
View Next Main Ranking Result (*view_main*)
Click Main Result (*click_main*)
Save Main Ranking Result (*save_main*)

Table 4.1: List of Simulation States and their aliases as we use throughout the remainder of the paper.

tionnaire that solicited opinions about the usefulness and quality of the sidebar, as well as their perceptions about their own performance.

4.3.2 Model Description

We begin by enumerating the list of states that our simulation might take. These are shown in Table 4.1. We chose behaviours that represent those that users perform for querying, examining snippets, clicking results, saving links for later, and reading documents. Because our system also includes two rankings, we separate the examination, clicking, and saving behaviours into those performed on the *main results* on the left of Figure 4.1, and the *sidebar* on the right.

In Table 4.2, we show the list of time-based costs we incorporate into the simulation. The costs of reading the next sidebar and main result snippets (*view_sidebar* and *view_main* respectively) are both approximately 4.52 seconds. To arrive at this time, we took inspiration from Smucker and Clarke [180] to use a Weibull distribution model for snippet judgements, but for simplicity we use the mean of the distribution and assume that because both rankings use similar snippets, the time necessary to judge them will be similar. The times to click a sidebar result and click a main result (alternatively *click_sidebar* and *click_main* respectively) were estimated as approximately the time between clicks in the main ranking as measured during our pilot; we assume, again for simplicity, that the documents in each ranking will take similar amounts of time to read.

Using the data from our pilot experiment, we calculated the probabilities of each state transition, which we show in Figure 4.2.

For the utility gained by our simulated users, we use three levels of relevance: not relevant ($R = 0$), relevant ($R = 1$), and highly relevant ($R = 2$). Table 4.3 shows the interaction

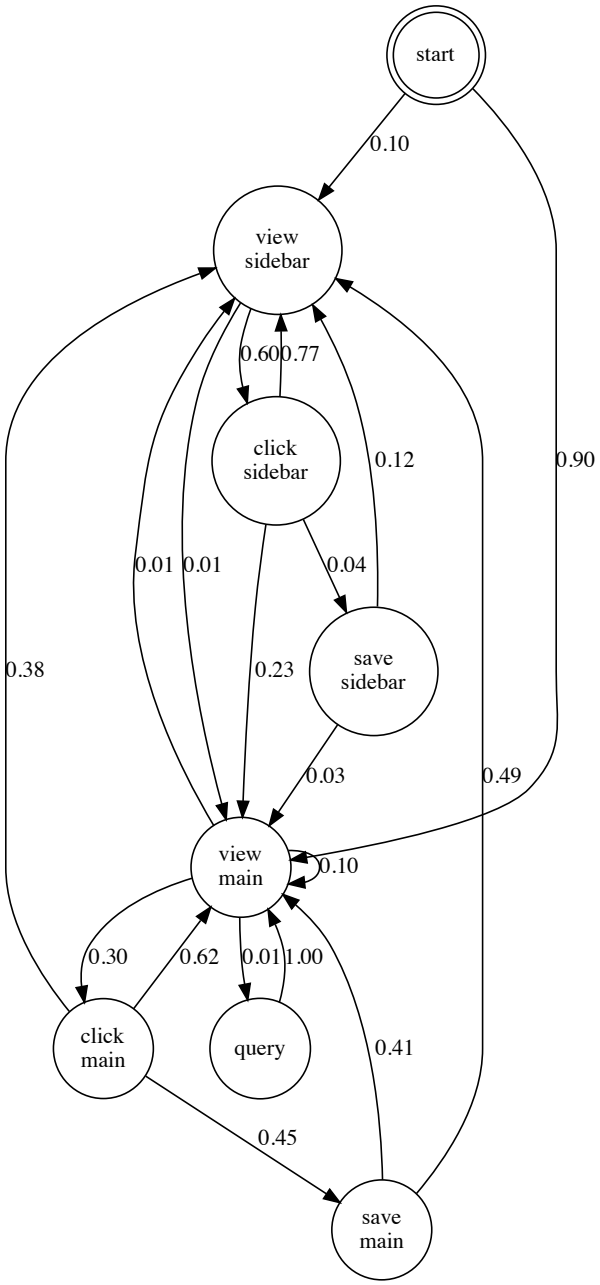


Figure 4.2: State transitions taken by our simulation. The transition probabilities are estimated from user data in our pilot experiment.

State	Cost (sec.)
View Next Sidebar Result	4.52
View Next Main Ranking Result	4.52
Click Sidebar Result	10
Click Main Result	10

Table 4.2: List of Time-based Costs in the Simulation.

State	Description	Value
$P(C = 1\|R = 2)$	Probability of clicking on a highly relevant snippet	0.3731343
$P(C = 1\|R = 1)$	Probability of clicking on a relevant snippet	0.4477612
$P(C = 1\|R = 0)$	Probability of clicking on a non-relevant snippet	0.1462687
$P(S = 1\|R = 2)$	Probability of saving on a highly relevant document	0.3924051
$P(S = 1\|R = 1)$	Probability of saving a relevant document	0.4367089
$P(S = 1\|R = 0)$	Probability of saving a non-relevant document	0.0664557

Table 4.3: List of Interaction Probabilities in the Simulation.

probabilities in the simulation – the probabilities of clicking or saving documents conditioned on their relevance. These values were in fact calibrated to [180], and adjusted to three grades by splitting the probability density of the lesser two grades. Because our click data was relatively sparse in our pilot, we opted for this estimation, and as we will see in Section 4.4.1, seems to give results similar enough to real users.

Sidebar–Main Results Correlation. We also add an optional pair of a pair of parameters to capture 1. the degree of correlation between the information in the sidebar and main results (μ_{corr}), and 2. the proportion of main results that are correlated with the sidebar (p_{corr}). These parameters are used when calculating the observed relevance rewards accumulated by a simulated user. Our expectation is that a large value for the correlation will reduce the benefit of examining the sidebar. The addition of these parameters gives us an additional dimension to explore when comparing the effects of changing various aspects

of the system. The effect of the correlation goes in both directions: clicking on an item in the sidebar after one has seen the same information from the default ranking should reduce the gain in relevance, *and vice versa*. We only consider the correlations between snippets in this work, and as such, only the rewards from clicking on search results will be affected. Extending this to correlations between the information on the pages or documents themselves would conceivably affect the reward that comes from saving after viewing those documents.

The the purpose of the simulation, during each run we randomly associate each result in the main ranking with a correlation, which approximately averages to our correlation parameter μ_{corr}. Furthermore, we associate each of these results with a result in the sidebar so that we can apply a discount factor equal to the correlation if one of these correlated results were clicked after the associated result had already been clicked in the opposing ranking. The number of main ranking results that have this association is determined by the second parameter p_{corr} mentioned above.

rel-AUC. For the rest of the paper we will refer to a measure that we call *rel-AUC*. We calculate this as the area under a time-relevance curve, as might be seen in Figure 4.5. This serves as a convenient measure, as it gives us a single number for comparison over the course of simulated sessions. We plot this simulated outcome *rel-AUC* as a factor of proportion of ranking correlated, and the correlation in Figure 4.3.

Considering the quality of the sidebar (each result may have one of three levels – very good, good, and not relevant), we look at the effect of all results having either poor or very good quality in Figure 4.4. This serves as to verify our expections, that as the quality of the results in the sidebar improves, so does the outcome of our simulation.

4.4 Results

4.4.1 Comparing Simulation to User Behaviour

As a sanity check, it is useful to ensure that our simulation is capable of realistic user behaviour. This sanity check will form the basis of our first research question: *How close is our simulation to the behaviour of actual users?*. For this question, we used the data from a pilot study of user indifference and the tradeoffs between risk and value.

To match the expected query behaviour of a user, we assume in the simulation that *on average*, each additional query issued is likely to increase the relevance of the results by at least a small amount. (The increase factor, chosen heuristically in this study, was 0.25.) Another assumption of note was that the probability of issuing a new query would increase as our simulated user moves further down the main ranking on the left by increments of

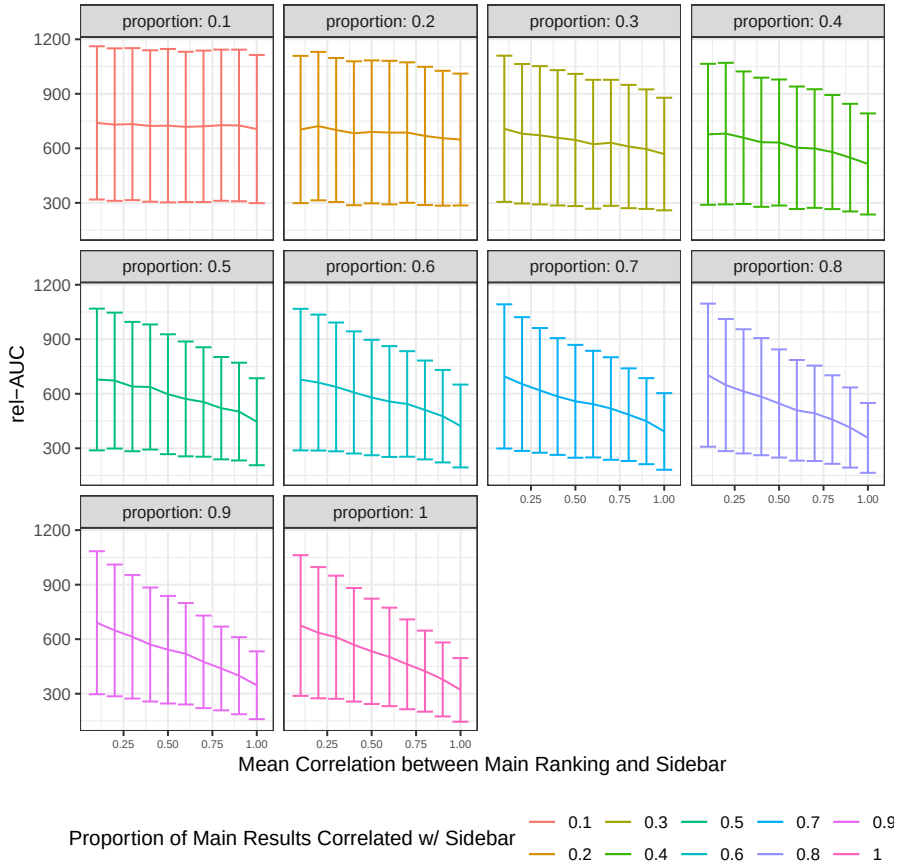
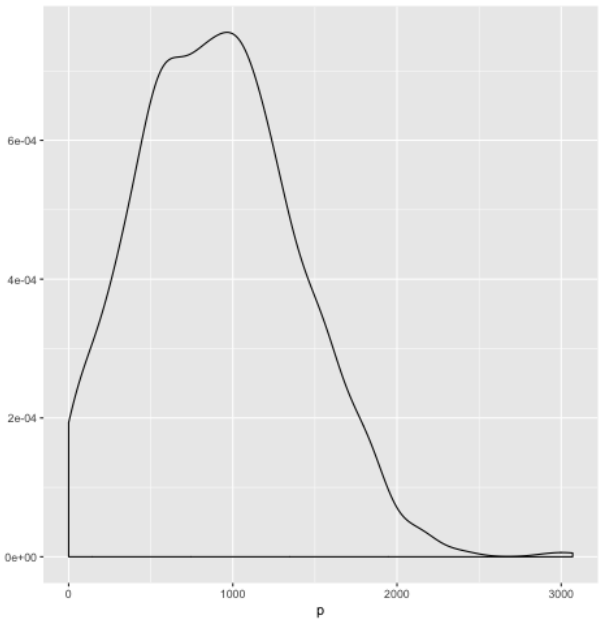
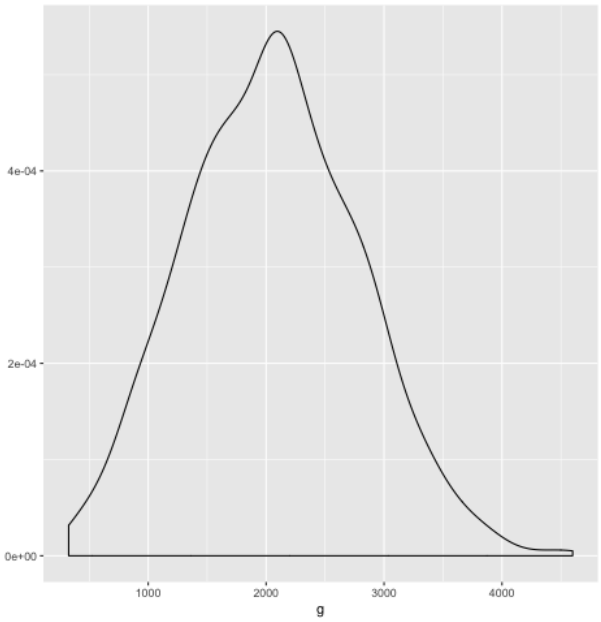


Figure 4.3: Our outcome measure *rel-AUC* decreases as correlation increases and the proportion of main results correlated with the sidebar increases in simulations with 10,000 runs each. Bars show the standard deviations.



(a) Quality: Poor



(b) Quality: Very Good

Figure 4.4: *rel-AUC* improves with sidebar quality during simulation, showing probability distribution of utility outcomes conditioned on low vs high-quality sidebar results

$\frac{n}{n_{max}}$, where n is the ranking position and n_{max} the number of results in the main ranking.

For the utility gained by our simulated users, we use three levels of relevance: not relevant, relevant, and highly relevant. The results in the sidebar were given relevance judgements by one of the authors, but the main results were evaluated for relevance via crowdsourcing on Amazon Mechanical Turk and Figure Eight[3]. Due to the nature of the tasks, as outlined in Section 4.3.1, we use a notion of relevance that more closely aligns with the multi-aspect nature of the type of search under study. More specifically, relevance is considered along the dimensions of whether a result 1. addresses the needs of the task, that is, if the search intent is satisfied; 2. the degree to which all aspects of the task appear in the result, and 3. the result is high-quality and correct. For more on the specifics of the crowdsourcing task, including the instructions given to crowdworkers, please see Appendix A.

To finally address ***RQ1***, we computed the average relevance curve across all six of our pilot users, as well as the average relevance curve across 1000 simulations using the rankings from the pilot runs that our users saw. We plotted these curves for comparison in Figure 4.5. As should be evident, the two curves track each other remarkably well. We also performed statistical tests, the first being a one-sample t-test on the differences between the two curves without the cumulative sum applied, which yielded a p-value of 0.22 ($t = 1.2245$). With the null hypothesis of the difference being equal to zero not rejected, this suggests that the curves are similar. We also performed an adaptive Neyman test for the difference between two curves. This test has been shown to be statistically sound for hypothesis testing between curves and involves applying a discrete Fourier transform to decorrelate stationary data into nearly independent frequency components [65]. The test statistic $T_{AN} = 5028.13$ with $\hat{m} = 87$, which resulted in a p-value < 0.01. We can therefore say that our simulation is capable of behaviour that closely matches the outcomes of real users in terms of the rate of relevance gain.

4.4.2 Differences in Real and Perceived Quality

For this section in which we address ***RQ2***, we are motivated by a hypothetical: *what happens if someone believes an interface element is more useful than it actually is?* We believe this counterfactual could serve an interesting purpose, and may be worth exploring as it has implications for *benevolent deception* [2] in design. As we mentioned in the Related Work (Section 4.2), this refers to a manipulation of belief in a user's mental model of a system for the benefit of a user as well as the developer. For the design of a search system, the implication here may involve signalling an overestimate of usefulness for an interface element

[3] https://www.figure-eight.com/

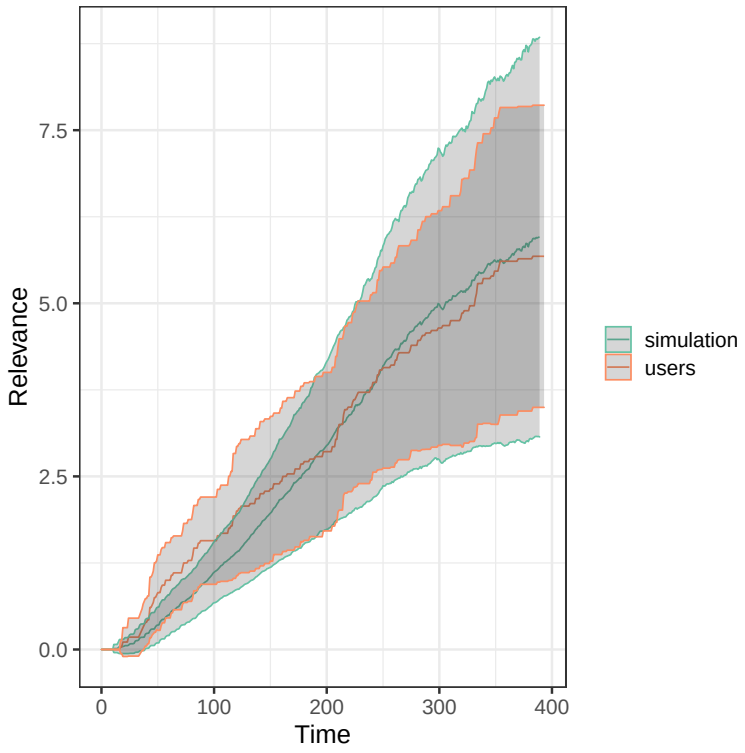


Figure 4.5: Mean cumulative relevance over time per task of our simulation using parameters estimated from our user pilot data and actual user outcomes. Relevance is achieved through clicks on relevant results. Simulation is averaged over 1,000 runs. Rates of relevance gain for the simulation and users are similar. Bands show the standard deviation.

that a user themselves *underestimates* the usefulness of. In the context of our experimental search system, the interface element is the sidebar of auxiliary high-quality results that our users may be hesitant to trust or engage with. Here, we know that users will benefit from the results we present, but to the users, this may not be clear.

To investigate this change, we make two concrete assumptions – that a change in a user's belief of the sidebar's usefulness manifests as a change in behaviour, and that a user that examines the sidebar will be able to accurately assess the quality of the results contained therein. These assumptions allow us to observe this change in preference as well as operationalise it in our simulation. The model describes the transition from the state "viewing the main results" to the state "viewing the sidebar" and vice-versa with probabilities of transitioning between states. We can cast these transition probabilities as *preferences* for one ranking or the other, such that a preference for the main results would be accompanied by a suitably low switching probability from the main results to the sidebar. Similarly, a preference for the sidebar would suggest a high switching probability from the main results to the sidebar.

Sidebar Quality. Closely related to relevance of the results contained in the sidebar, we define the notion of sidebar quality, which is simply the expected relevance value of results within the sidebar.

Figure 4.6 shows the results of simulating runs while manipulating the probability of switching from the main results to the sidebar, which we take as being proportional to the *perceived quality* of the sidebar. The outcome being measured here is the area under the cumulative relevance curve (*rel-AUC*). Analysis on these outcomes shows that as the quality of the sidebar results increases, the average cumulative relevance also increases, (from $rel - AUC = 180$ when $q = 0$ to $rel - AUC = 335$ when $q = 1$ and $rel - AUC = 484$ when $q = 2$). These changes are statistically significant as reported by an ANOVA ($p < 0.05$). The interaction between switching probability and quality is also statistically significant ($p < 0.05$). A post-hoc Tukey's pairwise comparison revealed a significant difference between qualities $q = 0$ and $q = 2$ ($p < 0.01$).

Figure 4.6 also shows some interesting trends regarding switching probability. When the sidebar quality $q = 0$ (and thus all results are non-relevant), the average cumulative relevance measured drops when the switching probability increases from $p_{switching} = 0$ to 0.1. Conversely, sidebar qualities of $q = 1$ (medium quality, or all results are relevant) and $q = 2$ (high quality, or all results are highly relevant) exhibit a sharp increase in cumulative relevance when the switching probability $p_{switching}$ increases from 0 to 0.1. Two-sample t-tests indicated that these differences are statistically significant ($p < 0.01$). Intuitively, this is understandable, as with limited time to complete a search, attention is taken away from

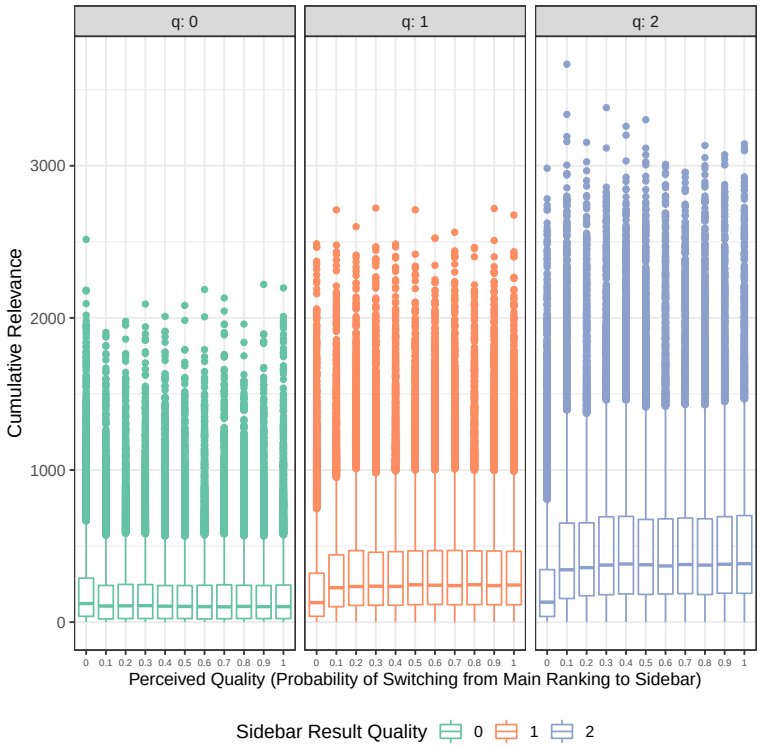


Figure 4.6: Relevance changes with real sidebar quality (low=0, medium=1, low=2) mediated by perceived quality (switching probability) with 5 sidebar results, averaged over 10,000 simulations. As sidebar quality increases, the median cumulative relevance increases as well as the variability in relevance measured by standard deviation. Baseline quality when sidebar is unused is where switching probability is zero.

a user's task towards an element that does not help. On the other hand, it reaffirms our base assumption that this element – the sidebar – may prove beneficial as long as relevant documents are present and users do in fact use it. In general, this finding suggests that it is worth exploring ways to draw attention to the benefit of an interface element such as our sidebar when it is able to provide said benefit, and perhaps conversely, to draw attention away from or de-emphasise the element when the benefit is absent.

4.4.3 Option Value of the Sidebar

We take this a step further by calculating win-loss distributions for these cases by considering $p_{switching} = 0$, where the user never switches from the main results to the sidebar to be alternative scenarios for each simulation. We subtract the simulated outcomes for each combination of p and q, which gives an idea of how beneficial switching would be to not switching. We show the win-loss distribution for $p_{switching} = 0.6$ in Figure 4.7.

