Beyond Ranking: Integrating User Interaction and Implicit Feedback in Optimising Search Engine Result Pages

PhD Thesis

Kanaad Pathak

iSchool Research Group

Computer and Information Science

University of Strathclyde, Glasgow

July 2, 2024

This thesis is the result of the author's original research. It has been composed by the author and has not been previously submitted for examination which has led to the award of a degree.

Signed: 

Date: July 2, 2024

Abstract

The process of searching for information through an information retrieval (IR) system is intrinsically interactive, involving users in a series of actions such as formulating queries and evaluating numerous result snippets and their corresponding documents to ascertain their relevance. These interactions, which can be regarded as costs, underscore a notable deficiency in conventional IR systems. The Probability Ranking Principle (PRP) states that ranking documents in decreasing order of relevance with respect to the user's query is the optimal way to maximise their expected utility from the results. However, the PRP fails to account for the nuances of result presentation and the inherent costs associated with interacting with the search engine results page (SERP).

Acknowledging the diversity in users' preferences, which can range from closely aligned to significantly divergent, poses a challenge in optimising the display of relevant results on a page to accommodate these varied inclinations. Different users will prefer distinct types of result presentations, and the layout of the result pages significantly influences their ability to interact with the system, discover relevant information, and, consequently, their overall satisfaction. This variability necessitates a nuanced approach to designing IR systems, one that goes beyond traditional ranking methods to consider the individualised ways users engage with and perceive the utility of search results. To address these challenges, the interactive probability ranking principle (iPRP), implemented via the Card Model offers a robust theoretical framework within the interactive IR space, enabling us to model the user interaction process while accounting for constraints such as presentation and screen space.

By incorporating users' implicit feedback, it is possible to assess these costs and pref-

erences towards certain result items. This assessment can then inform the calculation of the Expected *Perceived* Utility (EPU), thus offering a more nuanced understanding of user interaction with IR systems.

This thesis builds upon the card model, by expanding it to estimate the costs associated with user interactions in terms of time and to re-rank search engine results pages (SERPs), thereby raising three compelling questions; (1) how does the operationalisation of EPU affect result ranking, (2) what is the relationship between system-sided and user-side search costs and lastly (2) how do we optimise the presentation and what is its impact on user satisfaction. Each of these questions is explored through user studies centred on a news search task, designed to understand user preferences, presentation effects, and optimisation strategies.

Our first study operationalises the notion of EPU and examines the impact of different result presentation formats, revealing how presentation significantly influences user perception through metrics such as time spent and clicks, and how we can re-rank results beyond traditional ranking paradigms like the PRP. We find that changing the presentation of results significantly impacts system-side metrics such as DCG, RBO and TBG. In our second study, we uncover the dynamics of interaction costs, result presentation, and user satisfaction at both the query and session levels. Our findings indicate that while at the query level, user satisfaction is predominantly influenced by performance metrics such as nDCG rather than presentation, at the session level, satisfaction emerges from a complex interplay of factors, delineating a non-linear relationship with the presentation.

In the concluding study, we propose and evaluate a novel optimisation technique that synchronises ranking with presentation, tailored to individual user preferences. Although presentation optimisations lead to several behavioural changes in user interactions, they do not consistently align with user-reported satisfaction metrics, highlighting a subtle yet crucial gap between objective system enhancements and subjective user experience.

In this thesis, we operationalise and empirically validate the iPRP. The findings from this thesis advocate for a shift in the design of future interactive IR systems,

emphasising the need to personalise and dynamically display search results to enhance user experience. This research not only lays the foundational work for further exploration but also paves the way for validating the universal applicability of EPU-based result ranking across various interactive IR platforms and user demographics, thereby setting the stage for future studies in this vital area.

Acknowledgements

The acknowledgements section in one's thesis is, in my opinion, as important as the rest of the thesis. The path to success is not linear and nobody is successful all on their own. Therefore, rather unusually this acknowledgements section is rather long, feel free to skip it but I encourage you to read it. I feel that the people that have brought me so far deserve some recognition for my success.

I began my PhD right in the middle of the pandemic (COVID-19), which one might admit is a rather odd time to embark on such a journey. The lockdowns were tough, so firstly I would like to thank my parents for their love and affection and for always believing in me. I could not have done this without your support. So this one's for you mumma and baba, and not to forget my beloved grandmothers who I wished would be able to see this day.

In 2022 my (now) Fiancé moved to the UK, admittedly not near Glasgow, but I was super happy just to be near her. Smriti, you have been there for me through the highs and lows of the PhD (incl. all those paper rejections) and you stuck with me through thick and thin. Thank you. I love you.

Also, thank you to all my friends (Hari, Umair, Amogh, Ayah and Sophia) for all the good times along the way. I must also not forget those who pushed me to initially in my career, big thanks to Prakash uncle and my superstar mentor from P.E.S University, Arti Arya.

Well, obviously a PhD is simply impossible without a fantastic advisor. I am happy to say I had the best. Thank you, Leif, Martin and Michail. Thank you for investing in me and always pushing me to do better. Thank you for never giving up on me when I doubted myself the most. The learning in this PhD has been one of the most rewarding

experiences in my life, and I hope I was a good student to you.

Last but not least, thank you to all those involved in project DoSSIER and thank you to the EU for providing the funding for this PhD. So this one is dedicated to all of you, for believing in me, for pushing me, for never leaving my side. Thank you.

Dedicated to Family and Friends, I love you all.

Contents

List of Figures

List of Tables

List of Tables

Chapter 1

Introduction

The current era, often referred to as the *information age*, is defined by the swift transmission and dissemination of information via technology. A significant factor facilitating this rapid technological transformation is the ability to digitally store and transmit information across the world, enhancing accessibility. Central to this evolution has been the inception of the World Wide Web [3]. In the early days of the web, navigation was primarily via hyperlinks on websites or through directly accessing specific web pages. However, with the exponential increase in information and widespread adoption of computers, it became challenging to identify the specific functions of various websites. The emergence of search engines provided a solution, allowing users to sift through the vast amount of online content.

In today's information age, when users recognise a knowledge gap or need to verify facts, they experience what is termed as the anomalous state of knowledge (ASK) [4]. This perceived ***information need*** prompts users to articulate their requirements in the form of a ***query***. Typically, to address these queries, users resort to inputting their queries into search boxes on modern search engines, such as those developed by Google, Microsoft, and Duck Duck Go. Upon receiving a user's query, these search engines, which are essentially ***Information Retrieval*** (IR) systems, aim to retrieve web ***documents*** that are ***relevant*** to the expressed need. Typically, these results are ***ranked*** in decreasing order of relevance on a Search Engine Result Page (SERP). Users then proceed to evaluate these documents for their relevance to the information need

and gain some utility for each result they assess.

The interplay between users and the results involves multiple interactions, and understanding these dynamics is the focus of interactive information retrieval (IIR) [5, 6]. Within the scope of IIR, "interaction" extends to how users engage with the results presented on Search Engine Result Pages (SERPs). With the diversification of information modalities, the representation of results on SERPs has evolved, and this variation in quantity and style of presentation markedly impacts user interactions and satisfaction. Optimising how results are displayed on SERPs, therefore, emerges as a key factor in enhancing user experience, facilitating more rapid and effective retrieval of information.

Historically, research in IR and IIR has often relegated the presentation of results as a secondary problem, a downstream component following the prioritisation of document relevance and ranking. In this thesis, we argue for the intrinsic value of presentation in the IR process. We posit that the presentation of search results should not only follow decreasing order relevance but should be co-optimised, taking into account user preferences and the spatial constraints of the viewport that display the SERPs.

In this thesis, we address the challenge of this dual optimisation: harmonising the ranking of results with their presentation to maximise the utility gained within a given space on the screen. Through a series of studies, this thesis explores the multifaceted interaction process and how variations in result presentation formats can influence user satisfaction and perceived utility. We present novel methodologies that integrate user preference data into the ranking process, proposing a shift from static result presentation to a dynamic, user-responsive approach.

As the narrative of this research unfolds, we will explore as to why the consideration of result presentation is crucial in IIR and how an informed understanding of user preferences can aid in more user-centric IR systems.

1.1 Main Motivation and Context

Central to the evaluation of all modern IR research is the Cranfield paradigm [7]. The Cranfield paradigm describes a methodology which is characterised by the use

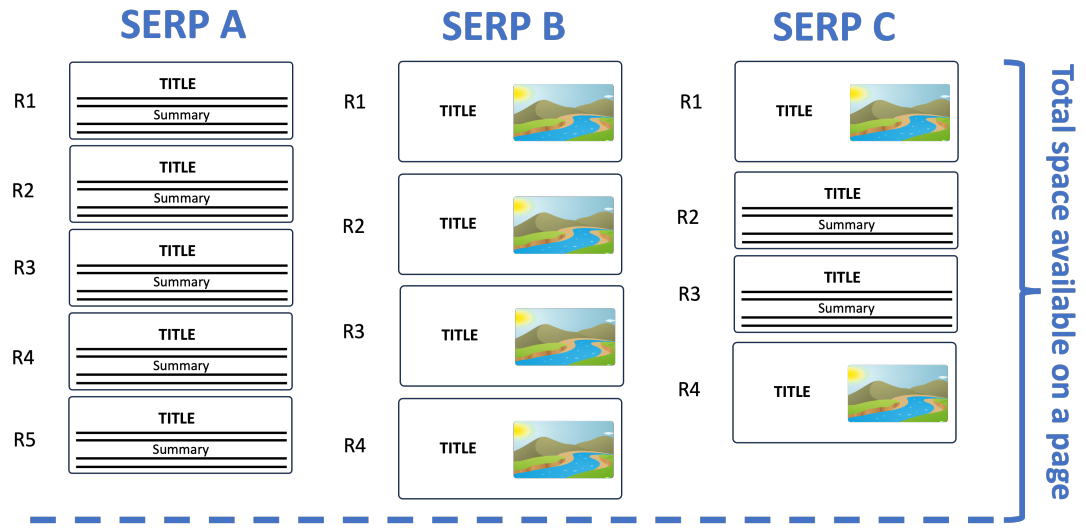


Figure 1.1: Example of displaying results in different formats on a single SERP. A SERP can homogeneously present results, as shown in SERP A and B or a heterogeneous combination of the result cards as shown in SERP C.

of a standardised test collection – which is a corpus of standard documents used in IR experimentation. While the paradigm has remained largely consistent since its establishment in the 1960s, it has evolved to incorporate new tasks and complexities.

The Cranfield paradigm abstracts the interactive retrieval process, simplifying the complex interactions between the searcher and the system. The prevailing experimental paradigms that are derived from Cranfield-style tests make assumptions that diverge from actual user behaviour during information search and interaction with results. These paradigms typically presume that the user will...

- ... issue a single query
- ... evaluate a fixed number of results (typically 1000 results in TREC evaluations) and,
- ... assess each result thoroughly.

While this approach offers simplicity, it fails to capture the nuanced interactions inherent in the interactive information retrieval process (IIR). Modern SERPs are complex; they contain results presented in different formats.

For example, Figure 1.1 shows three different ways of presenting results on a SERP. In the context of a news search task, these results may be presented with some combination of a title (headline), image or summary (lead sentences of article) on a result card.

Different presentation formats can lead users to interact with results in different ways. For example, while one user may prefer to view results as they are displayed in SERP A, another may prefer to view them as they are displayed in SERP B. Interacting with results in these different SERPs naturally leads to different interaction probabilities and processing times. It is generally observed that searchers aim to minimise their effort, seeking to reduce the expected rate of work over time, thus emphasising the importance of engaging result presentations to maximise information gain [8]. Therefore, engagingly presenting results so that the user can maximise information gain while expending minimal effort is crucial to an optimised interaction process. The optimal SERP may be a heterogeneous combination of the types of result cards shown in SERP C.

Presenting results in these formats means that all results cannot possibly be presented on a single SERP (unless the SERP contains the option to endlessly keep scrolling, or is paginated). It also means that users will most likely not browse thousands of results. Therefore, the number of results a user may assess per page is a function of not only the quality and quantity of the results but also the presentation of the result and the space it occupies. Changing the presentation of the results on a SERP means that the expected *perceived* utility a user derives from the result can change. Thus, the ranking of results is also likely to be affected as a result of this change. Traditional IR models do not account for the nuances in this interaction and therefore a better user model is required to describe the user-system interaction so that we can optimise the ranking of results taking into account the presentation.

From Optimising the Ranking to Optimising the Result List

Before looking at the optimisation of result lists, we must first define what constitutes an "optimal" list. Ideally, an optimal result list would facilitate users in finding relevant

information quickly and accurately, reflecting their informational consumption preferences. Some users, for example, are visual learners who may prefer image-rich result cards, while others may favour textual information cards. Capturing these nuances requires a thorough understanding of user behaviour in IIR, a task that is central to the design of effective search systems. Prevailing ranking principles, starting from the Probability Ranking Principle (PRP) [9] and subsequent developments in the interactive space (iPRP) have made significant strides in enhancing search rankings but often overlook the intricacies of user interaction with broad assumptions on user browsing behaviour [10–15].

These principles are generally predicated on the assumption that users interact with search results linearly and uniformly, which is increasingly misaligned with the dynamic and varied nature of modern search tasks.

Within this context, the concept of Expected Perceived Utility (EPU) from economic search theory emerges as a key construct. EPU measures the anticipated satisfaction a user derives from their search activities against the invested time and effort, which constitute the *benefits* and *costs* of the search process. This balance of benefits and costs is crucial as it heavily influences a user's decision to continue, modify, or stop their search strategies [16–18].

Meanwhile, empirical studies focusing on user interface layouts [17, 19–24] have highlighted the impact of interface design on user satisfaction. Other approaches have looked at direct optimisation of user interfaces [25–32].

However, in this thesis, we turn to the iPRP, a model rooted in a theoretically principled approach. We implement the iPRP via the Card Model. This approach diverges from traditional ranking principles by first conceptualising the search process as a dynamic series of user-system interactions. The Card Model simulates this process through a metaphorical card game, optimising a user's gain of relevant information with minimal effort by incorporating a blend of user actions, context, and constraints. These interactions are facilitated through a series of interface "cards" tailored to optimise the user's gain of relevant information with minimal effort, factoring in a comprehensive blend of user actions, context, and constraints such as screen space

The Interface Card Model builds upon the theoretical underpinnings of economic search theory to provide a framework that estimates these search costs. By incorporating elements such as action models and updating preferences, the model offers a fluid "sequential interaction" scheme that more accurately mirrors the actual behaviours of users navigating SERPs. While this model has been primarily evaluated analytically in controlled environments, the potential for a practical application in real-world SERPs suggests a promising direction for enhancing user experience and search efficacy on a broader scale.

The importance of optimisation cannot be overstated. Optimised interfaces can enhance user satisfaction by reducing search time and effort but can also potentially change the perceived relevance of the results presented. This, in turn, can lead to a more efficient and satisfying search experience, potentially increasing user trust and reliance on a given IR system.

The primary aim of this thesis is to bridge the gap between theory and practice, moving beyond the theoretical limitations of the Card Model to apply and refine it with a focus on optimising ranked lists in search engine results pages (SERPs) with consideration for presentation aspects. This research endeavours to modify the model to tackle the practical challenges encountered by contemporary search engines, with a special emphasis on the dynamics of user interaction and satisfaction. By applying theoretical insights in a practical context, the thesis aims to present results in a manner that accounts for users' costs and implicit preferences, thereby maximising their search performance and satisfaction.

1.2 High-Level Research Questions

To navigate the problem space within IIR delineated previously, our investigation will be directed by a series of high-level research questions (HL-RQx). These questions are designed to dissect the complex issue of user interface optimisation in IR systems.

The **T**ext **RE**trieval **C**onference (TREC) [33], is a bench-marking organisation renowned for employing the Cranfield evaluation model. The work encapsulated in this thesis is primarily based on data derived from TREC's endeavours, with a particular

focus on the news retrieval track. We use the TREC Washington Post corpus with a focus on the ad-hoc search task to ground our research questions. Below is an elaboration of each research question and its significance within the scope of our research:

HL-RQ1 How does incorporating users' interaction costs and implicit feedback influence the ranking of search results within the Card Model framework?

In addressing our first research question, we explored the impact of different result presentation formats on user perceptions by implementing a user study to measure and validate the operational definition of EPU. Through this empirical approach, we sought to quantify the variation in EPU across various result card presentations, focusing on elements such as click probabilities and the timing of interactions to identify significant differences among the result cards. This investigation served to assess the practicality and applicability of EPU as a metric within user experience contexts.

Our findings revealed notable differences in EPU across different result cards. We further discerned how variations in EPU—stemming from these different costs and preferences—alter the ranking of results, as evaluated by metrics such as Rank-Biased Overlap (RBO) and its subsequent effect on performance measures like Discounted Cumulative Gain (DCG) and Time-Biased Gain (TBG).

Given that EPU impacts these user-side performance metrics, our further exploration with our second high-level research question examines whether this impacts user satisfaction.

HL-RQ2 What is the influence of presentation style and performance metrics on user satisfaction during news search tasks?

To address the second research question, we adapted a methodology to elucidate the differentials in user satisfaction. From our first research question, we primarily collected annotation data. However, for this subsequent inquiry, we transitioned to situating the user within a more authentic search scenario. This approach permitted users to formulate queries, navigate through SERPs, and identify relevant documents. Our investigation uncovered that modifications in the presentation of results notably

influenced both user-side and system-side metrics. Crucially, we detected a direct correlation between query performance and user satisfaction. This discovery intimates the existence of an optimal method for presenting information that maximises user satisfaction. Consequently, the next high-level research question will explore the nature of this optimisation across different users and ascertain their perception of it, employing a comprehensive user study to gather empirical evidence.

HL-RQ3 In what ways does the optimisation impact user satisfaction and their ability to identify relevant information efficiently?

In addressing our third research question, we employed the methodology established in the second research question to investigate user perceptions of optimisations in ranked lists. This exploration aimed to discern whether optimisations, defined by either the rate of utility gained or the total utility gained, are perceptible to users in their interactions with SERPs. To validate our approach, we developed optimisers based on these criteria and anchored our findings in empirical data derived from a comprehensive user study. Our findings reveal that while these optimisers indeed influence user behaviour, notably altering the manner in which users navigate SERPs and evaluate documents, these modifications do not translate into a perceptible change in user experience. In essence, although the optimisers effectively refine the interface to theoretically enhance user interaction, users themselves do not consciously recognise these improvements in their search experiences.

These overarching questions guided the structure of our investigation. They are broad by design, to encapsulate the wide range of variables in play. Naturally, each high-level question begets a series of more granular, specific inquiries. These sub-questions will be addressed in dedicated chapters, providing a detailed exploration of each facet of the research.

1.3 Thesis Statement

The primary aim of this thesis is to bridge the gap between theoretical models and practical applications in IR by extending the Card Model for optimising the presen-

tation of SERPs. We thus explore the incorporation of user interaction costs and implicit feedback into the ranking process and also investigate their effects on user satisfaction and search performance. Through a series of empirical studies designed to capture interaction costs and understand user behaviour, we argue for the intrinsic value of considering presentation at the time of ranking rather than considering it to be a downstream component.

1.4 Main contribution of this thesis

In this thesis, we make substantial theoretical, methodological, and empirical contributions to the operationalisation of the EPU function within the Card Model framework and its practical implementation. We break these down by each of our HL-RQx

- **HL-RQ1** How does incorporating users' interaction costs and implicit feedback influence the ranking of search results within the Card Model framework?
  - **Methodological**
    - We describe a strategy for collecting user interaction data to operationalise EPU. This entails a detailed methodology for collecting and analysing data in ad-hoc news search tasks, facilitating the acquisition of necessary timing components to compute EPU.
  - **Empirical**
    - **Re-ranking SERPs:** We report on a study that demonstrates the re-ranking of Search Engine Result Pages (SERPs) based on the EPU of result cards and space constraints. This addresses the first high-level research question (**HL-RQ1**), illustrating the practical application of EPU in SERP optimisation.
- **HL-RQ2** What is the influence of presentation style and performance metrics on user satisfaction during news search tasks?
  - **Methodological**

    * We present a methodology to collect user interaction data for a user study which involves issuing queries, browsing SERPs and marking documents.

  - **Empirical**

    * **User Interaction with SERPs:** Further methodological advancements are showcased through a user study aimed at elucidating the second research question (**HL-RQ2**). This study reveals diverse user behaviours in interacting with SERPs, such as clicking and browsing patterns. We also shed light on the influence of presentation and performance on user satisfaction.

- **HL-RQ3** In what ways does the optimisation impact user satisfaction and their ability to identify relevant information efficiently?

  - **Theoretical**

    * We introduce an algorithm that extends existing work to accommodate the unique constraints of SERP presentation. This algorithm allows for the simulation of user behaviour to assess the effectiveness of different SERP configurations.

  - **Methodological**

    * We present a novel method for visualising and clustering user preferences based on the SERP layouts they encounter to answer **HL-RQ3**. This approach provides deep insights into behavioural shifts and interface satisfaction, contributing to a better understanding of user engagement with optimised search interfaces.

  - **Empirical**

    * We provide empirical findings from a user study used to find how different interface optimisations are perceived by users.

1.5 Structure of the Thesis

This thesis is organised as follows:

- In Chapter 2, the foundation of information retrieval is laid out, tracing its evolution towards IIR and exploring the diverse modelling methods employed in ranking retrieved results. A gap in the literature, which this thesis aims to address through subsequent investigations, is identified.
- In Chapter 3 we outline a general methodology adopted to implement the card model, including the estimation of its parameters. It details a methodological framework upon which our analysis is based, describing the data collection mechanism, the types of data utilised, storage and retrieval processes, and the experimental system architecture.
- In Chapter 4, we begin our investigation with the first high-level research question (**HL-RQ1**), examining how EPU varies across different result cards and the operationalisation of this concept.
- Chapter 5 advances into the second high-level research question (**HL-RQ2**), presenting findings from experiments aimed at understanding the influence of presentation and performance on user satisfaction.
- Chapter 6 discusses how to optimise the ranked list to include user preferences, enabling the re-ranking of items based on their EPU, which answers the third high-level research question (**HL-RQ3**).
- Finally, Chapter 7 concludes the thesis, summarising the findings, discussing the limitations of the methodologies employed, and suggesting directions for future research.

1.6 Publications

Parts of this thesis have been published at the following peer-reviewed conferences:

- Pathak, K., Azzopardi, L., Halvey, M. (2024). Ranking Heterogeneous Search Result Pages Using the Interactive Probability Ranking Principle. In: Goharian, N., et al. Advances in Information Retrieval. ECIR 2024. Lecture Notes in Computer Science, vol 14609. Springer, Cham. https://doi.org/10.1007/978-3-031-56060-6_7

- Pathak, K., Azzopardi, L., and Halvey, M., "The Influence of Presentation and Performance on User Satisfaction," in Proceedings of the 2024 ACM SIGIR Conference on Human Information Interaction and Retrieval, New York, NY, USA: ACM, 2024, pp. 77-86. doi: 10.1145/3627508.3638335.

Chapter 2

Background

The rapid growth in computing technology and increased global connectivity have significantly altered the way we access and share information. Today, millions of people can find and locate all kinds of information with ease from the comfort of their homes. This transformation is largely due to advancements in IR systems like Google Search, which have made finding information much more straightforward. In contrast to the past, where searching for information often involved visiting a library and looking through books, modern IR systems quickly provide relevant results, much like presenting a sorted list of what we need.

As the internet has evolved, so have the methods we use for information retrieval. Initially, the information online was presented in a straightforward manner – just text links on a web page. However, as the variety and modality of online content have increased, the way people interact with this information has also changed, leading to the development of interactive IR systems (commonly referred to as IIR). IIR focuses on how users interact with the information presented.

In this chapter, we will explore the evolution of IR systems, from their basic forms to more advanced versions that consider user interaction. We will examine how these systems have approached the task of ranking results. Our discussion will particularly highlight the card model and suggest improvements to better address our research questions. This chapter sets the foundation for our work, focusing on improving how results are presented and ranked in IR systems.

2.1 Historical Overview of IR

2.1.1 From Paper to Silicon

The pursuit of information or knowledge, whether it be through books or scriptures is a longstanding endeavour, deeply rooted in human history. In this section we will trace the advancements from early systems to the sophisticated IR systems of today, to set our work in a better context.

Historically, libraries have served as fundamental repositories of written knowledge. Their existence underscores the perennial need to organise and store vast quantities of textual content. A notable example of such an organisation is the Dewey Decimal System, an indexing method categorising information based on thematic relevance. This system represented an early attempt to streamline the retrieval of information [34].

Nevertheless, the process of retrieving information through these conventional methods was notably slow. To address this, mechanised technologies were developed to expedite the search process within these indexes, marking the nascent stages of what we now recognise as information retrieval.

The transition from mechanical (analogue) to digital methods was heralded by the invention of the transistor. Transistors were used to enable digital logic on a computer to make computations faster. As time progressed, these devices reduced their form factor and revolutionised computational power, thereby enhancing information storage and retrieval. The exponential growth in computing power, encapsulated by Moore's Law, has been a driving force in this evolution. Moore's Law observes that the density of transistors on microchips doubles approximately every two years, concurrently reducing the cost of computing [35]. This surge in processing power and storage capacity has enabled the cataloguing and retrieval of ever-growing information repositories with unprecedented efficiency.

2.1.2 The World Wide Web

A significant milestone in the evolution of IR was the emergence of The Web. The Web facilitated a global networking platform, dramatically altering the landscape of

information sharing and retrieval.

The web introduced a novel form of IR system, beginning with the early JumpStation system. This system utilised information on web pages such as anchor text and hyperlinks for document ranking [36]. However, the growing volume and diversity of online content necessitated a departure from such rudimentary methods. Previous retrieval models often overlooked the presentation aspect of search results, a gap that contemporary IR research aims to fill.

The evolution from physical libraries to the digital realm illustrates a critical shift in how information is accessed, stored, and retrieved. This historical perspective allows us to understand the complexities and challenges of modern IR systems. As we transition from the historical context to the present, it is important to examine the foundational elements of IR systems. This includes exploring how documents are stored in digital formats, the algorithms employed for efficient retrieval and ranking, and how these systems have adapted to the ever-changing landscape of digital information.

Accordingly, our next focus will be on IR fundamentals, starting with the architecture of digital storage and retrieval systems. We will explore the mechanisms that allow for the rapid processing of queries, the algorithms that determine the relevance of documents, and how these elements coalesce to present users with an efficient and ranked list of information. This examination will also consider the evolution of user interaction with these systems, acknowledging the shift from passive retrieval to more dynamic and interactive models.

2.2 But what is IR?

An IR system is designed to find and provide information that matches a user's search query. The effectiveness of these systems lies in their ability to return relevant results, ideally sorted from most relevant to least relevant, a concept pioneered by Luhn in 1957 [37] and later formalised by [9] as the Probability Ranking Principle (PRP) in 1977.

The PRP provided a solid formal proof based in probability theory to rank documents in decreasing order of relevance. IR systems primarily deal with retrieving and

presenting *documents*. The term "document" in IR can refer to a wide range of data objects, including text documents, images, audio, and videos. Document retrieval, a sub-field of IR, is defined as the matching of a user's query against a set of free-text records. These records, which constitute "documents" in this context, can be unstructured text such as newspaper articles, real estate records, or paragraphs in a manual. The user queries can range from multi-sentence full descriptions to just a few words.

A key function of an IR system is the ability to find and match relevant documents for a given query through the vast array of indexed documents. Elasticsearch, Lemur, Lucene, and the Terrier IR Platform are notable examples. The most popular of these is the engine from Okapi, with the BM25 matching algorithm. We will get into more detail about this algorithm later on in this chapter. BM25 is a commonly used baseline to benchmark IR engines.

The architecture of IR systems is underpinned by three essential components: the index, where documents are systematically catalogued; the retrieval model, which evaluates and scores documents for relevance; and the query, the user's input-seeking information. These components interact and influence each other, as illustrated in Figure 2.1, to produce a ranked list of documents most relevant to the user's query.