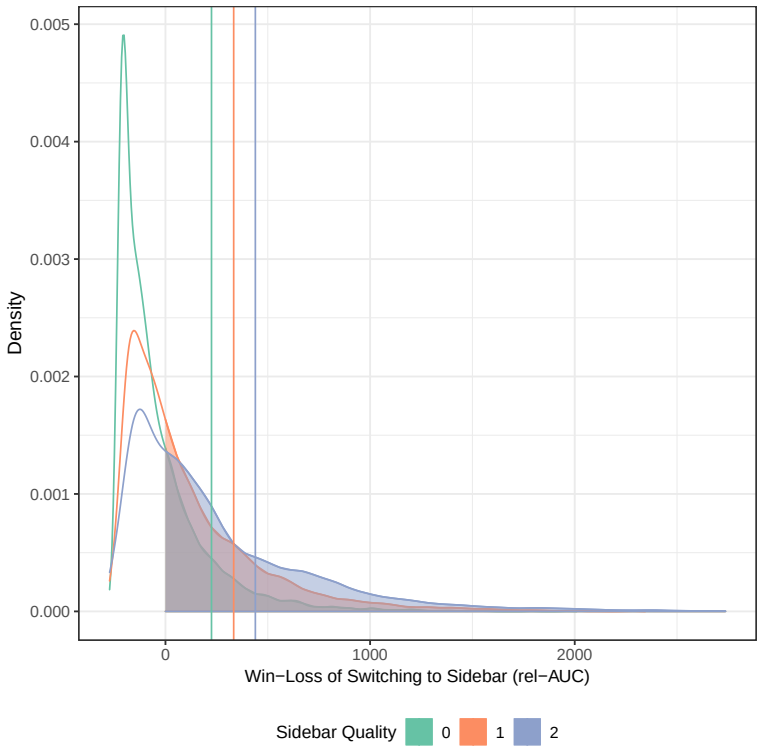


Figure 4.7: *rel-AUC* win-loss distibutions with real sidebar quality mediated by perceived quality (switching probability), where switching probability $p_{switching} = 0.6$. Density plots are for different measures of sidebar quality, from low (0) to high (2), and are shaded where the win-loss > 0. As the sidebar quality improves, we see fewer losses at the modes. The expected values given a win are shown as vertical lines corresponding to the sidebar quality; as quality increases so does the expected value given a win (225 when $q = 0$, 333 when $q = 1$, and 439 when $q = 2$.).

We can boil these distributions down to a number by using the Datar-Mathews real options approach to calculate a weighted average, $p(win) \times E[value|win]$. This number summarizes the likely future value of having a novel user interface element (the sidebar) available as an option to the user. Showing this in Figure 4.9 verifies the effectiveness of a high-quality sidebar. However, it also demonstrates the harm of a low-quality sidebar, where any degree of switching starts to show losses compared to not switching. We see that the most effective combination has $p_{switching} = 1$ and $q = 2$. When the sidebar quality is lower however, the option values are highest at switching probabilities $p_{switching} < 1$ – when $q = 0$, the option value is highest at $p_{switching} = 0.5$, and when $q = 1$, the option value is highest at $p_{switching} = 0.8$. This implies that there is value in having the option to switch between the main results and the sidebar to make up for any lack of quality in the sidebar.

We also consider the effect that the variance of the results in the sidebar might have on the option value, that is, this indicator of likely future value. A system designer may make the decision to prioritise an algorithm that produces relatively consistent results over an algorithm that might take risks in finding high-reward results. In Figure 4.8, we plot the option values from simulations with either a sidebar with results having high variance where results are specified by a random permutation of the quality vector $\langle 0, 0, 1, 2, 2\rangle$, or a sidebar with results having low variance and all results having a quality $q = 1$. From this, we can see that the higher variance sidebar always has a higher option value in comparison to the lower variance sidebar. In cases where we can give the option of using a ranking that has a higher variance, this could be beneficial to users.

We also look at the conditional value at risk, which is the expected shortfall, or the average loss for the worst $x\%$ of cases. When we investigated the of this calculation at $x = 50\%$, we saw that this corroborates the relative effects of real sidebar quality on the expected outcomes from use – the highest quality sidebar has the lowest average loss in the worst 50% of cases. However, we also see that this intersects with the perceived quality as it did when looking at the option value, in that the expected shortfall is the least with a high perceived quality or switching probability (80%–90%), but switching from the main ranking to the sidebar should not be exclusive here as well.

4.4.4 Sensitivity Analysis

To address ***RQ4***, characterizing user behavior associated with high vs. low performance, we performed a sensitivity analysis to increase our understanding of the relationships between a user's interaction behavior, as represented by the state transition probability parameters, and the *rel-AUC* outcome variable (i.e., the area under a time-relevance curve).

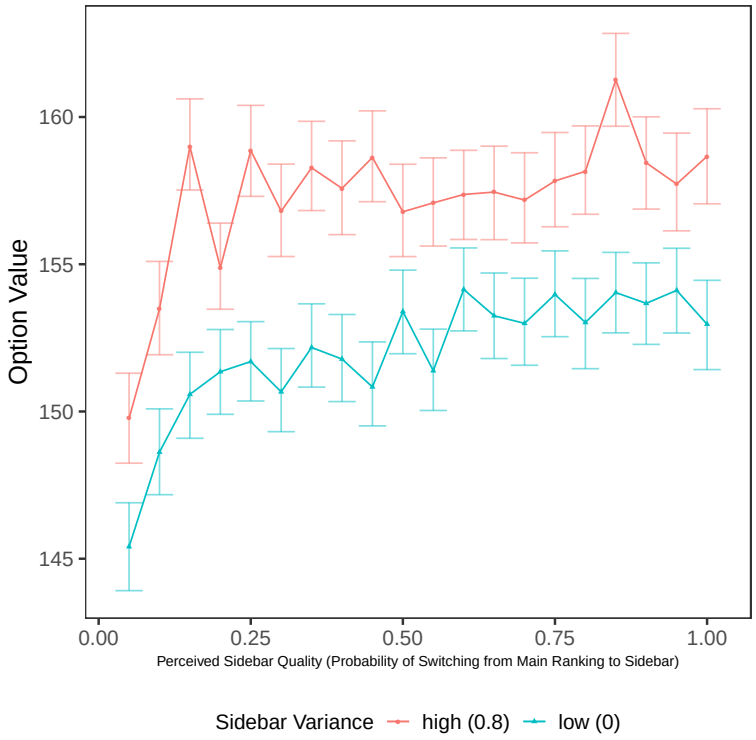


Figure 4.8: The option value of the sidebar to the user as a function of (1) the variance of the quality of the results in the sidebar (with fixed mean $q = 1$), and (2) perceived sidebar quality (switching probability from main results) resulting from 100,000 simulations at each combination of variables. Variance values were 0.8 (high) and 0 (low). Bars show the 95% confidence intervals. A high variance in sidebar quality leads to an increased option value over a low sidebar variance, which suggests a likely future value of a sidebar that exercises some degree of risk.

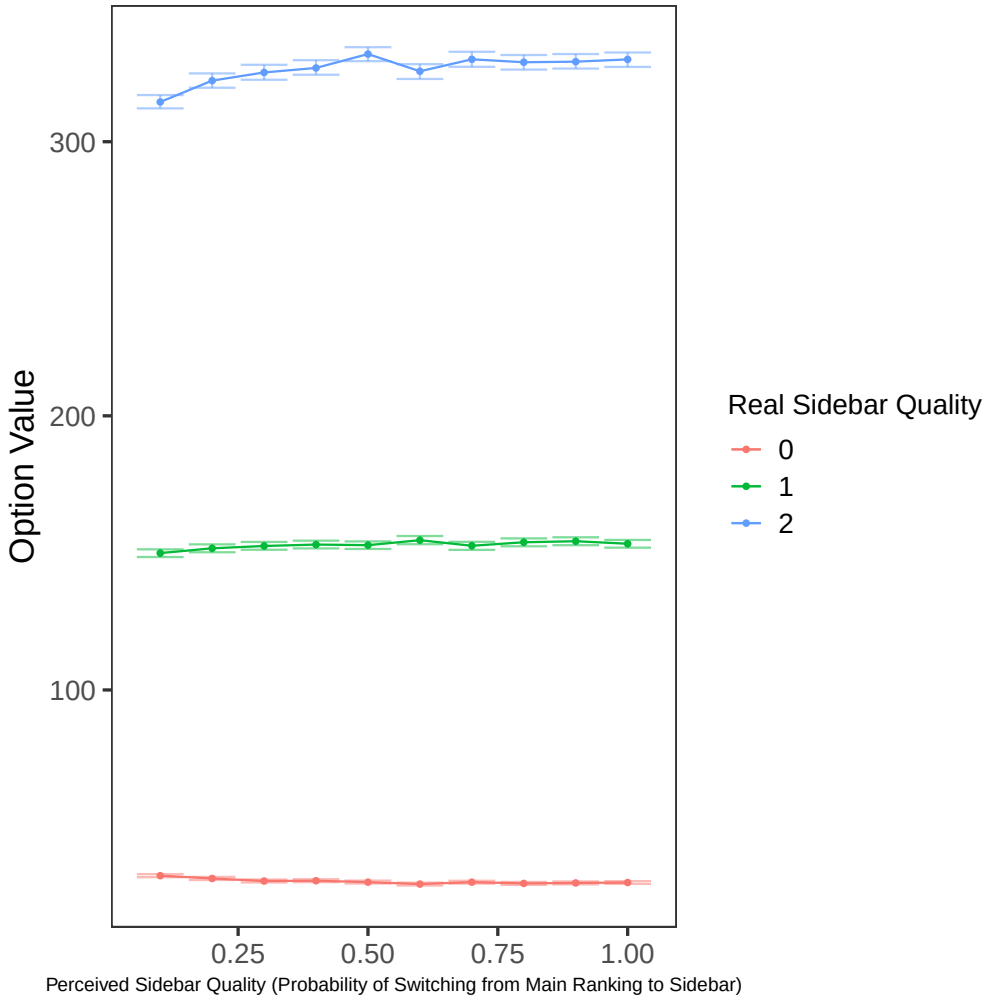


Figure 4.9: The option value of the sidebar with variance zero as a function of (1) the actual sidebar quality and (2) perceived sidebar quality (switching probability from main results) resulting from 10,000 simulations with each combination of variables. The sidebar is most effective when its results quality is high ($q = 2$) and the probability of switching to the sidebar $p_{switching} = 1$. For all values of sidebar quality, we see that the highest option value is reached when the sidebar is not preferred exclusively (option value is highest for other values of quality $p(main \rightarrow sidebar) = 0.2$ when $q = 0$ and $p(main \rightarrow sidebar) = 0.9$ when $q = 1$), which in these cases suggests value in switching to the main results in these cases.

4.4.4.1 Linear Model

Our first step for preforming our sensitivity analysis involved building a linear regression model with *rel-AUC* as the dependent variable, and model state transition probabilities as the independent variables, with parameters fitted with simulated data. This will enable us to analyse how changes in the independent variables (the predictors) were associated with changes in the *rel-AUC* outcome variable. Of the independent variables, *start→view_main*, *view_main→view_sidebar*, *view_main→view_main*, *view_main→click_main*, and *view_main→query* were positively associated of success ($p < 0.05$) and *click_main→view_main*, *click_main→view_sidebar*, and *click_main→save_main* were negatively associated with success ($p < 0.05$). This suggests that, for the system simulated, examining both rankings is a positive predictor of success (that is, time spent viewing the main or sidebar results). Clicking the main results after viewing them and issuing new queries also predict success. However, going back to re-examine either of the rankings after clicking a result is a negative predictor of success. One way to interpret this is as a manifestation of the position bias: clicking on a result means that it is likely that the next result examined will be less relevant than the result we already clicked. In this case, it makes sense to spend time examining results, clicking when relevant, and re-querying after as necessary.

With this linear model, we then performed our sensitivity analysis. A visualisation of this can be seen in Figure 4.10. This sensitivity analysis shows that a change to *view_main→view_sidebar* or *view_main→click_main* leads to the largest normalised changes in *rel-AUC*, where an increase would be expected to increase the *rel-AUC*. In the opposite direction, *click_main→save_main* and *click_main→view_main* would be the most negatively sensitive parameters, where an increase in the likelihood of moving from the sidebar to the main results, or the likelihood of continuing to use the main results after clicking, would be expected to decrease the *rel-AUC*.

As it relates to ***RQ4***, we would expect that a stronger focus on the main results is more beneficial when there are many more options in its ranking than in the sidebar. We will take a look at simulations of high and low performance in more detail in response to ***RQ5***.

4.4.5 Dimensionality Reduction of Simulation Space

To address ***RQ5*** – how to get from low to high user performance – we explore the simulation space in order to identify what changes might be necessary to the system and/or users' behaviours. As a reminder, the original high-dimensional simulation space in terms of the set of NN parameters representing the state transition probabilities shown in Figure 4.2 with the addition of the Sidebar–Main Results Correlation parameters in Section 4.3.2.

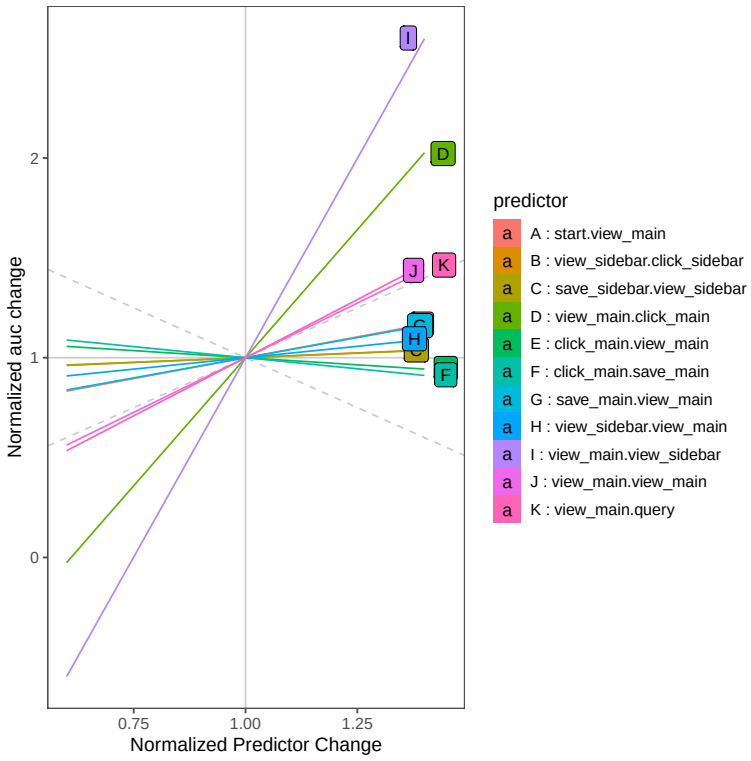


Figure 4.10: Sensitivity analysis showing how change in *rel-AUC* outcome is sensitive to changes in each of 11 independent variables representing user decision probabilities in the interaction model. Outcomes are averaged over 100 simulations. Positive change in *rel-AUC* was most closely tied to more exploration of main results and increased use of the sidebar (increased *view_main→view_sidebar* and *view_main→click_main*: I and D, top right corner). Decreases in *rel-AUC* were most closely tied to saving main page results and continued use of the main results (increased *click_main→save_main* and *click_main→view_main*: F and E, mid-right).

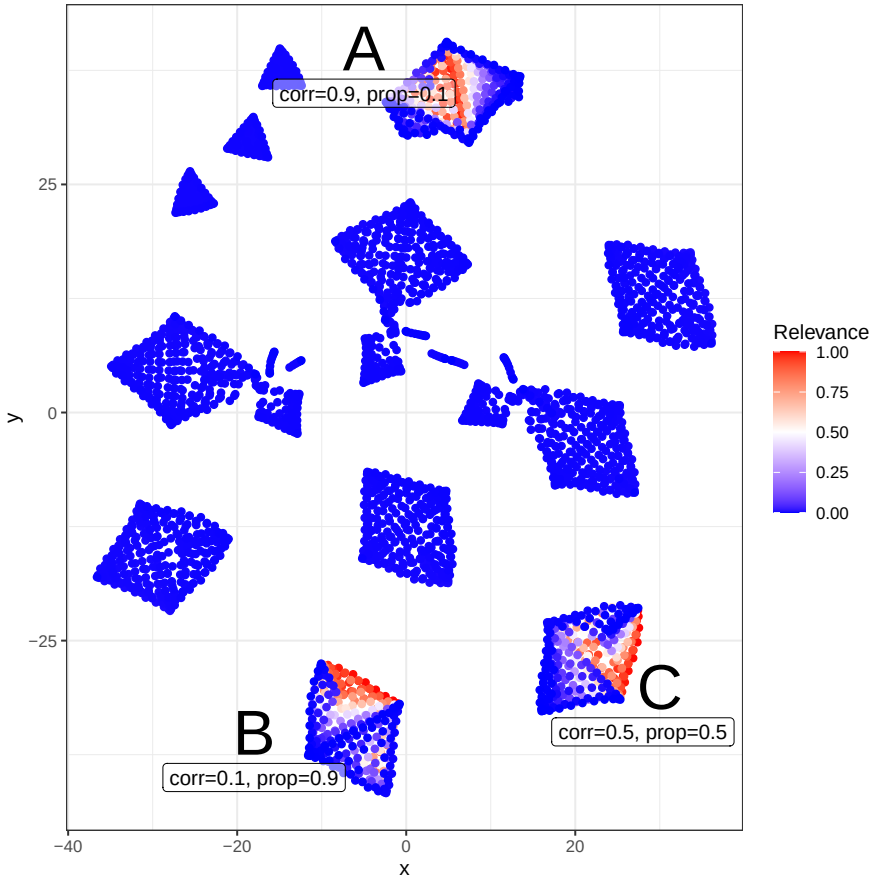


Figure 4.11: Dimensionality reduction of simulation space with t-SNE. Blue represents low *rel-AUC* and red represents high *rel-AUC*. Some structure is evident, such as the groups of parameter instantiations and areas of high/low effectiveness indicated by colour in the areas labelled "A", "B", and "C" and their correlation parameters. The *corr* parameter is the correlation between the sidebar and main results, and the *prop* parameter is the proportion of sidebar results with the given correlation.

Using the dimensionality reduction technique t-SNE [127], we project this NN-dimensional simulation space of state transition parameters into a 2-D representation. We show a visualisation of this reduction in Figure 4.11. In this figure, each point represents the average of simulation outcomes for a particular combination of simulation parameters; this means that with t-SNE applied, points close together have similar values for their parameters. We colour each point by its *rel-AUC* value using a non-linear scale described by [94, 121] to highlight the differences between high and low performance.

In the t-SNE plot, we can discern some structural characteristics of the space. There are areas in which high performance is seen, such as the area labelled "A" – this area of deep red (high performance) is surrounded by a points of low performance (the points in blue). We

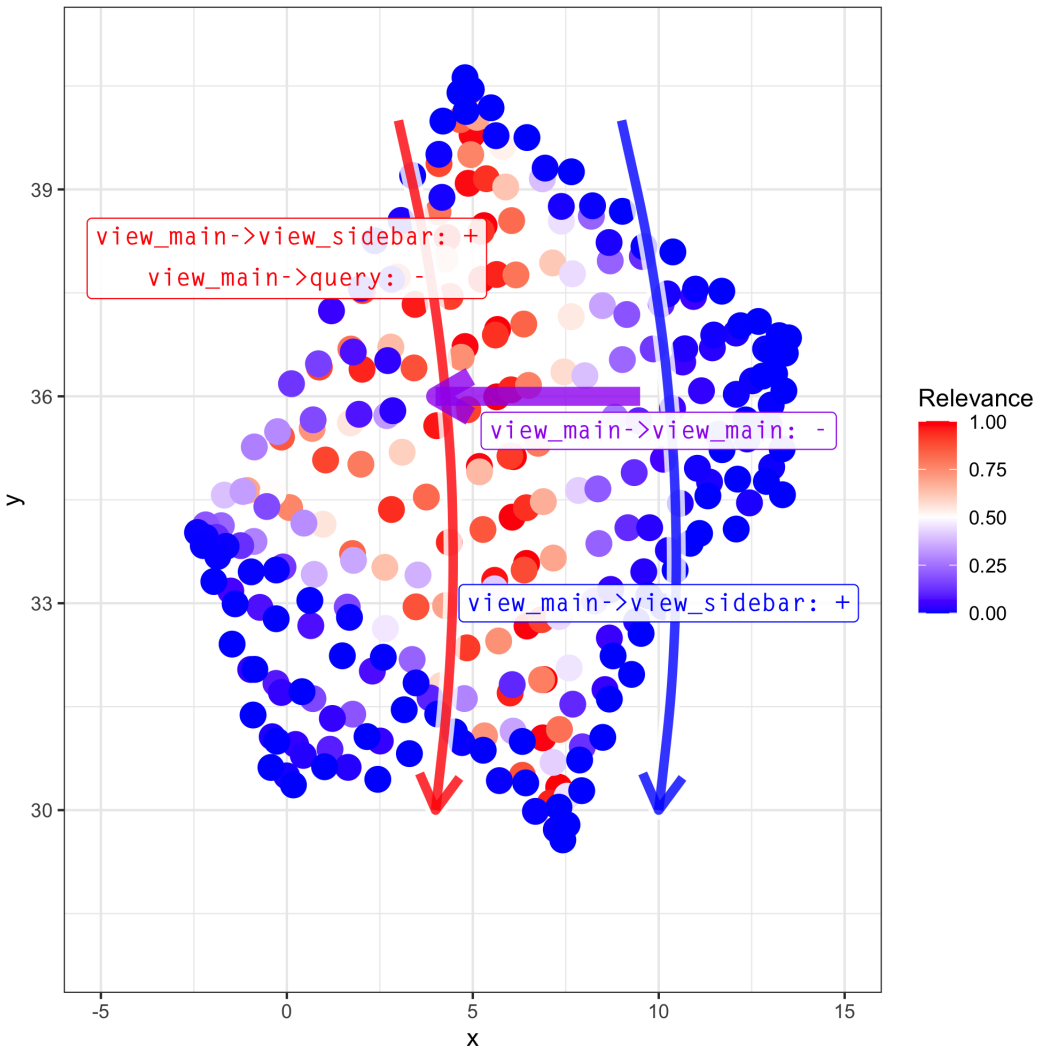


Figure 4.12: Dimensionality reduction of simulation space with t-SNE zoomed in to region A from Figure 4.11. See text for a description.

show this in greater detail in Figure 4.12, and will describe this particular figure in greater detail below. However, we also make note of similar regions labelled "B" and "C", which correspond to the three pairs of correlation parameters we simulated: $(0.9, 0.1)$, $(0.5, 0.5)$, and $(0.1, 0.9)$. This indicates that there are small differences between the combinations of correlation parameters such that, we can always find an area of high performance regardless of correlation and the correlation between the main results and sidebar plays a relatively minor role in leading users to high performance. Other areas have structure as dictated by the parameter space, but these are areas of low performance. Overall, this analysis suggests that, using this method, we can not only identify high-performing parameters, but we may also be able to guide users in low-performing areas to areas of higher performance.

State Transition	Average Distance
view_main→view_sidebar	0.1069
view_main→query	0.1049
view_main→click_main	0.1040
view_main→view_main	0.0829
save_sidebar→save_sidebar	0.0004
save_sidebar→view_sidebar	0.0002
save_sidebar→view_main	0.0001

Table 4.4: Most salient state transitions in region A from Figure 4.11. Average distance between points gives a rough estimate of how much the transition probabilities change within the region.

State Transition	Average Distance
click_main→view_main	0.1063
click_main→click_main	0.1050
click_main→view_sidebar	0.1050
click_main→save_main	0.1024

Table 4.5: Most salient state transitions in region B from Figure 4.11. Average distance between points gives a rough estimate of how much the transition probabilities change within the region.

Looking closer at region A in Figure 4.12, we see blue boundaries along the the edges with *rel-AUC* close to zero, but central points with high outcomes. When we investigate the transition probabilities in these points, we see that the region along the right boundary (marked by the arrow "A1") has the transition probability for *view_main→view_sidebar* increasing as we move down the region as indicated by the arrow. Slightly to the left, the region indicated with the arrow marked "A2" is one of high performance, and as we move down that region, the transition probability for *view_main→view_sidebar* increases while the transition probability for *view_main→query decreases*. Moving to this region from "A1" to "A2" entails a decrease in the probability of transitioning from *view_main→view_main* towards zero. In Table 4.4, we show the salient state transitions from region A, calculated by the average distance between all points in the region. We have excluded the transitions with average distances very close to zero. For comparison we also show the salient state transitions from region B in Table 4.5. The lack of overlap between these lists shows the reason for the discontinuity between regions in the low-dimensional space and illustrates the difference that the changes in certain transitions can make to outcomes.

Using the t-SNE reduction as a basis, we can find productive paths that lead from areas

of low performance to higher performance. The process by which this was done is as shown in Algorithm 4.1.

Algorithm 4.1: Find Productive Paths

```
Data: A list of points P from t-SNE dimensionality reduction
Result: A graph G consisting of paths from low to high performance
M ← DistanceMatrix(P);
foreach point p in P do
    N ← four nearest neighbours of p;
    foreach neighbour n in N do
        if distance between n and p < MeanDistance(p, N) then
            add adjacency (n, p) to graph;
        end
    end
end
```

This gives us an implicit graph of adjacencies, which we can traverse to determine directions of higher or lower performance. We perform a depth-first search along increasing edges to find the final network of productive paths.

Following from this, we perform this process for different three values of the correlation between the sidebar and main ranking (0.1, 0.5, and 0.9). This correlation captures the degree of overlap between the results in the main ranking and the sidebar as described in Section 4.3.2. This allowed us to identify the most "productive" paths as determined by the cumulative improvement along the path by *rel-AUC*. We show summaries of the top path for each correlation value in Figure 4.13, with the initial models followed by the cumulative changes in the models along the path and the final scores.

Figure 4.13 highlights the changes that it takes to get from a lower-performing model to a higher-performing one, with this example chosen as one of the models with the largest difference in its outcome. Here, the correlation between the sidebar and the main results and the sidebar is 0.9, and our analysis of other models has shown that with this high degree of correlation, there is a focus on switching to the main results and staying there for higher performance. This can be explained by the fact that many more results are present in the main ranking than the sidebar, and if they mostly overlap, there is little benefit to using the sidebar.

This analysis gives us a direction to pursue with regards to RQ4: ways to guide users towards better performance. Depending on the level of similarity between the two rankings, we can potentially employ different strategies – encouraging use of the sidebar when similarity is between the rankings is low and the quality of the sidebar is high. Furthermore, that we

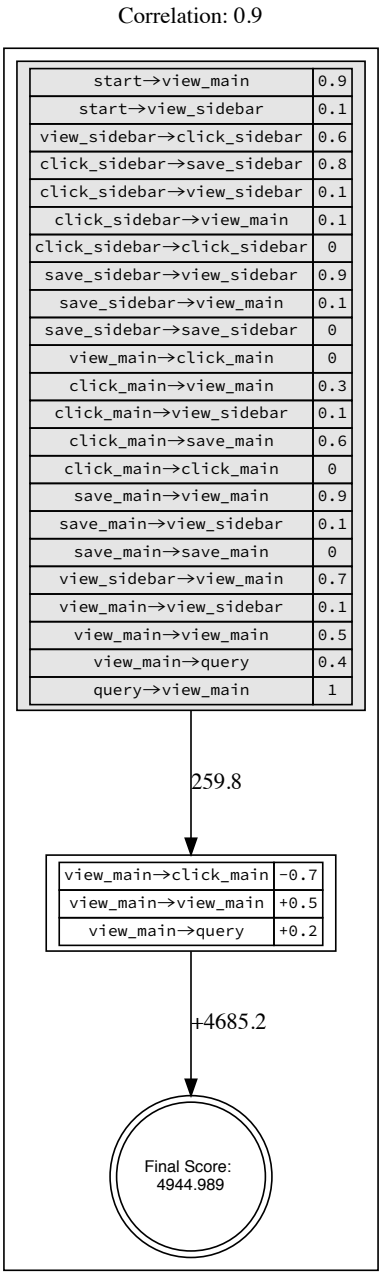


Figure 4.13: One of the most productive paths to highest cumulative gain, where each node represents the behaviour of a simulated searcher. The top node in grey is the starting model, and intermediate nodes in white are the changes to the model. Edges are labelled with the change in score resulting from the model change.

have devised a method for identifying these paths of increased performance means that we can further explore other system- and user-level simulation parameters and discover other means of guidance towards higher effectiveness.

4.5 Discussion

The changes between the starting model and subsequent increase resulting from the modifications suggest ways we might nudge users to change their mental models of the system to achieve better effectiveness. Most of these changes (the white, intermediate nodes in Figure 4.13) reflect behaviour changes in users' interactions with the sidebar and main results – whether these elements are *viewed*, *clicked*, or results within them *saved*.

It is difficult to estimate the *magnitude* of these potential interventions, but we might posit that varying magnitudes might be possible along a spectrum of *nudges* and *shoves*. Whereas a nudge preserves choice, a shove *removes* it [183]. These nudges could either involve explicit affordances, or else – along the lines of Information Foraging Theory [154] – more implicit cues in the interface or results content itself. *Highlighting elements* is an intervention that might serve as a nudge. To encourage a user to save a result after clicking it, we may highlight the save button for the result they just examined. Similarly, if we want to encourage a user to look at the results in the sidebar, we may highlight the entire sidebar. To encourage users to click on a particular result, we may highlight just that result.

Providing additional information may also be a way to introduce a nudge. The position in a search ranking gives relative information about a result being more relevant than another, but *how much more relevant* is often not conveyed. In a sense, the results in our sidebar are better at solving the tasks given by including more information in each result, but this might not be evident unless a user examines the documents explicitly and understands the abstract differences in content. A nudge to encourage usage of the sidebar may be to convey the benefit of the sidebar before the user starts using the interface, or to convey this information while they use the system, for example, by summarising the content (e.g., "this result addresses four of the five aspects of your task").

Disabling elements by contrast is an intervention that might serve as a shove. Because it removes the choice of using the element, we may be potentially frustrating users depending on their preferences on how to solve their task, or even inadvertently reducing their effectiveness if the intervention is actually inappropriate. However, if we know that there is only a cost to using an element, this shove could be beneficial. In our experiment, we essentially incorporate a shove by disabling the sidebar until the simulated sidebar "finds" a result worthy of showing in the sidebar. This discourages users from spending too much time trying to use the sidebar

before it would help.

With all this in mind, we believe we have demonstrated the value of session simulation as an exploratory tool. It pointed towards various avenues for future experimentation and analysis both as a result of the simulation outcomes and during the process of building and running the simulation. The simulation also enabled us to applying stochastic option pricing methods to compute the likely future value of having a novel user interface element as an option for the user. Investigating productive paths towards high performance was a goal of our analysis from the outset, but the ability to plot and explore a space of simulation parameters introduced a good deal of clarity and guided our methods in a way that would have been more difficult solely using log data.

We must also make note of the limitations of our analysis. Our simulation model was built largely in the interest of simplicity while being able to provide plausible behaviour, and as such is not as sophisticated as some other contemporary searcher interaction models. Different stopping behaviours [139], a model of relevance saturation, or browsing strategies [179] were not considered. Future simulation work may take these into consideration, as well as simulating other user phenomena such as fatigue, task difficulty, or reading level.

Finally, we note that the use of options pricing methods for estimating the likely future value of user interface components is quite a general idea: it could be applied, beyond search and recommender systems, to any interaction scenario where a user utility measure can be defined for an interface feature or affordance, along with a stochastic model of utility over time. Such a valuation measure could be used to evaluate the potential feature or affordance, as we do here with the sidebar, or it could be used as the objective to be optimized. We intend to explore theoretical and algorithmic aspects of this valuation framework more extensively in future studies.

4.6 Conclusion

We have presented the use of user session simulation as a tool for exploration in the design of information retrieval systems and for identifying interesting avenues for experimentation. We collected user interaction data from a small-scale pilot experiment, and showed that calibrating our simulation to this data was able to produce similar outcomes to those of real users. Using statistical analysis of a regression model, we were able to identify the state transitions of our model that characterise high and low performance, and applied principles of real options pricing to show that increasing the perceived quality of search results in our sidebar is benefitial to users when the real quality the search results is appropriately high. Furthermore, the simulation made it possible to identify ways to guide users from low

to high performance. Future work will investigate these avenues more directly, with more sophisticated interaction models and confirmatory experiments on real users.

4.7 Author Contributions

This work was prepared by Ryan Burton and Kevyn Collins-Thompson. Burton originated the state transition simulation design for modeling user behavior. Collins-Thompson provided the initial idea of using options pricing frameworks from finance to value user information sources, along with a mathematical valuation framework that would be suitable for this chapter's scenarios. Ryan Burton was the main contributor for the remainder of the work, having designed and conducted the experiments, as well as analyzing results and writing the manuscript. Kevyn Collins-Thompson also contributed revisions to the manuscript.

CHAPTER 5

Conversational LLM Assistance During Technical Reading

In previous chapters, we have shown the effect of users learning to use a new, novel search system and the value of simulation as a tool for exploration in the information retrieval system design space. In this chapter, we take these concepts a step further to a new domain – tools for vocabulary learning during technical reading and search. This work has been submitted to CHIIR 2024 and is currently under review.

For a student reading a document on a specialized topic, it may be difficult to penetrate the jargon contained within, even if the topic is in the student's area of study. As system designers, we can take a role in offering assistance to students engaging with resources to make learning and understanding a less onerous process. In this study involving data science student participants, we employ a chatbot assistant through two consecutive learning phases: one centered on document reading and the other involving a simulated search engine. The chatbot utilizes a contextual large language model (LLM) using a Retrieval Augmented Generation approach to provide responses to user questions about the documents and their associated keywords. We analyze log data, questionnaires, and interviews to identify usage patterns of the chatbot and to understand learning and interaction behaviors. Additionally, we assess the user's opinions on when a conversational assistant would be appropriate or helpful in a learning task. We find that trust is a recurring factor in users' opinions, shaping their perceptions of the assistant's usefulness. Tests of learning gains point to improvements over prior knowledge and vocabulary coverage before the study, but more extensive investigation is needed to provide conclusive results.

5.1 Introduction

The heavy use of technical terminology in learning resources can put unnecessary strain on students learning a new subject. A 1995 study of secondary school science textbooks found

that the number of new vocabulary terms was at least as much as that in a foreign language course [78]. The implications for learners are numerous – from the potential to memorise facts without developing a deep understanding of the material, to the perception of such subjects as "finished bodies of knowledge" with just facts to be absorbed [78, 83].

The primary goal for this study, therefore, is to investigate whether learning materials could be made more accessible by providing two key affordances: first, recommending important keywords to learn, using an approach that is personalised according to a student's level of prior familiarity with the keywords; and second, a conversational agent that is able to access and analyze the texts of the provided class resources, so that students can ask questions of the agent about these resources directly, including about how key technical concepts are used. To that end, we conduct a study to evaluate the effectiveness of an automated conversational assistant that a student can use not only for increasing their factual knowledge of key technical concepts, but for asking higher-level questions about how these concepts are used in the technical reading. Our analysis includes a focus on how a learner's knowledge evolves during the session and how individual aspects such as their previous domain knowledge and familiarity with AI chatbots are connected with features of their interaction behavior. In particular, we address the following research questions:

RQ1: What patterns of use can we characterise for learning from user interaction with an AI assistant during reading?

RQ2: Does a user's previous knowledge on a search topic affect their interaction behaviour during use and answers in the post-study test and interview?

RQ3: How does the use of an AI chatbot shape a user's experience including trust in the system?

In this work, we describe the implementation of a search interface to aid in vocabulary learning that uses a large language model for conversational assistance. This assistance is provided via a chatbot affordance, and this conversational agent is aware of the content of the page a user is reading during their search task and is able to answer questions about it. The chat is complemented by the use of keyword extraction from articles, where the keywords are adaptively presented on the search results page as interactive elements that a user may click to ask the chatbot for a definition. Therefore, the keyword recommendation on the search page is complemented by additional assistance on the articles for which we recommend these keywords. We present preliminary results of interaction patterns using trace data, as well as a qualitative analysis of in-depth interviews after users have completed the study.

We will now review related work for the aforementioned elements of the study, and will follow with a description of our study design in Section 5.3, review our results in Section 5.4,

and discuss the implications of our work in Section 5.5.

5.2 Related Work

This study combines elements of several research themes, including learning during information seeking, conversational assistants, and concept extraction.

5.2.1 Searching as Learning

With the heavy focus on relevance in search, learning has not always been a particular topic of explicit interest or study. Marchionini's (2006) depiction [132] of the three types of search activities ("Lookup", "Learn", and "Investigate") in his description of exploratory search, which encapulates the latter two, put the behaviours and needs that come with "learning searches" in stark relief to the other types of activities. These learning searches are iterative and require interpretation on the part of the user – intepretation that takes time and effort and calls for qualitative judgements.

Eickhoff et al. took steps to identify search sessions that seemed to boost users' learning within log data by building models of the users and their specific contexts [63]. Using metrics to characterise domain expertise from search behaviour such as domain and topic diversity, branchiness (tendency of returning to a previously visited page) and display time (experts spend less time reading a page within their domain of expertise), the authors were able to see how these metrics change within a session and whether learning effects "persisted" across session boundaries.

Although learning as a part of search has long been considered, it has only been more recently that there has been work to investigate the indicators of effective learning that comes as a part of the search process. Collins-Thompson et al. [50] looked at methods to assess learning at different stages of a simulated work task involving a search engine that proves intrinsically-diverse results, and found that both explicit and implicit measures such as perceived learning outcomes, interaction speed, and length of written responses to the given task served as indicators.

To investigate learning gains over time, Roy et al. [166] gave users a search task during which they were prompted every 20 minutes about their knowledge about the topic. Users who had some familiarity with a topic had the highest gains in learning, while users with no prior familiarity saw a sublinear increase in learning gains. In order to measure this, the authors gave participants four-level self-assessments of vocabulary knowledge.

Many learning assessments tend to involve administering pre- and post-tests, considering

the difference in score to indicate the gain in knowledge. Yu et al. [221] aimed to predict this difference with a supervised model using interaction features such as the maximum time spent per page and the average time per page. However, this process requires calibration for each topic, which may interfere with the prexisting knowledge levels we expect or want study participants to have. Otto et al. [149] took the idea further by adding multimedia, text, and structural features for the learning resources that participants used, and found a mixture of linguistic and multimedia features that proved salient in their random forest classifier, including the presence of classes of words (such as religious or certainty words), as well as document complexity, the presence of infographics, and the presence of a heading or menu bar.

In our present work, we take a search-inspired approach to our study design, presenting a set of documents relevant to the main topic of the task. We measure users' learning gains as they progress through the task at set stages, measuring vocabulary familiarity and topic knowledge. For simplicity we prevented users from issuing their own questions, but we see the evidence of iterative nature of learning in users' interaction with the chatbot.

5.2.2 Conversational Assistants

Conversational assistants have long been a "holy grail" of computer science research. With brief spikes of interest over the decades since the 1960s beginning most significantly with Weizenbaum's ELIZA [207], progress has been stilted due to the obvious limitations on the required syntax and semantics on the part of the user; however, the advent of large language models presents a new avenue for exploration with fewer obvious limitations than prior systems based on explicit pattern matching and rule application.

Large language models, which are currently enjoying the status of "foundation models" on which other language tools and systems may be built [26], may be distinguished from previous generations of language models not only by their size—they are typically trained on a large multimodal corpus of web documents consisting of text and code and result in models with billions of parameters—but also their training techniques which rely on dynamic pretraining techniques. These models are amenable to further fine-tuning to give better performance in specific domains [229]. ChatGPT [1] and LaMDA [192] are specific language models that have been designed for conversational use.

A conversational assistant, however, should be expected to be capable of more than carrying a conversation – it should also succeed in assisting the user with their task. A human assistant is expected to have domain competence, to be able to learn from their

[1] https://openai.com/blog/chatgpt/

client and adjust accordingly, know their own limitations, and handle inexact instructions, and we should aim for computational assistants to exhibit the same properties [89]. One system that focuses on assistance is Iris [66], which has the ability to combine commands in order to perform complex tasks beyond the single standalone commands that might have been included by the designer. To handle inexact instructions, it asks clarifying questions and understands dependent questions that relies on the answer to a subsequent request.

In the context of the present study, we consider the explicit capabilities given above aspirational, though the combination of a large language model with contextual information may simulate the capacity of an intelligent conversational assistant to some degree. As we describe in Section 5.3.6, we provide the large language model GPT-4 with contextual information and design the prompt such that the chatbot behaves as a helpful enough assistant for our limited degree of expected interactions. Although prompting a large language model can roughly serve a similar purpose to search, querying a large language model in the way we propose is more analogous to question-answering than retrieval, and as such, we use a chat affordance and keep the two modalities distinct.

Our work combines conversational assistance with information seeking, and therefore evokes some relationship to conversational search. Research into conversational search is relatively more recent, from Belkin et al. [21] applying case-based reasoning to create system-level scripts to Radlinski and Craswell presenting a theoretical model of desirable properties for a conversational search system to have [159]. In this case, the system should return a result set, but also request clarification if needed and accept feedback on the returned results. It is possible to view our system as conversational search but it does not quite fit because the response of our language model-based chatbot is synthesized rather than retrieved. There are relevant aspects of conversational search research that may be applicable, as work has been done to investigate its application to query reformulation [116], query understanding [163], and how to exploit relevance feedback [151]. However, this is outside the scope of our present study.

5.2.3 Vocabulary Learning

Vocabulary learning has been studied using a distinction between intentional and incidental learning. Swanborn and de Glopper [185] highlighted this difference, noting that intentional word learning occurs in situations where a student is instructed to derive or actively try to learn the meanings of unknown words in context while reading. In contrast, incidental learning "is not triggered by the reading task" (p. 262), that is, it occurs in a setting that is familiar to the reader, and when reading is not done for the purpose of learning or for

directing attention to unknown words. Although we will use findings from both intentional and incidental learning studies to inform our work, we are primarily interested in the task of intentional learning for our experiment.

5.2.3.1 Concept and Keyword Extraction

Besides the issue of learning concepts, we must consider on the system side the problem of determining what concepts exist in our corpus of documents. In the simple case, course concepts are provided by the instructor. We would expect these to be relatively high-quality, but possibly at a course level due to the effort required to come up with these concepts. We may also run across the issue of a concept being referred to by other names in a document besides the term provided by the instructor. It is for these reasons that concept extraction becomes worthwhile.

Keyword and key term extraction has been the subject of considerable study in natural language processing. A survey of the methods and challenges are presented by Firoozeh et al. [67] These methods are often concerned with identifying "key" lexemes within documents. For our work, we are interested in connected these key terms to broader concepts in order to disambiguate terms and recommend the user to learn a single concept covering multiple instances of the same idea. Considerable work has been carried out in applying information extraction techniques to this problem [77, 203, 54], and in our case we use the Wikifier service [30] for the task.

5.2.3.2 Prerequisite Relations

An issue that may arise for students learning within a particular domain is that it is often effective to learn about prerequisite concepts before learning about a target concept. Without this information, students may have to spend additional effort finding the required background knowledge on their own, and may not know what background knowledge is actually required. A course's structure is typically set up to introduce the information necessary to understand a concept before tackling the topic, but it is not always clear which concepts directly relate to each other.

Prerequisite relations may also be extracted from a document corpus. Particularly with the large-scale efficiency demanded of MOOCs, this has seen a good deal of research within that context [125, 152, 124, 167]. Hu et al. [90] used Wikipedia clickstream data to infer prerequisite relations among Wikipedia articles. They note that clickstream data is relatively sparse for any given relation, or pair of concepts, and as a result they also use the set of related concepts within Wikipedia to increase coverage. This points to the utility of clickstream

data for prerequisite relation mining, but relying primarily on clickstream data may only be appropriate for Wikipedia.

Techniques that utilise only the link structure between documents have also been explored. Liang et al. (2015) [123] proposed a measure of the pairwise relatedness of concepts called *reference distance*, or RefD. By modelling a concept in its semantic frame within vector space, they can use this single measure as a scalable method for predicting prerequisite relations from large hypertext document sets such as Wikipedia.

PreFace [198] is a faceted search system that determines the prerequisites of facets and balances the tradeoff between the relevance and diversity of each facet. At its core, it represents a facet as a language model based on a domain-specific corpus and knowledge base, and ranks them using a risk-minimization framework. This is perhaps the most similar work to our proposal but differs largely in purpose and therefore execution – PreFace is intended for retrieving interesting facets as well as the prerequisites for the aspects of a query (such as "implementation" or "application"), whereas our work is intended to provide an incremental vocabulary learning tool that aims to help students become familiar with the terms within documents through the duration of a course.

We use datasets from [124] and [125] to determine which concepts should be encountered before and after the concepts we extracted using Wikifier, and show the resulting collection as key concepts for each document on a search page as shown in Section 5.3.

5.3 Study Design

The objective of this study was to explore the role of a conversational AI assistant in facilitating user learning during technical reading. To that end, we designed a study protocol comprising multiple stages that involved masters-level data science students learning about a designated topic in data science. The stages included assessments of their knowledge of this topic before, during, and after the task.

Our study presented the task as two stages of a single learning session – a Reading stage and a Search stage – in that order. Users were given a total of 45 minutes to complete the task. During the Reading stage, which we conceived of to control for effects of the first document that a user would select, users were provided with a single *main document* to read and understand, while having access to a contextual chatbot powered by a large language model. When users finished reading the main document, they were able to elect to move to the Search stage by clicking a "Launch Search" button, provided they had time left in the task to do so. During the Search stage, within which a user essentially continues finding other documents as part of their search task, users were presented with a fixed list of five

documents mimicking the design of search engine results, and were given access to the same contextual chatbot that had access to the complete text of whichever document the user chose to read. As the user interacted with the system, keyboard and mouse interactions of the user with the page and the chat assistant were recorded.

After one week following completion of the study, users were interviewed to gather further insights into their experience with the system, and their learning. We now present a detailed description of the study's stages.

5.3.1 Study Workflow: Stages and Screens

The complete workflow of our study protocol is shown in Figure 5.1.

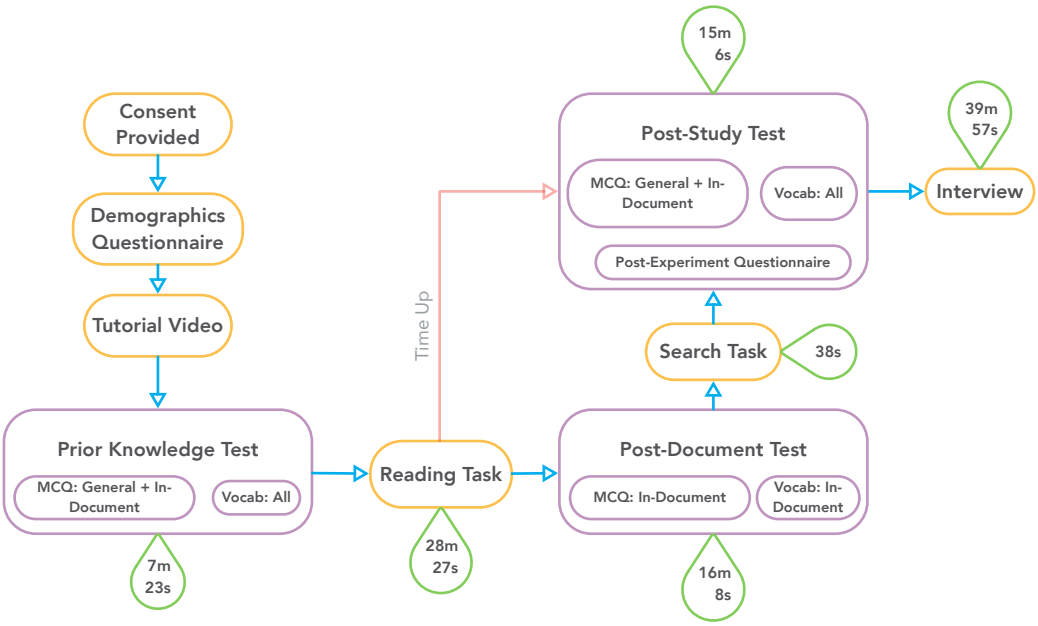


Figure 5.1: An overview of a user's progression through the study. Median times spent at each step are shown.