In the operational core of these systems is the matching process, a key mechanism where the retrieval model sifts through the index to align documents with the query. In this matching process, the retrieval model tries to retrieve relevant documents from the document index, which was generated from the indexing process. These components are typically associated with being system-sided aspects of the retrieval system

Examples of system-side costs incurred during this process include computational demands for indexing and retrieving documents. On the user side, costs may involve the cognitive and temporal investments in formulating queries and interpreting results and the *costs* associated with these components essentially being utilised are called system-side costs,

IIR scrutinises the user-side costs by analysing how individuals engage with the IR system's outputs. This analysis includes examining the duration users spend on results and their interactions, such as clicking, scrolling, and saving documents, which informs

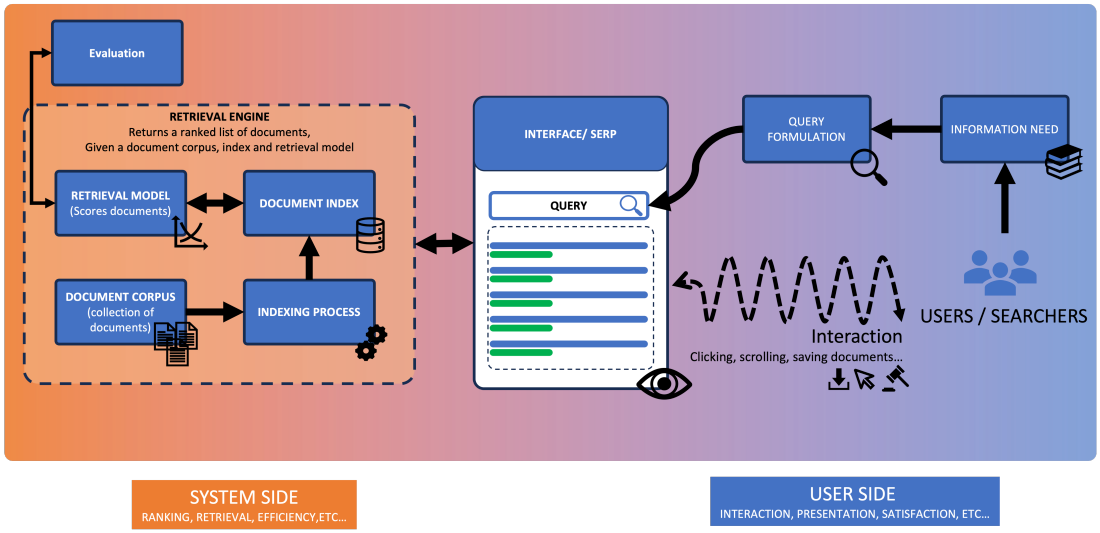


Figure 2.1: Key components and processes within an IR system contrasting system-side costs such, as indexing, retrieval and ranking; and user-side costs such as query formulation, interaction costs etc., The user's journey from recognising a need for information to interacting with the search results is outlined on the spectrum, with the SERP incurring both user-side and system-side costs.

on user satisfaction with the system.

To gain a comprehensive understanding of IR systems and how the system-side costs are perceived by users, one must first examine the system-side components. These foundational elements set the stage for a deeper investigation into the user-side elements, as the interplay between the two profoundly influences the overall search experience and system effectiveness.

So now, we will look into the system-sided costs and explain how different retrieval models use this for matching queries to relevant documents.

2.2.1 Retrieval Models

Given a document index (see §A.1 for more detail on how this is created), the subsequent step involves retrieval or matching the contents of the index with a user's query. Each retrieval strategy uses a specific model to represent documents. These models can be categorised based on their mathematical foundations and characteristics.

The creation of a document index is a preliminary step in the process of retrieval or matching the contents of the index with a user's query. In this context, a system with

access to a collection of documents, denoted as $\boldsymbol{D}$, aims to fulfil a user's information need. This need is communicated to the system in the form of a query $\boldsymbol{q_i}$. Mathematically, the matching function of a retrieval model can be conceptualised as employing a similarity function $\boldsymbol{sim}(d_i \mid q_i)$. This function matches the user's information need to a relevant document within the collection. A document $\boldsymbol{d_i}$ is considered relevant if it contains information that is valuable to the user who issued the query.

Broadly, these models fall into three categories: set-theoretic, algebraic, and probabilistic. These models then employ matching functions in different ways to match the query to a relevant document. Each category represents a step in the evolution of retrieval models, from simpler to more complex systems used today. However, this classification is not all-encompassing. Recent advancements in IR are centred around language models, which calculate a probability distribution over words. Additionally, current research is also focusing on neural approaches for ranking documents, although these do not consider user interactions and are solely concentrated on ranking. This thesis does not look into neural models, but it acknowledges that such models are already an active area of research.

Probabilistic Models

Central to this thesis are models derived from the probabilistic model, which has evolved from from set-theoretic and algebraic models. More details on the the evolution toward the probabilistic models can be found in §A.2

Probabilistic models in IR estimate the likelihood that a document is relevant to a given query, similar to vector space models. This thesis centres on probabilistic models that are grounded in a fundamental ranking principle, known as the Probability Ranking Principle (PRP) [9].

The PRP posits that an optimal ranking of documents is achieved by ordering them in descending probability of relevance or usefulness to the user's query. Usefulness is determined by the relevance of a retrieved document to the issued query. For instance, given two documents d_i and d_j, and if d_i is more relevant than d_j, we associate a cost C for retrieving a relevant document C_r and for a non-relevant one $C_{\bar{r}}$. Ideally, the

cost of retrieving a relevant document is less than that of a non-relevant document. Following this logic, the PRP can be articulated through a cost equation, leading to a ranking of documents by decreasing order of relevance probability:

$$\begin{aligned} P(d_i|r)\ C_r + (1 - P(d_i|r))\ C_{\bar{r}} &< P(d_j|r)\ C_r + (1 - P(d_j|r))\ C_{\bar{r}} \\ P(d_i|r)\ (C_r - C_{\bar{r}}) &< P(d_j|r)\ (C_r - C\bar{r}) \\ P(d_i|r) &> P(d_j|r) \end{aligned} \tag{2.1}$$

A straightforward implementation of the PRP is the Binary Independence Model (BIM). This model, introduced by [38, 39], ranks documents based on the odds of relevance, which is the division of the probability of relevance by the probability of non-relevance. It represents documents and queries as binary vectors, assuming statistical independence among terms. Accordingly, a document is represented as a product of term probabilities, and the model stipulates that terms not present in the query have equal frequencies in relevant and non-relevant documents.

2-Poisson Model In another approach, [40] presents a model that distinguishes the most informative terms of a document. This is based on two Poisson distributions and requires three parameters for each term in the vocabulary. While this model circumvents the need for a term weighting algorithm, estimating the parameters remains a challenge. Ranking is achieved by using a measure derived from the means of the Poisson distributions.

BM25 Deriving from both the 2-Poisson model and BIM is BM25, also known as "Okapi BM25" [41]. This model assesses a term's informativeness and a document's relevance by using term-frequency and inverse-term frequency. It has proven to be effective and popular in IR, despite not accounting for the inter-relationship between query terms within documents.

The BM25 scoring function is:

$$\text{sim}(d_i \mid q_i) = \sum_{w \in q_i} \text{IDF}\,(w, D) \cdot \frac{TF\,(w, d_i)\,(k_1 + 1)}{TF\,(w, d_i) + k_1 \left(1 - \beta + \beta \cdot \frac{|d_i|}{\text{avgdl}}\right)} \tag{2.2}$$

Where k_1 and β are free parameters related to the query and the collection that are often tuned on a training dataset. k_1 and β are usually set to 1.2 and 0.75 respectively[1]. Also, $|d|$ is the length of the document d_i measured in words, and $avgdl$ is the average document length in the text collection. Finally, the IDF(w) is computed as:

$$\text{IDF}\,(w, D) = \ln\left(\frac{N_D - N_{D_w} + 0.5}{N_{D_w} + 0.5} + 1\right) \tag{2.3}$$

where N_D is the total number of documents in the collection, and N_{D_w} is the number of documents containing the term w. This approach to ranking within probabilistic models embodies a balance between term frequency and document frequency, thereby determining the relevance of documents to a user's query with greater precision.

2.2.2 Evaluation Measures

Remember, that the aim of an IR system is to not only fulfil user's information need via their queries but to do so in a manner that ranks the retrieved results by their relevance, striving for an optimal ordering. This brings us to the question of how to gauge the quality of such ranked lists, especially when introducing the complexity of temporal factors. Though a detailed examination of all evaluation measures is beyond the scope of this thesis, we encourage looking at [42] for a more comprehensive summary of evaluation metrics.

The effectiveness of IR systems is fundamental, particularly in how they align with the user's intended task. Our discussion narrows to ad-hoc search, highlighting metrics that are apt for evaluating such tasks. At the core of our models is the concept of search economics, pointing to the necessity of measures that capture the utility or gain from information within ranked lists.

Cumulative Gain (CG)

We start with Cumulative Gain, a measure that sums the value derived from all relevant documents up to a specific rank, termed CG@k [43]. This measure is versatile, allowing for both a system-centric and user-centric perspective. It incorporates the TREC

[1]Throughout this thesis, all experiments utilising BM25 have used these values by default

relevance judgement scores, which range from 0 (irrelevant) to 2 (highly relevant) in increments of whole numbers, as a basis for valuing documents. These scores are then aggregated to arrive at the CG measure.

Cumulative Gain is mathematically represented as:

$$CG_k = \sum_{i=1}^{k} rel_i \tag{2.4}$$

Here, k signifies the rank position, and rel_i denotes the relevance score of a document at that rank.

However, while this is great, utility does not increase linearly forever, and comparing two ranked lists with different order of relevant documents is difficult. For example, two lists of rankings with one highly relevant document in rank 1 in one list, and rank 5 in the other, will yield identical CG values for CG@5. DCG is predicated on the understanding that the utility of information gained decreases as one progresses through a search result list. Thus, it introduces a mechanism to penalise documents of high relevance appearing lower in the list, adjusting their value logarithmically based on their position.

Discounted Cumulative Gain (DCG)

DCG further refines our evaluation by acknowledging the diminishing returns of relevance with deeper search result penetration. The formula for DCG at a particular rank k is:

$$DCG_k = rel_1 + \sum_{i=2}^{k} \frac{rel_i}{\log_2(i+1)} \tag{2.5}$$

However, the variability in search result list lengths poses a challenge for direct comparisons using DCG alone. This leads us to the concept of normalisation through the Ideal DCG (IDCG), facilitating a relative performance measure across different queries.

Normalised Discounted Cumulative Gain (nDCG)

To ensure fairness in comparison, DCG is normalised, resulting in the nDCG metric. This normalisation process involves calculating the Ideal DCG (IDCG) up to the same rank position k, offering a benchmark for evaluating the actual DCG performance.

Normalised DCG is computed as:

$$nDCG_k = \frac{DCG_k}{IDCG_k} \tag{2.6}$$

This method provides a comprehensive view of a search engine's ranking effectiveness, with perfect ranking algorithms achieving an nDCG of 1.0, denoting impeccable relevance ordering.

As we progress, it's crucial to acknowledge that while DCG and nDCG offer significant insights into ranking quality, they do not account for user engagement time with search results.

Time Biased Gain

Building upon the concept of Discounted Cumulative Gain (DCG), is the Time Biased Gain (TBG), which incorporates user interaction with ranked lists into our evaluation framework [44]. TBG is especially relevant for ad-hoc search tasks, acknowledging that user engagement with information retrieval systems unfolds over time, and not all interactions lead to the same level of gain. Specifically, TBG accounts for the diminishing returns of relevance as users spend more time searching and the likelihood of users engaging with content decreases over time. The user model for TBG assumes users navigate through the ranked list sequentially, one document at a time, which aligns closely with observed user behaviour. Data from Smucker and Jethani's study show that 94% of user interactions are directed towards documents of a lower rank, supporting this sequential navigation model.

The TBG for a ranked list L with a user model U, considering the list's length —L— and the time T(k) it takes a user to process the document at rank k to realise its gain g_k, is defined as follows:

$$TBG(L, U) = \sum_{k=1}^{|L|} g_k \cdot D(T(k)) \tag{2.7}$$

where D(t) represents the survival probability or the likelihood that a user continues to engage with the search results up to time t.

It is defined as follows:

$$D(t) = e^{-t \ln(2)/h} \tag{2.8}$$

where h represents the half-life, the time by which half of the population of users has ceased engaging with the search results. In practical terms, the half-life parameter h indicates the rate at which user engagement decays over time. For instance, if h equals 224 seconds, it suggests that after 224 seconds, the likelihood of a user continuing to interact with the search results drops by 50%.

This measure captures the expected utility a user derives from a ranked list, considering both the relevance of the documents and the user's willingness to continue engaging with the list over time. It provides a nuanced view of IR system effectiveness, factoring in user behaviour and the temporal dimension of information retrieval.

Incorporating TBG into our evaluation allows us to simulate a range of user interactions, from quick glances to deep dives, offering a richer understanding of how different ranking strategies might impact user satisfaction and information discovery in real-world scenarios.

Rank Biased Overlap

Following the exploration of Time Biased Gain, we turn our attention to another crucial measure for IR evaluation: Rank-Biased Overlap (RBO). Given that we are attempting to re-rank retrieved results, we require a measure to compute the differences in the ranked lists. This measure (RBO) focuses on the indefinite nature of rankings, a common scenario in real-world search systems where the full depth of rankings is often not observed or is impractically large for complete evaluation [45].

RBO introduces a novel approach to measuring the overlap between two ranked lists, L_1 and L_2, by applying a convergent series of weights to the proportional overlap at each

depth. This weighting scheme ensures that the "infinite tail" of the list—potentially unobserved portions—does not disproportionately influence the overall similarity score, allowing for a more balanced assessment of rank similarity.

The core concept of RBO is captured in the equation:

$$RBO(L_1, L_2, p) = (1-p)\sum_{d=1}^{\infty} p^{d-1} \cdot A_d \tag{2.9}$$

where A_d represents the agreement between lists L_1 and L_2 at depth d, and p is a parameter controlling the decline in weights, indicating the model's bias towards the top of the ranked lists. A lower p value emphasises the importance of the top-ranked items more strongly, reflecting scenarios where early results carry more weight in user satisfaction and utility.

RBO values range between 0 and 1—with 0 indicating no overlap (completely disjoint lists) and 1 signifying perfect overlap (identical lists)—allows for an intuitive interpretation of the similarity between rankings.

By addressing the challenge of indefinite rankings, RBO provides a tool for evaluating the similarity of ranked lists in a manner that realistically reflects user behaviour and preferences in IR systems. This measure's adaptability and the insightful weighting of overlaps make it a valuable addition to the toolbox for IR evaluation, particularly in settings where the depth of interest varies significantly across contexts or queries.

2.3 IIR

So far, in the preceding sections of this chapter, we presented an overview of the historical evolution of IR systems, highlighting the essential components of the IR process. Furthermore, we looked into a variety of key modelling methodologies that are instrumental in matching documents with user queries and also evaluating resulting ranked lists. It is important to note, however, that these mathematical constructs have not adequately accounted for user interactions within the IR framework. The fundamental aim of an IR system is to effectively satisfy the information needs of its users. This is examined by a distinct research domain that integrates user interaction

elements into the sphere of information retrieval, known as Interactive Information Retrieval (IIR). Before delving into IIR, it is essential to recognise the limitations of current evaluation paradigms, especially the importance of including user interaction in our models. Consequently, we will initially examine well-established experimental frameworks, with a particular focus on the Cranfield paradigm. In the next part of this chapter, we will describe modern SERPs and put into context how previous work has estimated the various costs associated with interacting with their components for ranking documents.

2.3.1 Experimental Paradigms

The methodological framework employed for the evaluation of IR systems is the Cranfield paradigm [46]. This paradigm aims to establish a consistent and controlled environment for evaluating the performance of various components of an IR system in isolation. It stipulates the use of a uniform set of documents and information needs, facilitating the comparative analysis of different systems based on performance metrics such as precision and recall, which are critical for assessing system effectiveness. Central to this paradigm is the concept of a test collection, a standardised set of document sets.

A test collection comprises three fundamental elements: the corpus (the set of documents), the topics (a collection of themes to segregate documents), and a set of relevance judgements. These relevance judgements dictate the set of documents which should ideally be retrieved by the system under evaluation. However, the paradigm operates under several simplifying assumptions. One key assumption is the uniform desirability of all relevant documents, leading to the notion of a static information need, implying that there is no evolution in the information requirement during the search process. Essentially, as the user gains more knowledge from the retrieved set, their information needs will remain the same. Additionally, it presupposes uniformity in information needs across all users, with the relevance judgements assumed to be representative of the broader user population. Consequently, this infers that when specific documents are retrieved, they will be deemed to be relevant by all users across

various systems, indicating a universal applicability of the findings.

The Text REtrieval Conference (TREC), an evaluation forum, originates from the Cranfield paradigm and delineates a variety of tracks to simulate distinct information retrieval tasks [33]. TREC establishes ground truth for documents within a topic through numerically-based relevance labels. In the context of this thesis, the TREC Washington Post corpus is utilised as the document collection [47]. Within this collection, relevance labels are assigned on an integer scale ranging from 0 to 2, where 0 signifies non-relevance and 2 indicates high relevance.

Participation in TREC requires that researchers first index the document collection for the track they are participating in, for example, the TREC Interactive Track. The aim is then to produce run files by subsequently utilising a set of queries to retrieve these documents from their systems. The efficacy of these systems is evaluated using metrics such as precision and recall based on these run files. Generally, this is an automatic process using an evaluation tool provided by TREC itself. However, a significant limitation of this approach is TREC's underlying assumption that all users exhibit uniform search behaviours, which is an overly generalised user model. It presupposes that users typically issue a single query, navigate to a predetermined depth in the search results, and inspect each document in the process.

This assumption implies a certain expectation regarding the presentation of document information. As will be elucidated further in this section, modern SERPs do not typically display full documents for user inspection. Instead, results are often presented in the form of result cards, which users leverage to decide whether to further inspect a document. Consequently, individual users may perceive and interact with this information differently, influencing factors such as interaction click probabilities, the depth of their search, and the number of queries they issue to satisfy their information needs.

These user interactions are inherently complex, and various studies have been conducted to analyse them from the system's perspective by incorporating time-based measures or probability-based click models. Notable among these are approaches like time-biased gain [44], which estimates the utility a user derives at each rank of retrieved results as a function of time and Markov-based models that aim to predict

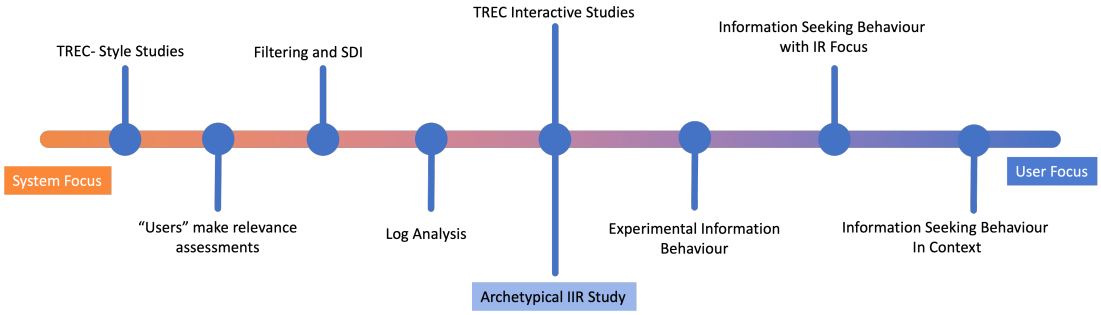


Figure 2.2: The IR / IIR Spectrum as envisioned by [1]

various interaction probabilities. IIR plays a crucial role in deciphering these complex interactions, paving the way towards developing models that account for these nuances based on presentation style. Such models can be utilised to generate ranked lists that are tailored to individual users' browsing preferences. We will now further dive into the IIR process to understand the main components of a SERP and then move on to how system-side and user-side models rank results on these SERPs.

2.3.2 The IR/IIR Spectrum

The essence of IIR lies in its multi-disciplinary approach, integrating insights from Information Retrieval, Library Sciences, Psychology, and Human-Computer Interaction. This convergence aims to deepen our understanding of the interplay between user behaviour and search systems. The focus of IIR extends beyond mere system functionality; it delves into optimising the user's experience and effectiveness in finding relevant information [1, 6].

[1] provides an intuitive spectrum on which IR and IIR are bridged, moving from the left which represents system-focused work to the right which looks at more user-focused work, as seen in Figure 2.2

This thesis examines the significant role of result presentation in ranking SERPs. We focus on understanding how the display of search results affects user interaction with the system and consider the potential implications for changing result rankings and enhancing the effectiveness of the search. Looking at the IR - IIR spectrum, our work focuses in on "Archetypical IIR studies". By grounding our work through a series

of experiments belonging to this category, we can draw more realistic and credible abstractions to describe user behaviour.

To fully grasp the aspects of these studies, it is essential to first understand the IIR process. The IIR process, illustrated in Figure 2.1, starts with a user identifying an information need, leading to a query in a retrieval system. The retrieval system then aims to bring back a set of results that are relevant to the user's query. Advances in modern web frameworks mean that these results can be presented in many different ways on the SERP, deviating away from the standard "10 blue links" of previous IR systems. Therefore, it is inevitable that the way results are presented on the SERP can significantly affect the user's interaction with the system. These interactions often depend on the searcher's intent. Recognising the user's intent is essential for determining when an information need is fulfilled and for designing experiments to evaluate aspects of SERPs effectively. In IIR, there are three types of user intents: navigational, transactional, and informational [48]

Understanding these intents is crucial for creating a theoretical interaction model that optimises the result list based on user interactions. The navigational intent means the user's need for information is typically satisfied by clicking a link or URL. The informational intent, however, involves a more extensive interaction with the SERP. This usually means the user will examine several items by looking at their result summaries and then inspecting them in detail to decide if the results meet their information needs or if they should reformulate the query to obtain different results. As the user interacts with the results, their understanding and knowledge also evolve.

Users generally allocate different portions of their time to examining results on the SERP, influenced by factors like the presentation format of these results. For instance, certain users might favour results accompanied by a summary and an image, while others might prefer results displayed solely as text links. Such varied interactions indicate that the search process can differ from one user to another. This understanding is critical for optimising the presentation of search results to align with user expectations and needs.

IIR concentrates on the interactions occurring on a SERP. An example of a SERP

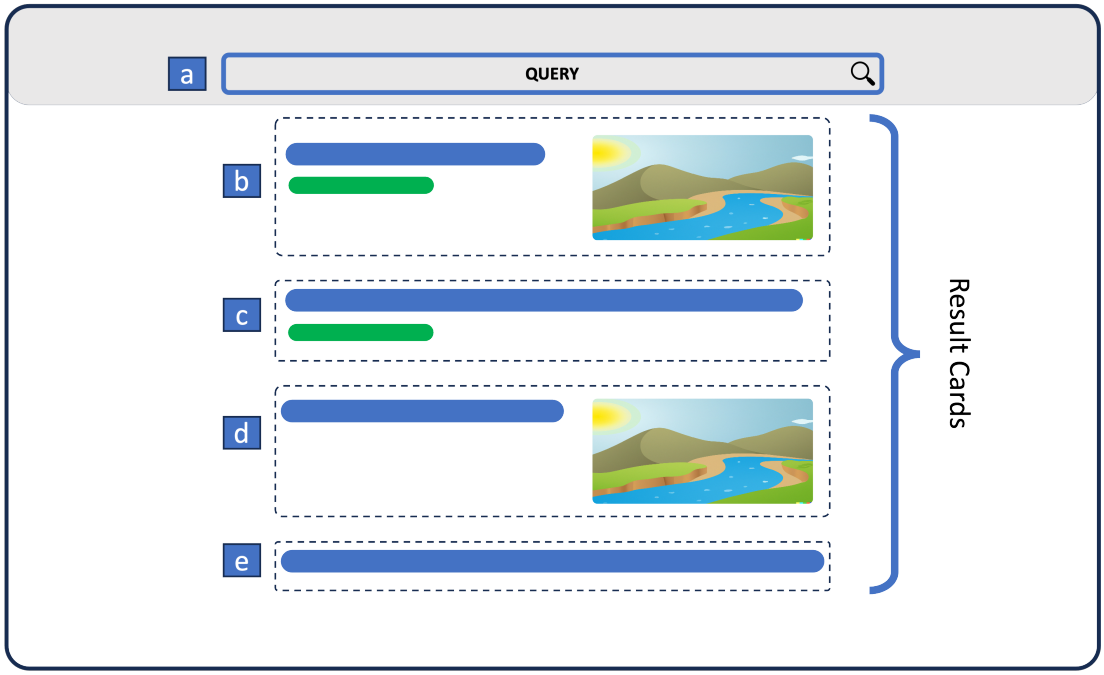


Figure 2.3: Abstracted example of a typical SERP stripped down its main components.

is depicted in Figure 2.3. This illustration is an abstract representation, reduced to the essential elements commonly found in a news SERP. In this SERP, two primary components are identified: the query [a], and the results [b, c, d, e]. The query, embodying the user's information need, is typically entered into a query box and is usually text-based. This thesis assumes text-based queries throughout. We acknowledge that modern IR systems are increasingly accommodating multi-modal inputs (including text, images, audio, video, etc.), which fall outside the scope of this thesis.

The results, labelled [b, c, d, and e], are displayed on what we shall now be referring to as "result cards". These result cards may feature various elements such as a title, summary, or image. Figure 2.3 illustrates some representative combinations of how these result cards might appear on a SERP. In this example, a blue line indicates a title, while a green line can signify a URL or result summary. Typically, the result summary includes the leading sentences of the main document to provide the user with an overview or a preview of the content behind the result card, aiding in assessing its relevance.

The types of these result cards are a particularly strong focus of this thesis, as we

examine how different types of result cards affect user behaviour and satisfaction to construct our model. It is also interesting to note that since these different result cards contain different information on them, they occupy different amounts of space on the screen. This means that if we wish to optimise the result list, we must also consider how many results we can fit within the same space so that user satisfaction remains high.

To comprehensively explore the dual optimisation of search results, we will initially look into existing work from both the system and user perspectives. We will thus gain a deeper understanding of existing models and methods for ranking and displaying documents on SERPs. We can then situate our proposed approach within the context of current research and practices in the field.

2.4 System Side (Ranking)

System-sided approaches have predominantly concentrated on the ranking of documents, with some consideration of user behaviour. Initial models, such as the PRP, did not incorporate user interaction. Subsequent models, evolving from the iPRP, began to recognise user interaction as a crucial component in the document ranking process. This section delves deeper into the mathematical foundations of these models and elucidates their methodologies for document ranking.

Central to this discussion is the concept of "expected utility" (EU), a term borrowed from economics and adapted to the field of information retrieval [18]. In economics, utility refers to the satisfaction or benefit derived by consumers from consuming goods or services. Translating this to information retrieval, the utility can be understood as the perceived value a user gains from a document. This value may encompass various factors such as the informativeness, relevance, or timeliness of the information presented.

While all of the ranking principles below do not explicitly mention the term "utility", we can adapt their approach to calculating this utility, enabling the ranking of documents based on the estimated utility they offer to the user. The exploration begins with the foundational PRP and extends to more sophisticated models that incorporate

user interaction in calculating this utility.

2.4.1 The Probability Ranking Principle

The PRP forms the foundation upon which probabilistic models are built in IR. Therefore it is imperative to understand the mechanism via which it ranks documents. Previously, we briefly examined the PRP, focusing on the underlying assumptions necessary for ranking one document above another. We will now look into a more detailed exploration of PRP, including its proof, particularly with costs and other factors involved in document ranking. Additionally, we will address some limitations of the PRP, highlighting developments that have been made to enhance its effectiveness.

To recapitulate, the PRP posits that the most effective method of retrieving documents involves ranking them in decreasing order of their probability of being relevant or useful, as stated by Robertson (1977) [9].

Let us denote $P(d_i|R)$ as the probability of a document ($\mathbf{d_i}$) being retrieved and relevant ($\mathbf{R}$), and $P(d_i|\bar{R})$ as the probability of a document ($\mathbf{d_i}$) being retrieved but non-relevant ($\bar{\mathbf{R}}$).

The expected utility (EU) of a list (L) of documents, where the are n documents in the list, can be estimated using Equation 2.10:

$$EU_{PRP}(L) = \sum_{i=1}^{n} P(d_i|R) \times 1 + P(d_i|\bar{R}) \times 0 \tag{2.10}$$

The PRP operates under the assumption of binary relevance, assigning a score of 0 to documents deemed non-relevant, and 1 for documents deemed as relevant. However, it is important to note that the probability $P(d_i|R)$ is an estimation, and the system must strive to calculate this value as accurately as possible to reflect the true relevance [49].

Furthermore, [49] illustrated that the PRP could optimise a specifically defined utility function.

Therefore, under the assumption of binary relevance, and referring to the proof in Equation 2.1, it becomes evident that the expected cost of retrieving document d_i

is lower than that of retrieving document d_j. Consequently, document d_i should be ranked higher than document d_j. This is based on the principle that retrieving non-relevant documents incurs a higher cost than retrieving relevant ones, meaning the system faces penalties for retrieving non-relevant items. As more relevant items are retrieved, however, there tends to be a decrease in returns.

The validity of the proof in Equation 2.1 depends on certain assumptions:

- The equation presupposes that relevance and usefulness are equivalent concepts.
- The application of the result to a set of questions encounters challenges, as the costs of retrieving relevant (C_r) and non-relevant ($C_{\bar{r}}$) items vary among users.

Limitations of the PRP

While the PRP was pioneering in modelling relevance in a probabilistic manner, it approached this task in a basic and abstract way. Subsequent research has identified and critiqued several assumptions inherent in the PRP model:

1. The PRP assumes the relevance of a document is independent of other documents in the collection. However, when retrieving subtopics, it becomes necessary to prioritise novel information over redundant content. Hence, previously retrieved documents should influence subsequent retrievals. Under PRP, if two highly relevant documents cover the same topic, they will both rank highly, which is not ideal as it fails to provide new information to the user.
2. The empirical evidence supporting the PRP is limited to individual queries. When considering a set of queries, the performance measure must be averaged across all these queries.
3. If the independence assumption of the PRP is not met, the resulting ranking is sub-optimal [49].
4. The PRP conflates relevance with usefulness. For instance, a search for "jaguar" might yield relevant results about the car, but these may not be useful if the user intended to find information about the jaguar animal.

The ranking principles discussed later in this thesis aim to enhance the basic framework of the PRP by extending it to interactive contexts and incorporating document dependence into the model.

2.4.2 The interactive Probability Ranking Principle

The Interactive Information Retrieval Probability Ranking Principle (iPRP) enhances the traditional PRP model to encompass interactive IR by creating a situation-based framework. This evolution involves modifying some of the PRP's foundational assumptions to better suit the iPRP context. Notably, the concepts of fixed information need and the independence of a document's relevance from previously seen documents are reconsidered and adjusted in the iPRP model [9, 50].

At the heart of the iPRP model lies the notion of "situations", each representing a specific state of user interaction. These situations evolve with the user's choices (clicking, skipping etc.), with the first positive choice leading to a transition into a new situation. To ensure progression and avoid repetitive loops in the interaction, the model implicitly assumes the existence of a final choice, although this terminal point is not explicitly addressed within the model's scope. Significantly, the iPRP acknowledges that a user's relevance judgement can shift the informational landscape; a document deemed relevant in one situation may lose its relevance in another as the user acquires new information.

In this model, a situation is denoted by s_i, where $s_i \in \mathcal{S}$ (set of situations), with each situation encompassing a set of choices, represented by documents $D = d_1, d_2, d_3, \ldots$. The probability of a document d_i being relevant in a specific situation, and hence being an acceptable choice to the user, is expressed as $P(d_i|R)$. This probability reflects the likelihood of a user selecting and retaining a particular choice from the result list. An illustration of the iPRP and situations is shown in Figure 2.4

The iPRP introduces the notion of effort (e_i), which encapsulates the amount of work, time, and cognitive resources a user expends to evaluate a document within an information retrieval system. This includes the physical actions (like clicking, scrolling, and reading) and mental processing (like understanding, assessing relevance,

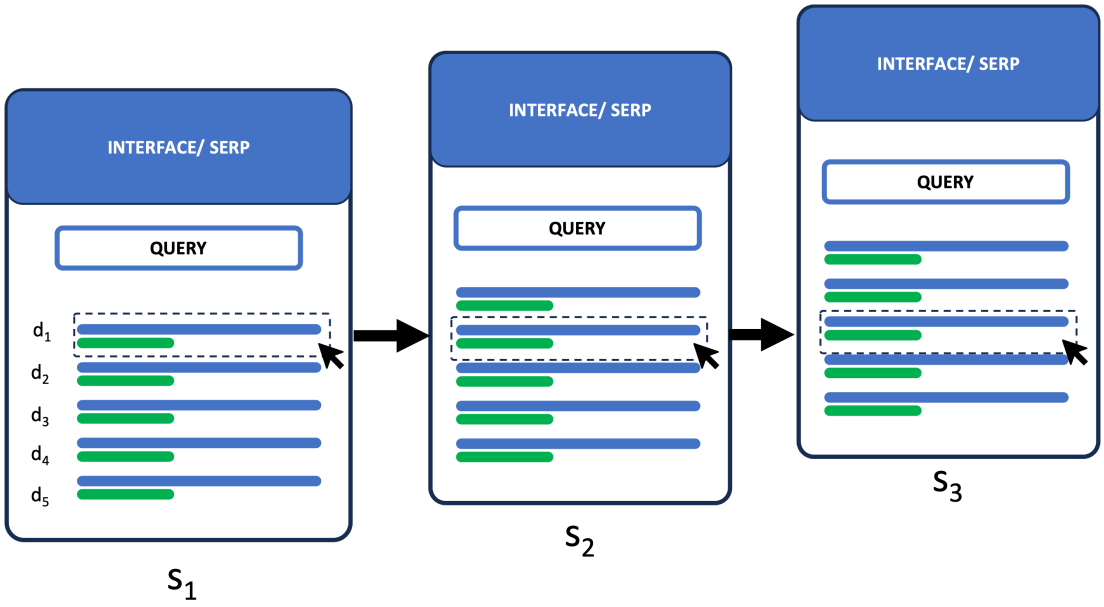


Figure 2.4: An illustration of the iPRP showcasing documents and situations

and decision-making) required to interact with and assess the information presented. While Fuhr's framework acknowledges these variations, the iPRP model does not explicitly quantify these effort differences. This assumption recognises the dynamic interaction nature but simplifies the model by not delving into the varying effort levels for each task.

The model posits that users make binary decisions (positive or negative) when interacting with IR systems. A positive decision, where a user sticks with their choice upon learning its consequences, is considered "correct", and this will yield a benefit. This is a key concept in the iPRP model, emphasising decision finality and the absence of backtracking or decision revision.

Only decisions that are deemed "correct" yield benefits (b_i), for the user. The user's benefit from choosing a document quantified as b_i, represents the comprehensive value gained from that selection. This includes not only the relevance and usefulness of the information but also how well it meets the user's specific needs in terms of satisfaction and applicability in the current context.

The user is assumed to evaluate the available choices in a sequential, linear order.

To optimise the benefits for the user within this model we must...

- ... minimise the effort ei required to evaluate a document d_i in the given situation s_i should be minimal and,
- maximise the selection probability ($P(d_i|R)$), i.e., the user needs to pick the document where ever appropriate.

In this context, selecting a particular document d_i leads to an average benefit denoted as a_i. Consequently, the expected benefit ($E[b_i]$) of choosing a particular document d_i in the iPRP model can be mathematically expressed as:

$$E[b_i, (d_i)] = e_i + a_i P(d_i|R) \tag{2.11}$$

This equation encapsulates the notion that a well-made decision can save the user time, whereas a poor decision results in time being wasted [50]. In a scenario where the user reviews a list of options linearly, their first positive decision propels them into a new situational context.

Building upon this understanding, Fuhr further derives the optimal ranking of a selection list L as:

$$EU_{iPRP}(L) = \sum_{i=1}^{n} \left(\prod_{k=1}^{i-1} P(d_k|\bar{R}) \right) E[b_i, (d_i)] \tag{2.12}$$

Ranking Criterion

In the iPRP, if we consider two subsequent documents d_i and d_{i+1} for ranking. By computing Equation 2.1 the documents d_i and d_{i+1} must be ranked according to the criterion as shown below:

$$a_i + \frac{e_i}{P(d_i|R)} \geq a_{i+1} + \frac{e_{i+1}}{P(d_{i+1}|R)} \tag{2.13}$$

where the ranking criterion for a document d_i is

$$EU_{iprp}(d_i) = a_i + \frac{e_i}{P(d_i|R)} \tag{2.14}$$

Suppose we take the following example as shown in Table 2.1 where the probability of d_2 being relevant is higher than d_1, but the average benefit of retrieving d_2 is higher than that of d_1

Then we can see that according to the PRP, d_1 must be ranked above d_2 under decreasing probability of relevance. But according to the iPRP, to rank the documents; the $EU_{iprp}(d_2)$ being higher makes it a better choice to be ranked above d_1. However, if we relax this assumption of average benefit and effort, then we get back the PRP ranking.

Document	$\boldsymbol{P(d_i\|R)}$	$\boldsymbol{a_i}$	$\boldsymbol{e_i}$	$\boldsymbol{E[b_i,(d_i)]}$	$\boldsymbol{EU_{iprp}(d_i)}$
d_1	0.50	10	-1	4	8
d_2	0.25	16	-1	3	12

Table 2.1: iPRP ranking example

Limitations of the iPRP model

The iPRP, grounded in solid theoretical underpinnings, nonetheless faces challenges in its formulation. It assumes the independence of user choices, overlooking how prior interactions could influence subsequent selections. The model primarily values positive, correct decisions, neglecting the insights that could be gleaned from choices users reject. This perspective misses the potential benefits of understanding why certain options are disregarded in a context where users face a vast array of choices. Additionally, the application of the iPRP is complicated by its relation to Markov models, especially when accounting for users' creative actions, such as introducing new query terms, which adds complexity to the model's predictive capabilities. Another significant challenge lies in the accurate estimation of crucial parameters like effort, selection probability, success probability, and benefit. This underscores the need for empirical research, potentially involving user studies and eye-tracking, to refine these estimates and enhance the iPRP's applicability and effectiveness in real-world interactive information retrieval scenarios. While we cannot address all the limitations of the iPRP in our work, we work toward refining the estimation of the parameters like benefits, efforts and selection

probabilities through user studies.

2.4.3 The Quantum Probability Ranking Principle

The quantum Probability Ranking Principle (qPRP), addresses the issue of document interdependence in the PRP model [51]. The qPRP draws an analogy from the double-slit experiment in quantum physics, which provides a vivid illustration of the wave-particle duality of electrons. Figure 2.5 shows an illustration of this through a popular meme.

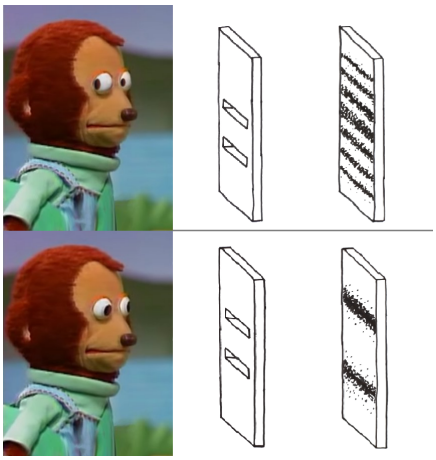


Figure 2.5: An interpretative depiction of the quantum double-slit experiment, humorously presented to elucidate the concept of particle-wave duality and observer effect. The top shows the electron being unobserved, and the bottom shows the electron being observed [2]

In the classic double-slit experiment of physics, an electron is projected towards a barrier with two slits. When it passes through these slits and is not directly observed, it demonstrates a wave-like behaviour, creating an interference pattern on a detector screen placed behind the slits. This pattern is characterised by alternating bands of high and low intensity, indicative of wave interference. However, if the electron's path is observed, meaning its passage through one of the slits is monitored, this wave-like interference pattern collapses, and the electron exhibits particle-like properties, producing two distinct and separate impact points corresponding to the slits.

Translating this concept to the realm of IR, the qPRP model likens the user and their information need to the electron. The documents in the retrieval system are anal-

ogous to the slits in the experiment. The observation or measurement in this analogy corresponds to the user's interaction with the documents. In the context of qPRP, when a user engages with a document (akin to the electron being observed through a slit), it influences the probability distribution of relevance for subsequent documents, just as the observation of an electron affects its interference pattern. When the particle passes through one of the slits (i or j), it hits a detector panel positioned at location x with the probability $P(d_{ij}|R)(x)$. Thus to maximise the user utility a suitable second slit size (document at rank 2) must be chosen while keeping the first slit (document at rank 1) a constant size. Therefore, in the qPRP model, the document ranked second (the second "slit") is selected not merely based on its standalone probability of relevance $P(d_i|R)$, but also considering the quantum interference term $Q_{ij}(x)$. This term encapsulates the interdependent effect of documents on each other's perceived relevance, akin to the interference pattern in the double-slit experiment.

If k is a document that induces stopping behaviour, then the expected utility (EU_{qPRP}) can then be defined as:

$$EU_{qPRP}(d_i) = t(P(d_i|R)(x) + P(d_k|R)(x) + Q_{ij}) + u(\bar{x}) \tag{2.15}$$

where, t describes the difference in utility $u(x)$ and $u(\bar{x})$, which is the utility of retrieving a document that stops the search and a document that doesn't stop the search respectively.

Ranking Criterion

The qPRP ultimately hypothesise that, to maximise the effectiveness of an IR system - the document that is ranked in the second position must be ranked after the set of documents already ranked and before any document in the list if and only if the difference in the value of the interference term is greater than the probability of ranking the document at the position x.

Limitations of the qPRP model

The qPRP is better suited to address diversity in the ranking of documents than the PRP since it takes into account the dependence of documents on one another by drawing analogies to the quantum interference effect. Whether or not modelling the dependence between documents explicitly is the best possible approach to achieve better ranking remains to be explored. Other models suggest that the overall dependence between documents is determined by the function that models the user and not the ranking principle.

2.4.4 Mean Variance Model

The mean-variance model, by Wang [11], establishes an approach to document ranking by drawing analogies to the modern portfolio theory (MPT) in finance by Harry Markowitz in 1952 [52]. MPT assists investors in constructing portfolios to maximise expected returns for a specified level of market risk. The cornerstone of this theory is diversification, advocating for a portfolio comprising a variety of asset classes. The underlying rationale is that not all assets will under-perform simultaneously, thereby reducing the overall risk. This principle finds a parallel in document ranking, where a diverse set of documents can mitigate the risk of irrelevant information retrieval.

Another key concept of MPT is the Risk-Return Trade-Off. It postulates that higher risk is linked with a greater likelihood of higher returns, and conversely, lower risk is associated with a higher probability of smaller returns.

A key objective of the model is to optimise the overall mean (relevance) while maintaining a predetermined level of variance (risk). The mean-variance model builds upon the Probability Ranking Principle (PRP), initially introduced by Robertson in 1977. It enhances this principle by accounting for the uncertainty involved in determining document relevance and the relationships between documents in the retrieval process. This model proposes a new approach to document ranking, likening it to the process of selecting a portfolio in finance. Here, instead of ranking documents individually, the model suggests choosing and ordering the top 'n' documents as a group. This method differs from the traditional practice of evaluating each document independently, thereby

overcoming one of the weaknesses of the PRP.

The mean-variance model calculates relevance as an overall average of relevance measures at each rank (EU_m)

$$EU_m(di) \equiv \sum_{i=1}^{n} w_i P(d_i|R) \tag{2.16}$$

Where w_i is a weighting term which determines the rank position importance for each relevant item and $\sum_{i=1}^{n} w_i$. This can be thought of as a penalty function that penalises late retrieval of relevant documents. The maximal EU_m can only be achieved when $w_1 > w_2 > \ldots > w_n$.

The PRP suggests maximising the overall mean. But if the overall variance is minimised then the "risk" can remain as low as possible. This is because the variance indicates dispersion from the expected relevance, representing the level of risky prospect. Following the same terminology from portfolio theory, if we vary the variance in an objective function, we obtain a set of efficient ranking solutions that can give us the maximal relevance value for a given variance (risk) level.

The expected utility $\boldsymbol{EU(d_i)}$ can be represented as a maximisation function

$$EU(d_i) = EU_m(d_i) - \alpha Var(d_i) \tag{2.17}$$

where α represents the upper bound of the risk level. Since directly optimising the objective in equation 2.17turns out to be computationally expensive, the ranking function must optimise the selection of each document, at each rank to maximise the objective function.

$Var(d_i)$ can be defined as

$$\text{Var}\,(d_i) = \sum_{i=1}^{n} \sum_{j=1}^{n} w_i w_j c_{i,j} \tag{2.18}$$

where, $c_{i,j}$ represents an element in the co-variance matrix, which determines the co-variance of the relevance measures between a document at position i and a document at position j

Ranking Criterion

Suppose we have four documents (d_1,d_2,d_3,d_4), and d_1 is fixed at position 1. The next document amongst the remaining documents must be chosen in such a way, that the overall increase in objective function is maximal. In other words, if d_2 is to be ranked second then, $EU(d_1) - EU(d_2) > EU(d_1) - EU(d_3)$ and $EU(d_1) - EU(d_2) > EU(d_1) - EU(d_1)$

Where, $EU(d_n) - EU(d_{n-1})$ can be generalised by expanding Equation 2.17 and written as

$$EU(d_n) - EU(d_{n-1}) = P\left(d_n|R\right) - \alpha w_n \sigma_n^2 - 2\alpha \sum_{i=1}^{n-1} w_i \sigma_i \sigma_n \rho_{i,n} \tag{2.19}$$

Where σ is the standard deviation i.e., $\sigma_i = \sqrt{C_{i,i}}$ and ρ is the correlation coefficient

$$\rho_{i,j} = \frac{c_{i,j}}{\sigma_i \sigma_j} \tag{2.20}$$

Limitations of the model

The mean-variance model, while innovative in its approach, encounters several challenges and limitations in its application to document ranking. The model adopts variance as the primary metric for assessing risk, drawing from financial concepts. However, in the realm of finance, other risk evaluation models might offer valuable perspectives, such as those considering "downside risk."

One significant challenge with the mean-variance model lies in the computational intensity required to calculate the covariance matrix, especially when dealing with a large set of documents. This complexity makes the direct optimisation of the objective function a daunting task. Furthermore, the model does not dynamically adapt to the evolving needs of users within their search session. It focuses on optimising the ranked list for a specific query, rather than considering the optimisation of rankings over the entire user interaction session.

Additionally, due to the complexity of its optimisation function, the mean-variance model faces difficulties in incremental learning from user feedback. Integrating new

information or feedback into the model would necessitate a complete re-optimisation of the objective function, which is not only computationally demanding but also time-consuming. This characteristic limits the model's responsiveness and adaptability in real-time information retrieval scenarios, where user preferences and needs can change rapidly.

2.4.5 Dynamic IR Model

The PRP and iPRP models can only deal with traditional ad-hoc query ranking and retrieval tasks. However, when search tasks are complex and exploratory, they are comprised of multiple stages and the information needs change over time. The iPRP and the PRP model cannot represent tasks over multiple stages [14]. To overcome this limitation, DIR tasks consist of three main characteristics (1) User feedback, (2) Temporal dependency and (3) Overall goal.

The DIR model first defines two frameworks to model static and dynamic retrieval.

Static IR Framework The static IR framework models interactions with information retrieval systems where each interaction is considered either as a single, isolated event or as multiple independent events with different search intents. This is typically applied in ad hoc ranking and retrieval systems.

The objective in a static IR framework is to select an action a_i (or a sequence of actions) that maximises a static utility function $EU_s(a, d_i)$ Here, "action" refers to a choice made by the system within a defined action space $\mathcal{A}$. This action could be, for example, a query suggestion presented to a user or the order in which a set of documents is ranked for retrieval. The utility function assigns value to each action based on its associated probability of relevance, which can be R or $\bar{R} \in \mathcal{R}$.

The expected value for the static utility function is defined as

$$E[EU_s(a_i, d_i)] = \sum_{R \in \mathcal{R}} P(d_i|R) EU_s(a_i, \mathcal{R}) \tag{2.21}$$

At this point, the static utility would be determined via direct computation of nDCG.

Interactive IR Framework The interactive IR framework extends the static IR framework and takes into account user feedback from the previous situation i, but cannot anticipate feedback from future interactions.

Dynamic IR framework The dynamic IR framework extends the interactive framework by being responsive to user feedback and optimising for it in advance. The DIR system has three elements that can distinctly define whether a system can be considered dynamic or not:

1. **Feedback**: An observation signal from the user (document being relevant or not)
2. **Temporal Dependency**: Multi-stage operation where each situation is inter-dependent on the previous situation
3. **Goal**: Objective across all situations

A dynamic system will find the optimal interactions that incorporate user feedback and temporal dependency, and then find an optimal sequence of interactions across all situations. As a result of this, the utility in one situation may be reduced so that benefits can be increased in a later situation.

The expected utility can be calculated as:

$$EU_{DIR}(d_i, i) = \max_{a_i \in A} \left[EU_s(a_i, d_i) + \omega(i) \sum_{d_i \in D} \sum_{R \in \{R, \bar{R}\}} P(d_i|\mathcal{R}) EU(\tau(a_i, d_i, \mathcal{R}), i+1) \right] \tag{2.22}$$

Here, $\tau(a_i, d_i, \mathcal{R})$, represents a transformation function based on action a_i, document d_i and relevance (R or $\bar{R}$) in $\mathcal{R}$. The DIR model also uses a discount function ω_i which helps to ensure a solution exists over a set of possibly infinite situations by giving greater weight to the utility by retrieving relevant documents early.

Limitations of the DIR model

The dynamic utility function can be shown to be PSPACE-Complete only for a small number of pages, but where the information space is potentially infinite, the action

space can be difficult to optimise. Also, although the authors talk about eye-tracking, the dynamic utility function is not guaranteed to be traceable thereby not ensuring an optimal solution. The authors also note that the data sets they have used (WT10g, AQUAINT and ClueWeb09) do not contain interaction data, which is important in an interactive setting. The dynamic utility function is a powerful tool, however, we reserve the use of this model to optimise ranked lists for future work. In our thesis, we are focused on an ad-hoc search task, for which the derivatives of the iPRP are sufficient. We only cover it here to cover ground on different utility measures across all search tasks.

2.4.6 Implicit User Model

The Implicit User Model (IUM) proposes an innovative approach to improve existing IR systems. It does so by modelling individual user behaviour [13]. Central to this modelling is the accurate representation of a user's information need. Given the challenge of acquiring explicit feedback from users in real-world scenarios, it is imperative to derive this information implicitly. Prior research has demonstrated the efficacy of capturing user interactions, such as mouse events and overall system engagement, to enhance user behaviour models. The IUM leverages the user's immediate search context to address their short-term information needs, a process termed "eager implicit feedback".

This model introduces a novel paradigm in IR by utilising two types of implicit feedback: firstly, the identification of similar previous queries and their outcomes to facilitate effective query expansion; and secondly, the utilisation of viewed document summaries to inform the re-ranking of as yet unseen documents. Every user interaction with the system, as recorded in log data, serves as a foundational element for the IUM to enhance subsequent system performance. Consequently, the optimisation challenge within the IUM framework is formulated as a decision-making task.

The IUM framework is conceptualised as a function dependent on four key elements:

1. User action, denoted as a_t.

2. System response, represented as $S(a_t)$.

3. The current user model, M.

4. A posterior probability that encapsulates both the user model and all observed user data up to time t.

The primary tasks for the system to generate an optimal response involve computing the current user model M, followed by selecting a system response $S(a_i)$ that minimises the composite loss function.

User Model Construction The user model in the IUM is composed of two principal components: (1) The user's information need, which is the system's inference about the user's interests, represented through a term vector and, (2) The record of documents already viewed by the user. The underlying rationale is that a document, even if relevant, may have been previously viewed by the user, thereby reducing its utility in subsequent searches.

Formulation of the Loss Function The loss function incorporates the probability of a user viewing a relevant document, denoted as $P(view|d_i)$. The PRP is thus expanded as follows:

$$EU_{IUM}(d_i) = \sum_{i}^{n} P(view \mid d_i)\ P(d_i|R) \tag{2.23}$$

The IUM was simulated for re-ranking documents by simulating various types of search agents. The main finding reported from this study was that, despite the search agent lacking control over the retrieval algorithm, it was feasible to re-rank the results by expanding the original query, especially when it was related to the current query.

Ranking Criterion: Suppose two documents d_1 and d_2 are to be ranked, if

$$P(view \mid d_1)\ P(d_1|R) > P(view \mid d_2)\ P(d_2|R) \tag{2.24}$$

then we can rank d_1 before d_2. Drawing an analogy to the PRP, this $P(view \mid D_i)$ is the same as C_r

Limitations of the model

The loss function integral to the IUM exhibits insensitivity to the internal ordering within the top n selected documents. Consequently, alterations in the ranking sequence of relevant and non-relevant documents do not affect the loss function's outcome. While the paper discusses the utility of displaying result summaries to facilitate the model's enhanced relevance assessment, it does not specify which types of summaries are most effective in aiding these assessments.

The empirical findings presented in the paper illustrate that the agent is designed to re-rank textual documents within Google search results. However, it is noteworthy that Google has evolved its Search Engine Results Pages (SERPs) to include snippets and images for certain prevalent queries. In this context, the efficacy of the IUM in handling picture-based documents remains unexplored and unverified.

Table 2.2 shows the different aspects and limitations that models evolving from the PRP address, arranged from system-sided aspects such as document interdependence to user-sided aspects such as implicit user feedback.

Table 2.2: Evolution of IR Models from System-Sided to User-Sided Aspects

Aspect	PRP	iPRP	qPRP	MV	DIR	IUM	Card
Document Interdependence			✓				
Uncertainty of Relevance				✓			
Dynamic Information Needs					✓		✓
User Interaction		✓			✓	✓	✓
Implicit User Feedback						✓	✓

2.5 User Side (Presenting)

While system-sided approaches in IR have been notably comprehensive, emphasising formal proofs for document ranking, user-sided approaches have concentrated on a distinct aspect of the retrieval process: the presentation of documents. In system-sided methodologies, the presentation is often considered a secondary factor, typically perceived as a downstream component that does not significantly influence ranking

algorithms.

However, understanding how the presentation of search results impacts user satisfaction enables the development of a more holistic utility function, one that extends beyond the system-sided Expected Utility (EU) and incorporates the user's perception. This leads to the formulation of Expected Perceived Utility (EPU). EPU encompasses both the intrinsic value derived from the information (as captured by EU) and the user's subjective experience influenced by the presentation of search results, thus the term "perceived". This dual consideration promises a more robust and user-centric approach in the modelling of utility functions within information retrieval systems.

In examining the presentation of search results, user-centred approaches in IR have identified several key factors warranting detailed investigation. These factors can be categorised as follows: (1) The type of layout employed, such as grid or traditional list formats. (2) The various types of result cards. Building on the insights gained from these initial categories, further exploration is conducted into (3) how distinct types of result cards, and (4) the design of result summaries, influence user satisfaction.

In addition to these elements, it is crucial to consider other factors integral to the user interaction process, such as the cost incurred by users to issue or reformulate a query and its subsequent impact on user satisfaction. This conceptualisation can be envisaged as a web page comprising multiple layers, with each component (layout, result card type, etc.) representing a specific layer. We will now "peel back" these layers starting with an examination of how different layouts affect user satisfaction and progressively increase in complexity, culminating in an understanding of user interactions with various queries on a fully constructed web page

2.5.1 Layouts

There has been limited work in studying the influence of the page layout on users. [53] investigated how users interact with list and grid interfaces in search results. Their findings indicated that list interfaces led to more uniform and linear viewing patterns, with a focus mainly on the top of the list. On the other hand, grid interfaces resulted in more varied viewing patterns and a more even distribution of attention across the

search results. Importantly, the study found that participants using a list interface, organised in ascending trustworthiness order, tended to pay more attention to less trustworthy results. They also selected trustworthy results less often compared to those using a grid interface. Therefore, Gerjets and Kammerer's research suggests that a grid interface might be more effective in helping users select trustworthy information sources. However, the unreliability of this study is limited due to its specific user group and task design, which was focused on informational tasks.