The study stages were structured in the following manner:

1. *Demographics questionnaire.* This questionnaire primarily gathered information about their experiences with search engines as tools for learning. We specifically asked about their existing familiarity with ChatGPT and ChatPDF which resembles our proposed chatbot in terms of conversational interaction with documents.

2. *Tutorial video.* Participants watched a short one-minute video giving a brief overview of their task and outlining how to use the chatbot. The video included examples

of questions they might be able to ask, such as, "What was the problem statement [outlined in the article]?" and interactions such as clicking on a keyword on the search page to get a definition.

3. *Pre-task questionnaire and prior knowledge test.* On this screen, users were introduced to the learning topic. They were asked open-ended questions about their existing knowledge of the topic and what they believe is missing in their knowledge about it. We assessed their interest and familiarity with the subject using a 5-point Likert scale. Additionally, we presented a set of multiple-choice questions (MCQs), vocabulary tests, and open-ended questions to assess their knowledge of the topic domain, across all cognitive learning levels defined by Bloom's taxonomy [25]. This initial assessment serves two primary objectives. First, it helps us position the user's prior knowledge about the topic. Secondly, it will enable us to later evaluate if their knowledge of the topic improves through interaction with the system and materials, acting as a benchmark for measuring their learning progress. The pre-knowledge assessment encompassed measurements across all levels of Bloom's taxonomy. This specifically involved the use of multiple-choice questions designed to assess the capacity to remember, as well as vocabulary questions aimed at measuring the ability to understand key concepts.

4. *Reading stage.* Participants were given the main document to read on the topic with access to the chatbot assistant.

5. *Post-document test.* Participants completed a secondary combination of knowledge and vocabulary tests to determine how much knowledge was gained from reading a single document with the assistance of the chatbot. Users were tasked with answering 9 multiple-choice in-document questions and 18 vocabulary questions without access to documents or the chatbot. This assessment aimed to quantify the knowledge acquired. We focused exclusively on in-document questions, as we did not expect users to gain knowledge on the general topic-related questions from reading the document

6. *Search stage.* Participants were given a mock search results page pre-populated with a query on the topic and a list of documents. Users were expected to use the remaining time to read the new documents in the list and learn about recommended keywords. They also had access to the chatbot assistant and the main document. The user was free to make use of the keywords in interacting with the chatbot, to ignore them, or to look for them manually in the provided documents.

7. *Post-task Knowledge and Vocabulary test.* Participants were given a third test at the end of the Search stage: the set of questions to reassess their knowledge on all levels.

We presented them with the same set of questions as the pre-knowledge test without access to documents or the chatbot. This step allowed us to measure any knowledge acquired or changed resulting from their participation in the study.

8. *Post-study questionnaire.* Participants completed a questionnaire to provide feedback on their experience of learning using the system.

9. *Interview.* Participants returned one week later for an interview.

In the following sections, we now describe how the specific learning topic and reading materials were selected for this study, how we measured learning and prior knowledge, and how we selected participants.

5.3.2 Topic and Document Selection

In this section, we describe the criteria for selecting the topic about which participants were expected to learn, the choice of the main document that participants were exposed to during the Reading stage, and the list of five documents shown on the results page during the Search stage.

5.3.2.1 The Topic: "What are the Netflix Prize and the SVD Machine Learning Techniques used by its winners?"

We chose the topic of the study with the intention of ensuring it would be accessible and engaging for our target audience of masters-level data science students. We chose the Netflix Prize as a topic because we expected current students to be relatively unfamiliar with it, presenting an opportunity to measure learning gains. We decided to center our study more precisely on the primary machine learning technique the winners were known for employing – namely Singular Value Decomposition (SVD). Users were presented with this topic and a description of their learning goals at the beginning of the study.

5.3.2.2 Main Document in the Reading Stage

For a suitable document to assign to users for reading and comprehension, we aimed for a document with a layout and design resembling that of a blog post, designed for accessibility and clarity. For the document's content, it was to be comprised primarily of text with a limited number of images, not exceeding four. Furthermore, a small number of mathematical formulas was expected to be present, with any formulas serving to clarify the specifics of SVD. Any documents with content that appeared too complex to understand in the limited

Article Title	Content Description
On the "Usefulness" of the Netflix Prize — by Xavier Amatriain — Medium	Presents implications of the prize.
Winning the Netflix Prize: A Summary	Gives specific techniques used by the winner using specific machine learning terminology.
Simple SVD with Bias for Netflix Prize — 叶某人的碎碎念	Gives an explanation of SVD with Python. Fairly technical.
Model Based Collaborative Filtering — SVD — by Cory Maklin — Medium	An explanation of SVD using intermediate-level formulas.
Recommendation System : Matrix Factorization with Funk SVD	More general explanation of matrix factorization using R.

Table 5.1: List of documents included in the *Search* stage.

task time or with content consisting of source code for implementing the techniques were excluded from consideration. Our final choice therefore was an article presenting lecture notes from the New Jersey Institute of Technology class *Introduction to Data Science* titled "The Netflix Prize and Singular Value Decomposition" [2]. This document was shown to users after completing the Pre-Task Test and could be revisited during the Search stage by clicking a link at the top of the results page.

5.3.2.3 Documents in the Search Stage

After reading the main document, we provided users with a mock search engine interface to explore additional content related to the topic. We restricted the set of documents to a list of five to constrain user choice, reducing potential sparsity of interaction, while also giving users the freedom to decide which to pursue and in which order. These documents were other potential candidates for the main document, found while the main authors were looking for articles that covered a high-level overview of the Netflix Prize and the winning team, SVD, recommender systems, or a combination of these topics. We also looked to have content covered that included the implications of the prize, the specific techniques used by the winner, and machine learning jargon. The final list of documents is shown in Table 5.1.

5.3.3 Search Interface Description

During the Search stage, the interface displayed a non-modifiable query "Netflix Prize SVD" set in place. This fixed-query approach served a dual purpose. The first was to provide every participant with access to the same set of documents, ensuring that later analysis

[2] https://pantelis.github.io/cs301/docs/common/lectures/recommenders/netflix/

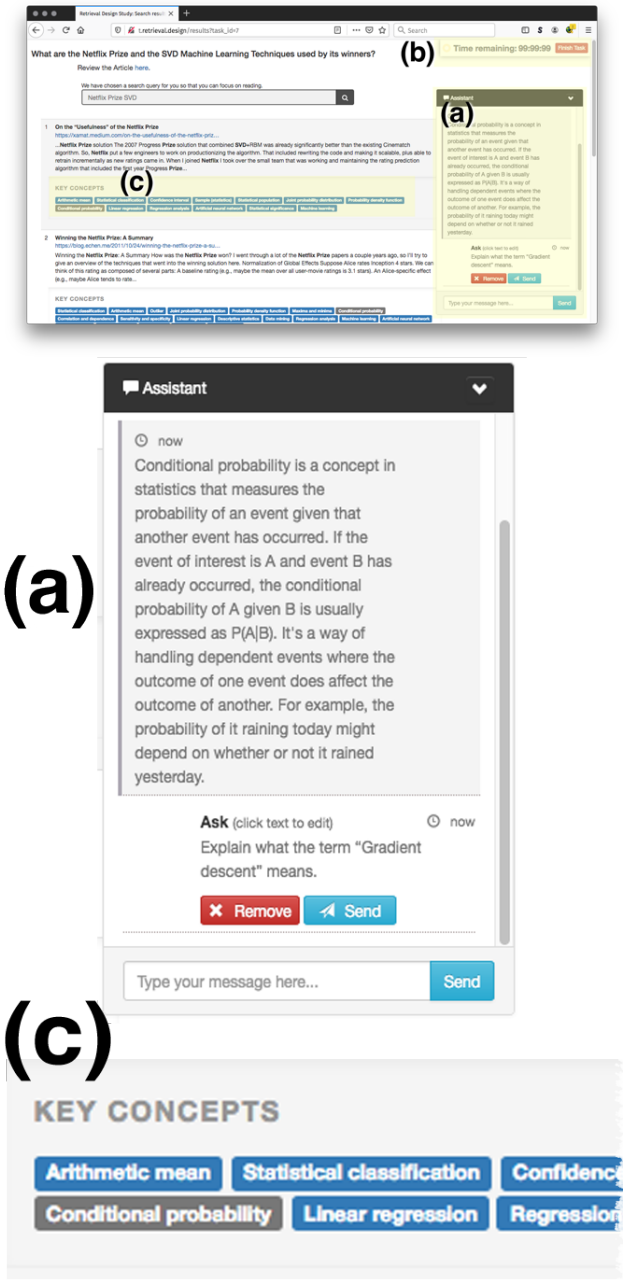


Figure 5.2: Search engine result page with suggested terms to learn within each document, with the chatbot highlighted as (a), the timer with "Finish Task" button highlighted as (b), and the key terms highlighted as (c). Each term is clickable, and populates the chatbot with an appropriate prompt (shown). A user may also ask an arbitrary question about the task.

would be unbiased by the content of the documents, and with a key source of traditional variation removed, namely, individual differences in users' query formulation ability. The second was to reduce the cognitive burden of formulating and reformulating queries, thereby allowing users to focus on interaction with the chatbot. Thus, in summary, we anticipated that freezing the query would enable participants to concentrate on the learning task rather than the searching process. Users still had access to the main document if needed. A mockup of the search engine result page is presented in Figure 5.2.

5.3.4 Knowledge Assessment and Learning Measurement

To explore the connection between chatbot usage and user learning, we needed to estimate the user's domain knowledge before and after exposure. We measured across all levels of Bloom's taxonomy [25], specifically focusing on the *Remember* and *Understand* levels, which represent the two lower levels of cognitive understanding.

The first level, *Remember*, measures users' ability to recall facts and basic concepts. The second level, *Understand*, assesses their capacity to explain ideas. We additionally asked users some open-ended optional bonus questions that encompassed the other levels of the taxonomy, for which answering correctly would net participants an additional $10. For the *Apply* level, we posed theoretical questions in which users demonstrated their capacity to apply SVD techniques to new situations. At the *Analyze* level, the question required drawing connections between two ideas: SVD and the reduction of data dimensions. For the *Evaluate* level, aimed at justifying a stand or decision, we asked them to assess two main strengths and two weaknesses of utilizing SVD for the Netflix Prize competition. Finally, at the highest level, *Create*, the question was meant to stimulate a high-level discussion about the Netflix Prize and the techniques used. No participants answered these bonus questions; therefore, we will not evaluate or discuss them in the remainder of this chapter. In a future study, we may alter the incentive structure to encourage participation. The full list of questions is included in Appendix B.

5.3.4.1 'Remember' Assessment: Multiple-Choice Questions

In line with conventional methods for evaluating users' knowledge online, we designed a questionnaire consisting of MCQs that addressed factual information about the Netflix Prize. Each question had a single correct answer, but participants also had the option to select "I don't know." We presented users with 18 questions, equally distributed between those related to the general topic and others specifically addressing content in the main document. These are referred to as "general topic-related" and "document-related" questions respectively.

Participants were asked to respond to these questions at multiple stages throughout the study.

An example of a document-related question is as follows: Here is an example of a MCQ asked during the assessment:

> What was the task of the Netflix Prize?
>
> - To identify users and films based on their ratings
> - To improve Netflix's own algorithm for predicting ratings
> - To predict ratings for films based on previous ratings without any other information
> - To award prizes to users who rated films most accurately
> - I don't know

An example of a general topic-related question is:

> Which of the following statements about the Netflix Prize Sequel is true?
>
> - The second Netflix Prize competition was never planned
> - No participant was declared the winner of the "Netflix Prize II" competition
> - The second Netflix Prize competition, known as "Netflix Prize II" competition was canceled
> - The Netflix Prize II differs from the original Netflix Prize competition by having a different evaluation metric on a smaller dataset size
> - I don't know

To assess the user's learning at the *Remember* level, we assigned a correctness score to each question. A user's test score is determined by the total number of correct answers. To measure the learning progress between two points, we calculate the difference between the test scores at these respective points.

5.3.4.2 'Understand' Assessment: Vocabulary Test

We designed a vocabulary knowledge test to evaluate the user's capacity to recall specific topic-related concepts. This test involved presenting users with a series of vocabulary terms and requesting that they assess their own familiarity with each term on a 4-point scale [166]. If participants indicated familiarity with a term, they were also required to provide a definition.

The test for a vocabulary term included the following options:

- I don't remember having seen this term/phrase before.
- I have seen this term/phrase before, but I don't think I know what it means.
- I have seen this term/phrase before and I think it means...
- I know this term/phrase. It means...

The tests consisted of a different number of vocabulary terms at each stage, ranging from 8 at the Pre-Task stage, to 18 at the Post-Document stage and 20 at the Post-Task stage. As we performed rounds of pilot testing, this appeared to be a reasonable tradeoff between coverage and effort, as based on participant feedback, higher numbers of terms proved to be discouraging. Therefore, with this design, we started with a small set of terms to probe basic, intermediate, and advanced prior knowledge in the topic, and increased the number in subsequent stages to include prior overlap with prior stages and appearance in the documents the participants read at the given stage.

The vocabulary terms selected were based on a list of candidates that was automatically extracted from the documents in our corpus about the Netflix Prize using the Wikifier service. Wikifier annotates a given text document with links to relevant Wikipedia concepts [30]. Using this list, we categorized each term based on difficulty, whether it is a prerequisite or post-requisite of another term, and whether it appears in the main document that the participant will be given to read. We used these criteria to arrive at our final lists to include in our tests, such that we achieved a balance of our criteria.

On the search results page, we excluded keywords that participants considered to be familiar in their vocabulary pre-test as a simple adaptive measure to potentially reduce the number of keywords we recommend learning and hence that the participant must consider. Because a participant is able to ask for definitions beyond the terms recommended, they may nonetheless choose to ask for a definition of a familiar term not shown for confirmation.

To evaluate participant responses, we coded the responses on the four-point scale given in [49]. Each response was considered to have multiple key aspects, and the score was dependent on how well the aspects were covered relative to the definitions in Wikipedia for the same term. As such, a definition that covers no aspects were given a score of zero, and one that covered all aspects were given a score of 3.

5.3.5 Study Participants

We conducted an online study with student subjects recruited through a masters program in data science. The study took approximately two hours to complete and we provided $30 in

compensation for each participant contingent upon completion of various stages of the study. To incentivize users to engage with the more advanced questions in the prior knowledge test, we provided a $10 bonus for users who answer them, independent of the correctness of their responses (which was assessed separately). Before participating in the study, all participants provided informed consent for data collection. Participants' confidentiality was maintained: all data collected were anonymized, removing any personally identifiable information. Seven participants completed our study; one was excluded due to technical issues experienced during the study.

The participants had diverse academic backgrounds, including computer engineering, data science, art and design, business, and microbiology. They all used search engines daily and had varying degrees of experience with ChatGPT, ChatPDF or other conversational tools designed for answering questions about documents.

While all participants had used ChatGPT at least once, their familiarity with it ranged from infrequent use (less than once per month) to daily use. By contrast, not all of them were familiar with conversational tools that chat with documents like ChatPDF. Two participants had never tried it, one had heard of it, and two had tried it and one used it regularly.

5.3.6 Chatbot Assisting the Learning Task

The chat was facilitated by a small messaging interface in the lower-right corner of the screen, fashioned after instant messaging. The bulk of the conversational functionality was provided by the OpenAI ChatGPT API with system-level prompting to set the "personality" of the agent as a helpful assistant. In addition, we provided the API with contextual information about the Web page that the user was reading including the title, an excerpt, and a snippet of context that we considered to be the most similar to the user query using word embeddings. Reviewing the chatbot responses after the experiment, we found that the chatbot rarely hallucinated in its answers. This will be discussed further in Section 5.4.2.

5.3.7 Interviews

One week after the study to coincide with a delayed post-test, we conducted interviews with participants to further understand their use of the system beyond our inferences from log data and their questionnaire responses. We asked for clarifications about participants' opinions about the study including the participants' perceptions of task clarity and the complexity of the topic, opinions about the chatbot including their levels of trust, and a verification of the information they provided while taking the study. Despite the potential opportunity to conduct an interview immediately after the study, we opted for a single interview one week

after to reduce the time and energy commitment needed from the participants. We may have potentially been able to gather insights when the experience was fresh in their minds, but participants may have had an increased opportunity to reflect on their experience in the intervening time. Participants were also asked to elaborate on their general preferences for search engines and conversational tools for learning, as well as the rationales behind their choices during the study. Interviews ranged in length from 28 minutes to 50 minutes, with a mean of 40 minutes.

5.4 Results

We now present our results, beginning with an inspection of the time spent on the learning task in Section 5.4.1, an analysis of user interaction in Section 5.4.2, the participants' responses to our tests and questionnaires in Section 5.4.3, an analysis of learning gains in Section 5.4.4, and ending with an examination of our post-task interviews in Section 5.4.5.

5.4.1 Time Spent on Task

Participants were provided with 45 minutes to complete the entire task in two stages. Our analysis of the data revealed a pattern in participants' time utilization, with all participants dedicating a substantial portion of their time to the reading part. Specifically, none of the participants exceeded a duration of 1.55 minutes for the search task. In comparison, the time spent during the reading stage ranged between 6.14 and 45 minutes. Later interview analysis revealed a misunderstanding regarding the distribution of parts and time allocation, which resulted in participant uncertainty about whether both tasks were required and how time was distributed. Half of the participants had no time left for the search phase, and the other half completed the entire task in fewer than twenty minutes. This behavior also influenced the interaction patterns seen during the chat analysis, justifying why nearly all questions to the chatbot were issued during the reading phase.

5.4.2 User Interaction

For user interaction, we now take a look at how participants used the chatbot and investigate how participants used the key concepts presented on the search page alongside each document snippet.

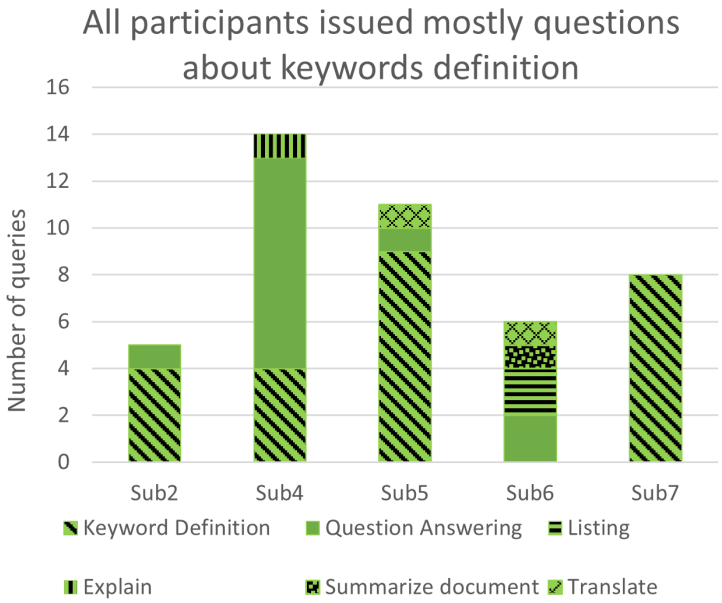


Figure 5.3: Distribution of different question types issued by each participant

5.4.2.1 Interaction With the Chatbot

We look at how users interacted with the chatbot by examining the characteristics of the questions they asked of it, and how the participants' prior topic knowledge and experience with chatbots interacted with their use.

Diversity of questions. To understand the user's interaction with the chatbot, we sought to categorize the questions they asked of it and identified six such categories: *Explain* (where the participant asks for an explanation or elaboration of a concept further than a prior response), *Keyword Definition* (where a participant asks the chatbot to define a keyword), *Listing* (where the chatbot is asked to list examples), *Question Answering* (where a more general question about a related, non-keyword concept is asked), *Summarize Document*, and *Translate* (where a participate asks the chatbot to translate a keyword or concept into another language). Keyword Definition was most frequent, followed by Question Answering, Translating, Listing, and Summarize Document in respective order.

One possible explanation for participant awareness of the chatbot's Keyword Definition capability is that the tutorial video, in demonstrating how to use the tool, included a recording of the chatbot being asked, "what is RMSE," and this became the expected interaction modality. It is also possible that users had prior experience submitting informational queries of this type to search engines or other conversational tools.

Number and length of questions. Users asked between 5 to 15 questions each, with

an average of 8.8 per user. The average length of questions ranged between 6 and 15 words per question, which is lengthier than the 2.3 word queries that users typically issue to a search engine [169].

Time between chat requests. Participants interacted with the chatbot with a frequency ranging between 1.47 to 6.68 minutes, with a mean of 3.56 minutes between successive requests. Additionally, the time taken for their first interaction ranged between 0.57 to 8.72 minutes for an mean of 4 minutes after having access to the chatbot in the main document.

Prior familiarity with chatbots. There was some evidence that prior chatbot familiarity may be associated with chatbot use. We noted that SUB6, who showed the highest familiarity with ChatGPT and ChatPDF, issued the lengthiest questions and was the fastest to initiate the interactions with the chatbot, with minimal intervals between questions of 1.47 minutes. In contrast, SUB2, the individual least familiar with chatbots, generated the fewest questions, totaling only 5, issued the shortest questions compared to other participants with an average of 6.4 keywords, and was one of the slowest to initiate the chat interaction, taking an average of 2.62 minutes to begin. Notably, the users who were most familiar with both ChatGPT and ChatPDF also tended to issue more diverse questions. For instance, SUB6 mentioned that they use ChatGPT on a weekly basis and have previously tried ChatPDF. They issued a total of 6 different types of questions. This participant was the only one to request a numbered list 'Give me a numbered list in simple English of the steps to...' and also asked for document summarization 'Summarize this document succinctly for me. Give me the big idea.'. This user employed the imperative tense, instructing the chatbot as if engaged in a real conversation.

Figure 5.3 displays the distribution of question types across subjects. SUB3 does not appear on the figure as they reported technical issues that may have prevented their questions from being captured despite their attempts.

We note however that we are unable to substantiate the exact nature of their technical issues; our server logs captured their interactions in clicking the "collapse" button of the chatbot, but no other chatbot interactions for this participant.

On the other hand, the least diverse set of questions was issued by SUB2, who is the least familiar with both tools. This user reported using ChatGPT once a year and had never heard of ChatPDF or similar tools. In contrast, the most familiar with ChatGPT alone, SUB7, claimed to use it every day but only issued one type of question, which was for keyword definitions. This was surprising, because we expected users familiar with ChatGPT to issue more complex prompts than just keyword definitions.

Prior domain knowledge. Regarding the user's prior knowledge of topic-specific factual

information, assessed through MCQs, we did not find a relationship with either the number of questions issued or their diversity. All users demonstrated limited knowledge of the topic. Even though they obtained varying scores on their pre-assessments, none of them were capable of answering more than half of the questions. To delve deeper into the impact of domain knowledge, it is advisable to contemplate a more diverse sample, encompassing participants with both high and low levels of familiarity with the Netflix Prize topic. For this study, the topic was intentionally selected to be significantly distinct from common user knowledge to provide a broader margin for learning.