Meanwhile, [20] explored the effectiveness of list and tabular interfaces in web search tasks using eye-tracking technology. The study, involving 16 participants, found no significant differences in task completion time, errors, or fixation duration's between the two interfaces. However, it revealed that users were more likely to make transitions within the same category of results when using the tabular interface compared to the list interface. This implies that the tabular interface may allow users to prioritise categories more efficiently. A major limitation of this study was the potential familiarity bias of participants towards list interfaces. This bias could have influenced how participants interacted with the tabular interface in the study. Furthermore, the small sample size raises concerns regarding the generalizability of the results. The limited participant number suggests that the findings may not accurately represent user behaviours and preferences in more diverse settings.

[54] evaluated seven interfaces for structuring search results using category information. They found that category interfaces (interfaces that were organised into grid-like structures) were consistently faster than list interfaces, even when the list presentation included category names and inline summaries, or the category presentation was degraded by removing category names or page titles. The best performance was achieved when both category names and page titles were available. Inline summaries were more effective than hover text summaries for both list and category interfaces. This study, involving 76 participants, underscores the critical role of interface design in search efficiency. However, the study was limited by a relatively small participant pool and possible bias in task selection, which might impact the applicability of their findings to broader user populations and search contexts.

[55] explored scrolling strategies with single-column (vertical scroll, essentially a list layout) and multi-column (horizontal scroll, essentially a grid layout) layouts in web browsers. They found that about one-third of participants preferred the horizontal-scroll (grid layout) layout for reading, while two-thirds favoured the vertical-scroll layout. The observed scrolling strategies were page scrolling, continuous scrolling, and region scrolling. Participants using the horizontal-scroll layout tended to use page scrolling, while those using the vertical-scroll layout often employed continuous scrolling. However, the findings from this study are limited due to the participant pool comprising only 24 students, who were mainly students.

From the synthesis of these studies, it is evident that the findings present some conflicting viewpoints and are constrained in terms of their unreliability. This limitation primarily arises from the small sample sizes and the considerable variation in the information-seeking tasks used in these studies. A critical aspect that builds upon the layout of search results is the presentation of these results to the user. Therefore, the next section will focus on examining how different designs of result cards influence user interaction and satisfaction.

2.5.2 Result Card Types

Positioned above the layout of search results is the result card, which, as previously defined, is a container resembling a box, capable of encompassing a variety of information elements including titles, images, URLs, and summaries. The diverse formats employed to display information on the SERP incur distinct interaction costs. These costs encompass the time needed to process or read a result and the satisfaction derived from the result and its presentation. Extensive prior research has delved into the impact of various result card designs on interaction costs and user behaviour, particularly in the context of navigational tasks during web browsing.

[21] studied interfaces comprising title, image only (thumbnail), and title+image+summary on an augmented Web SERP. This study involved two phases, 276 participants in Phase 1 and 197 in Phase 2, the study found no significant differences in page clicks across these interface types, but user satisfaction varied significantly.

In contrast, [22] focused on title, image only, and title+image interfaces, also by augmenting pre-existing web SERPs. This study involved 35 participants undertaking judgement tasks and revealed that users made more accurate decisions when presented with titles and images.

Meanwhile research from [23], compared interfaces including title, title+summary, title+image, and title+image+summary. Conducted with 24 participants, each with an average of 5 years of internet experience, the study found consistent user satisfaction across tasks, despite using various interfaces on the Web SERP Augmentation dataset.

Finally, [24] examined the impact of title vs. title+summary interfaces on the TREC WSJ dataset. Involving 20 participants in judgement tasks with 50 articles, the study observed a significant positive impact of summaries on relevance judgements.

Building on the previous discussion about the importance of result card summaries in relevance judgement, it is crucial to have a deeper look into understanding their effectiveness. This involves identifying the specific characteristics of these summaries, like their length and format, that make them valuable tools in assisting users during search-related activities. We now dive into understanding result summaries.

2.5.3 Result Summaries

Result summaries can be broadly classified along multiple dimensions, including the number of source documents, type of summarization, summary focus, and more [56]. However, in the scope of this thesis, we focus on extractive and abstractive techniques for single document summarization, looking at both generic and query-focused indicative summaries.

Pioneering work in automatic text summarization was initiated by [57], [58], and [59]. Advancements in this field have been significant, as evidenced by the work of [24] who found that *query-biased summaries* assist users in making more accurate relevance decisions in information retrieval tasks. This study utilised the TREC WSJ dataset and employed a methodology that assigns positive weight to terms appearing in the user's query, demonstrating the effectiveness of query-biased summaries containing relevant metadata.

Further developments in summarising techniques were presented by [60], who proposed a sophisticated model for sentence extraction in query-focused summarising. This was compared to simply extracting the lead sentence of a news article. More recently, [17] conducted an in-depth analysis of snippet length and informativeness. Their study, which used the TREC AQUAINT dataset, revealed that participants preferred snippets that were neither too short nor too long, highlighting the importance of optimal summary length in enhancing user experience.

In the domain of abstractive summarization, which aims to generate novel sentences that capture the essence of the original text, [56] provided an extensive survey covering recent advancements. This approach, requiring deep text comprehension, often employs techniques like paraphrasing, generalisation, and sentence fusion.

Comparing different summarization methods is a field of ongoing interest. For example, [61] found that automatic summarization methods performed comparably well against human-constructed paragraph extracts. Furthermore, the study by [62] concluded that there is no significant difference between extractive and generative methods in terms of user satisfaction.

Given the various methods of creating summaries, our research employs the technique of using the lead sentence for summarization. This decision is supported by the findings of [60], who demonstrated its effectiveness compared to more complex methods. The lead sentence approach offers a balance between informativeness and brevity.

The goodness of result summaries in search tasks is deeply intertwined with the quality of the initial queries. Poorly constructed queries can undermine even sophisticated summary techniques, highlighting the importance of query formulation and its associated costs. Thus, we will now explore how these costs impact user satisfaction.

2.5.4 Query Performance and User Satisfaction

Models developed on economic search theory by [50] provide a valuable framework for estimating the costs associated with user behaviour in search tasks. These models consider various factors, such as the length of the query, the number of viewed documents, and the interaction with clicked snippets. The implications of these costs on

user satisfaction have been further elucidated in studies such as those by [63] and [64], highlighting the direct correlation between search costs and user experience.

Moreover, [65] look into how the cost associated with query formulation and modification influences search behaviour. Their investigation reveals that higher query costs can lead to changes in user strategies, often resulting in less efficient search experiences. This insight is crucial in understanding the dynamics of user interaction with search systems.

Another significant contribution is made by [66], who examined the effectiveness of reading protocol software in interactive information retrieval experiments. Their findings provide a nuanced view of how users engage with search results, further emphasising the role of query costs in shaping search behaviours and outcomes.

Additionally, the work of [67] sheds light on the interplay between queries and search result quality. They propose that the quality of queries directly affects the relevance and usefulness of search results, thereby influencing user satisfaction. This perspective is particularly relevant in the context of how search engines adapt and respond to user queries.

These studies collectively underscore the significance of query formulation and modification in the broader landscape of user satisfaction in search tasks. They highlight the need for search systems to not only focus on generating relevant summaries but also to facilitate the creation of effective queries. By reducing the costs associated with querying, search systems can enhance user satisfaction and improve the overall search experience.

2.6 Direct Optimisation Models

In recent years, the focus on optimising the presentation of the SERPs directly has increased, with significant contributions from various researchers. Kicking off the exploration into SERP optimisation, [68] conducted a thorough survey of deep learning applications in this domain. Their work serves as a cornerstone, offering a broad perspective on how machine learning, particularly deep learning, has revolutionised SERP

optimisation. They critically analyse various models and techniques, setting a foundational context for understanding the subsequent, more focused studies. This survey underscores the transition from traditional search algorithms to more complex, data-driven approaches that are increasingly shaping the future of search engine technologies

[30] presented a framework for learning to rank whole-page web search results. This methodology goes beyond traditional approaches, aiming to optimise the entire SERP by integrating implicit costs into a reward function for Deep Reinforcement Learning (DRL). Their approach demonstrated the potential to outperform established search engines like Google and Bing. However, a significant limitation is the substantial data requirement for training these models, making it less practical in professional search domains like legal and medical fields, where interaction data is scarce.

[25] addressed the optimisation of two-dimensional search result presentations. They focused on improving traditional metrics like Discounted Cumulative Gain (DCG) and Rank-Biased Precision (RBP). While their approach brought a fresh perspective to SERP optimisation, it is important to note the inherent limitations in optimising solely for DCG and RBP. This narrow focus might overlook other crucial aspects of user experience and relevance.

Building upon the theme of SERP layout optimisation, [31] introduced techniques for ranking layouts, emphasising the need to optimise not just the individual elements but the overall structure of the SERP. Their research moves beyond the traditional list-based formats, experimenting with layouts that could potentially enhance user engagement and satisfaction. This work is significant in illustrating how the physical arrangement of information on a SERP can impact user interaction and decision-making processes, suggesting a more holistic view of search engine optimisation.

[26] offered a method to optimise search engine results using click-through data. This approach leveraged user feedback implicit in click patterns to enhance search algorithms. Despite its effectiveness, potential biases inherent in click-through data can limit the approach, as it might not fully represent the diverse range of user preferences and intentions.

[29] also expand this scope by enabling user click modelling beyond the traditional

"ten blue links" in SERPs. This research acknowledged the evolving nature of search pages and the need to interpret user clicks in a more nuanced manner. While this study advanced the field, it also faced limitations in comprehensively capturing user intent and the diverse ways in which different user segments interact with search results.

Further, [27] investigated user behaviour on SERPs that combine results from multiple specialised search engines, termed "verticals". This work focused on how the presence of these verticals, such as images and videos, alters user interaction patterns compared to standard text-based search results. Through a combination of large-scale log data analysis and eye-tracking studies, the researchers uncovered significant behavioural biases. It was observed that users engage differently with vertical results, exhibiting biases in examination and trust that impact the overall interaction with the SERP. Importantly, vertical results were found to have a higher probability of being revisited, indicating a distinct user engagement level.

Responding to these insights, Wang et al. proposed a new model, the Vertical-aware Click Model (VCM). This model is specifically designed to interpret user click behaviour more accurately on SERPs that include verticals. Similarly, [28] delve into this aspect by incorporating user models into the optimisation process. Their approach is pivotal in acknowledging that user interaction with search results is not homogeneous and varies significantly across different contexts and query types. By focusing on user models, they open up avenues for creating more personalised and contextually relevant search experiences, addressing a critical aspect often overlooked in earlier optimisation strategies.

Overall, these studies collectively represent a significant stride towards optimising SERPs. They showcase a shift from conventional ranking algorithms to more holistic and user-centric approaches. However, they also underline the challenges in balancing between algorithmic sophistication and practical applicability, particularly in domains with limited user interaction data.

2.7 The Card Model

So far, in our analysis of IR systems we established that there are broadly two approaches to study and optimise the user experience. We categorised them into system-side and user-side approaches. We shed light on various ranking principles that model the anticipated utility for ranking retrieved results (system-sided approaches). Consequently, there has also been an exploration of how different formats for presenting result cards and layouts impact user satisfaction (user-sided approaches). Among these ranking principles, the iPRP (§2.4.2) posits that the presentation format of displaying results can significantly influence user interactions with these results. Thus suggesting that altering the presentation of a result could lead to varying degrees of user engagement, such as attracting more or fewer clicks, necessitating more or less time to process the information, and occupying varying amounts of screen space. Consequently, this leads to the proposition that the ranking of results can be altered and optimised based on user interactions in accordance with the iPRP.

A specific implementation of the iPRP is the Card Model. The Card Model conceptualises the interaction process as a cooperative game between two participants: the system and the user. The goal of the game is to maximise the information gain, while trying to minimise the user effort. The model estimates the utility of a displayed card by considering both its presentation cost and the resulting user benefit (i.e., allowing us to both rank and present).

The Card Model is implemented as three distinct types of cards:

1. the interface card,
2. the plain card and,
3. the navigational card.

2.7.1 Interface Card

The interface card is the base from which the plain card and the navigational card can be derived as special cases. Consider a given result item **i** and a specified result card

"**card**". In this context, the user can perform actions, denoted as $\mathbf{A_{i,j}}$, within an action space $\mathcal{A}$. Here, j signifies the type of action: for example clicking, skipping scrolling etc.

Additionally $\mathbf{R_i}$, signifies the relevance of the result item, for example in a relevance space $\mathcal{R}$ of graded relevance, this can be relevant, partially relevant and non-relevant. These actions allow the user to transition to subsequent sets of choices within the retrieved results. Each action $\mathbf{A_{i,j}}$ undertaken by the user within this space comes with an associated expected benefit $\mathbf{B}(\mathbf{A_{i,j}})$ and a corresponding expected cost $\mathbf{C}(\mathbf{A_{i,j}})$, incurred from performing that specific action, considering the relevance $\mathbf{R_i}$ of the item. The Expected Perceived Utility (EPU) of a result card, for a result item $\mathbf{i}$, is thus generally formulated as:

$$EPU_{\text{card}}(i) = \sum_{R_i \in \mathcal{R}} \sum_{A_{i,j} \in \mathcal{A}} P(A_{i,j}|R_i)P(R_i)\Big(B(A_{i,j}|R_i) - C(A_{i,j}|R_i)\Big) \tag{2.25}$$

2.7.2 Navigational Card

The navigational card introduces the notion of blocks and tags. Where each card can be made of one or multiple blocks and each block contains tags. However, the original work has only described how a card with a single block can be used to optimise the mobile screen interface. The navigational card uses an information gain reward to model user preference with an entropy function $H(P) = -\sum_p P\ log\ P$. Where if the entropy is lower, the system knows more about the user's information need and can help them find more interesting information. The benefit for an action can incorporate this information need as context (x_i) and be rewritten as

$$B(A_{i,j}) = InfoGain(P(A_{i+1,j}|R_i), x_{i+1}), P(A_{i,j}|R_i), x_i)) \tag{2.26}$$

and in terms of the entropy function

$$B(A_{i,j}) = H(P(A_{i,j}|R) - H(P(A_{i,j}|R))) \tag{2.27}$$

There is also a capacity constraint of the total space the cards can occupy on the block, to model interfaces with relatively small capacity. However, since the capacity of the screen is assumed to be very small, the cost $C(A_{i,j})$ is assumed to be uniform. Which, cannot hold true during real interactions. Therefore we focus our attention to the **Plain Card** implementation of the card model, and estimate the various costs and benefits to estimate the optimal EPU given a fixed capacity constraint instead.

2.7.3 Ranking Criterion

The utility can be extended for a result list as $\boldsymbol{L}$ [15]:

$$\begin{aligned} EPU\left(\boldsymbol{L}\right) \quad &= \sum_{i=1}^{n}\left(\prod_{j=1}^{i-1}\left(1-P(R)_j\right)\right) EPU_{\text{card}}(i) \\ \text{subject to} \quad &1 \leq W \leq M \end{aligned} \tag{2.28}$$

Here, $P(R)_j$ represents the relevance probability of the result item, **W** is the space occupied by the result card and **M** is the total units of screen space available, and **n** is the total number of results in the list L.

2.7.4 Limitations of the card model

The Card Model presents a novel analogy, likening the IIR process to a card game played between a machine and a human user. The primary objective of this model is to maximise the user's benefits derived from the interaction. The Card Model posits that each piece of information presented to the user is akin to a card in a game, with the strategic presentation of these cards aimed at optimising the user experience.

In the context of this model, the concept of a "Plain Card" emerges as a significant element. It represents a simplification of the Card Model by relaxing certain assumptions related to user browsing behaviour, particularly the assumption of sequential browsing. However, a notable gap in the literature is the lack of a practical implementation of the Plain Card for document ranking, which hampers the ability to empirically estimate the various benefits and costs associated with it, particularly in terms of EPU.

To address this gap, the subsequent section proposes a methodological framework

for estimating these parameters, thereby facilitating a practical application of the Card Model. This approach involves a detailed exploration into the mechanics of estimating EPU within the context of the Card Model. By doing so, we not only operationalise the Card Model in a practical setting to re-rank documents but also to provide empirical insights that could answer overarching research questions in the field of Interactive Information Retrieval.

2.8 Summary

In the background section of this thesis, we started with a high-level exploration of the evolution of IR systems, tracing their origins from physical libraries to the advent of the web. This evolution catalysed the democratisation of information sharing and precipitated the development of search systems. A focal point of this section was the research dedicated to optimising result ranking to provide users with relevant information. We delved into the complexities of ranking, illustrating how the ostensibly simple concept of relevance became intricate in practice. The challenge of ordering results in descending order of relevance, in accordance with the PRP, was far from straightforward. Various interaction costs and assumptions regarding user interaction with search results are essential in developing an accurate model.

A critical assessment of the assumptions and limitations inherent in both system-side and user-side approaches led to the understanding that ranking principles must integrate the presentation of results to effectively optimise a ranked list. The multi-layered nature of result presentation, encompassing aspects from layout to query costs, was explored in detail. Each layer had been independently studied, shedding light on the complexity of this domain.

We emphasised the dual nature of optimising search results, underscoring the necessity to consider both ranking algorithms and the presentation of results. We examined methodologies from both perspectives - the system (ranking) and the user (presentation). This exploration revealed that while there were direct approaches to optimise search results, they were computationally intensive and reliant on substantial interaction data, which might not be universally available.

Subsequently, we introduced the Card Model. In the Card Model, the interaction process is modelled as a game and users incur benefits and costs while interacting with SERPs. We observed that the Card Model is a robust implementation of the iPRP, but the hurdle of estimating its parameters and ranking documents in a practical scenario remains. The iPRP, as suggested by Fuhr [10], included strategies like calculating benefits as "saved effort".

In upcoming sections, we will detail a methodology to estimate the parameters of the Card Model to use it for document ranking. We validate our assumptions through user studies to gather interaction data, subsequently applying this knowledge to optimise ranked lists, including presentation aspects, to tailor search results to user preferences.

Chapter 3

Experimental Design

Previously, we have examined different perspectives on modelling user interactions to optimise ranked lists. This examination included both system-based ranking principles and user-focused studies. From our findings, we have chosen to implement the Card Model, a specific application of the iPRP. In the current section, we will be demonstrating our methodology for estimating the parameters of the Card Model. This includes a detailed approach to calculating factors such as benefits and costs, which are essential for computing the expected perceived utility of displaying and ranking results through the Card Model. Following this, we will describe a general experimental setup. This setup is designed to solidify our theoretical findings, providing a practical framework for our user studies and ensuring the applicability of the Card Model to rank documents.

3.1 Implementing the Card Model

We now look at the implementation of the Card Model and how we can estimate EPU using it. We can recall the objective function of the Card Model from Equation 2.25 as

$$EPU_{\text{card}}(i) = \sum_{R_i \in \mathcal{R}} \sum_{A_{i,j} \in \mathcal{A}} P(A_{i,j}|R_i) P(R_i) \Big(B(A_{i,j}|R_i) - C(A_{i,j}|R_i) \Big) \tag{3.1}$$

We will now dissect the components of this equation in a step-by-step manner, elucidating how each component can be estimated. We begin with a typical search

scenario: a user is presented with a list of items displayed on distinct result cards. In this scenario, the user engages with each result card sequentially. The options available to the user at each card are twofold: (1) to **click** on the result card, thereby engaging with its content, or (2) to **skip** the result card and move on to the subsequent one. This sequential decision-making process aligns with the user model posited by the iPRP [10]. Furthermore, it is a foundational aspect of the Card Model [15] under the "plain card".

We use the following key symbols as shown in Table 3.1 and 3.2:

Symbol	Description
$\mathcal{A}$	Action space, containing actions: clicking, c and skipping, s.
$\mathcal{R}$	Relevance space containing R and $\bar{R}$.
$\mathbf{R_i}$	Relevance of an item i: relevant, R or non-relevant, $\bar{R}$.

Table 3.1: Description of Symbols in the Model (Part 1)

Symbol	Description
$\mathbf{P}(\mathbf{R_i})$	Probability of the relevance of item i.
$\mathbf{P}(\mathbf{A_{i,j}}\|\mathbf{R_i})$	Probability of taking action $A_{i,j}$ given R_i.
$\mathbf{B}(\mathbf{A_{i,j}}\|\mathbf{R_i})$	Benefit of taking action $A_{i,j}$ given R_i.
$\mathbf{C}(\mathbf{A_{i,j}}\|\mathbf{R_i})$	Cost of taking action $A_{i,j}$ given R_i.

Table 3.2: Description of Symbols in the Model (Part 2)

To estimate the EPU, we must first define the ***costs*** and ***benefits*** associated with each action, given the result card and associated document. We adopt the suggestion of [10], who proposed using *time* to represent both the benefit (time saved) and the cost (time spent). The rationale is that users invest their time to find relevant result items (a cost), and discovering relevant result items saves them time as they do not need to keep searching for the required information. The time taken for various actions is influenced by the relevance of the result items and the presentation of the result cards. This is just one method in which benefits can be computed. We can also incorporate other heuristics beyond dwell time such as mouse position and scroll behaviour and incorporate them into our action space for estimating the EPU [69], however, we leave that to be incorporated in the future.

Thus, we can calculate the expected cost and benefit based on the summation of the

item's relevance, $\mathbf{R}$ (relevant) or $\bar{\mathbf{R}}$ (non-relevant), given an action $\mathbf{A_{i,j}}$. The expected **benefit** of an action can be written as:

$$\mathbf{B}(\mathbf{A_{i,j}}) = P(A_{i,j}|R)B(A_{i,j}|R) + P(A_{i,j}|\bar{R})B(A_{i,j}|\bar{R}) \tag{3.2}$$

and the expected **cost** of an action can be written as:

$$\mathbf{C}(\mathbf{A_{i,j}}) = P(A_{i,j}|R)C(A_{i,j}|R) + P(A_{i,j}|\bar{R})C(A_{i,j}|\bar{R}) \tag{3.3}$$

Given our expressions for the expected cost and benefit, we can re-write the EPU of a card from Equation 2.25, where the action space is limited to clicking and skipping and the relevance is binary; for a given item as follows:

$$EPU_{\text{card}}(i) = \sum_{R_i \in \{R,\bar{R}\}} \sum_{A_{i,j} \in \{c,s\}} P(A_{i,j}|R_i)P(R_i)\Big(B(A_{i,j}|R_i) - C(A_{i,j}|R_i)\Big) \tag{3.4}$$

An open question now is: how to meaningfully estimate these costs and benefits in terms of time?

3.1.1 Estimating the EPU

Before we can exactly implement the iPRP via the Card Model, we still need to define how we estimate the costs and benefits of clicks and skips, as well as how we estimate the probability of relevance. Since we have two actions, clicking and skipping for both relevant and non-relevant items, we need to estimate the following, as seen in Table 3.3:

Action	Relevance	Benefit	Cost
Click	Relevant	$B(c\|R)$	$C(c\|R)$
Click	Non-Relevant	$B(c\|\bar{R})$	$C(c\|\bar{R})$
Skip	Relevant	$B(s\|R)$	$C(s\|R)$
Skip	Non-Relevant	$B(s\|\bar{R})$	$C(s\|\bar{R})$

Table 3.3: Benefits and Costs Terms

We will use the time spent to estimate the different costs and benefits. We will denote this time spent as T (measured in seconds), therefore, the

- cost to click a relevant item is $\boldsymbol{T(c|R)}$ and non-relevant item is $\mathbf{T}(\mathbf{c}|\bar{\mathbf{R}})$, and,
- the cost to skip a relevant item is $\boldsymbol{T(s|R)}$ and non-relevant item $\mathbf{T}(\mathbf{s}|\bar{\mathbf{R}})$.

For the benefits, we need to map the gain from a result, to be in the same units as the cost (i.e., in units of time). Therefore, we assume that users derive no benefit from choosing to "view" or "click" a non-relevant result. Also, skipping over a result whether it is relevant or not should yield no benefit. This is the assumption made in the iPRP. The logic behind it is that a benefit is estimated as time saved, and only a positive correct decision yields a benefit. We further elaborate on this in §3.1.1.

This leaves the final case when a user clicks on a relevant result. We consider that the time spent reading a relevant result $T(read|R)$ facilitates information acquisition, and thus aligns with the concept of time well spent [44]. We define our benefit from a relevant click to be the time required to read the result $\mathbf{B}(\mathbf{c}|\mathbf{R}) = \mathbf{T}(\mathbf{read}|\mathbf{R})$ (we describe below how to estimate this from our measurements) There are potentially other ways to map the gain of information to time or vice versa (e.g., [50,70]), however, we leave exploring such avenues for future work. Now, our updated table with the costs and benefits with our new time-based notation becomes:

Action	Relevance	Benefit	Cost
Click	Relevant	$T(read\|R)$	$T(c\|R)$
Click	Non-Relevant	0	$T(c\|\bar{R})$
Skip	Relevant	0	$T(s\|R)$
Skip	Non-Relevant	0	$T(s\|\bar{R})$

Table 3.4: Updated Benefits and Costs

We can now expand Equation 3.4 as:

$$\begin{aligned} EPU_{card} =& P(c|R)P(R)\Big(B(c|R) - C(c|R)\Big) + P(c|\bar{R})P(\bar{R})\Big(B(c|\bar{R}) - C(c|\bar{R})\Big) \\ &+ P(s|R)P(R)\Big(B(s|R) - C(s|R)\Big) + P(s|\bar{R})P(\bar{R})\Big(B(s|\bar{R}) - C(s|\bar{R})\Big) \end{aligned} \tag{3.5}$$

We can further cancel out the 0 terms per Table 3.4 as:

$$EPU_{card} = P(c|R)P(R)\Big(B(c|R) - C(c|R)\Big) + P(c|\bar{R})P(\bar{R})\Big(\cancelto{0}{B(c|\bar{R})} - C(c|\bar{R})\Big)$$
$$+P(s|R)P(R)\Big(\cancelto{0}{B(s|R)} - C(s|R)\Big) + P(s|\bar{R})P(\bar{R})\Big(\cancelto{0}{B(s|\bar{R})} - C(s|\bar{R})\Big) \quad (3.6)$$

Further, substituting the terms from Table 3.4, our final expanded EPU equation can be written as:

$$EPU_{card} = P(c|R)P(R)\Big(T(read|R) - T(c|R)\Big) - P(s|R)P(R)T(s|R)$$
$$-P(c|\bar{R})P(\bar{R})T(c|\bar{R}) - P(s|\bar{R})P(\bar{R})T(s|\bar{R}) \quad (3.7)$$

We can observe that we still need to estimate some probabilities, such as the probability of relevance ($P(R)$) and also the probability of clicking or skipping a relevant and non-relevant item ($P(c \text{ or } s|R \text{ or } \bar{R})$)

Estimation of time

For each of the card types, given the relevance, we can calculate the average time (in seconds) to click or skip the card.

- $\mathbf{T(c|R)}$ and $\mathbf{T(s|R)}$: For each result presented on the SERP, we can measure the time from when the result appeared until the user either clicked the "view" or "skip" button next to the result, respectively. This will work while collecting annotations one by one, however, while the results are presented on a SERP, can use the time spent looking at the item via mouse hovers. We provide a method and justification for this decision later in this section on the experiment design. We can then compute the average time across all results and users for each result card type, taking into account its relevance.

- $\mathbf{T(read|R)}$: For each user, given a relevant result and a card type; we can measure the time spent reading the result. Then, for a given user we compute the maximum reading time for each card type. For computing the average user, this

can be further averaged out by computing the average maximum reading time across all users for a given card type. This approach of maximum reading time ensures that spending less time on a relevant result does not negatively impact the utility value (via lowered benefits). For example, if different results are presented in the same card type, depending on the density of the information in the result it may take longer or shorter amounts of time to read it. This could potentially mean that longer reading times would give more benefits. Therefore, if we cap and fix the benefit per card type to the maximum time to read the result, quickly reading a relevant result will not give a small benefit to one card type or vice versa. Thus, in our benefit computation, we account for different reading speeds, lengths and comprehension of information in the result by taking the maximum time to read the result per card type.

A deeper understanding of Benefits

In the iPRP, the concept of "benefit" is described as the time saved by a user in locating relevant information. This is predicated on the assumption that when a user identifies an item of relevance, they effectively acquire the sought-after information, thereby circumventing the need for further search. This acquisition, consequently, manifests as a benefit, optimising the user's efficiency in information retrieval.

A seemingly paradoxical aspect emerges when considering the time spent on relevant documents to estimate this benefit. Conventional logic might suggest that prolonged engagement with a document is not necessarily *more* beneficial. Contrarily, in this context, extended interaction with a relevant document is classified as "time well spent". This designation arises from the fact that such engagement, albeit time-consuming, significantly contributes to the user's primary objective of information acquisition. Thus, the time invested in reading and understanding relevant content is not merely expended but is rather an integral component of the user's information-seeking endeavour.

The intrinsic value of the information, in this framework, remains a constant entity. The crux of the matter lies in the relevance of the information to the user's specific needs. When information aligns with these needs, the duration involved in its acqui-

sition is regarded as beneficial. Conversely, time spent on irrelevant information gives the users nothing, thus yielding them zero benefit, and instead costing them the time they spent on the item

This approach acknowledges the diversity in users' reading speeds, comprehension levels, and depth of interest in the information. Such differences do not result in penalisation within this model. Instead, the model accommodates varying durations of engagement with relevant information, ensuring that the end goal, which is successful information acquisition, is not undermined by these variations.

Furthermore, it is important to address the differentiation in benefits accorded to various card types. In our benefit computation for a relevant item, we impose an upper limit on the quantifiable benefit, which varies across different card types. This variation is essential because distinct card types engender differing user expectations. For instance, a title card may only provide a cursory overview, prompting further exploration of the document for substantive value. In contrast, a detailed card (thereby needing more time to read the card itself) could offer comprehensive information at the outset, potentially diminishing the necessity for extensive document perusal. The time spent by users on these different card types, therefore, serves as an indicator of the perceived utility of the information. Extended engagement with a detailed card might signify the exceptional utility of the content, whereas prolonged interaction following a title card could imply a need for additional time to contextualise and comprehend the content due to the initial lack of detailed information.

Thus, the card type influences the user's initial perception, expectation, and approach to the content. Depending on the card type, the time spent reading can reflect different levels of engagement, effort, and perceived value, leading to varying benefits, between different users, which cannot be compared directly.

Estimating Probabilities

We can observe that we still need to estimate some probabilities, such as the probability of relevance ($P(R)$) and also the probability of clicking or skipping a relevant and non-relevant item ($P(c \ or \ s|R \ or \ \bar{R})$). To estimate the interaction probabilities, we can

count the number of times an item was shown and how often it was clicked or skipped, given its relevance score and card type. We can then use the maximum likelihood estimation to calculate the probability of each type of interaction occurring for each card type.

To estimate the probability of relevance in our analysis, we employ the BM25 retrieval function with $\beta = 0.75$. Given that BM25 yields an unbounded retrieval score, it is necessary to convert it to a probability. Following the approach from a related study [17], we use a set of previously submitted queries used on our indexed test collection (TREC WaPo) and issued them to our retrieval engine. For every query variation, the top 50 documents are selected. This way, we get a large range of BM25 scores, and then we can later use a function to map them to a probability of relevance. Then, across all the documents retrieved (per query), the BM25 scores are normalised using z-normalisation, and subsequently mapped to a range of [0-1] through a logistic curve transformation (for all documents across all topics). A regression model can then be constructed, to predict the probability of relevance based on a BM25 score. However, it is worth noting that perfecting this model is not the primary focus of this thesis. While we acknowledge that there are other methods to directly get a probability estimates from for example, neural rankers, we do not focus on those in this thesis due to their high variability

To estimate the interaction probabilities, we will count the number of times an item was shown and how often it was clicked or skipped, given its qrel value and card type. We can then use a maximum likelihood estimation to calculate the probability of each type of interaction occurring for each card type, for each user.

3.1.2 Limitations to Our Implementation

While our implementation of the Card Model in this thesis offers a novel approach to quantifying user benefit for EPU, it is designed with certain simplifications. Notably, the model currently operates on a structured framework where the benefit is fixed based on card types, with a capped maximum benefit for each type. This approach addresses potential non-linearities in how different users derive value from time spent on various

card types and the associated content to some extent.

By capping the benefit for each card type, the model acknowledges that different documents, regardless of their complexity and information density, can provide the maximum benefit if they are presented within the same card type. For instance, a user spending an extended period on a high-density document or a shorter period on a more straightforward, less dense document can receive the same level of benefit, provided the content is relevant and effectively engages the user. This system ensures a level of equity in benefit distribution across different document types and user interactions.

However, it is important to note that while this approach simplifies the benefit calculation and addresses some aspects of non-linear benefit acquisition, it may not capture all the subtleties of user interactions with different types of content. The focus on a standardised maximum benefit for each card type is a deliberate choice to maintain simplicity and manageability in the initial stages of testing the concept of perceived utility. Future iterations of this model could explore more nuanced approaches that further account for the complexities of user behaviour and information utility in IR systems.

3.2 Proposed Data Collection Mechanism

In the previous section, we laid out the key parameters that we need to estimate for implementing the Card Model. Therefore, now we will offer a comprehensive overview of the methodology that will be employed in conducting the experiments presented in this thesis. Using this general methodology we will address our overarching research questions, grounding them in empirical evidence derived from user interaction data obtained through user studies. We built our retrieval system based on news search, below we describe how we implemented it for collecting our data, we will dive into details of several aspects of our methodology such as:

1. The development and components of the news search system, encompassing the dataset and the topics integrated within the system.

2. An exploration of the variety of result cards utilised in the context of news search.

3. A detailed framework of the user study, segmented into two distinct parts: The first segment details the methodology for gathering annotation data pertinent to the initial research question, including the explanation of the interface and the progression of the user study. The subsequent segment elucidates the system and procedure developed for the second and third user studies, which are directed towards the remaining research questions.

4. An analysis of the extracted data, focusing on user behaviour, system performance, and user satisfaction metrics.

3.2.1 The News Search System & Interface

We anchor our experimental framework within the context of news search. To this end, we utilise a corpus comprising widely accessible news articles from the Text Retrieval Conference (TREC), specifically, the Washington Post corpus (WaPo). This choice is predicated on the relevance of this corpus to the context of news search. In line with the work of [71], who demonstrated the efficacy of simulated work tasks in closely mimicking real-world search scenarios, we adopt a similar approach. These simulated tasks are designed to emulate the search patterns in journalistic search environments, thereby providing realistic interaction contexts with the retrieval system.

Our experimental design includes two primary types of work tasks tailored to the research context. The first task addresses our initial research question (**HL-RQ1**), which seeks to collect interaction data to discern differences among various types of result cards with respect to EPU. Here, participants are tasked with evaluating the relevance of documents based on how the information is presented, related to predefined topics and queries. The second and third research questions (**HL-RQ2** and **HL-RQ3**) delve deeper into the relationship between presentation and performance, and strategies for optimising ranked lists based on EPU. To investigate these aspects, we introduce a more sophisticated work task. Participants are instructed to assume the role of journalists, tasked with identifying and saving documents that are relevant to the selected topics. We describe this task in greater detail below.

Document Corpus

Given the context of news search for our experiments, we used the TREC Washington Post Corpus (WaPo) collection from the TREC Common Core 2018 track [1]. The WaPo collection consists of 608,180 news articles and blog posts published between January 2012 and August 2017 categorised into 50 topics for information retrieval tasks. This collection provides a diverse range of topics for analysis and experimentation, allowing us to explore the effectiveness of our proposed approach across different topical themes.

The Retrieval System

The indexing of the Washington Post (WaPo) corpus was conducted using a pre-existing program that incorporates the Whoosh Information Retrieval (IR) toolkit to index documents [72]. This program, incorporates several optimisation steps for efficient indexing. These include the implementation of Porter stemming for stop-word removal and the capability to download and re-scale images within the document collection. During the indexing phase, certain challenges were encountered, notably the presence of duplicate documents. These duplicates were identified by identical titles and contents, differing only in their document identifiers (doc ids). To refine the dataset, duplicates were systematically removed (using cosine similarity score of > 0.95), with preference given to retaining the versions exhibiting more recent timestamps.

For the purposes of our user studies, which necessitated ground truth relevance judgements (derived from the Qrels file), the indexing process was confined to documents corresponding to the 50 topics featured in the Qrels file and possessing relevance judgements. This resulted in the creation of an index approximately 1.5 GB in size, with the downloaded and re-scaled images contributing an additional 8 GB of storage space. In total, 1321 documents were indexed, all retrievable via the Whoosh system. A specialised framework, developed by [73] was utilised for query issuance and document retrieval. This framework is grounded in the Whoosh IR toolkit, employing the BM25 ranking algorithm with the parameters $\beta = 0.75$ and $k_1 = 1.2$ In terms of query formulation, all terms were conjoined using the logical AND operator, as is common

[1] https://trec-core.github.io/2018/

in many retrieval systems which implicitly apply this conjunction. This approach is exemplified in platforms like Google News, where a search for "tropical storms" results in a query structured as "tropical" AND "storms"

Topics

From the 50 topics available in the TREC WaPo collection, we selected four topics for our studies:

1. **Topic 341:** Airport Security,
2. **Topic 363:** Transportation Tunnel Disasters,
3. **Topic 367:** Piracy at Sea and,
4. **Topic 408:** Tropical Storms.

Table 3.5 shows the distributions of documents in our indexed documents. The four topics, along with a short description of what constitutes a relevant document, are listed below. These summaries are derived from the TREC WaPo topic descriptions.

1. **Topic 341: Airport Security** This topic concerns the effectiveness of efforts to better scrutinise passengers and luggage on all flights, particularly international ones.

   Relevant Document Criteria: A document is relevant if it reports on steps taken by airports worldwide to improve passenger and luggage scrutiny on domestic and international flights. Articles should discuss increased airport security measures in response to terrorism concerns, specifically those that go beyond normal passenger and carry-on screening methods. Examples of new steps include additional personnel, automated screening processes, sophisticated monitoring and screening devices, whole body imaging techniques, and extraordinary measures for screening luggage in the baggage compartment.

2. **Topic 363: Transportation Tunnel Disasters** This topic focuses on disasters that have occurred in tunnels used for transportation.

Relevant Document Criteria: A relevant document identifies a disaster in a tunnel used for trains, motor vehicles, or pedestrians. The disaster could be caused by fire, earthquake, flood, or explosion, whether accidental or planned. Documents discussing tunnel disasters during construction are also relevant if lives were threatened. However, incidents involving wind tunnels or tunnels for wiring, sewage, water, oil, etc., are not considered relevant.

3. **Topic 367: Piracy at Sea** This topic addresses modern instances of old-fashioned piracy, involving the boarding or taking control of boats.
   Relevant Document Criteria: Documents discussing piracy on any body of water are relevant. However, documents discussing the legal taking of ships or their contents by a national authority, or clashes between fishing vessels over fishing rights, are not relevant unless one vessel is boarded.

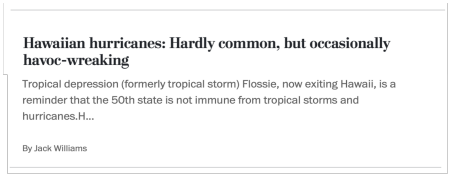
4. **Topic 408: Tropical Storms** This topic examines tropical storms (hurricanes and typhoons) that have caused significant property damage and loss of life.
   Relevant Document Criteria: Documents are relevant if they detail the date of the storm, the area affected, and the extent of damage/casualties. Documents that describe the damage caused by a tropical storm as "slight", "limited", or "small" are not relevant.

Topic	**Total**	**Non-Relevant**	**TREC Relevant**		
			Somewhat	**Definitely**	**Total**
341	390	102	160	128	288
363	357	113	55	189	244
367	276	74	106	96	202
408	298	99	20	179	199

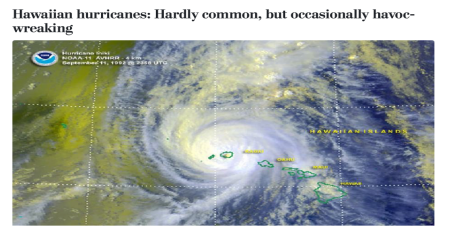
Table 3.5: Document relevance distribution across selected TREC WaPo topics

Card Types

In our thesis, we experimented with six different layouts for news result cards, which are shown in Figure 3.1. These layouts vary not only in design but also in the type and



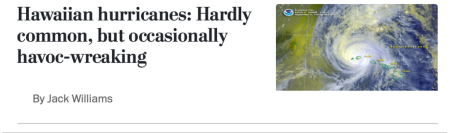
(a) Title+Summary (TS)



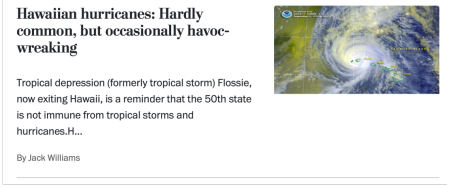
(b) Title+Image+Summary (TIS)

(c) Title Only (T)

(d) Title+Image (TI)



(e) Title+Summary (Google)



(f) Title+Image+Summary (WaPo)

Figure 3.1: The six different card types used in our experiments

amount of information they display. For instance, some layouts include images, while others contain summaries. This leads to each layout occupying a different amount of screen space. We used the Bootstrap framework to help estimate the column widths for these cards. While Bootstrap is useful in estimating column widths, it has no set definition for row space. Therefore we approximate how space to be approximately 100px for a single row. For styling, we followed the Washington Post's CSS guidelines, setting title fonts at 14pt and summary fonts at 12pt. In keeping with the Washington Post's format, our result summaries were limited to the first 250 characters, mirroring the approach used by the news site at the time of our study. This decision was made to maintain a realistic and relevant user experience in our experiments. We assume that images are relevant to the content of the document since these images are directly pulled from the Washington Post.

The card types illustrated in Figure 3.1 is a result displaying a document from Topic 408: Tropical storms, using the query "Tropical Storms". The images in the result have

been re-scaled to different sizes. We can observe two cards have been repeated, the TS card and the TIS card. We replicated these cards to fit the standard cards displayed on the actual Google News and Washington Post websites to use as comparisons to different row sizes and to also have a comparable baseline later on for our user studies.

3.2.2 Experimental Flow & Interfaces

This subsection is structured into two distinct parts, aligning with the bifurcated nature of our experimental approach. The first part (Annotation App) delineates the development of an annotation interface, instrumental in addressing **HL-RQ1**. The second part, conversely, elucidates a general methodology to facilitate the experimental flow, aiding in the resolution of **HL-RQ2** and **HL-RQ3**. While we later detail specific adaptations of this methodology for **HL-RQ2** and **HL-RQ3**, the foundational approach remains consistent throughout.

Annotation App

In the first part of our study, we developed an interface for the collection of annotations to answer **HL-RQ1**. The design of the study was a within-subjects design. A within-subjects design allows us to expose one participant to all conditions while also requiring fewer participants. We need every participant to see every condition since we are interested in calculating the EPU on a per-user basis. Prior to commencing the experiment, participants were required to sign an electronic consent form and were offered a practice topic to familiarise themselves with the experimental procedure.

The result cards utilised in this study were, 1: TS, 2: TIS, 3: T and 4:TI. After completing the practise topic, participants were presented with results displayed in various result card styles, one at a time in a sequential manner. This is a simplified browsing assumption, however, it is the assumption used in the iPRP and will help us to isolate the timing to the presentation more accurately. Upon seeing a result, participants had the option to either click the “view” button to indicate that they would read it or click the skip button to indicate that they were “non-relevant”. In either case, clicking the buttons moved the participant to the next result to annotate.

The selection of both results and result cards was randomised without replacement from the TREC document pool, aiming to minimise any potential order effects. After every 10 annotations, participants were asked if they wished to continue with the experiment or to exit. This process is explained in Figure 3.2.

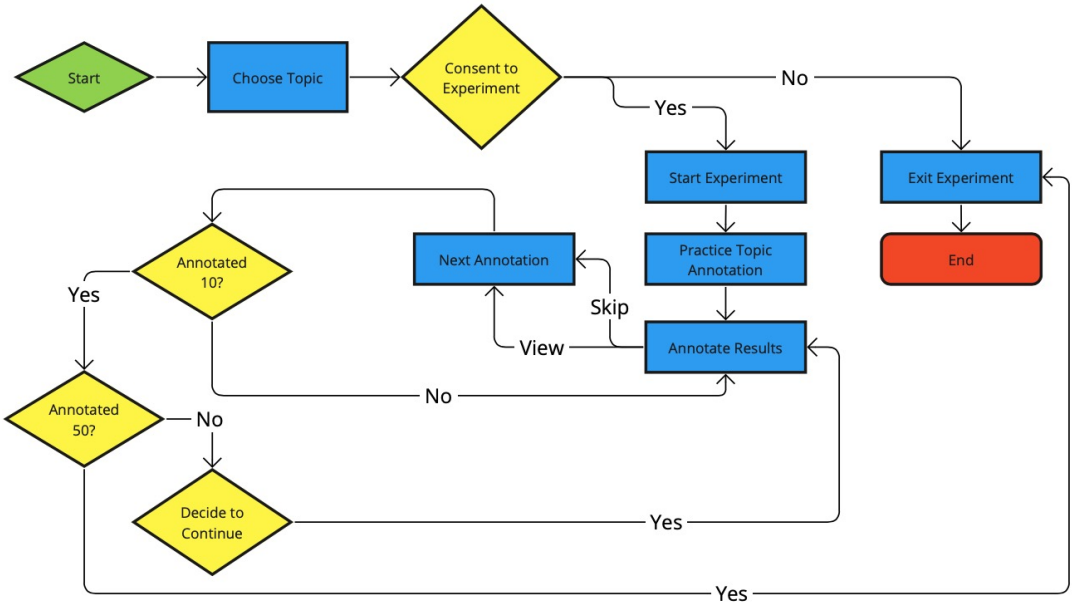


Figure 3.2: Annotation study general procedure

Figure 3.3 provides a visual example of the interface used by participants during the annotation process. This interface primarily allowed users to judge the relevance of documents based on their presentation. The figure includes an example of the instructions displayed to participants above the news results. Since we are implementing the Card Model with respect to time, we track the time taken for various actions, such as clicking the view or skip buttons and reading time. Data collected during this phase was systematically stored in a database and subsequently downloaded as a CSV file for comprehensive analysis.

News Search App

To address **HL-RQ2** and **HL-RQ3**, we devised a novel experimental search interface. To answer the RQs and, we needed to gather more realistic interaction data and thus,

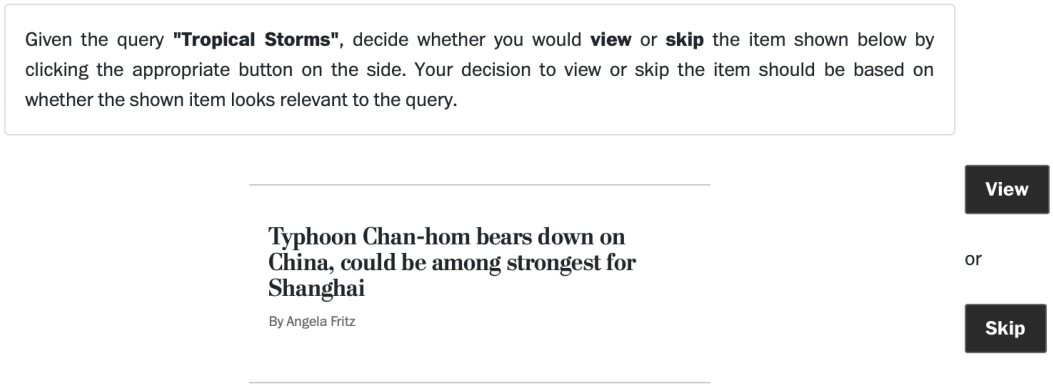


Figure 3.3: Example of the annotation interface

this interface was essential for allowing users to select a query, view the SERP, and interact with, as well as save, relevant documents. A significant aspect of this study involved gathering user satisfaction data, which would enable to study the user-sided aspects

The designed SERP interface encompasses multiple components: a query view, the SERP view (comprising result cards in various layouts), and a document view. For our experiments, we utilised the following result card formats:

1. Title + Image + Summary [TIS]
2. The Washington Post Style, Title + Image + Summary [TIS WaPo],
3. Google News Style, Title + Image[TI]
4. Title only [T]

These result cards could then be arranged into distinct layouts. Each layout could either comprise a single type of result card or a random combination of all types. Prior to participating in the study, participants were presented with an on-screen information sheet detailing the study procedure and required to give their informed consent.

The experimental objective was to conduct a news search task centred around specific topics. Participants were instructed to identify and select documents pertinent to a

predetermined topic by utilising a set of predefined queries. To develop these queries, we employed the methodologies delineated in [17], resulting in generating multiple queries for every topic. We computed the distributions of nDCGs for every topic and based on the observed nDCG distributions across all topics, we chose to stratify the queries into three tiers. The categorisation of queries was as follows: low (nDCG ranging from 0.1 to 0.2), medium (nDCG between 0.2 and 0.6), and high (nDCG exceeding 0.6).

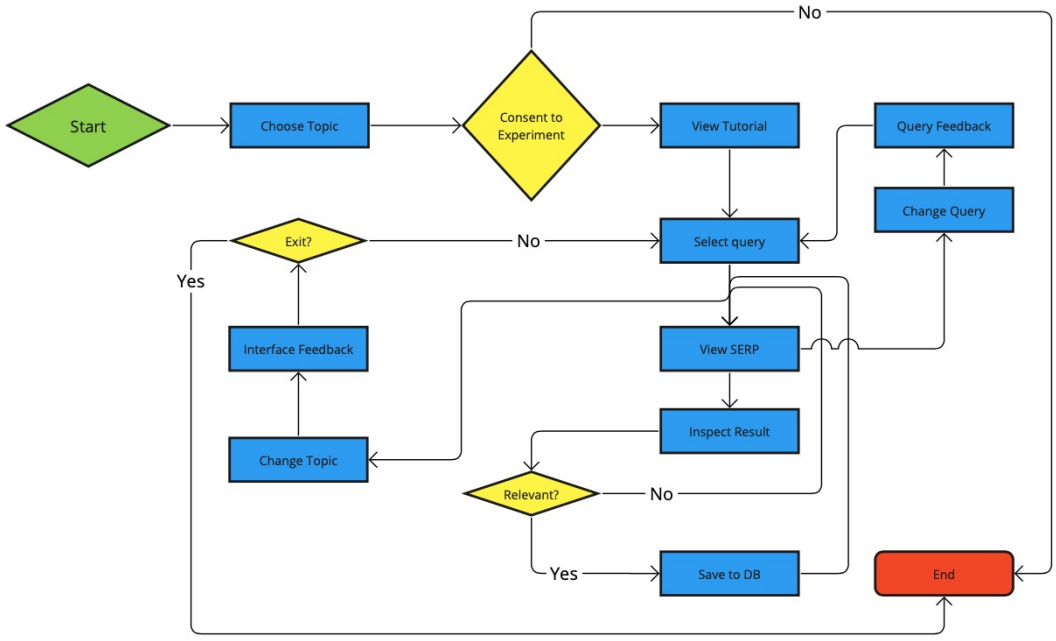


Figure 3.4: General study procedure for experiment 2 and 3

Same as before and observing Figure 3.4, after providing consent and prior to initiating the experimental tasks, participants were acquainted with a tutorial. This tutorial was designed to elucidate the task methodology and to navigate the user interface of the study. Upon starting the experiment, participants were presented with a 3x2 grid of six queries corresponding to their selected topic, arranged in a randomised order, as illustrated in Figure 3.5. Participants could choose any query to inspect and explore the associated results with that query to find the relevant documents.

The relevance evaluation criteria for documents were displayed in a floating instruction box on the left side of the screen, detailed in § 3.2.1. Upon selecting a query,

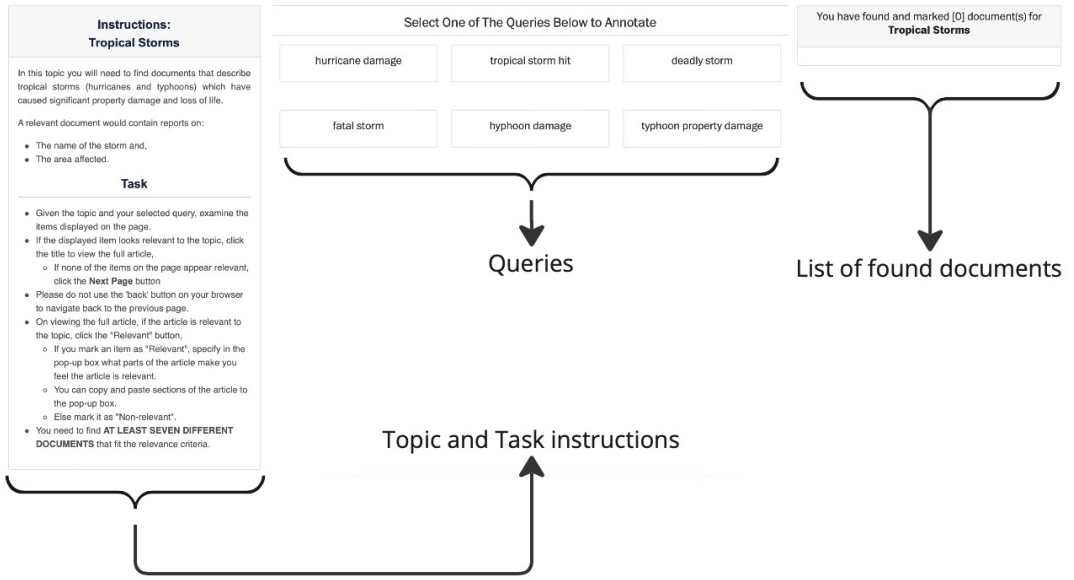


Figure 3.5: Example of the query view, where participants are annotating documents for the topic: "Tropical Storms"

participants viewed the documents associated with that query in one of the predefined or generated SERP layouts, as seen from Figure 3.6

Participants could inspect individual result cards for detailed document information, as shown in Figure 3.7. Marking a document altered the title colour of the corresponding result card to purple, indicating that the document had been inspected, as seen in Figure 3.8

Data on query satisfaction was collected as participants navigated between different queries within the same topic. The query selection view also featured a sidebar displaying titles and sections of documents marked as relevant, aiding participants in tracking their progress. Upon completing a topic, participants provided feedback on interface satisfaction, as elaborated in § 3.2.4.