5.4.2.2 Key Term Use During Search

Despite being provided with a list of key terms that participants could click on during the search phase to more easily ask the chatbot about a given term, no participants opted to use this capability. Instead, they chose to explicitly ask the chatbot questions, including a case of one participant doing so for definitions of key terms. This participant, SUB2, asked questions such as "what is overfitting" and "what is a sparse matrix" on the article pages themselves, which may indicate that it could be more useful to provide the key term breakdown on the article page at least in addition to, if not instead of, on the results page. For SUB2, these requests for definitions were for terms that on the vocabulary test prior to the stage during which they asked the question, they indicated that they were unfamiliar with the term. Perhaps in this case an adaptive keyword list provides value, but this warrants further investigation with more participants.

5.4.3 Test and Questionnaire Responses

Increase in vocabulary definitions. At each stage, we asked the participants to indicate the vocabulary terms with which they were familiar, which we described in Section 5.3.4.2. As shown in Table 5.2, the number of definitions entered tended to increase in subsequent testing stages.

Few definitions were revised. With a total of twenty vocabulary terms, we might expect that participants who entered a definition at an earlier stage may revise a definition at a later stage after they understand more about the topic or after they refresh their memories. We see that there were cases of documents being refined slightly (where we consider a refinement to be a change of more than one character), but as Table 5.2 shows, these changes are relatively sparse. Furthermore, the changes are relatively minor in ways that either correct what could have been a typographical mistake (such as correcting "Root Mean Square Evaluation" to "Root Mean Square Error"). The most substantial correction

Subject	Pre-Task Definitions	Pre→Post-Doc Edits	Post-Doc Definitions	Post-Doc→Post-Task Edits	Post-Task Definitions
SUB2	1	0	6	1	9
SUB3	6	1	17	1	17
SUB4	1	0	7	0	7
SUB5	3	0	10	0	12
SUB6	0	0	0	0	5
SUB7	8	3	11	1	14

Table 5.2: Summary of total definitions entered, and edited, at each testing stage. The number of definitions entered tended to increase from one stage to the next. Two out of seven participants revised definitions after reading the introductory document, and three out of the seven participants revised a definition at the end of the task. The two participants who revised at the post-document stage also revised definitions at the post-task stage.

could have been based on a slight misunderstanding of item-item collaborative filtering, where a participant first describes feature-based similarity and revises it in the post-task response to more closely resemble the correct definition.

5.4.4 Learning Gains

We analyze the learning gains by considering two parts of the learning task as well as the overall gain from completing it. This section will address the learning gain by comparing users' knowledge at three points: the pre-knowledge assessment, post-document knowledge, and post-task assessment.

Learning at the 'Remember' level of Bloom's Taxonomy. Before the search task, participants could answer approximately half of the MCQs, typically achieving only 1 to 2 correct responses out of the 18 questions. As shown in Fig. 5.4, by the end of the task, participants exhibited an improvement in their ability to recall facts related to the topic, successfully answering an average of 9 to 10 questions correctly. This represents a 38% increase in their knowledge of the subject.

We aimed to examine the progression of user knowledge in each phase of the learning process. As previously mentioned, the 18 MCQs were divided into two sets: 9 directly related to the document's content, which were part of all three assessments, and the other 9 related to the general topic, featured in both the pre-and post-task assessments. Figure 5.4 illustrates a general improvement in performance across all questions before and after the study.

A notable pattern emerged in the document-related questions. All users exhibited a sim-

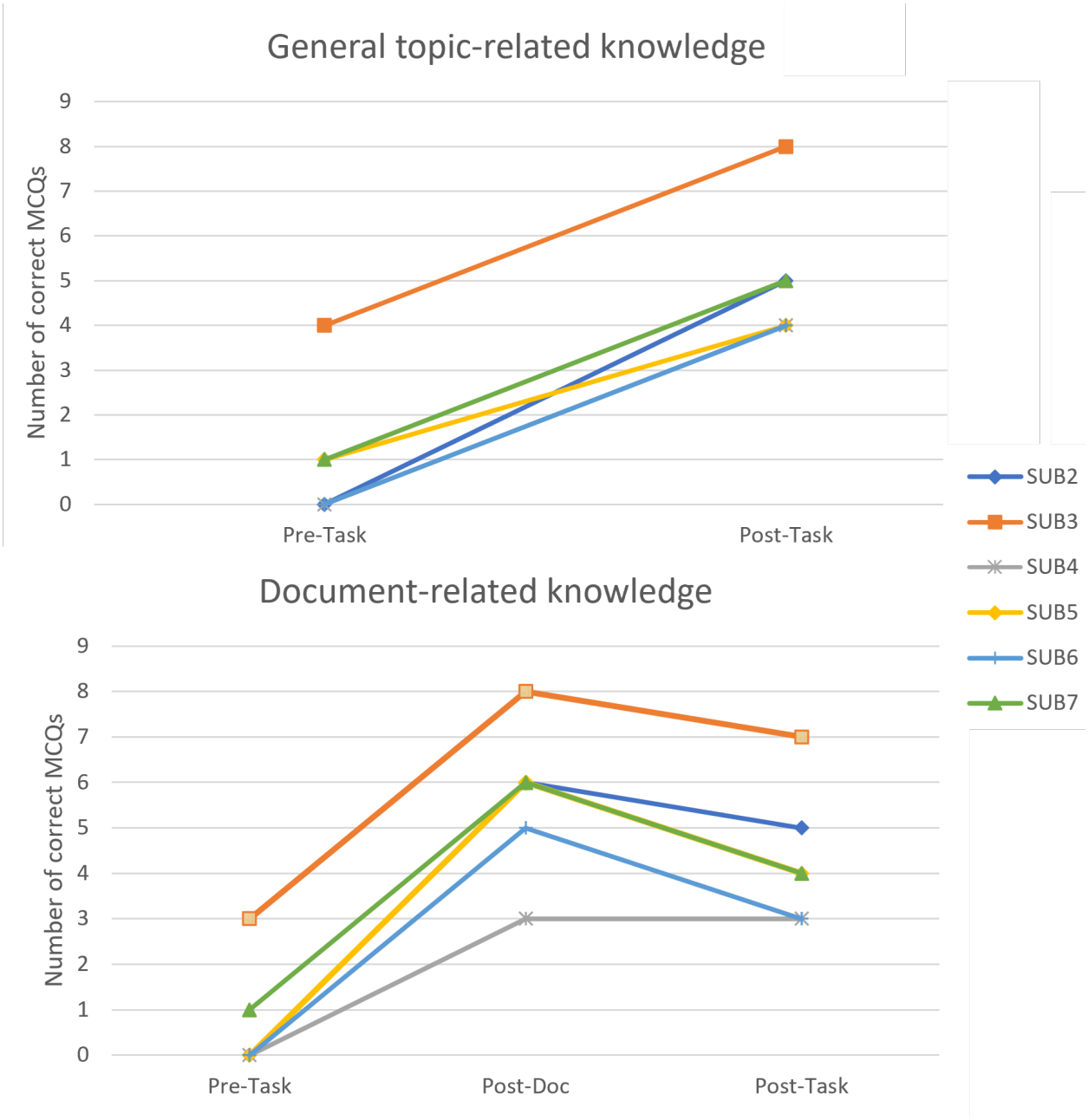


Figure 5.4: Mean knowledge change of participants over the multiple-choice question set.

ilar trend, showing a slight 5.5% decrease in their knowledge between the post-document and post-task assessments, equivalent to a drop of correctness in one question. It's worth noting that half of the users did not progress to the search stage or only accessed it briefly. Therefore, the decline in user knowledge is less likely attributed to exposure to other information sources and more likely due to the participants' fatigue towards the end of the study. Subsequent interview analysis confirmed that users found the study lengthy and were fatigued by the point of its conclusion.

Progression of vocabulary scores. We observed a similar trend to that of the drop in the knowledge questions in the result of vocabulary scores. We present a table of averages in Table 5.3. Although the mean score increases from 1.65 to 1.73 out of a maximum of 3 from the Pre-Task to Post-Document stages, the mean falls back to slightly below the Pre-Task level at the end in the Post-Task stage. We saw previously in Table 5.2 that participants were nearly consistently adding new definitions at each stage. However, because participants tended to spend most of their time reading the main document instead of on learning in the subsequent Search phase, it is likely that for new terms introduced in the post-task test – terms that primarily occur in the Search phase – the participants' understanding of these terms were less robust as they simply spent less time with them.

Pre-Task	Post-Document	Post-Task
1.65	1.73	1.61

Table 5.3: Average vocabulary definition scores at each testing stage (out of 3). There is a drop in the score at the Post-Task stage to below that at the Pre-Task stage.

5.4.5 Interview Analysis

Familiarity with search engines and conversational agents. The participants provided nuanced insights into their usage of search tools and conversational tools like ChatGPT for learning purposes. Compared to search engines, participants highlighted the dynamic nature of the experience, describing it as "wild" due to the necessity of vetting resources and validating data. Perceptions of ChatGPT were generally positive, with users valuing its speed, convenience, and ability to provide specific answers. However, there were reservations about the tool's potential to give inaccurate information, creating an illusion of correctness. Despite differences in familiarity levels with these tools, all participants were able to distinguish between conversational tools and traditional search engines, accurately listing the benefits and drawbacks of each. Some participants considered ChatGPT useful for simpler questions, but not necessarily suitable for more complex information needs where Web

search engines and scholarly platforms were deemed more reliable. Our interviews suggest that users navigate between these tools based on the nature and depth of their learning objectives and the strengths and limitations of each. When asked about their trust in the chatbot's answers, the users showed a split in attitudes towards reliability.

Study feedback. Participants generally comprehended the main objectives of learning about a specific topic and answering related questions, but they expressed confusion in using the system's features such as the clickable key concepts on the search page and navigating the different stages within the time limit. Confusion about when to switch from the Reading stage to the Search stage might have led to perceptions of the study being too long, when many participants spent the full 45 minutes on the Reading stage. Testing added to the perceptions of lengthiness, with up to 38 questions as many as up to three times. Users reported vocabulary questions being more difficult, which may be due to the potential open-ended answers required if users were familiar with a term compared to MCQs with predefined options.

The article's length was satisfactory to participants, but some found the content difficult due to technical jargon. This was intended to an extent. Three participants, particularly those less familiar with advanced data science concepts, struggled with specific technical terms such as SVD, sparse matrices, and mathematical content. Nevertheless, three participants proactively claimed that they conducted further searches for the Netflix Prize and rececommender systems after the study, which we think indicates that the study spiked their curiosity.

Behavioral analysis. In shifting from the Reading stage to the Search stage, only three participants manage to do so. When questioned about their reason for transitioning, they mentioned that they did so when they felt they had sufficiently understood the main document. Notably however, despite this, the interactions with the chatbot during the second stage were minimal, and participants ultimately opted to proceed to the post-assessment test.

On question and term selection, users gave positive feedback in instances where the chatbot provided answers in a relatable, human-like manner. Participants found that the chatbot could be a valuable resource in comprehending the article, acting as a supportive guide akin to consulting a librarian. The chatbot was able to help participants in overcoming obstacles like unfamiliar terms. Two participants in particular emphasized the positive influence of the chatbot's human-like responses on their learning experiences, saying *"asking the AI was like asking a librarian"*, and "the chatbot answered some of the questions in a very human way, so it was helpful".

Interaction with the chatbot. While two participants found value in using the chatbot for straightforward tasks like obtaining definitions, others noted its limitations in more

nuanced questions. A participant mentioned the assistance provided by the chatbot in conjunction with the article enhancing their understanding. Concerns about the chatbot's reliability surfaced as participants reported instances where it failed to adequately understand requests, prompting skepticism about its overall trustworthiness. Interestingly, there was a preference among most participants to validate information using a search engine, indicating a reliance on traditional methods for fact-checking. Other subjects relied on a more subjective assessment of the coherence of the chatbot's responses. Additionally, two participants actively engaged in validating responses within articles, reflecting a more discerning approach to ensuring the accuracy of the chatbot's output.

The lengthy duration of the study may have exacerbated user frustration when faced with system-level issues. With an imposed timeout of 30 seconds on our chatbot interface, there were cases in which OpenAI's overloaded servers took longer than expected to respond leading to some time wasted on the part of the participant as they needed to resubmit their question. Furthermore, a large language model is not truly "intelligent" – it can easily get confused by the way a question is posed, or by the complete prompt not being robust enough. There were instances of these issues as well, where one participant, SUB5, asked about matrix factorization followed by "what is mes" [sic], a misspelling of "MSE". This confused the model, leading to an answer primarily about matrix factorization. SUB3 had a similar issue, where the model was confused about the previous and current requests about "RMSE" and "SVD" respectively, conflating the question into a comparison between the two. Challenges such as these may have contributed to participants perceiving the chatbot as unreliable.

Learning and knowledge gain. Four out of six participants noted that they could grasp the article's general content but found it challenging to understand the technical intricacies. One also noted attempting to grasp the mathematical concepts by employing a traditional approach with paper and pen, jotting down information from the article in an effort to comprehend the concepts. However, they expressed difficulty in achieving a profound understanding of the material using this method. The collective findings from both the study and interviews suggest that while the chatbot proved valuable for summarizing the article's content, participants were not able to adequately use it to facilitate a deep understanding of the article's technical details.

5.5 Discussion and Implications

Our study set out to understand how users interacted with a chatbot that has the contextual awareness to answer questions about the page a user was reading and to investigate

the learning effects that might emerge from exposure to both this intelligent assistant and information about the key concepts in a set of articles. We discuss our main findings below, as well as potential implications.

The chatbot was used to explain vocabulary terms for comprehending articles (RQ1). In interviews, we received feedback that the chatbot was not only useful for defining unfamiliar terms, but that it was able to provide information about the articles in a helpful way. For users who tried it in this manner, they considered it to be "like a librarian". Interaction logs and the transcript verify this type of use; its conversational nature meant that a participant was able to ask for the definition of a term in English, and then ask for the term's translation to Mandarin.

Users' prior knowledge did not affect their learning (RQ2). We expected that that prior knowledge would play a role, and perhaps it in fact does. However, our participants had the same levels of knowledge gain. A likely factor is that all participants tended to have lower levels of prior knowledge, as they were unable to answer most of the knowledge multiple-choice questions. A larger-scale study with a more diverse subject pool might be needed to understand this characteristic.

User trust in chatbots was low, and remained low (RQ3). Users saw the potential of the chatbot to provide quick and helpful answers, but understood that large language models are prone to hallucination and misunderstanding requests. These expectations were confirmed during the study – although we saw no instances of hallucinations in our logs when reviewed by domain experts, there were cases of the model misunderstanding which parts of the chat transcript was the history and which parts was the request. These cases tended to be when both the transcripts and requests are short, and when there are grammatical or spelling errors in the request. During interviews, users expressed a preference for access to a search engine to verify the output of the model; we disallowed this as the focus of the study is on chatbot interaction, but for a production system search integration may prove helpful. We left the idea of trust up to interpretation by interviewees, but a limitation of this approach was that we failed to capture the complexity of the phenomenon. For instance, [120] distinguishes between cogitive trust and emotional trust when customers engage in electronic commerce with software agents. Cognitive trust, in this case, is the "rational expectation" that the agent is competent and can be relied upon. Emotional trust, in contrast, is the degree to which the user feels secure and comfortable that the agent is reliable. It would be interesting to qualitatively study the rational components of trust in comparison to the emotional aspects by controlling familiarity, demeanor, propensity to explain responses, and expertise personalized to users' prior knowledge. For new users, we may also examine *swift trust* – a presumptive form that sees trust being formed between team members with no

prior relationships [141]. It may also be the case that trust is continously negotiated in a scenario such as ours when an assistant is used in various conditions and for different tasks – in this instance, understanding how trust is formed and re-formed would have implications for design.

Further findings in our results show that users have a good understanding of the capabilities and limitations of large language models, even if they are not familiar with using them. They were able to articulate the difference between the chatbot and a search engine in fact, which tells us that there is a place for both and it may be a mistake for system designers to try to conflate the two.

Participants' recognition of the contrast between the chatbot's singular responses and the search engine's multiple results underscores the importance of adaptability in the learning process. From these observations, we can infer that learners benefit from a dynamic and flexible learning strategy that combines both specific, focused responses (as provided by a chatbot) and a broader exploration of multiple resources (as facilitated by a search engine). This suggests that a balanced and adaptive learning approach, leveraging different tools for their respective strengths, contributes to a more comprehensive and nuanced understanding of the subject matter.

We also saw that user who was most familiar with ChatGPT or systems like it showed more expertise in interacting with the chatbot. They issued more and longer questions, and had smoother conversations with faster turnaround. This is to be expected, however, we cannot confidently confirm the exact nature of the relationship between familiarity and frequency of interaction. From our results it is not a linear relationship, but a larger-scale study would be needed to confirm and quantify this.

An unexpected phenomenon seen in our results was a drop in learning outcomes – on both MCQs and vocabulary – at the end of the study, despite an initial increase after reading the main article. We suspect that this is due to fatigue, and might indicate that our study protocol needs revision.

The adaptive nature of our key concept extraction is quite simple due to the scope of our study, but there is much room for improvement. Earlier vocabulary tutoring systems tended to focus on question generation, as seen in work by Brown et al. [32] We have also seen work in reading support for second language documents that uses a classifier to predict unfamiliar words, presenting a pop-up with the meaning when hovered over [61]. This is closer to our work, but we present keywords and definitions before users encounter them in the document and rely on the simple method of excluding concept candidates that users indicated they were familiar with in a prior knowledge test. A planned future work will employ eyetracking to not only understand users' reading patterns, but also as a source of data to predict word

familiarity [93, 178, 24]

The results of our study show the potential of large language models as a conversational assistant for vocabulary learning and reading. Users were not only satisfied with the capabilities of our chatbot assistant, but they also exhibited learning gains beyond their initial prior knowledge during the study. Limitations in the design and implementation of our study protocol such as OpenAI integration, study length, and instruction clarity might have hampered users' potential learning gains and added to their frustration and fatigue. Our integration with OpenAI's APIs left us unable to control the timing of the responses which led to timeouts and reduced user confidence. The length of the study was likely too long, and the effort it required was a common source of complaints during the interviews. Finally, perhaps due to unclear instructions, most users did not move from the reading to the search stage, which limited our data collection and prevents us from addressing all of our research questions. These issues can be remedied, and a future larger-scale study will allow us to more confidently quantify these effects. One of the aims of this study was to explore the usability of chat as an interface for reading assistance. Although we added interactive interface elements such a clickable keywords to the search results page, chat remained the primary modality for using the assistant while reading. It remains to be seen if chat is the most appropriate interface – a future study may explore other modalities using large language models such as generating assistive information on keywords alongside article text. Finally, our pre- and post-tests have a high degree of overlap, which may have affected user behavior by encouraging participants to look for specific answers to these questions while reading. We do not think this is necessarily a major shortcoming, but different questions on the same corresponding concepts at each stage might be preferrable.

We may potentially expand or refine a few aspects of our experiment and protocol in future iterations. As an example, we intended to administer a delayed post-test one week after the study to coincide with our interviews, but the test was not given due to an oversight. Giving this test would allow us to measure retention and the effect that prior knowledge might have on retention. Furthermore, the work presented here did not implement an experimental manipulation in order to gather interaction data and identify trends in our interviews. However, we intend to compare knowledge gain among chatbot users in comparison to a control group without a chatbot by providing standard definitions for keywords on the search page. Additionally, we may also choose to introduce a manipulation based on prior knowledge in which we dynamically personalize keyword recommendations and adjust the chatbot's prompt to answer with more or less sophistication depending on a user's expertise.

5.6 Future Work

The work we present in the following sections will serve as the first steps towards a retrieval framework for optimising the set of actions a user may take towards maximising the potential utility of a document. Here, we propose a search algorithm that performs joint optimization of search rankings and words that a particular student should learn.

5.6.1 Keyword Optimisation

With the increased interest in search as learning, there has been a fair amount of recent work to place retrieval within a framework of optimisation for learning outcomes. In 2017, Syed and Collins-Thompson [186] incorporated a cognitive learning model into their ranking objective and showed using a crowdsourced study that personalising in this manner led to increased learning gains for words read. This work built upon an optimisation algorithm proposed by Raman et al. [161] to re-rank search results to provide a user with intrinsically diverse results.

The intrinsic diversity re-ranking algorithm by Raman et al. does its job by performing greedy optimisation on a *diversity function* that incorporates query suggestions to give a ranking that jointly optimises the combination of documents and related queries. The work by Syed and Collins-Thompson modified this algorithm by consider the *aspects* of a topic by extracting subheaders from Wikipedia articles on a query's topic and incorporating a new sub-objective term to represent a user's effort in reading documents based on keyword density. This current proposal takes inspiration from these works, by reformulating the original intrinsic diversity algorithm by Raman et al. [161] to jointly optimise the final search result ranking with a set of keywords to learn for each result, as well as incorporating a user's familiarity with the set of keywords. The aim of this optimisation step is to present a set of actions to users that, when taken, should maximise the user's utility. In this case, the actions are the keywords we present for users to learn (which, when clicked, serve as resources to learn more about the particular term through either a video or definition), and utility is gained through each result.

5.6.2 Optimisation Algorithm

We present a sketch of the algorithm as Algorithm 5.1, which we will use as a starting point to obtain the rankings of documents and words to learn for each document in response to a search query. This algorithm is an extension of that proposed by Raman et al. [161] which jointly optimises the relevance of documents with the diversity of topics to which the

Clustering problems and *clustering algorithms* are often overly sensitive to the presence of *outliers*: even a handful of points can greatly a ect the structure of the *optimal* solution and its *cost*. This is why many algorithms for *robust clustering* problems have been formulated in recent years. These *algorithms* discard some points as *outliers*, excluding them from the *clustering*. However, *outlier* selection can be unfair: some categories of *input* points may be disproportionately affected by the *outlier* removal *algorithm*.

Figure 5.5: Terms identified by Wikifier in a scientific paper abstract (italicised, text from https://doi.org/10.1145/3488560.3498485)

documents belong. For our purposes, our changes revolve around two primary goals:

1. Return a document set that, for each document, has an associated set of keywords that were jointly optimised with the relevance of the document during re-ranking, and

2. Incorporate the vocabulary of the document as well as related terms that would help in learning this vocabulary in the re-ranking as potential words to learn.

We outline a ranking objective function that satisfies these goals, along with the following:

1. The documents in the final ranking should be relevant to the submitted query q

2. Keywords should be related to the initial query q.

3. Ideally there should be diversity in coverage of the concepts.

With these Wikipedia concepts extracted, we use a dataset of prerequisites for the domain of data mining collected by Hu et al. (2021) [90] to determine the prerequisites of each concept. An example of the prerequisites of the term "cluster analysis" from the abstract in Figure 5.5 for example would include "data analysis", "correlation and dependence", and "arithmetic mean", and "standard deviation".

A user's familiarity with concepts will also serve as a factor. Ideally, we would like to not recommend terms that a user is already familiar with or their prerequisites. In our running example, "arithmetic mean" and "standard deviation" are prerequisites of both "cluster analysis" as well as one of its prerequisites, "correlation and dependence". We show this relationship diagrammatically in Figure 5.6. Assuming the user is already familiar with "correlation and dependence", we would avoid presenting this term as well as the shared prerequisites "arithmetic mean" and "standard deviation". Thus, for this concept, we would only present "data analysis" as a potential prerequisite.