3.2.3 User Recruitment & Ethics Considerations

Participants for our experiments were recruited from the Prolific platform, with a specific focus on individuals who indicated proficiency in English and were residents of the

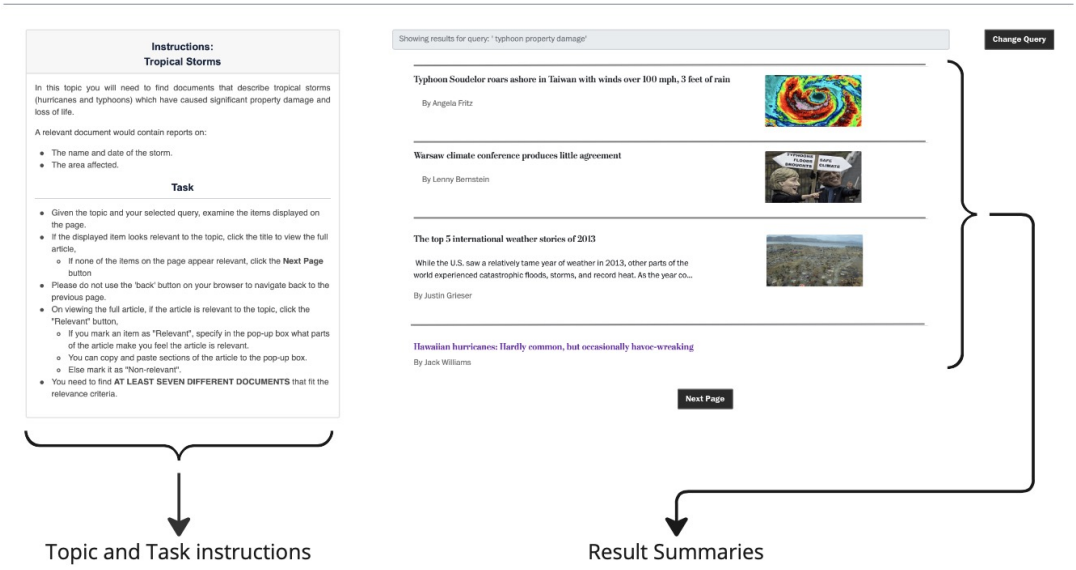


Figure 3.6: Example of the SERP view

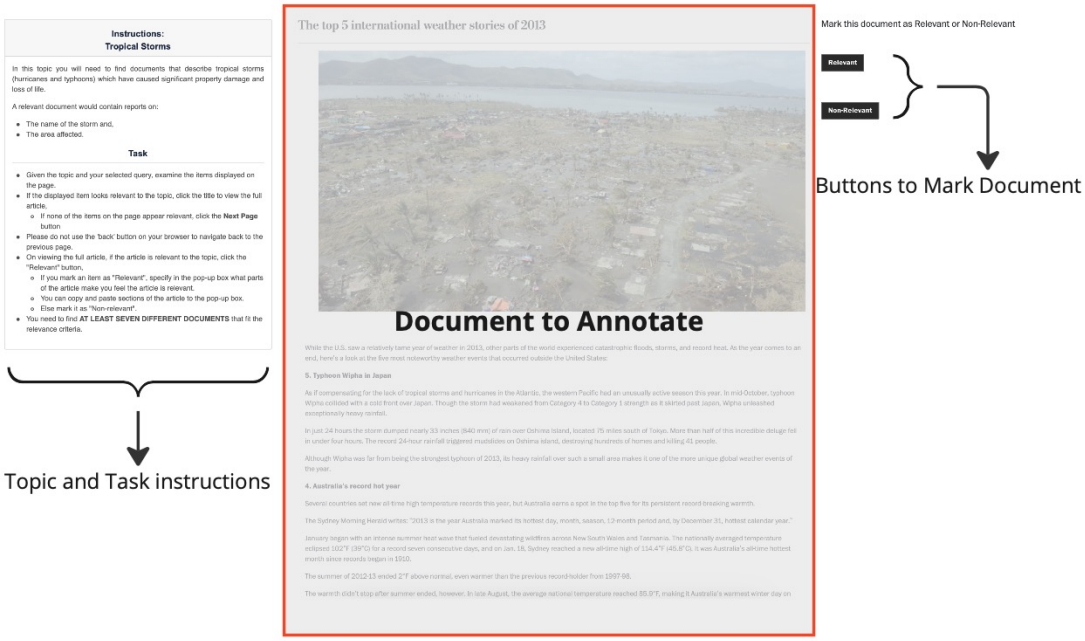


Figure 3.7: Document view example

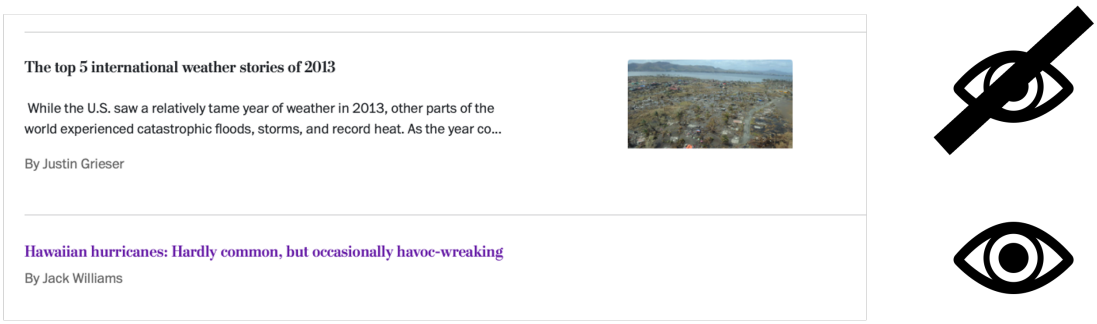


Figure 3.8: An example of a previously un-viewed vs viewed document. The document that has been viewed has its title colour change to purple.

United Kingdom or the United States. This criterion was established to ensure consistency and to mitigate the potential confounding variable of varied English language interpretation in our user study. Additionally, we set forth technical requirements for the devices used by participants. Specifically, participants were required to use a desktop or a laptop equipped with a mouse. This requirement was important as our study was conducted remotely, and we utilised mouse movements as a proxy measure for eye-tracking.

To recruit the appropriate number of participants, we refer to this power analysis chart from Figure 3.9.

3.2.4 Data Extraction

We split the dependent variables in our studies for **HL-RQ2** and **HL-RQ3** into three main categories: (a) search behaviours, (b) search experience and (c) performance:

Search Behaviours: To provide insights into user search behaviours we logged the number of...

1. ...queries clicked
2. ...pages viewed
3. ...documents viewed
4. ...documents saved

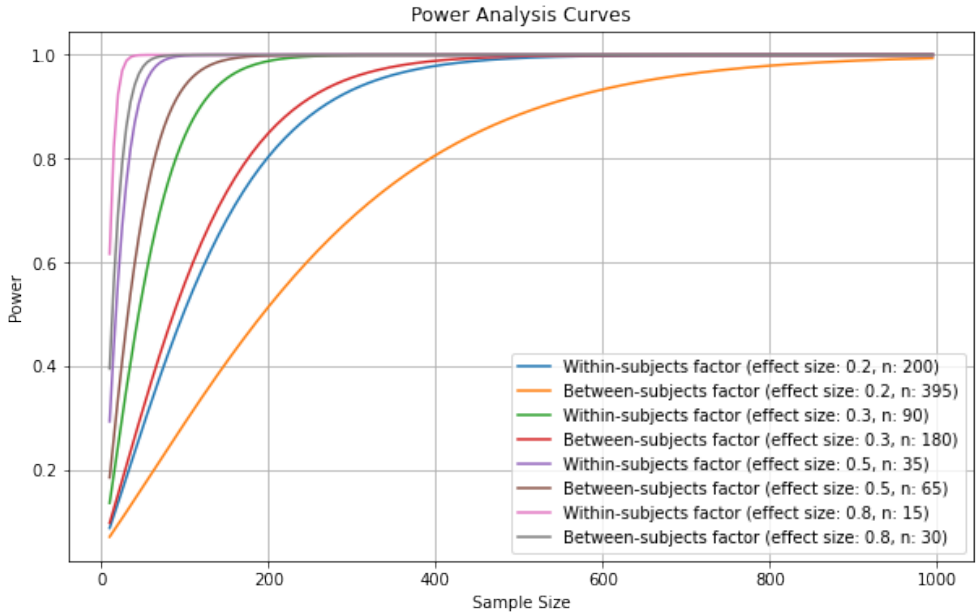


Figure 3.9: Power Analysis for within-subject and between-subject designs for different effect sizes

For relevant documents saved, we instructed the participant to record the relevant bits of the document into a text field that popped up if the user clicked on the "relevant" button on the document view page and saved it to our database. However, for **HL-RQ3**, we removed this requirement as, we increased the number of topics for the participants to complete and it would take substantially longer to complete the task by including this aspect, leading to fatigue and higher cognitive load. From the interaction logs, we could also compute the following time-based measures, including the time spent...

1. ... to complete the task
2. ... per result card (snippet)
3. ... on a relevant document
4. ... on a non-relevant document

The relevance and non-relevance of a document were obtained using the TREC WaPo Qrels for the retrieved documents. One thing to note is that, in our document index,

we only indexed documents which had TREC relevance judgements. For time spent on a snippet, we use aggregated mouse hover times as a proxy for eye gaze [74–78] computed with a modified lightweight JavaScript code [79].

Search Experience: We measured the search experience of the participant through a user satisfaction score. We collected user satisfaction at two levels: (a) the query level (collected after changing a query) and (b) the interface level (collected after every topic/task). For (a) query satisfaction, we collected data using a 6-point Likert scale by asking participants how satisfied they were with the results for that given query (with 1 being very dissatisfied to 6 being very satisfied). For (b) interface satisfaction, we asked participants whether they ...

1. ...felt **productive** using the system
2. ...found the interface layout to be **mentally taxing**
3. ...found the interface layout to be **engaging**
4. ...found the interface layout to be **distracting**
5. ...were **satisfied** with the interface layout,

on a 6-point Likert scale with 1 being strongly disagree and 6 being strongly agree, as can be seen from Figure 3.10

Performance: By using the TREC Common Core 2018 relevance judgements, we were also able to provide an estimate of search performance at the (a) system side and (b) user side. On the system side, for each query that was submitted by a participant, we evaluated the query's nDCG@10, Total gain on the Page and Precision@k (see §5.3.2 for further detail on total gain of page). For the user-side performance measures, given all of the documents that participants clicked on and saved, we could use the aforementioned relevance judgements as ground truth, allowing us to compute the accuracy of a participant's searching ability. This was summarised as the proportion of correctly identified relevant items saved (i.e., documents that are identified as relevant in the relevance judgements) vs. the total number saved.

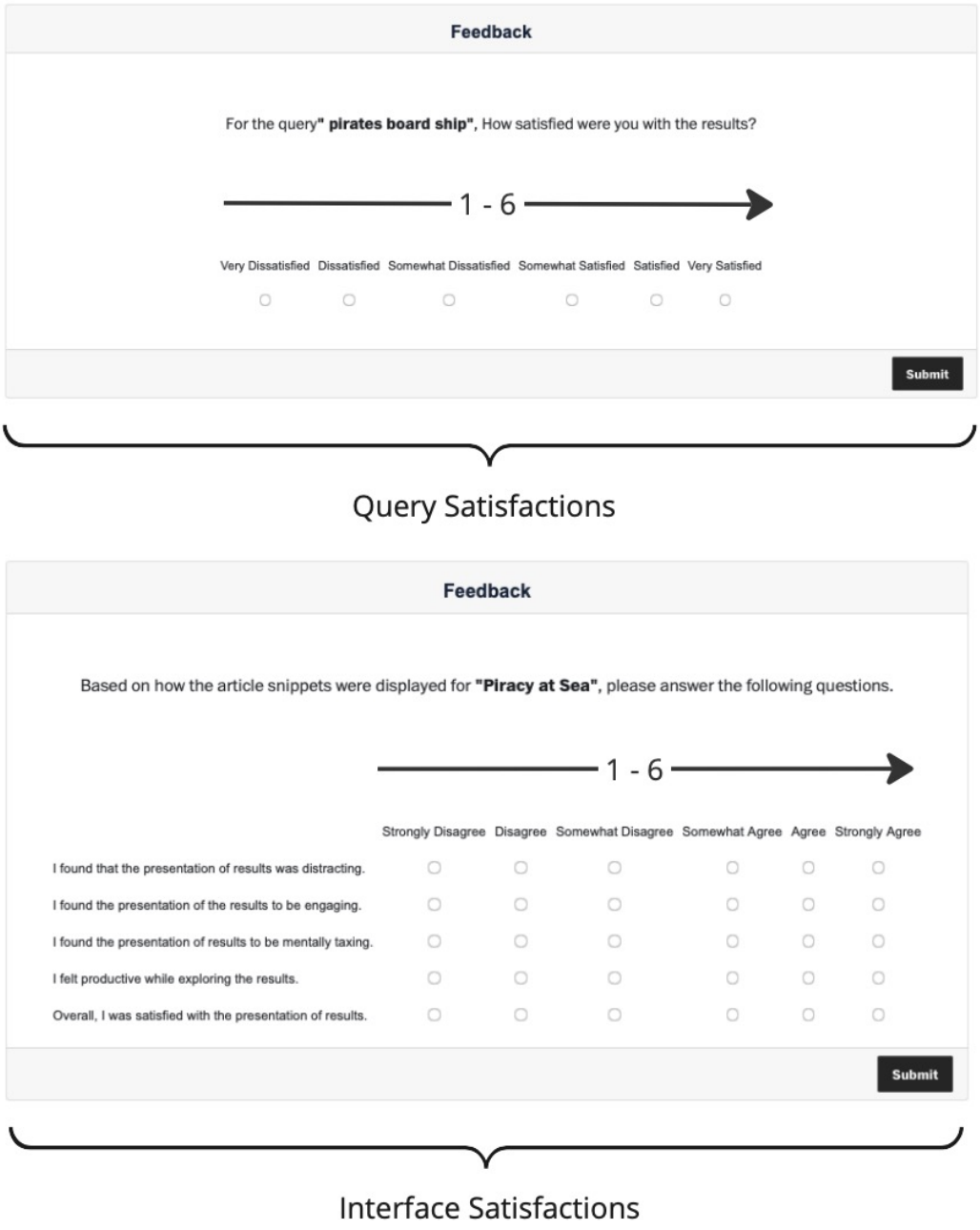


Figure 3.10: Example for the query and interface feedback within the topic of "Piracy at Sea"

3.3 Summary

In this methods section, we initially delineated the underlying assumptions of the Card Model, followed by a rigorous mathematical derivation of the EPU. We thoroughly examined the potential actions and features within our model, determining that benefits should only be computed for actions involving relevant results, underpinning our stance that user engagement with non-relevant items yields no benefits. To harmonise the framework, we proposed the use of time as a metric for EPU estimation.

Addressing our first high-level research question, we developed a methodology focused on discerning the differences between various card types, thereby establishing a simplified approach for initial data annotation to validate our assumptions. To tackle the more complex second and third research questions, we formulated a methodology to capture nuanced user interactions on SERPs. We established the framework for a user study to capture both user-side costs such as satisfaction and also system side costs by situating real users in a simulated work task.

For our user studies we also meticulously outlined the topics and tasks selected from the Washington Post (WaPo) corpus for the user study and included details on how we indexed them. We described in detail the different result presentation formats that were evaluated. This comprehensive groundwork sets the stage for the subsequent chapter, where we aim to explore the impact of result presentation on EPU and examine whether such perceived changes in utility could precipitate shifts in document ranking.

Chapter 4

Ranking Heterogeneous Search Results Pages Using the iPRP

4.1 Introduction

Remember that, in the information seeking and retrieval process, that the primary objective of search engines is to facilitate users in locating documents that are relevant to their information needs, as illustrated in Figure 2.1. This process typically involves the submission of multiple queries, the examination of numerous documents, and the assessment of the relevance of the documents retrieved [17]. To augment the search experience, it is imperative that the results are displayed in a manner that enables users to efficiently discern relevant information [23].

From the perspective of the user, investigation into studies reveals that variations in the design of result cards lead to differing levels of user satisfaction (see §2.5). Bearing this consideration in mind, it is noted that traditional result cards are characterised by a title, image, and summary. However, SERPs of the present day exhibit a diverse array of card types, encompassing images, data, and recommendations, among others (see §2.5.2)

Research centred on the user indicates that the configuration of result cards as well as the overall layout of SERPs exert a substantial impact on user interactions, as well as on their satisfaction and effectiveness in search tasks (see §2.5.1 and §2.5.3). This

body of work underscores the significance of thoughtful design in the presentation of search results, highlighting the role it plays in enhancing the overall search experience.

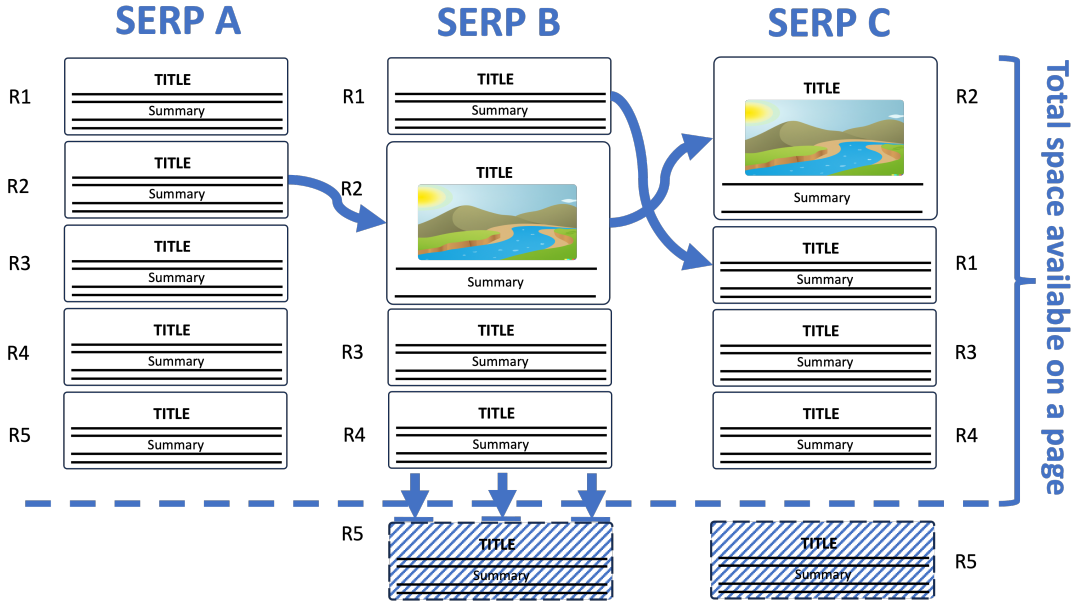


Figure 4.1: Compared to SERP A, only four cards can be shown above the fold (dotted horizontal blue line) on SERP B and C. However, changing the card type (e.g., TS to TIS) may also lead to changes in the ranking under the iPRP.

For example, in Figure 4.1, we can see three different SERP layouts: A, B, and C. In SERP A, all results {R1, R2, R3, R4 and R5} are presented using a title and summary (TS). So changing R2 to be presented with a title, image, and summary (TIS) means that it may:

- attract more (or less) clicks, thus changing its interaction probabilities,
- take more (or less) time for a user to decide if they want to click the result, or skip over it, thus changing its cost of interaction, and,
- occupy more (or less) screen space, resulting in a different number of results being displayed above the fold.

We saw at a high level that the Probability Ranking Principle (PRP) ranks results based purely on the decreasing order of their relevance [9] (see §2.4). Whereas, the iPRP incorporates interaction probabilities and the cost of processing each result card [10].

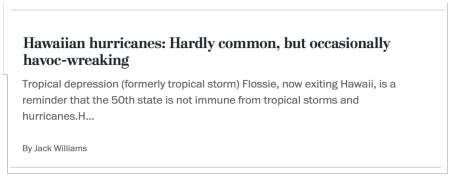
Such considerations might cause R2 to rank higher than R1 in terms of "Expected Perceived Utility" (EPU) under the iPRP, as demonstrated in SERP C. Moreover, the type of result cards can significantly influence the overall utility presented to users. Due to space constraints, different card types can alter the number of results displayed on the results page or above the fold, as exemplified by the 5 results in SERP A versus the 4 in SERP B and C. Consequently, adjusting the combination and type of result cards within a SERP introduces trade-offs between EPU, overall utility, and the number of results shown.

Given the potential variations in EPU, which can be attributed to different formats of result cards; this paves the way for addressing the primary research question (see **HL-RQ1**) regarding the assessment of EPU variation across different result card types. The differentiation in utility provided by distinct result cards catalyses the emergence of intricate sub-questions. Therefore, we dive into the potential shifts in rankings and performance outcomes precipitated by the application of the iPRP, particularly in the context of heterogeneous SERPs that include a broad spectrum of result card types. We conducted a within-subjects user study to answer the following research questions that will feed into understanding **HL-RQ1**:

- RQ1: What is the impact of different result cards on user behaviour?
- RQ2: How do the rankings from iPRP in heterogeneous SERPs contrast with those generated by the PRP?

4.2 Methodology

To explore the impact of the iPRP on ranking heterogeneous search engine result pages, we experimented with four different result card types (see Figure 4.2. These cards represent typical variations on SERPs. To ground our analysis, we gathered timing data and click data on these result cards across three topics from the TREC WaPo collection, employing 150 annotators. Following this, we utilised the annotation data to estimate interaction probabilities and timing components of EPU, leading to the determination of rankings under the iPRP using the Card Model, as detailed in § 4.3.



(a) Title+Summary (TS)

(b) Title+Image+Summary (TIS)

(c) Title Only (T)

(d) Title+Image (TI)

Figure 4.2: The four different card types used in this experiment.

(a) **TIS**: These cards displayed the article's Title, Image and Summary (representing the most attractive and informative result card, often used for promoted news articles).

(b) **TI**: These cards displayed the article's Title along with the Image (representing a similar result card to what is used on Google News).

(c) **TS**: These cards displayed the title and summary (representing the default result cards used by the Washington Post).

(d) **T**: These cards only displayed the title of the article (representing the sufficient headlines result card).

4.2.1 Topics

From the WaPo collection, as outlined in the methods chapter, we selected three topics for annotation (341: Airport Security, 363: Transportation Tunnel Disaster, and 408: Tropical Storms). See §3.2.1 for more details. This ensured we had a sufficient mixture

of relevant/non-relevant items to annotate and that we could render all card types. The images were downloaded and re-scaled so that images were of the same height and width.

4.2.2 Annotations

We used the interface we developed and described in §3.2.2 to collect annotations. We followed the same broad method to collect annotations where, given a description of the topic, the annotators were shown results styled as one of the different result cards. They were then given the option to click the "view" button (if they thought it was relevant), or skip the card (if they thought it was unlikely to be relevant). We recorded participants' actions (e.g., clicking, skipping) and the time taken to perform these actions.

4.2.3 Participant Demographics

In total, we collected 6,052 annotations from 150 participants (approx. 40 annotations per participant, approx. 10 annotations per result card per topic per participant.) The study sample comprised a near-equal gender distribution of 77 males, 73 females and two participants preferring not to identify with either gender. Participants ranged in age from 21 to 75 years, capturing a broad spectrum of adult age groups. Within this cohort, a minority of 14 individuals (9.3%) identified as students, while the majority, 136 participants (90.7%), were non-students.

A significant proportion of the participants, 83 individuals (55.3%), were engaged in either full-time or part-time employment. The remaining 67 participants (44.7%) were not involved in paid employment at the time of the study, which includes groups such as homemakers, retired, or disabled individuals.

Ethics approval (no. 1643) was granted via the departmental ethics committee at the University of Strathclyde for this task, and participants were compensated in line with national working wage guidelines in the UK, at the time of study (circa. 2021).

4.2.4 Estimating the EPU

Given our estimates of the different components (see §3.1.1 and §3.2.2), we calculated the expected perceived utility for a given result list (**EPU**(**L**)) (see Equation 2.28) for a list layout using different results card types for each result to determine how the rankings between the iPRP and PRP vary.

4.3 Results

Table 4.1 presents the timings, probabilities, and Expected Perceived Utility (EPU) for each card type, with timings measured in seconds. We use the methodology used for estimating the components from §3.1.1. From this data, we crafted a typical user profile, grounding it in the observed timings and interaction probabilities. For every TREC topic, we used the topic title as a query to fetch the top 20 results using BM25 ($\beta = 0.75$). We then calculated the EPU values for each card type, which are displayed in the table as reflective of our average user profile. To detect variations among the card types, we used one-way ANOVA tests. Post-hoc analysis was done with Tukey's HSD test. In the table, the mean $\pm$ standard deviation of the timings and probabilities are shown and significant differences ($p < 0.05$) are emphasised using superscripts.

4.3.1 RQ1: What is the impact of different result cards on user behaviour?

Our observations indicate that there are distinct variations in the timings and probabilities associated with different card types, and these differences are statistically significant. Referring to Table 4.1, we can see that integrating an image with a T card diminishes its EPU (T: 10.90, TI: 6.4). This trend implies that the addition of images can potentially divert or mislead users, thereby compromising their ability to swiftly discern relevant details. The interaction probabilities bolster this argument as they display a decreased probability of interacting with a relevant item when an image is included ($P(c|R)$ for TI: 0.78 ± 0.26).

Conversely, augmenting a T card with a summary enhances its EPU (T: 10.9,

Table 4.1: Components of the Utility Function, Probabilities, and Expected Perceived Utility (EPU) for All Card Types. Significant differences in values between the card types are indicated by a,b,c, or d in superscript. Where c=click and s=skip.

(a) Timing Data (T) for All Card Types.

Card Type	$T(s\|\bar{R})$	$T(c\|\bar{R})$	$T(c\|R)$	$T(s\|R)$
a. TS	$4.63 \pm 2.42^{c,d}$	4.41 ± 3.31	4.13 ± 2.32^{d}	5.49 ± 2.81
b. TIS	4.40 ± 2.19^{d}	$5.15 \pm 3.61^{c,d}$	$4.38 \pm 2.70^{c,d}$	$5.86 \pm 3.53^{c,d}$
c. T	3.58 ± 1.62	3.86 ± 2.42	3.64 ± 2.05	4.48 ± 2.23
d. TI	3.80 ± 1.67	3.72 ± 1.95	3.42 ± 1.64	4.43 ± 2.21

(b) Interaction Probabilities (P) and Expected Perceived Utility (EPU) for All Card Types.

Card Type	$P(s\|\bar{R})$	$P(c\|R)$	EPU_{card}
a. TS	0.69 ± 0.27	0.81 ± 0.25	$11.78 \pm 4.64^{c,d}$
b. TIS	0.73 ± 0.27	0.82 ± 0.23	$11.77 \pm 4.66^{c,d}$
c. T	0.68 ± 0.30	0.80 ± 0.25	10.90 ± 4.14^{d}
d. TI	0.73 ± 0.25	0.78 ± 0.26	6.40 ± 2.89

TS: 11.78). Summaries, especially when coupled with titles and images, potentially supply crucial context, enabling users to better assess the accompanying image. They also present an overview of the result's content, which aids users in determining its relevance and deciding about further engagement. Hence, while T cards enriched with summaries do incur a slightly higher processing time (T: 3.71 ± 1.95, TS: 4.23 ± 2.36), they demonstrate a reduced probability of mistakes, corroborated by the heightened $P(c|R)$ values (T: 0.80 ± 0.25, TS: 0.81 ± 0.25).

In the context of EPU, we discerned variations in the average EPU across card types. Explicitly, TIS and TS cards exhibit a superior EPU compared to T and TI cards. Our ANOVA test showed a statistically significant difference in EPU ($F(3, 3996) = 406.33, p < 0.05$). Post-hoc analysis via Tukey's HSD Test showed that TIS, TS, and TI cards possess an EPU surpassing that of T cards. Moreover, TIS and TS cards outperformed TI cards in terms of EPU. However, the distinction between TS and TIS in EPU was not statistically significant ($p = 0.129$).

While our study observed differences in EPU, timings, and interaction probabilities across card types, the probability of a user clicking or skipping a relevant item, intrinsic

to their behaviour, remained unaffected by the card type. This finding is consistent with [21], which reported no variance in click behaviour across diverse interfaces. However, user satisfaction did differ, highlighting individual differences in satisfaction preferences. Although the card type does not consistently alter click probabilities across all users, the EPU can encapsulate these individual variations by incorporating additional context such as the time required to process and read items. For instance, even if a particular card type inherently takes longer to process, it could still be more effective for some users due to their personal preferences or cognitive strengths. Such advantages, like lower error rates (higher $P(c|R)$), could counterbalance the longer processing times, resulting in a higher EPU for specific card designs, such as TS cards, for certain users. Given that these individual card features play a role, an overarching metric, the EPU presents a more holistic view. Our findings underscore that the card type significantly affects user interactions with search results. This raises a subsequent question of how mixing card types on a search results page influences the overall rankings, since changing the card type can change its EPU.

4.3.2 RQ2: How do the rankings obtained from heterogeneous SERPs differ compared to the PRP (in terms of performance)?

Table 4.2: Comparison of RBO, DCG of Page, and TBG for different card type combinations. Results show a statistically significant difference in RBO between different groups of combinations after running a one-way ANOVA of (F(7,31841)=2517.66, $p < 0.001$). "$\sim$" shows that there is no statistically significant difference with that row.

Combination Type	RBO	DCG of Page	TBG of Page
a. Baseline	1.000 ± 0.000	3.137 ± 1.625	3.073 ± 0.095
b. T or TI	$0.952 \pm 0.135^{\sim c}$	2.437 ± 1.405	1.960 ± 0.482
c. TIS or TS	0.951 ± 0.136	$2.437 \pm 1.407^{\sim g,b}$	$1.962 \pm 0.478^{\sim b}$
d. TIS or T	0.762 ± 0.251	$2.614 \pm 1.649^{\sim g}$	2.381 ± 1.130
e. TS or T	0.741 ± 0.222	$\mathbf{3.588 \pm 2.029}$	$\mathbf{4.363 \pm 0.916}$
f. Random	0.637 ± 0.291	$2.640 \pm 1.646^{\sim d}$	$2.413 \pm 0.784^{\sim d}$
g. TS or TI	$0.505 \pm 0.321^{\sim h}$	$2.525 \pm 1.636^{\sim b}$	2.318 ± 0.215
h. TIS or TI	0.501 ± 0.385	2.024 ± 1.384	1.595 ± 0.249

To explore the differences between ranking results by EPU and by EU (ordering with

the PRP), we ran a simulation using all 50 TREC WaPo topics. In this simulation, we used the EU from retrieved results with BM25 ($\beta = 0.75$) as our baseline, by retrieving the top 20 documents from every topic using the topic title as the query . We assumed that the default result card type was 'TS' and that a page could display up to 12 rows. Thereby creating a baseline ranked order similar to "n blue links". The core of our simulation involved altering this baseline according to the space constraint. Specifically, we selected every result in the list and changed its card type randomly to one of two possibilities, as illustrated in Table 4.2. For example, the first result might change to TIS, the second to TS, and so on. Since the result page is constrained to 12 rows, a page containing TIS and TS cards can have cards in the following combinations – TIS, TS, TS or TIS, TIS or TS, TS, TS etc. We repeated this random alteration 100 times for each result list combination type to observe how such changes impacted the ranking order. After applying these changes, we re-ranked the documents in the altered result list in decreasing order of EPU and then compared this new order with our baseline EU ranking. We used the Rank Biased Overlap (RBO) metric [80] to measure any changes in ranking order. Additionally, we looked at the DCG (of the page) and Time Biased Gain (TBG) metrics ($h = 224$) to see how different SERP layouts affected search result effectiveness.

Our results, presented in Table 4.2, show that adjusting the presentation of results via different card types to construct heterogeneous SERPs can change document ordering. The RBO metric can quantify this change, however, we acknowledge that RBO is opaque in the sense that it cannot tell us if the change was positive or negative. At this stage we left the exploration of this to future studies that would collect user satisfaction scores to quantify this.

In analysing DCG scores for our altered result pages, we found that some SERP layouts influenced both RBO and DCG scores similarly. Post-hoc tests using Tukey's HSD Test revealed no significant difference in RBO for certain combinations of card layouts such as T, TI and TIS, TS or TS,TI and TIS or TI. Notably, for DCG, there wasn't a significant difference among several combinations of card layouts, despite the differences in card type mixes.

TBG accounts for the time spent by the users and their attention on retrieved results [44]. We can observe with the TBG how the costs associated with reading each item in the result list affects the gain of the page. For example the TBG of Page is significantly higher when we combine T cards with TS cards for a result list, where as combining TIS cards with TI cards has a significantly lower TBG compared to the baseline. These results emphasise the role of card types and their arrangement in influencing search result effectiveness and how the time spent assessing the results will affect users' gain. This underscores the need to carefully consider both the presentation and number of search results to optimise user experience (space-utility trade-off).

Our observations show how the alteration of presentation influences the order of ranked list for the iPRP compared to the PRP. In our implementation of TBG, we have implemented a simplistic user model that assumes linear browsing, like the iPRP. In further work, we aim to explore how changing the presentation affects other complex browsing models.

4.4 Summary

In this chapter, we began chipping away at our high level research questions, specifically, **HL-RQ1**. We aimed to understand how the EPU can vary across different presentation formats by collecting real world timing data to ground our notion of EPU. We had further split our high level research question to understand how the ranking of results is affected by this change in EPU. Our study examined whether the iPRP, implemented via the Card Model, significantly affects the ranking of search result pages based on the relevance of items and their presentation. We aimed to understand the impact of presentation when ranking heterogeneous result pages with four common types of result cards under the iPRP. We framed the iPRP/Card Model as producing the expected perceived utility of each result presented, factoring in different interaction probabilities and decision-making times for various result card types. Our method contrasts with the original PRP, which only considers item relevance for ranking. Our research focused on two main questions, exploring the EPU of different result card types, and the impact of ranking results by EPU on performance with respect to Rank Biased Overlap (RBO),

Discounted Cumulative Gain (DCG) and Time Biased Gain (TBG).

Our findings indicate that in the context of ad-hoc news search for the TREC WaPo dataset, result cards using a title, image, and summary (TIS) or title and summary (TS) yield the highest EPU, which is in line with previous research that finds users tend to be more satisfied with a Title and Summary or a Title and Image [20–24]. However, these card types also limit the number of cards that can be displayed on the screen, creating a trade-off between space and utility. We found that this trade-off is crucial, as the choice of result card type can significantly affect SERP effectiveness through the re-ordering of documents at higher ranks, as evidenced by RBO, DCG and TBG measurements.

Moreover, we showed how altering the result card type on a SERP changes the ranking of items on the SERP (and also the DCG and TBG of the page) compared to a homogeneous result card format. This suggests that when ranking heterogeneous result pages, it may be possible to manipulate the presentation of results to demote or promote items in the ranking, given the differences in how people engage with different card types. This can raise some ethical concerns as manipulating the presentation can be used to bias users toward specific results.

This study underscores the importance of considering the presentation of search results when designing ranking algorithms. The perceived relevance of items can change the ranking of documents depending on the presentation of results. We have established that presentation matters when ranking, and that presentation effects can be encoded within a theoretical framework to estimate the expected "perceived" utility.

So far, we laid the groundwork by extending the Card Model model to assess the EPU for ad-hoc search tasks within the TREC Washington Post dataset. We found the potential for customising heterogeneous SERPs to enhance user satisfaction and efficiency in finding relevant information through re-ranking results by their EPU. Given this finding, we aim to further examine whether improvements in SERP presentation could further optimise these outcomes. To this end, our next steps involve a detailed analysis of user interactions with SERPs and the impact of interface variations on user satisfaction and query performance. We thus propose to blend the system-focused

and user-centric strategies to determine if refined SERP presentations can contribute to the effectiveness of searches beyond the benefits of more accurate queries. Our findings so far suggest re-ranking could positively influence search results, prompting us to investigate how such modifications might enhance the overall search experience. We will now try to understand the influence of presentation and performance on user satisfaction.

Chapter 5

The Influence of Presentation and Performance on User Satisfaction

5.1 Introduction

In our previous experiment, annotation data was collected to quantify the variations in EPU among different result cards. Based on the insights from the data analysed, we determined that the TIS and TS result cards provide comparable utility levels (i.e., adding or removing an image from a card already containing a summary does not do much). Furthermore, we hypothesised that the distinct dimensions of these result cards on a display would result in varying amounts of total expected utility within a given space. This hypothesis, however, remains speculative at this stage, and our initial study's limitations, such as restricting users from selecting queries and viewing all results on a SERP, were acknowledged. In practical scenarios, documents are typically presented to users on SERPs, with each document represented by a result card. An effective result card should aid users in making more informed decisions about whether to explore a document further by including key information like a title, image, or a summary.

We have seen how previous works from [20, 21, 53] and [23] have studied how the

presentation of these result cards affects user satisfaction (see §2.5.2). The broad consensus from these analyses is that incorporating visual elements like images, links, and text summaries can strongly influence user satisfaction and perceptions of relevance.

However, it is not solely the presentation that drives user satisfaction. Users also spend their time creating queries so that the system may retrieve and present them with appropriate relevant documents for their information need. The performance of these queries is typically measured by system-side metrics such as Cumulative Gain (CG), Discounted Cumulative Gain (DCG), normalised Discounted Cumulative Gain (nDCG) etc (see §2.2.2 for more detail). Work from [81] has explored how *some* these metrics affect user satisfaction, finding that there is a strong correlation in most query performance metrics, such as CG, DCG etc.

During the interaction process, users can also perform other actions such as inspecting various result cards, saving the documents behind the cards etc. These actions come with inherent costs to perform them and further work such as [82] have developed formal models to estimate the costs in the interaction process (such as cost to query, examine cards etc). Given this formal framework, further research such as [63, 65, 83] have studied how costs such as the cost to query affect user satisfaction (see §2.5.4)

Given that the presentation of the result cards can also affect user satisfaction, it is unclear how changing the presentation can affect both the system side costs (query costs) and user side costs (user satisfaction). Take, for example, two result lists with slightly differing nDCGs for a given query, presented in the same result card type (all titles). Findings from [63, 81, 83] would suggest that spending longer examining results for a query with a higher nDCG will lead to more satisfaction. However, if we modify the presentation of the result list with a lower nDCG to be presented with, say all titles and images (TI), users may now spend more time examining results in this list due to their changed presentation and thus feel a similar amount of satisfaction as that obtained from a result list with a higher nDCG.

This segues into our second high-level research question (**HL-RQ2**), aimed at investigating the relationship between query effectiveness, presentation format, and user satisfaction. To anchor our investigation, we discuss findings from a crowd-sourced user

study focused on an ad-hoc news search task. This study evaluated five interface layout designs, varying in the number of results displayed per page, across four distinct result card types. Leveraging topics, queries, and documents from the TREC Washington Post 2018 corpus, participants were charged with identifying and marking documents relevant to two topics. User satisfaction ratings were gathered for each query and for each result card layout upon completion of a topic. Consequently, we address the following research questions to elucidate **HL-RQ2**:

(RQ1) How do the quality of search results (as measured by query performance) and the interface layout impact user satisfaction in information retrieval tasks?

(RQ2) What are the effects of different interface layouts on user satisfaction as measured by overall satisfaction, the likeability of the engine, productivity, and mental effort?

5.2 Methodology

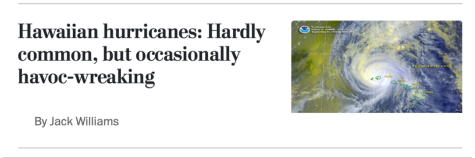
To explore how query performance and result card layouts influence user satisfaction, we conducted a between-subjects study using a simulated ad-hoc search task [84]. To position the information-seeking process within a structured context, participants were presented with a series of six pre-picked queries, which were grouped into three categories based on their nDCG@10: low, medium, and high. Each category contained two queries. The task involved participants engaging in an exploratory search session, examining various queries and documents to find and pinpoint relevant examples within relevant documents related to the given topic. The between-group variable in our study was defined by five distinct interface layouts. These layouts prominently featured cards consisting of titles, images, and summaries of news articles.

5.2.1 Collection and System

We used the TREC Washington Post Corpus (WaPo) collection that we previously indexed (§3.2.1). We presented results on a SERP, as shown in Figure5.2(b). Our

(a) Title+Image+Summary (TIS): ∼6 rows

(b) Title+Image+Summary (TIS WaPo): ∼3 rows



(c) Title+Image (TI Google): ∼2 rows

(d) Title Only (T): ∼1 row

Figure 5.1: Example of the different result card types, with an approximation of the number of rows each card type occupies.

SERP view consisted of result cards in presentation formats of two major news sources (The Washington Post and Google News).

We chose five different types of interface layouts to show the participant, with the four different result cards shown in Figure 5.1.

1. Title + Image + Summary [TIS]
2. The Washington Post Style, Title + Image + Summary [TIS WaPo],
3. Google News Style, Title + Image[TI]
4. Title only [T]
5. Random, a combination of the four above.

We consider our viewport to have a fixed amount of space (6 columns using bootstrap column widths and 12 rows, computed using approximately 100px per row). Thus, the total number of results shown on the page depended on the type of card and the

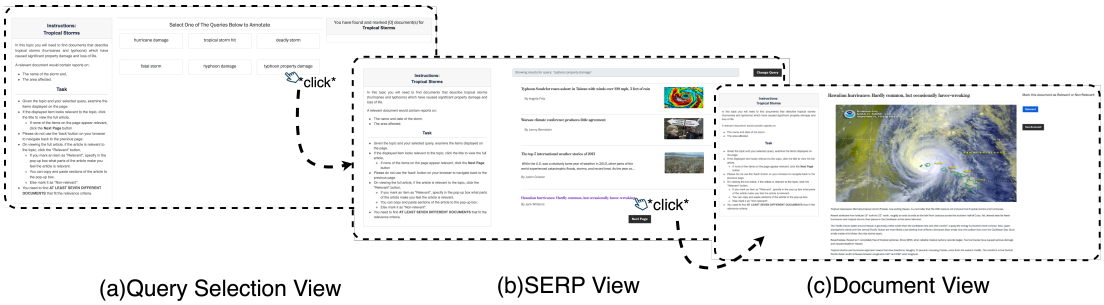


Figure 5.2: An example of the user interface presented to participants for collection of annotations. Sub-figure (b) shows an example of a SERP layout with a random arrangement of cards.

number of rows it occupied. For example, on a single result page layout with our page constraint, there could either be approximately 12 T, 2 TIS cards, 6 TI Google Cards or 4 TIS WaPo cards.

5.2.2 Search Topics and Tasks

We use all available topics from our indexed TREC collection (see §3.2.1):

1. **Topic 341:** Airport Security,
2. **Topic 363:** Transportation Tunnel Disasters,
3. **Topic 367:** Piracy at Sea and,
4. **Topic 408:** Tropical Storms.

Participants were instructed to find and save – different and relevant documents that they felt suited the relevance criteria for the given topic by exploring as many queries as necessary. For example, in topic 408 (see instructions in Figure 5.2(a),(b) and (c) on the left side), participants were asked to find a number of different tropical storms that caused widespread destruction and loss of life. Examples requested for the other topics were:

- Topic 341 the airport and security measures employed;
- Topic 363 the name of the tunnel and the cause of the disaster and,

- Topic 367 instances of piracy where vessels were boarded.

We generated 6 queries per topic using the techniques outlined in [17], and then stratified the resulting queries into three tiers based on their nDCG scores as described in §3.2.2. Specifically, the queries were grouped into low (0.1-0.2), medium (0.2-0.6), and high (0.6+) nDCG categories.

5.2.3 Measures

We split the dependent variables in our study into three main categories: (a) search behaviours, (b) search experience and (c) performance as highlighted in §3.2.4

5.2.4 Procedure

Participants were recruited from the online crowd-sourcing platform Prolific. Participants were also pre-screened based on their first language; all participants indicated a native speaker proficiency in English at the time of undertaking the experiment. This was done to maintain consistency across the participant's ability to carry out the task accurately. Four different pages were created on Prolific to fill participants for each topic. Each page contained the link to complete the task using the topic specified for that page. Before starting the study, participants were presented with an on-screen information sheet detailing the procedure of the study. They were required to provide their informed consent before proceeding with the study. Upon successful completion of the study, participants received the equivalent of USD$7 for their time, which fell in line with minimum payment requirements (at the time of the experiment). Each participant was randomly allocated one of the five layouts when they began the experiment.

The goal of the experiment was to complete a news search task based on the chosen/given topic. Participants were asked to find and mark documents relevant to one of the selected topics by exploring a set of pre-defined queries as described in §5.2.2. When the participants began the experiment, they were presented with a list of six queries in a 3x2 grid that corresponded to the topic they selected. The order in which these queries were presented was randomised. An example of this query selection grid can be observed in Figure 5.2(a). Participants were instructed to choose any query to

inspect and explore the associated results with that query to find the relevant documents. Participants were asked to evaluate the relevance of the documents based on the criteria provided on the left of the screen in a floating instruction box. An example of this floating instruction box can be seen in Figure 5.2(a),(b) and (c). These instructions were continuously visible during the process of the experiment. Once a participant picked a query, they were shown all the documents associated with that query in one of the layouts, in the style of a SERP.

The ordering of the relevance of results was random in all layouts. In figure 5.2(b) we can see on the SERP how results were presented for a random layout, we can observe the results presented as TI, T, TIS WaPo and T. Pagination was made available via a button at the bottom of the screen to move to the next set of results for a query. The participant could click on any result card to inspect the document behind it in further detail. Upon inspecting a card, the full contents of the document were displayed on a new page, this can be seen from Figure 5.2(c). If a participant inspected a result card and found the document relevant, they were asked to provide instances of the document that made it relevant in a pop-up text area. Participants were asked to provide at least one instance per relevant document.

When a participant moved between queries of the same topic, we collected the query satisfaction. In the query selection view, on the right side, we displayed the titles of the documents that the participant had marked as relevant, along with what section they marked within that document so that participants could quickly glance at their task progression. The participant needed to inspect at least two queries and find seven different relevant documents before finishing the topic. When participants finished one topic we collected the interface satisfaction as described in §5.2.3, and then the second topic was randomly assigned to them (from the pool of three remaining topics) with the same result layout as the first topic.

5.2.5 Participant Demographics

Participation was completely remote, with the researchers not interacting with any participant in any capacity. Participants directly interacted with the web application

designed to collect interaction data.

The user study involved 164 participants, most of whom fell in the age range of 20 to 40 years old, with a mix of students (27) and non-students (137). The majority of participants were employed, with 122 reporting full-time or part-time work, while the remaining participants were not engaged in paid work, such as homemakers, retired, or disabled individuals.

5.2.6 Ethics Approval

Before conducting the study we obtained ethics approval from the department ethics committee (ethics no 2027) at the University of Strathclyde. We strictly followed ethical guidelines and ensured that every participant gave informed consent. All participants received a thorough explanation of the study's procedures, potential risks, their rights, and the option to leave at any point. The consent form also provided a link to the ethics application approval.

5.3 Results

5.3.1 Summary of Search Behaviours

Comprehensive data analysis examined differences in task completion rates, interaction times, and other user metrics, such as the number of queries, clicks, and time spent across various interface layouts. Welch's ANOVAs was used to assess whether significant differences existed between the conditions and the measures under investigation. The primary effects were analysed at a significance level of $\alpha = 0.05$. Pairwise Games-Howell tests were utilised for post-hoc analyses. For the reported tests, the F-score, p-value, and effect size η_p^2 are presented to two decimal places. The ranges of η_p^2 values correspond to small (< 0.06), medium (0.06 - 0.14), and large (> 0.14) effect size [85]. The $\pm$ values reported in the tables denote the mean and standard deviation.

Table 5.1 reports the average search behaviours of users for each interface layout, detailing the number of actions performed per topic, per query. Incorporated in this analysis is the accuracy measure, highlighting how well participants identified relevant

Table 5.1: Search behaviours, with the mean number of actions performed per user, per topic, per query. Here, $\mathcal{Q}$ denotes Queries, *Docs* denotes documents. R and $\bar{R}$ denote relevant and non-relevant. Highest accuracy values are bolded.

(a) Query submissions and page navigations.

Interface Layout	$\#\mathcal{Q}$	**#Pages**
a. TIS	4.12±2.19	1.05±0.22
b. TIS WaPo	4.34±3.00	1.12±0.39
c. TI Google	4.17±1.83	1.01±0.00
d. T	3.91±2.33	1.05±0.24
e. Random	4.17±2.04	1.02±0.13

(b) Document saving behavior and accuracy.

Interface Layout	**#Docs...**			**Accuracy**
	viewed	**saved**	**relevant**	
a. TIS	4.71±4.03	2.71±2.24	2.19±1.69	0.79±0.27
b. TIS WaPo	4.65±3.73	3.31±2.43	2.69±2.01	**0.83±0.22**
c. TI Google	4.94±5.14	2.97±2.25	2.33±1.65	0.79±0.26
d. T	5.22±4.67	2.94±2.15	2.40±1.74	0.75±0.28
e. Random	4.63±4.31	2.89±2.96	2.38±2.21	0.78±0.25

documents from the non-relevant (i.e., the proportion of relevant documents saved versus the total number saved).

Considering the varied interface layouts, there is evident consistency in user behaviours. Across the board, for any topic, participants on average clicked to view 3 to 4 queries. Notably, participants examined on average only a single page for every query they issued. This is despite the fact that users could examine more pages within the same query. For every query viewed, participants clicked and viewed between 4 to 5 documents. They saved about 3 of the viewed documents, and out of these, they correctly identified around 2 as relevant. The accuracy of judgements fluctuated between 0.75 to 0.83.

We found that with the TIS WaPo layout (when all results were presented with TIS WaPo cards) participants were able to more accurately identify and mark relevant documents, achieving a peak accuracy of about 0.83, which was significantly more than other layouts ($F(4,441.837) = 2.51$, $p = 0.04$, $\eta_p^2 = 0.01$). This is possibly due to TIS

Table 5.2: Average timings for various search behaviours actions during the study, per user, per topic, per query. The timing data is in seconds. Asterisks (*) denote a significant difference between all groups($p < 0.05$)

Interface Layout	**Task**	**Time per ...**		
		Snippet	**R Doc**	**$\bar{\text{R}}$ Doc**
a. TIS	1345.23 ± 670.13	$2.25 \pm 1.23^*$	44.13 ± 41.90	37.92 ± 30.28
b. TIS WaPo	1469.89 ± 1138.38	$2.09 \pm 1.25^*$	52.34 ± 45.74	36.94 ± 33.55
c. TI Google	1442.04 ± 931.42	$1.82 \pm 1.18^*$	41.24 ± 36.48	34.44 ± 36.26
d. T	1519.73 ± 1027.33	$1.95 \pm 1.18^*$	42.36 ± 40.77	52.29 ± 65.62
e. Random	1367.29 ± 882.38	$2.11 \pm 1.27^*$	42.97 ± 43.16	36.00 ± 29.78

WaPo cards providing useful information in the form of a summary that helped users to click and accurately mark them as relevant. However, it is interesting to note that this was significantly higher than the TIS layout, in which the result cards contained the same information but occupied more space. We hypothesise that this occurs due to the ability to view more cards containing summaries within the same space, potentially expanding the context window of users viewing the result cards. However, on average, per query and topic, we found no statistically significant differences in the search behaviours of participants across any layout.

Due to the synthetic nature of our queries and the controlled nature of our study, we hypothesise that these behaviours may be specific to our study and that in a more naturalistic search scenario, where users can type out queries, they may tend to issue queries differently to find relevant information. Also, in real-world scenarios, images for documents may not be relevant to the document content. This could further impact other factors such as the time spent examining documents and inspecting pages on the SERP.

Table 5.2 offers a comprehensive look at the average timings for the search behaviours (in seconds) participants took for various actions during their search sessions. Firstly, in general, we observe that there is no statistically significant difference between the times that users took to complete the task (topic). Participants took on average approximately 20 minutes to annotate a topic. The time spent on a snippet in a layout was computed as the amount of time users spent hovering over results. We found sig-

nificant differences between all layouts ($F(4,9579.212) = 34.306$, $p < 0.001$, $\eta_p^2 = 0.01$). We found no notable differences in the time required to read and make a decision for a relevant or non-relevant document for any given interface layout. Participants spent an average of 43 seconds to read the document and decide the relevance.

5.3.2 RQ 1: How do the quality of search results (as measured by query performance) and the interface layout impact user satisfaction in information retrieval tasks?

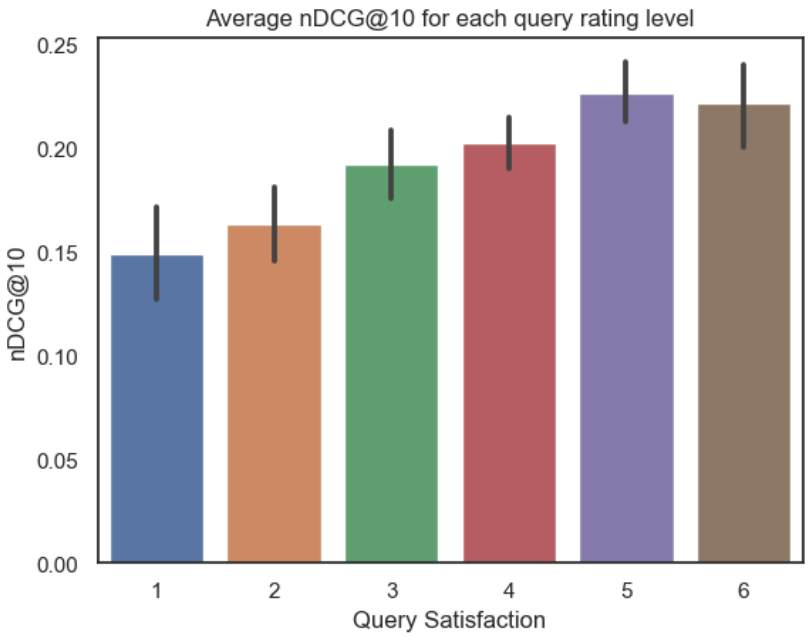


Figure 5.3: The relationship between query satisfaction and nDCG@10

We ran two ordered models on 1,398 observations from 164 participants to scrutinise the association between query performance, presentation and user satisfaction. We were concerned with examining two main metrics to measure query performance. (1) nDCG@10 and (2) Total gain of the first result page.

Within our models, we also explored several interaction effects, such as the interplay between the topic and interface layout, the sequence in which the topic was completed (referred to as "topic order", meaning if the topic was completed as the first or second), and the relationship between the interface layout and query performance. Equation 5.1

shows the independent variables in our ordered model alongside the coefficients β for the predictor Y_{ij}. For analyses focusing on the total gain on the first page, we adjusted the equation by substituting the β_1 coefficient.

$$\begin{aligned} Y_{ij} = \beta_0 &+ \beta_1(\text{nDCG@10})_{ij} + \beta_2(\text{Topic Order})_{ij} \\ &+ \beta_3(\text{Topic ID})_{ij} + \beta_4(\text{Interface Layout})_{ij} \\ &+ \beta_5(\text{Topic ID} \times \text{Topic Order})_{ij} \\ &+ \beta_6(\text{Topic ID} \times \text{Interface Type})_{ij} \\ &+ \beta_7(\text{Interface Type} \times \text{nDCG@10})_{ij} \\ &+ b_{0j} + (1|\text{user})_j + \epsilon_{ij} \end{aligned} \tag{5.1}$$

Table 5.3: Results of the Ordered Model analysis on query satisfaction, where p-value was statistically significant for the β parameter. The category differences were all significant.

Beta Parameter	**Coeff.**	**SE**	**z-value**	**p-value**	**95% CI**	
β_1(nDCG@10)	2.3015	0.897	2.566	0.010	0.543	4.06
β_5(TOPIC 408 × Order = 2)	0.8355	0.289	2.891	0.004	0.269	1.402
β_6(TOPIC 408 × Interface Layout = Random)	1.5771	0.452	3.490	< 0.001	0.691	2.463
β_1(Total Gain on Page 1)	0.2002	0.079	2.542	0.011	0.046	0.355
β_3(TOPIC 408 × Interface layout = Random)	1.5491	0.451	3.434	0.001	0.665	2.433
β_5(TOPIC 408 × Order = 2)	0.8471	0.289	2.936	0.003	0.282	1.413
β_7(Interface Layout = TIS WaPo × Total Gain on Page 1)	0.5173	0.186	2.778	0.005	0.152	0.882
β_7(Interface Layout = TI Google × Total Gain on Page 1)	0.3024	0.152	1.990	0.047	0.005	0.600

As we can observe from the top half of Table 5.3, we found a significant positive relationship between nDCG@10 and user satisfaction, which can also be observed from Figure 5.3. We observed no interaction effects between the nDCG@10 and the interface layout which could have affected the query satisfaction. This signifies that a poorer nDCG of a query cannot increase user satisfaction to match the same level as that of a higher nDCG if we change the presentation of results. However, we observed a significant effect on query satisfaction when users attempted Topic 408 with the Random layout as the second topic.