We expect to represent a user's familiarity as a function of the reading time of a keyword or concept:

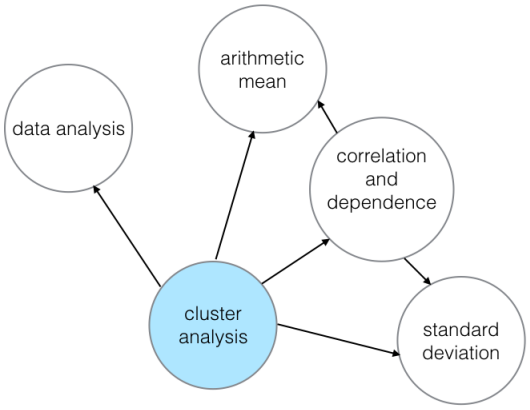


Figure 5.6: Prerequisites of "cluster analysis", a term identified in Figure 5.5.

$$Familiarity(U, k_i) = \sigma(-reading\_time(k_i))$$

This formulation is inspired by [72], which found an inverse relation between reading time and exposure to a word. This formulation is also similar to that outlined in Item Response Theory [35], where the probability of a correct answer is predicted by a logistic function of the difference between the learner's skill an the difficulty of the resource. Collecting the reading time of a word will be facilitated by the use of an eye tracker.

For the joint ranking objective of documents and keywords, we use the following formulation:

$$argmax_{(d_1,K_1)\ldots(d_n,K_n)} \sum_{i=0}^{n} Rel(d_i|q) \times e^{\beta \cdot Div(d_i,K_i) - \gamma \cdot Familiarity(U,K_i)} \quad (5.1)$$

where K_i is the set of keywords and their prerequisites recommended for document d_i and $Div(\cdot)$ is an MMR-like diversity function defined as:

$$Div(d_i, K_i) = \eta \cdot Sim(K_i, d_i) - (1 - \eta) \max_{j<i} Sim(K_i, K_j) \quad (5.2)$$

.

This formulation is due to [161]. $\eta \in [0, 1]$ is a parameter that controls the tradeoff between the relatedness of the keywords to the documents and diversity. By satisfying this tradeoff and preserving the value of diversity, we can present documents and keywords that represent various aspects of the query; for example, for a query on machine learning, we may wish to present results covering both the statistical and algorithmic aspects of the topic. This not only helps in disambiguating an ambiguous query, but also has the potential to give results on subtopics the user might not have considered explicitly searching for. Intrinsic diversity [161], or the condition of search results covering a range of related subtopics in a single ranking, has been shown to provide some additional benefit for factual and conceptual knowledge gains [50]. The complete algorithm is presented as Algorithm 5.1.

Algorithm 5.1: Words-to-Learn Algorithm that jointly ranks documents and vocabulary for learning.

Input: Query q
Input: User Model U
Result: A document set D consisting of pairs of documents and words to learn

$D \leftarrow \emptyset$;
$baseD \leftarrow$ `Query(`q`)`;
for $i = 1, ..., |baseD|$ **do**
  $d \leftarrow baseD_i$;
  $K_i \leftarrow$ `Wikifier(`d`)`;
  $bestS \leftarrow -\infty$;
  foreach *term* k *in* K_i **do**
    $K_i \leftarrow K_i \cup Prerequisites(k)$;
  end
  $v \leftarrow Rel(d) \times Sim(d, r) \times e^{\beta \cdot Div(d,K_i) - \gamma \cdot \sum_k^{K_i} Familiarity(U,k)}$;
  if $v > bestS$ **then**
    $bestS \leftarrow v$;
    `// When adding keywords, remove those and their prerequisites that the user is already familiar with`
    $K_i \leftarrow K_i \setminus \{k \in K_i | Familiarity(U, k) > \alpha \vee k \in \bigcup_j^{K_i} \{Prerequisites(j) | Familiarity(U, j) > \alpha\}\}$;
    $D \leftarrow D \cup \{(d, K_i)\}$;
  end
end
return D;

Keyword Evaluation. To evaluate the effectiveness of the algorithm's ability to surface useful keywords and their prerequisites, we may create a dataset for concept map extraction and use this dataset for evaluation – both to compare the algorithm's results to a baseline

and also to adjust its parameters.

The first step may involve collecting the top ten results of a set of search queries based on the topics covered in various course syllabi in the field of data science. Because we rely on Wikifier for keyword extraction, we may exclude Wikipedia pages from each list of results. We may then extract the text of the pages, extract candidate terms, then manually rank the terms to identify the most salient "key concepts", and then for each pair of key concepts, we manually label whether term A is a prerequisite of term B, whether term B is a prerequisite of term A, or if there is no prerequisite relationship. Using 10 queries may initially give us a seed set that we can use as gold standard data – we may then use the large language model GPT-4 from OpenAI to label the remainder by giving it the document and asking it to extract the key concepts and relationships with a few-shot prompt.

Following this step, we may use Precision@n, where $n = 1, 3, 5$ as our evaluation metric. We assume that finding the most relevant terms and prerequisites is most important for our task of recommending words to learn.

5.6.3 Potential Optimisation Study Design

In this study we would investigate the effect of our proposed algorithm on participants' learning performance for domain-specific vocabulary words. The following research questions would be useful to address:

RQ1: Does a system that optimises relevance jointly with vocabulary learning provide increased effectiveness for learning?

RQ2: Does incorporating a user's word familiarity in the optimisation objective improve learning effectiveness?

RQ3: Do reading time and exposure effectively facilitate domain-specific vocabulary learning?

We therefore propose a between-subjects experiment design, conditioned on our algorithmic intervention. Therefore, there might be three conditions based around algorithmic interventions:

- **Full**: Joint optimisation of documents and words to learn with personalisation based on word familiarity
- **No Familiarity**: Joint optimisation of documents and words to learn *without* personalisation based on word familiarity
- **Baseline**: No optimisation of documents and words to learn

Motivated by the aforementioned research questions, we may test the following hypotheses:

- As familiarity increases, users exposed to words to learn will give less attention to the recommended words while reading
- Users exposed to an algorithmic ranking of documents (**Full** or **No Familiarity**) will exhibit improved learning outcomes compared to those in the **Baseline** ranking condition (RQ1)
- Users in the **Full** condition (exposed to an optimised ranking of words to learn) will show a faster rate of learning than those in the **No Familiarity** condition (RQ2)
- Users exposed to words to learn will give additional attention to the recommended words during searching [63] (RQ3)
- Users exposed to words to learn with low familiarity will give additional attention to the recommended words while reading [62] (RQ3)

5.7 Conclusion

We reported on a user study that investigated three research questions about user interaction with a contextually-aware chatbot assistant during reading for technical learning. Using log data, knowledge tests, and interviews, we characterized usage patterns, investigated learning gains, and examined user trust in large language models as information sources. We found that users employed the chatbot assistant to explain unfamiliar terms and to help in understanding the articles they were reading (RQ1), that despite users showing learning gains we could not confirm that their prior knowledge was a factor (RQ2), and that users saw the potential of chatbots but remained skeptical of the accuracy of its output (RQ3). A future larger-scale study will explore additional means of assessing familiarity with technical terms and will seek to quantify effects after addressing the shortcomings of the present study.

5.8 Author Contributions

This study was a collaborative effort between Ryan Burton, Dima El Zein, Arpitha Ghanate, Kevyn Collins-Thompson, and Célia da Costa Pereira. Ryan Burton and Dima El Zein jointly designed the study, Arpitha Ghanate recruited participants and conducted interviews, and Kevyn Collins-Thompson and Célia da Costa Pereira contributed revisions to the manuscript. Ryan Burton implemented the study design and served as the lead author.

CHAPTER 6

Discussion and Conclusion

Web search has been stuck in a local maximum of "ten blue links" for the past few decades. Despite lamentations from researchers, professionals, and commentators, escaping this comfort zone has been difficult. There are numerous reasons for this, particularly on the part of users – expectations, inertia, and difficulty in evaluation are primary among them. It is my hope that this dissertation, in which we have explored both novel search systems as well as a simulation framework to explore the design space of interface affordances, might inspire more consideration of new interface additions, algorithms, and potentially new search systems.

6.1 Interface Additions

Chapter 3 showed the design and implementation of a new sidebar added to a conventional Web search engine result page. This sidebar introduced flexibility in the tradeoff between time and result quality – by waiting an additional amount of time, users would be provided with high-quality results relevant to their current search task. Results showed that users were willing to wait for these higher quality results and that they ultimately used less time in completing their task. However, there was some reluctance in users engaging with this new interface element, which led to the simulation framework in Chapter 4, which would would enable us to systematically ensure that "users" would be willing to use this new element, as well as change aspects of this element and explore outcomes. Regardless, slow search seems to be a good fit for decomposable tasks with subtasks that can be tackled in parallel by both the user and the system – users in our experiment tended tended not to submit new queries in parallel to slow search, choosing instead to explore the already-retrieved set of results. Future work in slow search may investigate alternative interface interventions, visualizations to convey predictions of expected future value, and various means for revealing intermediate state to users.

6.2 Algorithms

Query relevance has been the dominant metric for search algorithms, but with greater time flexibility, there is room to explore other avenues for improvement. Ideally, this may be done in conjunction with interface changes in the interest of operational transparency, but we could potentially see algorithms to improve the intrinsic diversity of results, to summarize and organize results, or to use crowdsourcing to augment algorithms. In the final sections of Chapter 5, I propose an algorithm that jointly optimizes the ranking of results and the words that a user would be expected to learn in order to get the most out of a document. This is also something that the simulation framework of Chapter 4 could be used to evaluate.

6.3 New Search Systems

For the first time in decades, Web search users outside the academic community are eager for new types of search. Spurred by the recent interest in generative AI and the perceived drawbacks of popular search engines [31, 153, 82, 122], users, especially technology enthusiasts, are interested to see what's next for Web search. Web search presents an outsized influence on users' perceptions and mental models of other search systems [115]. Therefore, as system designers, we are, for better or for worse, tied to the directions of major online search providers. I will use the remaining space to discuss how this dissertation ties into future directions of search.

Reasearch in conversational search has sought to combine the benefits of incremental assistance with interactive information seeking [107, 7]. The explosive popularity of ChatGPT however, enabled in part by the Transformer architecture [200] and abundant content available on the Internet, has led to a renewed interest in chatbots and other forms of conversational artificial intelligence (AI) as the primary modality for information seeking. Microsoft has invested in OpenAI, the company behind ChatGPT, and is introducing new AI-based products including the conversational Bing Chat search engine as well as actively iterating on existing ones such as the programming assistant GitHub Copilot [57]. Google, whose researchers invented the Transformer architecture, and who felt at risk at being left behind in AI-powered conversational search [165], rushed to develop Bard, its competitor to Bing Chat [64]. Both thus far have been unsuccessful in capturing many users despite the fanfare and media coverage, perhaps due to the penchant for generative pretrained transformers (GPTs) to "hallucinate" [215]. This does not necessarily mean that conversational search, with or without AI, is a dead-end; what it does perhaps mean however, is that generative AI is likely an unsuitable substrate for search, which centers on retrieval. Retrieval and the

types of interactions afforded by generative AI are distinct, and should likely remain as such.

Shah and Bender support this viewpoint with their paper "Situating Search" [177]. This work argues for systems that preserve the interactive aspects of information seeking while remaining transparent and accountable. These current affordances lead to the benefits we see in using search such as the capacity for information verification and serendipity. My work falls squarely along these lines – augmenting interactive information retrieval to lead users towards optimal behavior and outcomes. Operational transparency was a major factor in my design of slow search in Chapter 3. The use of generative AI in Chapter 5 in the form of a chatbot assistant was less focused on this aspect, but it I considered it important to preserve a degree of interactivity, autonomy and control within this design. Feedback from subjects showed that trust was nonetheless an issue, with the participants highlighting their relative mistrust of generative AI as a reason for their lack of interaction with the chatbot assistant we provided. This is another case in which visually conveying the value of this new interface element may increase trust and hence its usefulness. For future work, we may ask users not only about trust, but also about verification – in which situations it would be paramount versus a nice-to-have, how users currently verify information from a large language model, and how a system might help to facilitate the verification process. Qualitative coding and affinity mapping would serve as useful techniques to facilitate this analysis.

6.4 Significance of this Work to Information Retrieval

More broadly, I believe that this dissertation may contribute to potential changes in perspective in how IR is approached in the future. Firstly, as Chapters 3 and 4 highlight, there is a gap between how users may use a system and the optimal value they may derive from it. This gap deserves more attention, and I believe could be a fruitful avenue for future research and application. Chapter 4 in particular investigates a method for exploring this gap and how it might be reduced, and not only would new methods and metrics be valuable new additions to the field, but so would scalable evaluation systems in production systems be beneficial to users.

Additionally, we investigated real option pricing as a technique for evaluating time-biased gain in Chapter 4, and saw the effect of the risk/reward tradeoff in a system in which an additional sidebar ranking is provided as an option, similar to the system in Chapter 3. In particular, we observed that such an augmented system with a high-variance option can result in greater overall likely future value. This demonstrates the utility of not only a system that provides such optional components, but also of the application of real options to investigate value in dynamic interactive IR scenarios. I believe that both of these aspects

will be the subject of additional interest.

Finally, I believe it is worth highlighting the additional tradeoff of time and quality that was the focus of Chapter 3. More exploration of this tradeoff is perhaps worthwhile in the face of time- and energy-intensive AI algorithms potentially being integrated into search, but more generally I believe that there is much room for analyzing other forms of system flexibility such as the aforementioned risk and reward from Chapter 4, and of diversity [158, 161] and relevance.

6.5 Conclusion

The studies presented here were motivated by a focus on human-computer interaction in search. They contribute to the design space of search systems with the first implementation of slow search, towards understanding how users interact with novel interfaces, and in proposing ways to measure and improve user outcomes through the simulation of user interaction. Despite this, it must be noted that a primary limitation of this dissertation is that of evaluation – the studies presented here are on a small scale, and future work to investigate the effects of these interventions on large populations, especially with regards to the systems presented in Chapters 4 and 5 would allow us to test hypotheses such as the effect of the conversational learning assistant or the degree to which the findings of the simulation framework presents benefits to users exposed to system-level changes.

APPENDIX A

Crowdsourcing Relevance Judgements

The instructions given to crowdworkers for judging the relevance of search results from Section 4.4.1 are shown verbatim below in Section A.1. An example task as seen by crowdworkers is shown in Figure A.1.

A.1 Instructions

> If someone is doing a search for the task "**Find the five most influential professors in the United States in the field of sociology**", how well would this item answer the question of the task?
>
> You should be assessing the results based on the *ability of the result to address the needs of the task*.
>
> - If a person is searching for "Five Ways to Cook Bacon in the Oven", the ideal result has as many ways to cook bacon in the oven as we need (at least five).
> - If the result was "Ten Ways to Deep Fry Bacon" the ability to answer the question would be poor as the user needs ways to cook bacon in the oven and not in the deep fryer.
> - If the result was "The Best Way to Cook Bacon in the Oven", the ability to answer the question would be good since it provides one way, though the user might have to search for more.

We also provided more detailed instructions, with examples:

> **Overview**
>
> Help us determine how relevant results are to search tasks.
>
> **Steps**

1. Read and Understand the Search Task
2. Review the Search Result made during the Task
3. Select the level of relevancy for each result to the Task

Rules Tips

- You should be assessing the results based on the **ability of the result to address the needs of the task**.
  - Ability to answer the question is the purpose of the search
    - If a person is searching for "Five Ways to Cook Bacon in the Oven", the ideal result has as many ways to cook bacon in the oven as we need (at least five).
    - If the result was "Ten Ways to Deep Fry Bacon" the ability to answer the question would be poor as the user needs ways to cook bacon in the oven and not in the deep fryer.
    - If the result was "The Best Way to Cook Bacon in the Oven", the ability to answer the question would be good since it provides one way, though the user might have to search for more.
- Review the information we've provided before making your decision.
- If you aren't sure what the result is saying, open the link provided to see the full page and decide whether the content addresses the needs of the task.

Relevancy Definitions

- You should choose **Off-Topic** if:
  - The task cannot be addressed by the result.
  - The results are irrelevant to the task
  - *"Why is this item even being returned?"*
- Choose **Poor** if:
  - The ability to address the needs of the task is poorly matched.
  - The result is somewhat related to the query, but it not a good match.
  - *"I see why this is returned but it's definitely not everything I need – I probably would need to search a lot more to find what I need"*
- Choose **Good** if:

  - Matches most of the conditions of the task - or the most important parts of the task.
  - Technically, most of the task are satisfied but result doesn't provide a full, clear and complete answer to what is needed.
  - *"This broadly matches what I need, but it's not a perfect match. I might need to search a bit more"*
- Choose **Excellent** if:
  - The search intent is clearly satisfied.
  - **All specifics** of the Task appear in the Result
  - *"This is exactly what I need to finish the task"*

Examples:

Task	**Find five popular toys for girls aged 7-10 in 2015**
Result	**Title:** Uttermost Stockton White Rescued Denim Rug (5' x 8')
Relevance Score	**Off-Topic:** • *"Why is this item even being returned?"* • The intent of the query was not matched • The results are irrelevant to the search query

Task	**Find five cell phones on AT&T have the highest quality cameras**
Result	**Title:** Phone Review: Samsung Galaxy Note 3
Relevance Score	**Poor/Good:** • *"I see why this is returned but it's definitely not everything I need – I probably would need to search a lot more to find what I need"* • The result only gives information about one cell phone, but we need information for five • The ability to fully answer the question: Is this phone on AT&T? Does it have a good camera? This might depend on what is on the full page and bring the rating from Poor to Good

Task	**Find five I.T. companies in Los Angeles with fewer than 50 employees**
Result	**Title:** Meet the Hottest Startups in L.A. of the Year
Relevance Score	**Good/Excellent:** • *"This broadly matches what I need, but it's not a perfect match. I might need to search a bit more"* • Matches most of the conditions of the task - or the most important parts of the task • The ability to fully answer the question: do these companies have fewer than 50 employees? Do we have 5 companies? This might depend on what is on the full page and bring the rating from Good to Excellent

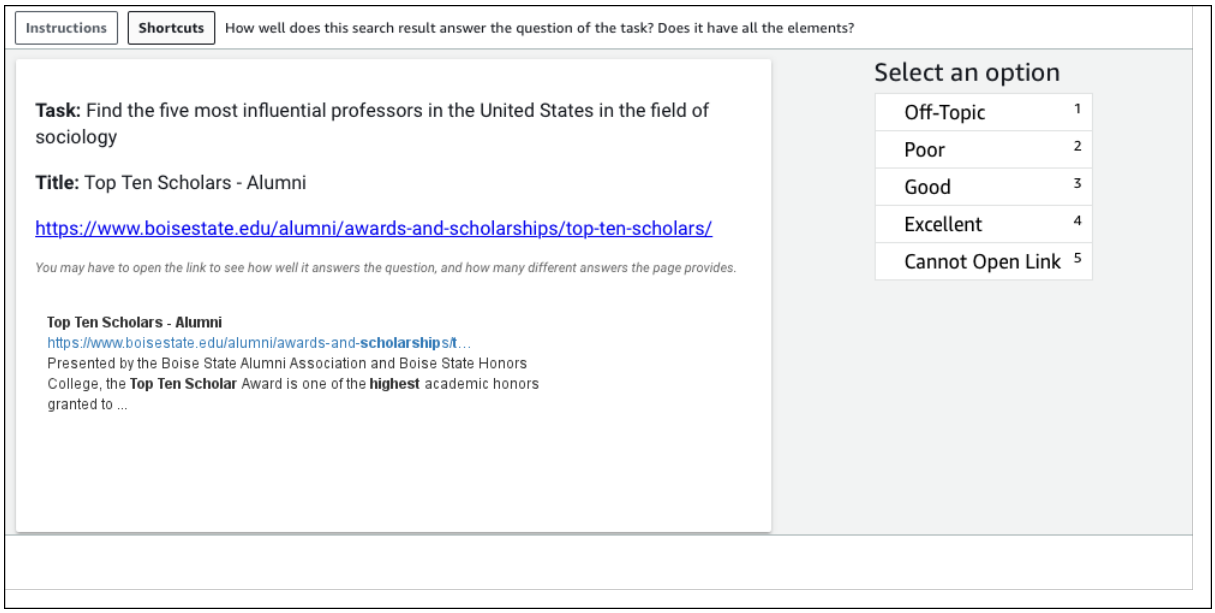


Figure A.1: An example of the task crowdworkers were given to complete on the Amazon Mechanical Turk platform.

APPENDIX B

Knowledge Assessment Questions

The following are the questions used for knowledge assessment before, during, and after the study in Chapter 5. Refer to the study design in Section 5.3 for details of when and how users were exposed to them.

B.1 'Remember' Assessment: Multiple Choice Questions

B.1.1 Document-Related Questions

For the following questions, participants were able to find the answers within the main document given during the Reading stage.

- What was the task of the Netflix Prize? (*Pre-Task*, *Post-Document*, *Post-Task*)
  - To identify users and films based on their ratings
  - To improve Netflix's own algorithm for predicting ratings
  - To predict ratings for films based on previous ratings without any other information
  - To award prizes to users who rated films most accurately
  - I don't know
- How was the movie rating data represented for prediction? (*Pre-Task*, *Post-Document*, *Post-Task*)
  - As a graph/network
  - As a list

  - As a matrix
  - As a set
  - I don't know

- Which of the following explains why the Netflix Prize matrix was considered sparse? (*Pre-Task*, *Post-Document*, *Post-Task*)
  - The matrix had high dimensions with many rows and columns
  - The matrix had a significant number of missing values
  - The matrix had an imbalance in dimensions with the number of rows being significantly lower than the number of columns
  - The matrix had unequal dimensions, with a smaller number of rows in comparison to the number of columns
  - I don't know

- Which of the following statements is true about SVD and matrix factorization? (*Pre-Task*, *Post-Document*, *Post-Task*)
  - They increase the dimensionality of a matrix
  - They reduce the dimensionality of a matrix
  - They have no effect on the dimensionality of a matrix
  - They only work on square matrices
  - I don't know

- What is the role of SVD decomposition in revealing latent features of a data matrix? (*Pre-Task*, *Post-Document*, *Post-Task*)
  - To generate new data points based on existing data
  - To reduce the dimensionality of the data matrix
  - To cluster similar data points together
  - represent the data matrix in terms of its latent features
  - I don't know

- How are rating predictions obtained using matrix factorization techniques like SVD in movies recommendation systems? (*Pre-Task*, *Post-Document*, *Post-Task*)

  - By averaging the user ratings in the user latent feature matrix
  - By adding the user latent features and movie latent features matrices
  - By subtracting the user latent features from the movie latent features matrices
  - By multiplying the user latent features and movie latent features matrices
  - I don't know

- Which statement accurately describes the use of the Stochastic Gradient Descent (SGD) method in Singular Value Decomposition (SVD) decomposition? (*Pre-Task*, *Post-Document*, *Post-Task*)
  - It maximizes mean square error to determine optimal parameters
  - It minimizes mean square error to identify optimal parameters
  - It randomly selects parameters to minimize error
  - It has no involvement in SVD decomposition
  - I don't know

- Which of the following statements accurately defines the variables in the Singular Value Decomposition (SVD) formula: $M = U\Sigma V^t$? (*Pre-Task*, *Post-Document*, *Post-Task*)
  - M represents the original data matrix, U represents the left singular matrix, Σ represents the diagonal matrix of singular values, and V represents the right singular matrix
  - U represents the diagonal matrix of singular values, Σ represents the left singular matrix, M represents the right singular matrix, and V represents the original data matrix
  - U represents the original data matrix, Σ represents the diagonal matrix of singular values, V represents the left singular matrix, and M represents the right singular matrix
  - M represents the diagonal matrix of singular values, U represents the left singular matrix, V represents the right singular matrix, and Σ represents the original data matrix
  - I don't know

- How many matrices result from SVD factorization? (*Pre-Task*, *Post-Document*, *Post-Task*)

  - Two
  - Three
  - No fixed size
  - The size is a parameter that is calculated during error minimization
  - I don't know

The following questions are more general, but still topic-related. Their answers are not necessarily in the main document, but are more evident in the set of documents on the search page.

- What was the main goal of the Netflix Prize? (*Pre-Task*, *Post-Task*)
  - To improve the user interface of the Netflix website
  - To improve the accuracy of Netflix's movie recommendations
  - To increase the number of subscribers to Netflix
  - To reduce the cost of producing original content for Netflix
  - I don't know
- What metric was used to evaluate the performance of the models in the Netflix Prize competition? (*Pre-Task*, *Post-Task*)
  - Mean squared error (MSE)
  - Mean absolute error (MAE)
  - Root mean squared error (RMSE)
  - Precision
  - I don't know
- What was the winning team's approach to the Netflix Prize? (*Pre-Task*, *Post-Task*)
  - Collaborative filtering with matrix factorization
  - Content-based filtering with decision trees
  - Item-based collaborative filtering
  - Association rule mining
  - I don't know

- Why did Netflix not adopt the winning algorithm from the Netflix Prize? (*Pre-Task*, *Post-Task*)
  - The algorithm was too computationally expensive to use in production
  - The algorithm was too difficult to implement with Netflix's existing technology
  - The algorithm did not result in a significant improvement in recommendation accuracy
  - The algorithm was not scalable to larger datasets
  - I don't know
- Cinematch is: (*Pre-Task*, *Post-Task*)
  - The name of the grand prize winner
  - The name of the algorithm to improve
  - The name of the dataset
  - I don't know
- Which of the following is a true statement about the grand prize-winning team in the Netflix Prize competition? (*Pre-Task*, *Post-Task*)
  - The grand prize was given to a team that used a completely different algorithm than SVD
  - The team that submitted their results first was declared the winner
  - Two teams were able to reach the benchmark; the winning team was the one that submitted their results first
  - No team was able to reach the competition's benchmark, and the grand prize was awarded to the team with the highest score
  - I don't know
- Which of the following statements is true about the algorithm used by the winning team of the Netflix Prize competition? (*Pre-Task*, *Post-Task*)
  - The winning algorithm was eventually found to be impractical for use by Netflix
  - Netflix chose not to implement the winning algorithm and had no intentions of doing so
  - The winning algorithm is currently being utilized by Netflix for recommendations

  - Netflix combined the algorithms of the winning teams to create an improved version that is now being used for recommendation
  - I don't know

- Which of the following statements about the Netflix Prize Sequel is true? (*Pre-Task*, *Post-Task*)
  - The second Netflix Prize competition was never planned
  - No participant was declared the winner of the "Netflix Prize II" competition
  - The second Netflix Prize competition, known as "Netflix Prize II" competition was canceled
  - The Netflix Prize II differs from the original Netflix Prize competition by having a different evaluation metric on a smaller dataset size
  - I don't know

- What was the main privacy concern raised by the release of the Netflix Prize dataset? (*Pre-Task*, *Post-Task*)
  - Competition datasets are not subject to privacy policies since they do not contain personal information about users, and no terms were violated
  - Netflix violated the terms of service by users who shared their accounts
  - Netflix data was collected and published without the users' consent
  - The possibility of re-identification attacks that could link anonymous user data to real-world identities
  - I don't know

B.2 'Understand' Assessment: Vocabulary Test

The vocabulary test's terms, which consists of a set of key terms of varying difficulty selected by their dependencies on other key terms, is shown in Table B.2. Words were given at different stages of the study, which is also shown in the table. Participants were asked to indicate their familiarity with terms by selecting on of the following options for each term:

- I don't remember having seen this term/phrase before.
- I have seen this term/phrase before, but I don't think I know what it means.

- I have seen this term/phrase before and I think it means...
- I know this term/phrase. It means...

If participants indicated familiarity with a term, they were also required to provide a definition.

Term	Dependency			Difficulty				Stage		
	Pre-	In-	Post-	S	E	M	H	Pre-	Doc	Post-
Collaborative Filtering Algorithm		•		•					•	•
Sparse Matrix	•			•				•	•	•
Item-item Filtering		•		•					•	•
Regularization		•		•				•	•	•
Implicit Feedback		•			•			•	•	•
Latent Factor	•					•		•		•
Normalization	•					•			•	•
Matrix Factorization		•		•					•	•
Singular Value Decomposition		•					•		•	•
Gradient Boosted Decision Trees		•					•		•	•
RMSE		•		•					•	•
Overfitting		•		•					•	•
Artificial Neural Network		•			•				•	•
Backpropagation		•		•					•	•
K Means Clustering		•		•					•	•
Gradient Descent		•		•				•	•	•
Ensemble Learning		•		•						
Boosting		•					•		•	•
Conditional Probability		•			•			•	•	•
Expectation Maximization			•				•	•		•

Table B.2: Vocabulary Terms which participants indicated familiarity with at various stages of the study. Dependencies are *Pre-Requisite*, *In-Document*, and *Post-Requisite* respectively. Difficulties are *Standard*, *Easy*, *Medium*,and *Hard*. Stages are *Pre-Test*, *Post-Document*, and *Post-Test* respectively.

B.3 Bonus Questions

B.3.1 Bloom's Taxonomy Level 3: 'Apply'

Suppose you were provided with a dataset for a music recommendation system. What would you consider as the rows and columns of the corresponding matrix? Additionally, how

might the matrix change when Singular Value Decomposition techniques are applied to this dataset? (4 points)

B.3.2 Bloom's Taxonomy Level 4: 'Analyze'

What makes Singular Value Decomposition (SVD) a dimensionality reduction method? Do you think it is possible that the reduction of dimensions using SVD results in a loss of data? (2 points)

B.3.3 Bloom's Taxonomy Level 5: 'Evaluate'

Assess two main strengths and two weaknesses of utilizing SVD for the Netflix Prize competition. (2 points)

B.3.4 Bloom's Taxonomy Level 6: 'Create'

List the steps required to transform a matrix using SVD. (2 points)

BIBLIOGRAPHY

[1] Introducing Deep Search. https://blogs.bing.com/search-quality-insights/december-2023/Introducing-Deep-Search, December 2023.

[2] Eytan Adar, Desney S Tan, and Jaime Teevan. Benevolent deception in human computer interaction. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1863–1872. ACM, 2013.

[3] Mikhail Ageev, Qi Guo, Dmitry Lagun, and Eugene Agichtein. Find it if you can: a game for modeling different types of web search success using interaction data. In *Proceedings of SIGIR 2011*, pages 345–354. ACM, 2011.

[4] Denise E. Agosto. Bounded rationality and satisficing in young people's web-based decision making. *Journal of the American Society for Information Science and Technology*, 53(1):16–27, 2002.

[5] Christina Aperjis, Bernardo A Huberman, and Fang Wu. Human speed-accuracy tradeoffs in search. In *System Sciences (HICSS), 2011 44th Hawaii International Conference on*, pages 1–10. IEEE, 2011.

[6] Benjamin Arai, Gautam Das, Dimitrios Gunopulos, and Nick Koudas. Anytime measures for top-k algorithms. In *Proceedings of the 33rd international conference on Very large data bases*, pages 914–925. VLDB Endowment, 2007.

[7] Sandeep Avula, Gordon Chadwick, Jaime Arguello, and Robert Capra. Searchbots: User engagement with chatbots during collaborative search

[8] David Azari, Eric Horvitz, Susan Dumais, and Eric Brill. Actions, answers, and uncertainty: a decision-making perspective on web-based question answering. *Information processing & management*, 40(5):849–868, 2004.

[9] Leif Azzopardi. The economics in interactive information retrieval. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 15–24. ACM, 2011.

[10] Leif Azzopardi. Modelling interaction with economic models of search. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pages 3–12. ACM, 2014.

[11] Leif Azzopardi. Cognitive biases in search: a review and reflection of cognitive biases in information retrieval. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*, pages 27–37, 2021.

[12] Leif Azzopardi, Diane Kelly, and Kathy Brennan. How query cost affects search behavior. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 23–32. ACM, 2013.

[13] Leif Azzopardi and Guido Zuccon. Two scrolls or one click: A cost model for browsing search results. In *Advances in Information Retrieval*, pages 696–702. Springer, 2016.

[14] Ricardo Baeza-Yates. Bias on the web. *Commun. ACM*, 61(6):54–61, May 2018.

[15] Simon Baron-Cohen. *Mindblindness: An Essay on Autism and Theory of Mind*. The MIT Press, February 1997.

[16] Simon Baron-Cohen. The extreme male brain theory of autism. *Trends in Cognitive Sciences*, 6(6):248 – 254, 2002.

[17] Simon Baron-Cohen. *Essential difference: Male and female brains and the truth about autism*. Basic Books, 2004.

[18] Simon Baron-Cohen and Sally Wheelwright. The empathy quotient: An investigation of adults with asperger syndrome or high functioning autism, and normal sex differences. *Journal of Autism and Developmental Disorders*, 34(2):163–175, 2004.

[19] Nicholas J Barrowman and Ransom A Myers. Raindrop plots. *The American Statistician*, 57(4):268–274, 2003.

[20] Feza Baskaya, Heikki Keskustalo, and Kalervo Järvelin. Time drives interaction: Simulating sessions in diverse searching environments. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 105–114, New York, NY, USA, 2012. ACM.

[21] Nicholas J Belkin, Colleen Cool, Adelheit Stein, and Ulrich Thiel. Cases, scripts, and information-seeking strategies: On the design of interactive information retrieval systems. *Expert systems with applications*, 9(3):379–395, 1995.

[22] Jens Bengtsson. The value of manufacturing flexibility: real options in practice. In *3rd Annual Real Options Conference*, 1999.

[23] Paul N. Bennett, Ryen W. White, Wei Chu, Susan T. Dumais, Peter Bailey, Fedor Borisyuk, and Xiaoyuan Cui. Modeling the impact of short- and long-term behavior on search personalization. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 185–194, New York, NY, USA, 2012. ACM.

[24] Nilavra Bhattacharya and Jacek Gwizdka. Measuring learning during search: Differences in interactions, eye-gaze, and semantic similarity to expert knowledge. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*, CHIIR '19, pages 63–71, New York, NY, USA, 2019. ACM.

[25] Benjamin Samuel Bloom. Taxonomy of educational objectives: The classification of educational goals. *Cognitive domain*, 1956.

[26] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[27] Christine L. Borgman. The user's mental model of an information retrieval system. In *Proceedings of the 8th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '85, pages 268–273, New York, NY, USA, 1985. ACM.

[28] Christine L Borgman. All users of information retrieval systems are not created equal: An exploration into individual differences. *Information processing & management*, 25(3):237–251, 1989.

[29] D. Scott Brandt and Lorna Uden. Insight into mental models of novice internet searchers. *Commun. ACM*, 46(7):133–136, July 2003.

[30] Janez Brank, Gregor Leban, and Marko Grobelnik. Annotating documents with relevant wikipedia concepts. *Proceedings of SiKDD*, 472, 2017.

[31] Dmitri Brereton. Google search is dying. https://dkb.blog/p/google-search-is-dying, February 2022.

[32] Jonathan Brown, Gwen Frishkoff, and Maxine Eskenazi. Automatic question generation for vocabulary assessment. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 819–826, 2005.

[33] Jake Brutlag. Speed matters for Google web search. http://services.google.com/fh/files/blogs/google_delayexp.pdf, June 2009.

[34] Ryan W Buell and Michael I Norton. The labor illusion: How operational transparency increases perceived value. *Management Science*, 57(9):1564–1579, 2011.

[35] Sahan Bulathwela, Maria Perez-Ortiz, Emine Yilmaz, and John Shawe-Taylor. Truelearn: A family of bayesian algorithms to match lifelong learners to open educational resources. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 565–573, 2020.

[36] Ryan Burton and Kevyn Collins-Thompson. User behavior in asynchronous slow search. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, pages 345–354, New York, NY, USA, 2016. ACM.

[37] Georg Buscher, Susan T Dumais, and Edward Cutrell. The good, the bad, and the random: an eye-tracking study of ad quality in web search. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 42–49. ACM, 2010.

[38] Georg Buscher, Ryen W White, Susan Dumais, and Jeff Huang. Large-scale analysis of individual and task differences in search result page examination strategies. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 373–382. ACM, 2012.

[39] Stefan Büttcher, Charles LA Clarke, and Ian Soboroff. The TREC 2006 Terabyte track. In *TREC 2006 Notebook*, volume 6, page 39. NIST Special Publication, 2006.

[40] Paula J Caplan, Mary Crawford, Janet Shibley Hyde, and John TE Richardson. *Gender Differences in Human Cognition. Counterpoints: Cognition, Memory, and Language Series.* ERIC, 1997.

[41] John M Carroll, Nancy S Anderson, Judith Reitman Olson, et al. *Mental models in human-computer interaction: Research issues about what the user of software knows.* Number 12. National Academies, 1987.

[42] John M. Carroll and John C. Thomas. Metaphor and the cognitive representation of computing systems. *IEEE Transactions on Systems, Man, and Cybernetics*, 12(2):107 – 116, 1982.

[43] Jay Chen, Saleema Amershi, Aditya Dhananjay, and Lakshmi Subramanian. Comparing web interaction models in developing regions. In *Proceedings of the First ACM Symposium on Computing for Development*, page 6. ACM, 2010.

[44] Sherry Y Chen and Robert Macredie. Web-based interaction: A review of three important human factors. *International Journal of Information Management*, 30(5):379–387, 2010.

[45] Benoit Chevalier-Roignant and Lenos Trigeorgis. *Competitive Strategy: Options and Games.* MIT press, 2011.

[46] Charles LA Clarke and Mark D Smucker. Time well spent. In *Proceedings of the 5th Information Interaction in Context Symposium*, pages 205–214. ACM, 2014.

[47] Cyril W Cleverdon, Jack Mills, and E Michael Keen. Factors determining the performance of indexing systems, (volume 1: Design). *Cranfield: College of Aeronautics*, 28, 1966.

[48] Mikael Collan. Thoughts about selected models for the valuation of real options. *Acta Universitatis Palackianae Olomucensis. Facultas Rerum Naturalium. Mathematica*, 50(2):5–12, 2011.

[49] Kevyn Collins-Thompson and Jamie Callan. Automatic and human scoring of word definition responses. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 476–483, 2007.

[50] Kevyn Collins-Thompson, Soo Young Rieh, Carl C. Haynes, and Rohail Syed. Assessing learning outcomes in web search: A comparison of tasks and query strategies. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval*, CHIIR '16, pages 163–172, New York, NY, USA, 2016. Association for Computing Machinery.

[51] M. Correll and M. Gleicher. Error bars considered harmful: Exploring alternate encodings for mean and error. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2142–2151, Dec 2014.

[52] Daniel Wayne Crabtree, Peter Andreae, and Xiaoying Gao. Exploiting underrepresented query aspects for automatic query expansion. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 191–200. ACM, 2007.

[53] Anita Crescenzi, Diane Kelly, and Leif Azzopardi. Time pressure and system delays in information search. In *Proceedings of SIGIR 2015*, pages 767–770, New York, NY, USA, 2015. ACM.

[54] Fu-Rong Dang, Jin-Tao Tang, Kun-Yuan Pang, Ting Wang, Sha-Sha Li, and Xiao Li. Constructing an educational knowledge graph with concepts linked to wikipedia. *Journal of Computer Science and Technology*, 36:1200–1211, 2021.

[55] Vinay T Datar and Scott H Mathews. European real options: An intuitive algorithm for the black-scholes formula. *Available at SSRN 560982*, 2004.

[56] Jérôme Dinet and Muneo Kitajima. Draw me the web: impact of mental model of the web on information search performance of young users. In *23rd French Speaking Conference on Human-Computer Interaction*, page 3. ACM, 2011.

[57] Thomas Dohmke. Universe 2023: Copilot transforms github into the ai-powered developer platform. https://github.blog/2023-11-08-universe-2023-copilot-transforms-github-into-the-ai-powered-developer-platform/, November 2023.

[58] M Dörk, P Bennett, and R Davies. Taking our sweet time to search. In *Proceedings of CHI 2013 Workshop on Changing Perspectives of Time in HCI*, 2013.

[59] Shaddin Dughmi and Haifeng Xu. Algorithmic bayesian persuasion. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 412–425. ACM, 2016.

[60] Yuka Egusa, Masao Takaku, and Hitomi Saito. How concept maps change if a user does search or not? In *Proceedings of the 5th Information Interaction in Context Symposium*, pages 68–75. ACM, 2014.

[61] Yo Ehara, Nobuyuki Shimizu, Takashi Ninomiya, and Hiroshi Nakagawa. Personalized reading support for second-language web documents. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 4(2):1–19, 2013.

[62] Carsten Eickhoff, Sebastian Dungs, and Vu Tran. An eye-tracking study of query reformulation. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, pages 13–22, New York, NY, USA, 2015. Association for Computing Machinery.

[63] Carsten Eickhoff, Jaime Teevan, Ryen White, and Susan Dumais. Lessons from the journey: A query log analysis of within-session learning. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, WSDM '14, pages 223–232, New York, NY, USA, 2014. Association for Computing Machinery.

[64] Jennifer Elias. Google CEO issues rallying cry in internal memo: All hands on deck to test ChatGPT competitor Bard. https://www.cnbc.com/2023/02/06/google-ceo-tells-employees-it-needs-all-hands-on-deck-to-test-bard.html.

[65] Jianqing Fan and Sheng-Kuei Lin. Test of significance when data are curves. *Journal of the American Statistical Association*, 93(443):1007–1021, 1998.

[66] Ethan Fast, Binbin Chen, Julia Mendelsohn, Jonathan Bassen, and Michael S Bernstein. Iris: A conversational agent for complex tasks. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–12, 2018.

[67] Nazanin Firoozeh, Adeline Nazarenko, Fabrice Alizon, and Béatrice Daille. Keyword extraction: Issues and methods. *Natural Language Engineering*, 26(3):259–291, 2020.

[68] Nigel Ford, TD Wilson, Allen Foster, David Ellis, and Amanda Spink. Information seeking and mediated searching. part 4. cognitive styles in information seeking. *Journal of the American Society for Information Science and Technology*, 53(9):728–735, 2002.

[69] Donald J. Foss, Mary Beth Rosson, and Penny L. Smith. Reducing manual labor: An experimental analysis of learning aids for a text editor. In *Proceedings of the 1982 Conference on Human Factors in Computing Systems*, CHI '82, pages 332–336, New York, NY, USA, 1982. ACM.

[70] Steve Fox, Kuldeep Karnawat, Mark Mydland, Susan Dumais, and Thomas White. Evaluating implicit measures to improve web search. *ACM Trans. Inf. Syst.*, 23(2):147–168, April 2005.

[71] Fay Fransella, Richard Bell, and Don Bannister. *A manual for repertory grid technique*. John Wiley & Sons, 2004.

[72] Aline Godfroid, Jieun Ahn, Ina Choi, Laura Ballard, Yaqiong Cui, Suzanne Johnston, Shinhye Lee, Abdhi Sarkar, and Hyung-Jo Yoon. Incidental vocabulary learning in a natural reading context: an eye-tracking study. *Bilingualism: Language and Cognition*, 21(3):563–584, 2018.

[73] Sreenivas Gollapudi and Aneesh Sharma. An axiomatic approach for result diversification. In *Proceedings of the 18th international conference on World wide web*, pages 381–390. ACM, 2009.

[74] Robert Gould. Bootstrap hypothesis testing. *Stats 110A*, 2002.

[75] Miriam Greis, Thorsten Ohler, Niels Henze, and Albrecht Schmidt. *Investigating Representation Alternatives for Communicating Uncertainty to Non-experts*, pages 256–263. Springer International Publishing, Cham, 2015.

[76] Ulrike Gretzel and Daniel R Fesenmaier. Persuasion in recommender systems. *International Journal of Electronic Commerce*, 11(2):81–100, 2006.

[77] 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*, pages 661–670, 2009.

[78] Fred H Groves. Science vocabulary load of selected secondary science textbooks. *School Science and Mathematics*, 95(5):231–235, 1995.

[79] T. Gschwandtnei, M. Bögl, P. Federico, and S. Miksch. Visual encodings of temporal uncertainty: A comparative user study. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):539–548, Jan 2016.

[80] Qi Guo, Ryen W. White, Yunqiao Zhang, Blake Anderson, and Susan T. Dumais. Why searchers switch: Understanding and predicting engine switching rationales. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 335–344, New York, NY, USA, 2011. ACM.

[81] Frank G. Halasz and Thomas P. Moran. Mental models and problem solving in using a calculator. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '83, pages 212–216, New York, NY, USA, 1983. ACM.

[82] Scharon Harding. Protests broke Reddit hack for useful Google search results—and Google knows it. https://arstechnica.com/gadgets/2023/06/google-admits-reddit-protests-make-it-harder-to-find-helpful-search-results/, Jun 2023.

[83] Janis M. Harmon, Wanda B. Hedrick, and Karen D. Wood. Research on vocabulary instruction in the content areas: Implications for struggling readers. *Reading & Writing Quarterly*, 21(3):261–280, 2005.

[84] Stephen P. Harter. Psychological relevance and information science. *Journal of the American Society for Information Science*, 43(9):602, Oct 01 1992. ISBN: 0002-8231; Last updated - 2013-02-24.

[85] Marti Hearst. *Search user interfaces*. Cambridge University Press, 2009.

[86] Lucy Holman. Millennial students' mental models of search: Implications for academic librarians and database developers. *The Journal of Academic Librarianship*, 37(1):19–27, 2011.

[87] Eric Horvitz. Lumiere project: Bayesian reasoning for automated assistance. *Decision Theory & Adaptive Systems Group, Microsoft Research. Microsoft Corp. Redmond, WA*, 1998.

[88] Eric Horvitz, Jack Breese, David Heckerman, David Hovel, and Koos Rommelse. The lumière project: Bayesian user modeling for inferring the goals and needs of software users. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, UAI'98, pages 256–265, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc.

[89] Peter Hoschka. *Computers as assistants: A new generation of support systems*. CRC Press, 1996.

[90] Cheng Hu, Kui Xiao, Zesong Wang, Shihui Wang, and Qifeng Li. Extracting prerequisite relations among wikipedia concepts using the clickstream data. In Han Qiu, Cheng Zhang, Zongming Fei, Meikang Qiu, and Sun-Yuan Kung, editors, *Knowledge Science, Engineering and Management*, pages 13–26, Cham, 2021. Springer International Publishing.

[91] Jessica Hullman. Why evaluating uncertainty visualization is error prone. In *Proceedings of the Beyond Time and Errors on Novel Evaluation Methods for Visualization*, pages 143–151. ACM, 2016.

[92] Jessica Hullman, Paul Resnick, and Eytan Adar. Hypothetical outcome plots outperform error bars and violin plots for inferences about reliability of variable ordering. *PLoS ONE*, 10(11):1–25, 11 2015.

[93] Aulikki Hyrskykari, Päivi Majaranta, Antti Aaltonen, and Kari-Jouko Räihä. Design issues of idict: a gaze-assisted translation aid. In *Proceedings of the 2000 symposium on Eye tracking research & applications*, pages 9–14, 2000.

[94] Young Hyun. Nonlinear color scales for interactive exploration. https://www.caida.org/~youngh/colorscales/nonlinear.html, 2008.

[95] Mathias Jesse and Dietmar Jannach. Digital nudging with recommender systems: Survey and future directions. *Computers in Human Behavior Reports*, 3:100052, 2021.

[96] Di Jiang, Kenneth Wai-Ting Leung, and Wilfred Ng. Context-aware search personalization with concept preference. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, pages 563–572, New York, NY, USA, 2011. ACM.

[97] Thorsten Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, pages 133–142, New York, NY, USA, 2002. ACM.

[98] David H Jonassen and Philip Henning. Mental models: Knowledge in the head and knowledge in the world. In *Proceedings of the 1996 international conference on Learning sciences*, pages 433–438. International Society of the Learning Sciences, 1996.

[99] Susan Joslyn and Jared LeClerc. Decisions with uncertainty: The glass half full. *Current Directions in Psychological Science*, 22(4):308–315, 2013.

[100] Susan L Joslyn and Jared E LeClerc. Uncertainty forecasts improve weather-related decisions and attenuate the effects of forecast error. *Journal of experimental psychology: applied*, 18(1):126, 2012.

[101] Malte F. Jung, David Sirkin, Turgut M. Gür, and Martin Steinert. Displayed uncertainty improves driving experience and behavior: The case of range anxiety in an electric car. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 2201–2210, New York, NY, USA, 2015. ACM.

[102] Daniel Kahneman and Amos Tversky. Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–291, 1979.

[103] Mika Käki and Anne Aula. Controlling the complexity in comparing search user interfaces via user studies. *Information processing & management*, 44(1):82–91, 2008.

[104] Vaiva Kalnikaitė, Jon Bird, and Yvonne Rogers. Decision-making in the aisles: informing, overwhelming or nudging supermarket shoppers? *Personal and Ubiquitous Computing*, 17(6):1247–1259, 2013.

[105] Gloria Yi-Ming Kao, Pei-Lan Lei, and Chuen-Tsai Sun. Thinking style impacts on web search strategies. *Computers in Human Behavior*, 24(4):1330–1341, 2008.

[106] Abhijith Kashyap, Vagelis Hristidis, and Michalis Petropoulos. Facetor: Cost-driven exploration of faceted query results. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, pages 719–728, New York, NY, USA, 2010. ACM.

[107] Abhishek Kaushik, Vishal Bhat Ramachandra, and Gareth JF Jones. An interface for agent supported conversational search. In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*, pages 452–456, 2020.

[108] Alan Kay. User interface: A personal view. *The art of human-computer interface design*, pages 191–207, 1990.

[109] Matthew Kay, Tara Kola, Jessica Hullman, and Sean Munson. When (ish) is my bus? user-centered visualizations of uncertainty in everyday, mobile predictive systems. In *Proceedings of the 34th Annual ACM Conference on Human Factors in Computing Systems*, CHI '16, 2016.

[110] Matthew Kay, Dan Morris, mc schraefel, and Julie Kientz. There's no such thing as gaining a pound: reconsidering the bathroom scale user interface. In *Ubicomp '13*, 2013.

[111] Diane Kelly. Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends in Information Retrieval*, 3(1—2):1–224, 2009.

[112] Diane Kelly and Leif Azzopardi. How many results per page?: A study of serp size, search behavior and user experience. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, pages 183–192, New York, NY, USA, 2015. ACM.

[113] George A Kelly. *The psychology of personal constructs. Vol. 1. A theory of personality. Vol. 2. Clinical diagnosis and psychotherapy.* WW Norton, 1955.

[114] Willett Kempton. Two theories of home heat control. *Cognitive Science*, 10(1):75 – 90, 1986.

[115] Michael Khoo and Catherine Hall. *What Would 'Google' Do? Users' Mental Models of a Digital Library Search Engine*, pages 1–12. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[116] Johannes Kiesel, Xiaoni Cai, Roxanne El Baff, Benno Stein, and Matthias Hagen. Toward conversational query reformulation. In *DESIRES*, pages 91–101, 2021.

[117] Yubin Kim, Kevyn Collins-Thompson, and Jaime Teevan. Using the crowd to improve search result ranking and the search experience. *ACM TIST: Special Issue on the Crowd in Intelligent Systems*, 7(4):50, 2016.

[118] SA Knight and Amanda Spink. *Toward a Web search information behavior model.* Springer, 2008.

[119] David A Kolb. *The Kolb learning style inventory.* Hay Resources Direct Boston, MA, 2007.

[120] Sherrie Xiao Komiak and Izak Benbasat. Understanding customer trust in agent-mediated electronic commerce, web-mediated electronic commerce, and traditional commerce. *Information Technology and Management*, 5(1):181–207, 2004.

[121] Y. K. Leung and M. D. Apperley. A review and taxonomy of distortion-oriented presentation techniques. *ACM Trans. Comput.-Hum. Interact.*, 1(2):126–160, June 1994.

[122] Amanda Chicago Lewis. The people who ruined the internet. https://www.theverge.com/features/23931789/seo-search-engine-optimization-experts-google-results.

[123] Chen Liang, Zhaohui Wu, Wenyi Huang, and C Lee Giles. Measuring prerequisite relations among concepts. In *Proc. 2015 Conf. Empirical Methods in Natural Language Processing*, pages 1668–1674, 2015.

[124] Chen Liang, Jianbo Ye, Shuting Wang, Bart Pursel, and C Lee Giles. Investigating active learning for concept prerequisite learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[125] Chen Liang, Jianbo Ye, Zhaohui Wu, Bart Pursel, and C Lee Giles. Recovering concept prerequisite relations from university course dependencies. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[126] Tracy Xiao Liu, Jiang Yang, Lada A Adamic, and Yan Chen. Crowdsourcing with all-pay auctions: A field experiment on Taskcn. *Proceedings of the American Society for Information Science and Technology*, 48(1):1–4, 2011.

[127] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of machine learning research*, 9(Nov):2579–2605, November 2008.

[128] Stephann Makri, Ann Blandford, Jeremy Gow, Jon Rimmer, Claire Warwick, and George Buchanan. A library or just another information resource? a case study of users' mental models of traditional and digital libraries. *Journal of the American Society for Information Science and Technology*, 58(3):433–445, 2007.

[129] Yazdan Mansourian and Nigel Ford. Search persistence and failure on the web: a "bounded rationality" and "satisficing" analysis. *Journal of Documentation*, 63(5):680–701, 2007.

[130] James G March and Herbert A Simon. *Cognitive limits on rationality*. Wiley and Sons, New York, 1958.

[131] Gary Marchionini. *Information Seeking in Electronic Environments*. Number 9. Cambridge University Press, New York, NY, USA, 1997.

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

[133] Takahiko Masuda and Richard E Nisbett. Attending holistically versus analytically: comparing the context sensitivity of japanese and americans. *Journal of personality and social psychology*, 81(5):922, 2001.

[134] Scott Mathews, Vinay Datar, and Blake Johnson. A practical method for valuing real options: The boeing approach. *Journal of Applied Corporate Finance*, 19(2):95–104, 2007.

[135] Michael L Mauldin. Retrieval performance in FERRET: A conceptual information retrieval system. In *Proceedings of SIGIR 1991*, pages 347–355. ACM, 1991.

[136] David Maxwell and Leif Azzopardi. Stuck in traffic: how temporal delays affect search behaviour. In *Proceedings of the 5th Information Interaction in Context Symposium*, pages 155–164. ACM, 2014.

[137] David Maxwell and Leif Azzopardi. Agents, simulated users and humans: An analysis of performance and behaviour. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, CIKM '16, pages 731–740, New York, NY, USA, 2016. Association for Computing Machinery.

[138] David Maxwell and Leif Azzopardi. Information scent, searching and stopping. In *European Conference on Information Retrieval*, pages 210–222. Springer, 2018.

[139] David Maxwell, Leif Azzopardi, Kalervo Järvelin, and Heikki Keskustalo. An initial investigation into fixed and adaptive stopping strategies. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 903–906. ACM, 2015.

[140] Joanna McGrenere, Ronald M. Baecker, and Kellogg S. Booth. An evaluation of a multiple interface design solution for bloated software. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '02, pages 164–170, New York, NY, USA, 2002. ACM.

[141] Debra Meyerson, Karl E Weick, Roderick M Kramer, et al. *Swift trust and temporary groups*, volume 166. Sage, Thousand Oaks, CA, 1996.

[142] George D Montañez, Ryen W White, and Xiao Huang. Cross-device search. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 1669–1678. ACM, 2014.

[143] Thomas P Moran. The command language grammar: A representation for the user interface of interactive computer systems. *International journal of man-machine studies*, 15(1):3–50, 1981.

[144] Jack Muramatsu and Wanda Pratt. Transparent queries: Investigation users' mental models of search engines. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '01, pages 217–224, New York, NY, USA, 2001. ACM.

[145] Isabel Briggs Myers. *The Myers-Briggs Type Indicator*. Consulting Psychologists Press Palo Alto, CA, 1962.

[146] Philip M Napoli and Jonathan A Obar. The emerging mobile internet underclass: A critique of mobile internet access. *The Information Society*, 30(5):323–334, 2014.

[147] Richard E Nisbett and Yuri Miyamoto. The influence of culture: holistic versus analytic perception. *Trends in cognitive sciences*, 9(10):467–473, 2005.

[148] Donald A Norman. Cognitive engineering. *User centered system design: New perspectives on human-computer interaction*, 3161, 1986.

[149] Christian Otto, Ran Yu, Georg Pardi, Johannes von Hoyer, Markus Rokicki, Anett Hoppe, Peter Holtz, Yvonne Kammerer, Stefan Dietze, and Ralph Ewerth. Predicting knowledge gain during web search based on multimedia resource consumption. In Ido Roll, Danielle McNamara, Sergey Sosnovsky, Rose Luckin, and Vania Dimitrova, editors, *Artificial Intelligence in Education*, pages 318–330, Cham, 2021. Springer International Publishing.

[150] Antti Oulasvirta, Janne P Hukkinen, and Barry Schwartz. When more is less: the paradox of choice in search engine use. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 516–523. ACM, 2009.

[151] Haojie Pan, Cen Chen, Chengyu Wang, Minghui Qiu, Liu Yang, Feng Ji, and Jun Huang. Learning to expand: Reinforced response expansion for information-seeking conversations. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 4055–4064, 2021.

[152] Liangming Pan, Chengjiang Li, Juanzi Li, and Jie Tang. Prerequisite relation learning for concepts in moocs. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1447–1456, 2017.

[153] Jay Peters. Google is getting a lot worse because of the Reddit blackouts. https://www.theverge.com/2023/6/13/23759942/google-reddit-subreddit-blackout-protests, June 2023.

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

[155] Liz Poirier and Lyn Robinson. Informational balance: slow principles in the theory and practice of information behaviour. *Journal of Documentation*, 70(4):687–707, 2014.

[156] David Premack and Guy Woodruff. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, 1(04):515–526, 1978.

[157] Philip Quinn and Andy Cockburn. Loss aversion and preferences in interaction. *Human–Computer Interaction*, 0(0):1–48, 2018.

[158] Filip Radlinski, Paul N Bennett, Ben Carterette, and Thorsten Joachims. Redundancy, diversity and interdependent document relevance. In *ACM SIGIR Forum*, volume 43, pages 46–52. ACM, 2009.

[159] Filip Radlinski and Nick Craswell. A theoretical framework for conversational search. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*, CHIIR '17, pages 117–126, New York, NY, USA, 2017. Association for Computing Machinery.

[160] Davood Rafiei, Krishna Bharat, and Anand Shukla. Diversifying web search results. In *Proceedings of the 19th international conference on World wide web*, pages 781–790. ACM, 2010.

[161] Karthik Raman, Paul N. Bennett, and Kevyn Collins-Thompson. Toward whole-session relevance: Exploring intrinsic diversity in web search. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, pages 463–472, New York, NY, USA, 2013. Association for Computing Machinery.

[162] Bryon Reeves and Clifford Nass. *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*. Cambridge University Press, New York, NY, USA, 1998.

[163] Gary Ren, Xiaochuan Ni, Manish Malik, and Qifa Ke. Conversational query understanding using sequence to sequence modeling. In *Proceedings of the 2018 World Wide Web Conference*, pages 1715–1724, 2018.

[164] Yvonne Rogers, Andrew Rutherford, and Peter A Bibby. *Models in the mind*. Academic Press, 1992.

[165] Kevin Roose. Bing (yes, bing) just made search interesting again. https://www.nytimes.com/2023/02/08/technology/microsoft-bing-openai-artificial-intelligence.html, February 2023.

[166] Nirmal Roy, Felipe Moraes, and Claudia Hauff. Exploring users' learning gains within search sessions. In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*, CHIIR '20, pages 432–436, New York, NY, USA, 2020. Association for Computing Machinery.

[167] Sudeshna Roy, Meghana Madhyastha, Sheril Lawrence, and Vaibhav Rajan. Inferring concept prerequisite relations from online educational resources. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 9589–9594, 2019.

[168] Floyd Leon Ruch and William W Ruch. *Employee aptitude survey: Technical report*. Psychological Services Los Angeles, CA, 1963.

[169] Daniel M. Russell and Mario Callegaro. How to be a better web searcher: Secrets from google scientists, 2019. Accessed: October 2023.

[170] Daniel M. Russell, Mark J. Stefik, Peter Pirolli, and Stuart K. Card. The cost structure of sensemaking. In *Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems*, CHI '93, pages 269–276, New York, NY, USA, 1993. ACM.

[171] G. Salton. Evaluation problems in interactive information retrieval. *Information Storage and Retrieval*, 6(1):29–44, 1970.

[172] Tefko Saracevic. Relevance reconsidered. In *Information science: Integration in perspectives. In Proceedings of the Second Conference on Conceptions of Library and Information Science*, volume 1, pages 201–218, 1996.

[173] Tefko Saracevic, Paul Kantor, Alice Y Chamis, and Donna Trivison. A study of information seeking and retrieving: 1. background and methodology. *Readings in Information Retrieval. San Francisco: Morgan Kaufmann*, pages 175–190, 1997.

[174] Robert M. Schumacher and Mary P. Czerwinski. *Mental Models and the Acquisition of Expert Knowledge*, pages 61–79. Springer New York, New York, NY, 1992.

[175] Barry Schwartz and Andrew Ward. Doing better but feeling worse: The paradox of choice. *Positive psychology in practice*, pages 86–104, 2004.

[176] N Sebe, I Cohen, TS Huang, and Th Gevers. Human-computer interaction: a bayesian network approach. In *International Symposium on Signals, Circuits and Systems, 2005. ISSCS 2005.*, volume 1, pages 343–346. IEEE, 2005.

[177] Chirag Shah and Emily M Bender. Situating search. In *Proceedings of the 2022 Conference on Human Information Interaction and Retrieval*, pages 221–232, 2022.

[178] John L Sibert, Mehmet Gokturk, and Robert A Lavine. The reading assistant: eye gaze triggered auditory prompting for reading remediation. In *Proceedings of the 13th annual ACM symposium on User interface software and technology*, pages 101–107, 2000.

[179] Mark D Smucker. An analysis of user strategies for examining and processing ranked lists of documents. *Proc. of 5th HCIR*, 2011.

[180] Mark D. Smucker and Charles L. A. Clarke. Stochastic simulation of time-biased gain. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 2040–2044, New York, NY, USA, 2012. ACM.

[181] Mark D Smucker and Charles LA Clarke. Time-based calibration of effectiveness measures. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 95–104. ACM, 2012.

[182] Hanna Stelmaszewska and Ann Blandford. Patterns of interactions: user behaviour in response to search results. In *Proceedings of the Joint Conference on Digital Libraries Workshop on Usability*, pages 29–32, 2002.

[183] Cass R Sunstein. Nudges vs. shoves. *Harv. L. Rev. F.*, 127:210, 2013.

[184] A. Sutcliffe. Towards a cognitive theory of information retrieval. *Interacting with Computers*, 10(3):321–351, June 1998.

[185] M.S.L. Swanborn and K. de Glopper. Incidental word learning while reading: A meta-analysis. *Review of Educational Research*, 69(3):261–285, 1999.

[186] Rohail Syed and Kevyn Collins-Thompson. Retrieval algorithms optimized for human learning. In *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval*, pages 555–564, 2017.

[187] Jaime Teevan, Kevyn Collins-Thompson, Ryen W White, and Susan Dumais. Slow search. *Communications of the ACM*, 57(8):36–38, 2014.

[188] Jaime Teevan, Kevyn Collins-Thompson, Ryen W White, Susan T Dumais, and Yubin Kim. Slow search: Information retrieval without time constraints. In *Proceedings of HCIR 2013*, page 1. ACM, 2013.

[189] Andrew Thatcher. Information-seeking behaviours and cognitive search strategies in different search tasks on the WWW. *International Journal of Industrial Ergonomics*, 36(12):1055–1068, 2006.

[190] Andrea Lockerd Thomaz. *Socially guided machine learning*. PhD thesis, Massachusetts Institute of Technology, 2006.

[191] Andrea Lockerd Thomaz and Cynthia Breazeal. Socially guided machine learning: Designing an algorithm to learn from real-time human interaction. In *NIPS 2005 workshop on Robot Learning in Unstructured Environments*, 2005.

[192] Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*, 2022.

[193] Peter M Todd, Yvonne Rogers, and Stephen J Payne. Nudging the trolley in the supermarket: How to deliver the right information to shoppers. *International Journal of Mobile Human Computer Interaction (IJMHCI)*, 3(2):20–34, 2011.

[194] Dimitrios Tsekouras, Ting Li, and Izak Benbasat. Scratch my back and i'll scratch yours: The impact of the interaction between user effort and recommendation agent effort on perceived recommendation agent quality. *Available at SSRN 3258053*, 2018.

[195] Phil Turner and Emilia Sobolewska. Mental models, magical thinking, and individual differences. *Human Technology*, 5(1):90–113, 2009.

[196] Amos Tversky and Daniel Kahneman. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, 5(4):297–323, 1992.

[197] Kazutoshi Umemoto, Takehiro Yamamoto, and Katsumi Tanaka. Scentbar: A query suggestion interface visualizing the amount of missed relevant information for intrinsically diverse search. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, pages 405–414, New York, NY, USA, 2016. ACM.

[198] Prajna Upadhyay and Maya Ramanath. Preface: Faceted retrieval of prerequisites using domain-specific knowledge bases. In Jeff Z. Pan, Valentina Tamma, Claudia d'Amato, Krzysztof Janowicz, Bo Fu, Axel Polleres, Oshani Seneviratne, and Lalana

Kagal, editors, *The Semantic Web – ISWC 2020*, pages 601–618, Cham, 2020. Springer International Publishing.

[199] Saúl Vargas and Pablo Castells. Rank and relevance in novelty and diversity metrics for recommender systems. In *Proceedings of the fifth ACM conference on Recommender systems*, pages 109–116. ACM, 2011.

[200] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. *Advances in neural information processing systems*, 30, 2017.

[201] Suzan Verberne, Maya Sappelli, Kalervo Järvelin, and Wessel Kraaij. User simulations for interactive search: Evaluating personalized query suggestion. In *European Conference on Information Retrieval*, pages 678–690. Springer, 2015.

[202] Jun Wang and Jianhan Zhu. Portfolio theory of information retrieval. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 115–122. ACM, 2009.

[203] Shuting Wang, Chen Liang, Zhaohui Wu, Kyle Williams, Bart Pursel, Benjamin Brautigam, Sherwyn Saul, Hannah Williams, Kyle Bowen, and C Lee Giles. Concept hierarchy extraction from textbooks. In *Proceedings of the 2015 ACM Symposium on Document Engineering*, pages 147–156, 2015.

[204] Xuerui Wang, Wei Li, Ying Cui, Ruofei Zhang, and Jianchang Mao. Click-through rate estimation for rare events in online advertising. *Online Multimedia Advertising: Techniques and Technologies*, pages 1–12, 2010.

[205] Rick Wash. Folk models of home computer security. In *Proceedings of the Sixth Symposium on Usable Privacy and Security*, SOUPS '10, pages 11:1–11:16, New York, NY, USA, 2010. ACM.

[206] Ingmar Weber and Carlos Castillo. The demographics of web search. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 523–530. ACM, 2010.

[207] Joseph Weizenbaum. ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45, 1966.

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

[209] Ryen W White. *Interactions with search systems*. Cambridge University Press, 2016.

[210] Ryen W. White, Wei Chu, Ahmed Hassan, Xiaodong He, Yang Song, and Hongning Wang. Enhancing personalized search by mining and modeling task behavior. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pages 1411–1420, New York, NY, USA, 2013. ACM.

[211] Ryen W White and Steven M Drucker. Investigating behavioral variability in web search. In *Proceedings of the 16th international conference on World Wide Web*, pages 21–30. ACM, 2007.

[212] Ryen W White and Eric Horvitz. Belief dynamics and biases in web search. *ACM Transactions on Information Systems (TOIS)*, 33(4):1–46, 2015.

[213] Ryen W White, Gary Marchionini, and Gheorghe Muresan. Evaluating exploratory search systems: Introduction to special topic issue of information processing and management. *Information Processing & Management*, 44(2):433–436, 2008.

[214] Ryen W. White, Ian Ruthven, Joemon M. Jose, and C. J. Van Rijsbergen. Evaluating implicit feedback models using searcher simulations. *ACM Trans. Inf. Syst.*, 23(3):325–361, July 2005.

[215] Matteo Wong. AI search is a disaster. https://www.theatlantic.com/technology/archive/2023/02/google-microsoft-search-engine-chatbots-unreliability/673081/, Feb 2023.

[216] Tingxin Yan, Vikas Kumar, and Deepak Ganesan. Crowdsearch: exploiting crowds for accurate real-time image search on mobile phones. In *Proceedings of the 8th international conference on Mobile systems, applications, and services*, pages 77–90. ACM, 2010.

[217] Grace Hui Yang, Marc Sloan, and Jun Wang. Dynamic information retrieval modeling. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 8(3):1–144, 2016.

[218] Michael Yeomans, Anuj Shah, Sendhil Mullainathan, and Jon Kleinberg. Making sense of recommendations. *Journal of Behavioral Decision Making*, 2017.

[219] Emine Yilmaz, Manisha Verma, Nick Craswell, Filip Radlinski, and Peter Bailey. Relevance and effort: an analysis of document utility. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 91–100. ACM, 2014.

[220] Richard M. Young. The machine inside the machine: users' models of pocket calculators. *International Journal of Man-Machine Studies*, 15(1):51 – 85, 1981.

[221] Ran Yu, Ujwal Gadiraju, Peter Holtz, Markus Rokicki, Philipp Kemkes, and Stefan Dietze. Predicting user knowledge gain in informational search sessions. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, pages 75–84, New York, NY, USA, 2018. Association for Computing Machinery.

[222] Lisl Zach. When is "enough" enough? modeling the information-seeking and stopping behavior of senior arts administrators. *Journal of the American Society for Information Science and Technology*, 56(1):23–35, 2005.

[223] Vasilios Zarikas. Modeling decisions under uncertainty in adaptive user interfaces. *Universal Access in the Information Society*, 6(1):87–101, 2007.

[224] Xiangmin Zhang and Mark Chignell. Assessment of the effects of user characteristics on mental models of information retrieval systems. *Journal of the American Society for Information Science and Technology*, 52(6):445–459, 2001.

[225] Yan Zhang. The influence of mental models on undergraduate students' searching behavior on the web. *Information Processing & Management*, 44(3):1330–1345, 2008.

[226] Yan Zhang. Undergraduate students' mental models of the web as an information retrieval system. *Journal of the American Society for Information Science and Technology*, 59(13):2087–2098, 2008.

[227] Yan Zhang. Dimensions and elements of people's mental models of an information-rich web space. *Journal of the American Society for Information Science and Technology*, 61(11):2206–2218, 2010.

[228] Yan Zhang. The impact of task complexity on people's mental models of medlineplus. *Information Processing & Management*, 48(1):107–119, 2012.

[229] Ce Zhou, Qian Li, Chen Li, Jun Yu, Yixin Liu, Guangjing Wang, Kai Zhang, Cheng Ji, Qiben Yan, Lifang He, et al. A comprehensive survey on pretrained foundation models: A history from BERT to ChatGPT. *arXiv preprint arXiv:2302.09419*, 2023.

[230] Xiaojin Zhu. Machine teaching: An inverse problem to machine learning and an approach toward optimal education. In *AAAI*, pages 4083–4087, 2015.