For the total gain on the first page, we examined the effectiveness of each query within the context of the first page of results, utilising the metric NDCG@k. Recognising the dual significance of result relevance and quantity, we used a "total gain"

measure for the first page of results. This measure was calculated by multiplying the NDCG@k score, which evaluates the relevance of the documents on the first page, by the number of results (k) displayed on that page. By doing so, this 'total gain' measure accounts for both the quality and quantity of results, providing a more holistic assessment of query performance on the first page of results for across multiple queries. This allows us to factor in the varying number of results displayed by different interface layouts, and understand how these layouts perform not just in terms of relevance per document (as captured by NDCG@k), but also in terms of total relevance gain for the user across multiple queries. We can also observe this similarly positive relationship for the total gain from Figure 5.4

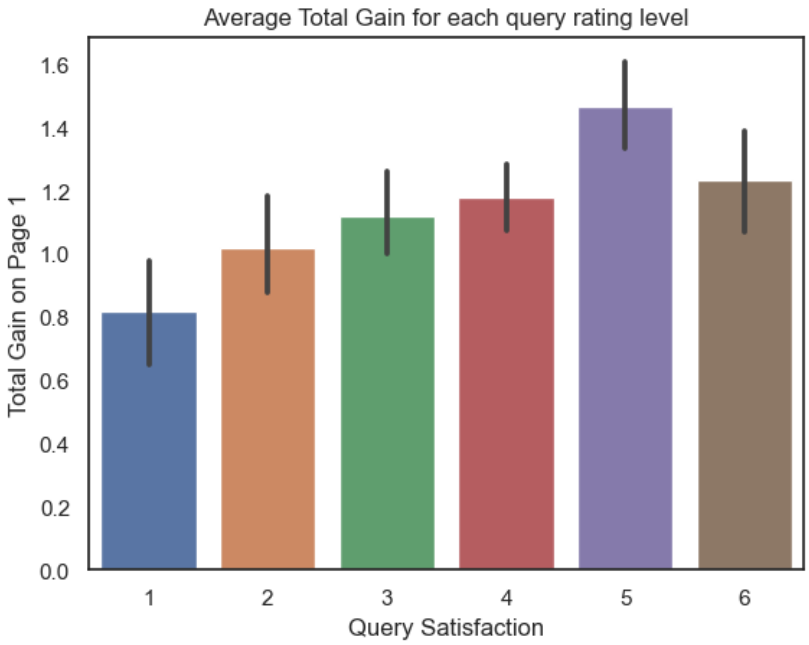


Figure 5.4: The relationship between query satisfaction and Total Gain on Page 1

The second ordered model was run with the formula defined in Equation 5.1, but with the β_1 parameter being substituted for the total gain on page 1. The results from this second model (as shown in the bottom half of Table 5.3) showed that the interaction effects between the total gain and the interface layouts consisting of TI Google and TIS WaPo cards played a significant effect ($p < 0.05$) on the query satisfaction. Same as with our first ordered model, we observed a significant effect on query satisfaction when

users attempted topic 408 with the random layout as the second topic. Our findings from this model essentially indicate that for total gain on the first page, the presentation of results can affect user satisfaction (i.e., by modifying the interface layout, a layout with a lesser total gain on the first page can attain user satisfaction comparable to a layout with a higher total gain.)

Our results on the link between nDCG@10 and user satisfaction diverge slightly from past studies, such as [81], which found only weak ties between nDCG and satisfaction[1]. We identified strong linear correlations between nDCG@10 and query satisfaction. Additionally, we noted distinct gains on the first page for two layouts, TI Google and TIS WaPo, revealing an interplay between result presentation, total gain, and query satisfaction. In conclusion, while nDCG@10 effectively predicts user satisfaction, no direct linear relationship exists between result presentation and user satisfaction for metrics like nDCG@10. However, metrics like total gain on the first page do influence presentation and satisfaction.

5.3.3 RQ 2: What are the effects of different interface layouts on user satisfaction as measured by overall satisfaction, the likeability of the engine, productivity, and mental effort?

Looking at Table 5.4, we see the average satisfaction scores at the interface satisfaction for each aspect we considered. When we directly compare the layouts based on these individual metrics, the Welch ANOVA test reveals that there is no statistically significant difference between them. Since the differences might be more subtle or complex, to gain a better understanding, we used a MANOVA test.

Our analysis revealed significant differences across the different layouts. For the test statistics, including Wilks' lambda, Pillai's trace, Hotelling-Lawley trace, and Roy's greatest root, we found $F(5, 318) = 131.647, p < 0.001$. The observed effect sizes (η_p^2) ranged from medium (0.065 for Wilks' lambda and 0.135 for Pillai's trace) to large (0.414 for both Hotelling-Lawley trace and Roy's greatest root).

[1]We also compared other metrics from the [81] study such as precision and CG and confirmed that precision and CG are strongly correlated to user satisfaction ($p < 0.05$) but the interface layout did not affect the user satisfaction.

Table 5.4: Results of Interface Satisfaction. No statistically significant differences were found between any of the measures for a given interface layout.

(a) User engagement and mental effort.

Interface Layout	**Felt Productive**	**Mentally Taxing**	**Liked Engine**
TIS	3.71±1.47	3.52±1.51	3.78±1.24
TIS WaPo	4.05±1.58	3.19±1.31	4.00±1.22
TI Google	4.06±1.18	3.01±1.40	3.80±1.12
T	4.00±1.13	3.32±1.30	3.81±0.97
Random	3.78±1.33	3.07±1.38	3.85±1.13

(b) Distraction levels and overall satisfaction.

Interface Layout	**Distracting**	**Overall Satisfaction**
TIS	2.95±1.31	3.94±1.37
TIS WaPo	2.81±1.34	3.92±1.35
TI Google	2.79±1.27	4.07±1.17
T	2.91±1.06	4.16±1.05
Random	2.90±1.40	4.10±1.17

Given these differences exist, we try to separate the contributing components to each interface layout via Linear Discriminant Analysis (LDA). The coefficients from the LDA, which are provided in Table 5.5, represent the standardised contribution of each user satisfaction metric to the discriminant of the interface layouts.

From Table 5.5, the LDA coefficients underscore that variations in interface designs subtly impacted user perceptions and experiences, culminating in different satisfaction levels, productivity perceptions, and cognitive demands. For instance, the T layout was predominantly associated with high overall satisfaction (0.296) and cognitive load (0.106). In contrast, the random layout interface layout was characterised by higher overall satisfaction (0.312) and lower cognitive load (-0.167), but lower productivity (-0.341), revealing a potential trade-off between user satisfaction and perceived productivity.

The explained variance ratios from the LDA show the proportion of variance captured by each discriminant function. Specifically, the first discriminant function accounts for approximately 54.8% of the variance, highlighting its significance in distinguishing between the interface types. This is followed by the second, third, and

Table 5.5: Coefficients of the Linear Discriminant Analysis (LDA) for distinguishing between different interface layouts based on the features captured in the interface feedback. Each row represents the coefficients for a specific interface type.

(a) Coefficients for user engagement and engine preference.

Interface Layout	**Felt Productive**	**Mentally Taxing**	**Liked Engine**
TIS	-0.180	0.209	0.030
TIS WaPo	0.260	-0.013	0.349
TI Google	0.162	-0.120	-0.215
T	0.016	0.106	-0.237
Random	-0.341	-0.167	0.014

(b) Coefficients for distraction levels and overall satisfaction.

Interface Layout	**Distracting**	**Overall Satisfaction**
TIS	-0.105	0.133
TIS WaPo	-0.001	-0.559
TI Google	-0.003	-0.033
T	-0.002	0.296
Random	0.119	0.312

fourth functions, which capture 26.4%, 16.6%, and 2.2% of the variance, respectively. Based on these ratios of the discriminants, Figure 5.5 shows a visualisation of these two discriminants in separating the different interface layouts.

Our assessment reveals that although satisfaction metrics are interconnected, they do not completely linearly differentiate the interface layouts and that there are small overlaps between the layouts (even though some clustering-like behaviour is observed), as seen from Figure 5.5. While the layouts exhibit distinct characteristics, their differences are not solely driven by individual satisfaction metrics. Instead, a collective, non-linear interaction of these metrics influences the differences observed across interface layouts.

In our study, we have found some interesting insights. Based on the findings from RQ1, it is evident that nDCG@10 acts as a robust predictor for user satisfaction at the query level, with interface layout being influential when considering total gain on page 1. With RQ2, our exploration extends into understanding how these layouts influence user satisfaction when users complete a task (session level). While there

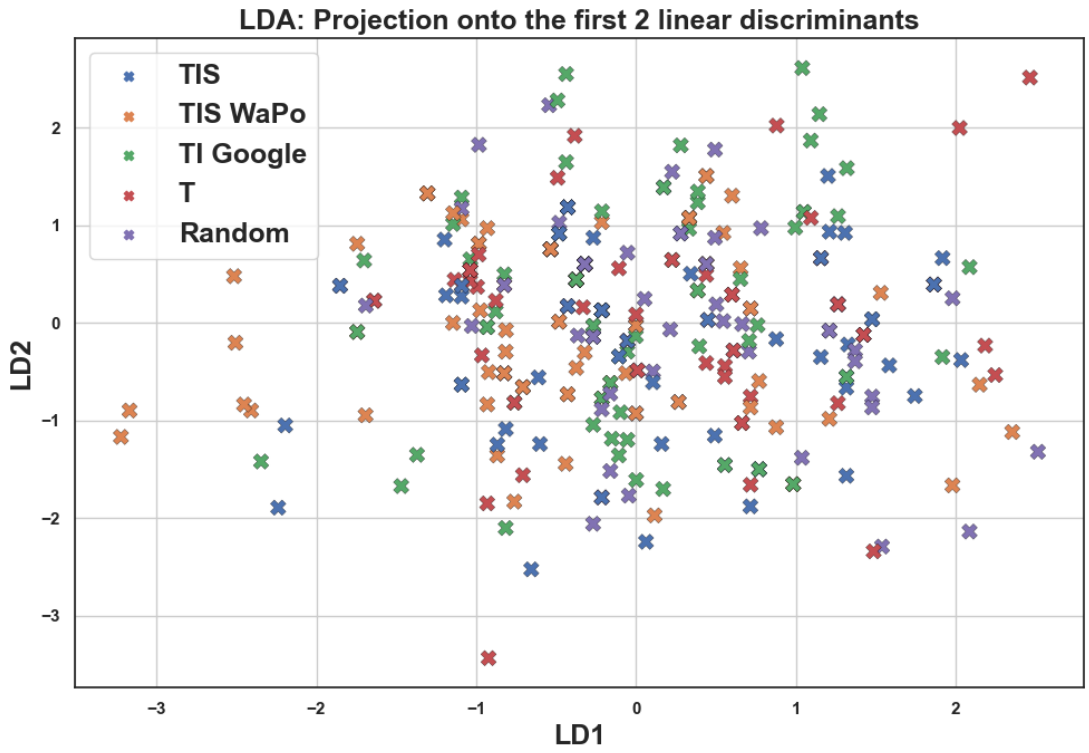


Figure 5.5: Visualisation of the first two Linear Discriminants (LD1 and LD2), for different interface layouts

exist differences in layout preferences and perceptions, our analysis using LDA revealed that the connection between satisfaction metrics and interface layout satisfaction is intricate and layered, deviating from a straightforward relationship. The layouts did not differ on any one specific metric of user satisfaction. These nuanced differences uncovered by LDA demonstrate that users' satisfaction with interface layouts is multi-factorial, influenced by various combinations of satisfaction metrics. By integrating the insights from both research questions, we discern that optimising user satisfaction in IR systems is not solely about enhancing query performance or refining the presentation of results. It requires a harmonious synchronisation of both elements, considering the subtle intricacies in user preferences and satisfactions, offering a pathway to building more user-centric and adaptive Information Retrieval systems.

5.4 Summary

In this chapter, we explored the correlation between query performance, specifically marked by nDCG@10 scores, presentation and user satisfaction, with a user study consisting of 164 participants in an ad-hoc news search task. We aimed to bridge the gap between query performance, presentation and user satisfaction, venturing beyond independent studies such as [20, 21, 23, 53, 63, 81, 83] to encapsulate the nuances of presentation impact.

Our analysis revealed a strong and significant correlation between nDCG@10 scores and user satisfaction at the query level, deviating in findings from [81], where only a weak correlation was observed. However, we observed no direct relationship between the presentation (interface layouts), user satisfaction and query performance (with nDCG@10). Signifying that, while interface modifications impact user interactions and perceptions, they do not intrinsically augment the effectiveness of the queries for metrics such as nDCG@10, however, it does lead to changes with respect to other metrics such as the total gain on the first page. This means that with respect to presentation, the number of results and the space they occupy play a role in user satisfaction. Despite the absence of a direct correlation between interface layouts and query performance, the presentation can still impact user satisfaction metrics—such as pro-

ductivity, cognitive load, likeability, distraction, and overall satisfaction, falling in line with all previous work such as [20, 21, 23, 53, 63, 83] which reports that users perceive different result cards in different ways. We further assert that the differentiation in user satisfaction across interface layouts is complex, stemming from a multi-factorial combination of user satisfaction metrics. It is imperative to acknowledge that the interface's structure holds substantial weight in shaping user satisfaction, even though it does not directly impact query performance.

This study, therefore, serves as a catalyst for a more nuanced understanding of the dynamics between search performance, result presentation, and user satisfaction. We underscore the importance of interface layouts, stressing the role it plays in altering user interaction and satisfaction without directly altering search performance metrics such as the nDCG@10. Having established that various result cards possess distinct EPUs (see §4) and their presentation in different layouts influences not only user satisfaction but also query performance, we proceed to address the third and final research question of this thesis. This question seeks to extend the card model by applying it practically to a ranking task, taking into consideration the presentation of results. In the forthcoming chapter, we will examine how users perceive different optimisations, assess whether these optimisations are realised, and explore whether user preferences vary or if a universal interface benefits all users equally.

Chapter 6

Optimising Ordering of Results Based on Presentation

6.1 Introduction

So far, in this thesis we explored the influence of various presentation formats and layouts, on user satisfaction, establishing that result cards displayed on a page are imbued with distinct perceived utilities (EPUs). Consequently, incorporating this consideration into result list ranking, in conjunction with existing ranking principles, presents a notable challenge. The PRP and its interactive counterpart iPRP do not make any assumptions about the presentation of a SERP. Given that modern SERPs present results in varying formats and that this variation can lead to certain results being pushed outside the user's viewport (assuming a fixed viewport size, with no scrolling).

Extensive research has been conducted on both system-side factors (including query performance, information retrieval, and ranking algorithms) and user-side behaviours (such as query formulation, query length, and presentation preferences) to optimise the retrieval process and enhance user satisfaction. The focus on the presentation of various visual elements has been well-documented in the literature [20,21,23,53], leading to the development of models that estimate interaction costs through comprehensive studies [18, 63, 65, 83].

Furthermore, pioneering works [25–29] have laid the groundwork for SERP optimi-

sation through the integration of user interaction data, such as click patterns, into the design of 2-D SERPs. Subsequent research [30–32] has explored the application of deep reinforcement learning (DRL) to SERP optimisation, encoding the space and utility of each result item into the DRL reward function. While the adoption of DRL models holds promise for advancing research in this area, the computational expense and substantial prerequisite of interaction data for model training pose significant challenges, particularly outside the domain of web search (see §2.6 for more detail).

Acknowledging the unique EPU associated with each type of result card, its spatial constraint on the screen, and the insights from our initial experiment demonstrating the impact of presentation changes on result ranking, it becomes imperative to optimise the ranked list based on user preferences. This chapter aims to elucidate our approach to examining the effects of such optimisations on user satisfaction and performance metrics.

Given the constraints of fixed-page presentations typically used for displaying search results, this leads to intriguing possibilities for optimisation. We can either optimise the presentation of results currently viewed (per page) by the rate of utility gained or the total utility. This raises fundamental questions about the effectiveness and potential benefits of such optimisations.

Therefore, in this chapter, we propose a general optimisation algorithm that accounts for space-utility constraints and simulate the influence of varying user behaviours on these optimisations. By correlating the outcomes of these simulations with actual user behaviour data, we aim to address further research questions related to how presentation-based ordering influences user experiences to answer **HL-RQ3**, specifically:

(RQ1) How do differing optimisation strategies impact the resulting user interface configurations, and to what extent do these strategies diverge in accommodating various user behaviours?

(RQ2) How do user satisfaction and cognitive load metrics evolve as the user interface is iteratively optimised across multiple topics within a search engine environment?

(RQ3) To what extent do user preferences converge towards a unified SERP configuration, and what are the cognitive load variations associated with different SERP optimisation strategies across tasks?

6.2 Methodology

Having established that the presentation of results on a SERP can be optimised through two distinct methodologies, namely: the rate of utility gained per page and the total utility gained per page— it is imperative to investigate whether fundamentally different SERPs emerge from the algorithms when applied to varying user behaviours. To address this question, this section outlines a general approach to optimising a SERP, taking into consideration the utility value derived from a result card and the space it occupies within the constraint of total space available on a single SERP. Subsequently, we will simulate the performance of the optimisation algorithm across users with extreme behaviours to ascertain the variance in the SERPs generated.

6.2.1 Optimisation Algorithm

In the context of presenting a set of documents on a SERP, each document can be conceptualised as being represented by a result card type. This card occupies a defined number of rows and possesses an associated EPU value, engendering a trade-off between the space consumed and the utility delivered to the user.

In addressing this optimisation problem, two distinct strategies emerge. The first is a greedy algorithm that optimises on an individual basis but cannot guarantee a globally optimal solution. This limitation stems from its design to maximise the EPU from only the initial set of documents, potentially relegating further documents to lower positions or even to subsequent pages. This approach risks documents being overlooked due to their positioning, undermining the efficacy of the search results.

The alternative strategy employs dynamic programming (DP) to guarantee a globally optimal solution. The algorithm proposed is analogous to the knapsack problem [86–90]. Its objective is to maximise the total EPU of documents within the

confines of a fixed number of rows, rather than maximising the value of items within a knapsack of a fixed capacity. The inputs for this algorithm include a set of documents, each represented by a tuple that encapsulates the row size, the utility of a row configuration for the document, and the maximal number of rows available for use. This approach aims to finely balance the allocation of space with the maximisation of user utility, ensuring an optimised presentation of search results. However, we need to modify the standard knapsack problem to include the additional dimension of one item having multiple presentation formats. Therefore this is a multi-bounded variant of the knapsack problem.

Algorithm 1: Multi Bounded Knapsack for Optimal Document Set(MBKDS)

```
Input  : n, max_page_size, documents
Output: (max_utility, optimal_documents)
dp ← [[0] * (max_page_size + 1) for _ in range(n + 1)];
selected ← [[[] for _ in range(max_page_size + 1)] for _ in range(n + 1)];
for i ← 1 to n do
    for w ← 1 to max_page_size do
        document_utility ← 0; document_selected ← []; for
         j ← 0 to len(documents[i − 1]) − 1 do
            (size, utility) ← documents[i − 1][j]; if size ≤ w then
                new_utility ← dp[i − 1][w − size] + utility; if
                 new_utility > document_utility then
                    document_utility ← new_utility;
                     document_selected ← selected[i − 1][w − size] + [(i − 1, j)];
        if dp[i − 1][w] > document_utility then
            dp[i][w] ← dp[i − 1][w]; selected[i][w] ← selected[i − 1][w];
        else
            dp[i][w] ← document_utility; selected[i][w] ← document_selected;
max_utility ← dp[n][max_page_size]; optimal_documents ← []; i ← n;
 w ← max_page_size; while i > 0 and w > 0 do
    if selected[i][w] == selected[i − 1][w] then
        i ← i − 1;
    else
        document ← selected[i][w][−1];
         optimal_documents.append(document); i ← document[0];
         w ← w − documents[document[0]][document[1]][0];
return (max_utility, optimal_documents);
```

The Multi Bounded Knapsack algorithm, as outlined in Algorithm 1, effectively resolves the challenge of selecting the optimal set of documents that maximises the total EPU, within the constraints of maximum page size. This algorithm iteratively assesses each document against all possible page sizes to determine the maximum utility achievable for each specific document-card configuration. It employs dynamic programming

to systematically construct a table with dimensions $(n+1) \times (W+1)$, where n is the number of documents and W is the upper limit on page size. The value at each table entry, $dp_{i,w}$, signifies the highest utility that can be attained using up to the first i documents within a page size of w. A secondary table, *selected*, is maintained in parallel to track the documents selected for achieving each utility value recorded in *dp*. The entries in $selected_{i,w}$ contain lists of tuples representing the documents chosen for $dp_{i,w}$. Following the completion of these tables, the algorithm retraces the selections recorded in the *selected* table to compile the optimal document set that attains the maximum utility, conforming to the page size restriction. This process not only ensures the optimisation of utility but also efficiently manages the spatial constraints of the SERP.

The problem can be represented mathematically as:

Maximise:

$$\sum_{i=0}^{n-1} \sum_{j=0}^{|d_i|-1} \mathbf{EPU_{i,j}} \cdot selected_{i,j} \tag{6.1}$$

Subject to:

$$\sum_{i=0}^{n-1} \sum_{j=0}^{|d_i|-1} \mathbf{size_{i,j}} \cdot selected_{i,j} \leq W \tag{6.2}$$

Where n is the number of documents, W is the maximum page size, D is the list of n documents, where each document d_i is a list of $(\mathbf{size}, \mathbf{EPU})$ tuples representing the size and utility of each document in the retrieved set, $selected_{i,j}$ is a binary decision variable that indicates whether or not to include the jth card type of the ith document in the optimal set, and $\mathbf{EPU}i, j$ and $\mathbf{size}i, j$ are the utility and size of the jth card type of the ith document, respectively.

The space complexity of the Multi Bounded Knapsack algorithm is $\mathcal{O}(n \cdot \mathbf{M})$, which is the size of the DP and selected tables. The algorithm's time complexity is $\mathcal{O}(n \cdot \mathbf{M} \cdot k)$. Initialising the ***dp*** and selected tables takes $\mathcal{O}(n \cdot \mathbf{M})$ time. The main dynamic programming loop then iterates over each document and each possible page size, and for each page, it iterates over each possible subset of sizes of the card for the document. Each of these iterations takes $\mathcal{O}(k)$ time, so the total time complexity of the dynamic

programming loop is $\mathcal{O}(n \cdot \mathbf{M} \cdot k)$.

The backtracking step to obtain the optimal set of documents takes $\mathcal{O}(n+m)$ time, where m is the number of pages in the optimal set. This is because the backtracking step iterates over each selected page in the optimal set and looks up its corresponding document and page index.

Therefore, the overall time complexity of the Modified Bounded Knapsack algorithm is $\mathcal{O}(n \cdot \mathbf{M} \cdot k)$, and the space complexity is $\mathcal{O}(n \cdot \mathbf{M})$. However, in practice, the algorithm often runs much faster than the worst-case time complexity because if we fill in the DP table by ordering the documents in decreasing order of relevance, it guarantees that at least one of the n card types is arranged in decreasing order of their EPU, thus enabling those sets of documents to be evaluated greedily.

6.2.2 RQ1: How do differing optimisation strategies impact the resulting user interface configurations, and to what extent do these strategies diverge in accommodating various user behaviours?

To answer the first research question of whether the different optimisation strategies produce fundamentally distinct SERPs for identical user behaviours, we propose a simulation as a cost-effective alternative to the prohibitively expensive user studies. The foundation of our simulation is based on an optimisation algorithm aimed at maximising the utility of all retrieved documents. However, it is important to note that our interest lies in optimising results on a per-page basis. This necessitates a slight modification of the algorithm to optimise for content that can be displayed on a single page, repeating this process recursively until all retrieved documents have been considered.

This modification is straightforward, as shown in Algorithm 2. Given that each result type occupies a predetermined amount of space and that the page has a fixed total space (for example, see§5.2.1), we can calculate the maximum number of documents of a specific card type that can fit on a page. For instance, if a card type occupies 4 rows of space and the total space on the page is 12 rows, we can fit a maximum of 3 results if all results were to be presented in this one card type alone. The utility for documents that cannot be presented on a page is set to zero. Consequently, we construct a matrix

where rows represent documents and columns represent the card types in which the document can be presented. Each cell in this matrix indicates the utility that the user can gain from the respective card type. Feeding this matrix through our algorithm yields a list of documents to be selected to maximise the page's utility, alongside the card types in which they should be presented.

After applying the algorithm to determine the optimal set of documents for a page, we update the pool of available documents by removing those that have been selected for display. This process of optimisation, selection, and removal continues until there are no remaining documents. This ensures that each page's layout is individually optimised based on the remaining documents and the fixed space constraints, thereby maximising the utility of each SERP page within the limitations of page size and document availability.

We will additionally transform the EPU of each item with Equation 6.3, to prevent negative values, since our optimiser cannot handle negative values.

$$y = \begin{cases} e^{EPU_{card}(i)} & \text{for } EPU_{card}(i) \neq 0 \\ 1 \times 10^{-10} & \text{for } EPU_{card}(i) = 0 \end{cases} \tag{6.3}$$

While computing the rate of utility, we need to inject zeroes where the $EPU_{card}(i)$ is equal to 0, since the rate of gain at 0 is 0. Otherwise, this may cause significant undesirable ranking changes, such as (1) lower utility cards being ranked higher since the rate of increase would be significantly higher, and (2) lower-ranked items surfacing to the top. It is broadly unrealistic for a bottom-ranked item to be suddenly ranked at the top purely due to presentation effects, thus we need to incorporate this scaling.

Our simulation explores two approaches: one optimises for total utility per page, and the other for the rate of utility gained, which is calculated as the utility gained for a given result and card type divided by the total utility a user would have gained if all results were presented in that card type.

To test the optimiser, we simulate specific user behaviours based on interaction data, particularly timing data and interaction probabilities required to compute the EPU of the card types. We categorise users into two dimensions: speed and preference,

Algorithm 2: Per-Page Optimisation for SERPs

Input : $documents, max\_page\_size, total\_documents$
Output: List of $optimal\_documents\_per\_page$
$optimal\_documents\_per\_page \leftarrow []$;
$current\_documents \leftarrow documents$;
while $len(current\_documents) > 0$ **do**
  $(max\_utility, optimal\_documents) \leftarrow$
   MBKDS($len(current\_documents), max\_page\_size, current\_documents$);
  $optimal\_documents\_per\_page$.**append**($optimal\_documents$);
  $current\_documents \leftarrow$ remove($current\_documents, optimal\_documents$);
return $optimal\_documents\_per\_page$;

resulting in groups such as speedy clickers and slow clickers, as well as users who prefer only one type of result card, regardless of their speed, as seen in Table 6.1. The simulation process is simple: we retrieve the top 20 documents from our index for a given topic and query (tropical storms in the case of this example). We then calculate the probabilities of relevance using the BM25 scores, as discussed in §3.1.1, and compute the utilities for all results given the selected card types. This data is then fed as input into our algorithm to observe the variations in the resulting pages.

Table 6.1: Simulation User Behaviour Categorisation

User Type	**Speed**	**Preference**
Speedy Clickers	Fast	No specific preference
Slow Clickers	Slow	No specific preference
Type-Focused Fast	Fast	Prefers a specific card type
Type-Focused Slow	Slow	Prefers a specific card type

We largely initialise the probabilities and timings to be similar to those of the average user as found in § 4, but we tweak them according to the type of user. For example, a fast user would have very low interaction times, whereas a slow user would have very large interaction times. The users who are type-focused have their probabilities set as follows: if the user prefers a specific type of card, for example, TIS, they will have their interaction probability of clicking the TIS card to be four times as likely as the other cards. We operate on a similar principle for the timing data.

To better represent outputs from our optimisers, we convert the output list of card

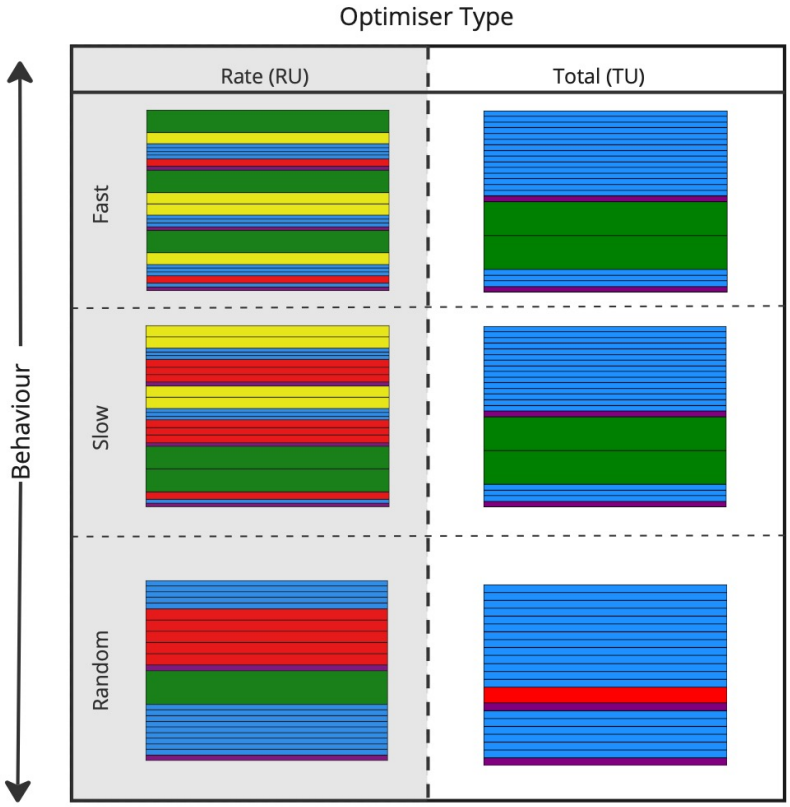


Figure 6.1: Optimisation convergence for the fast, slow and random click behaviour users for the RU and TU optimisers

types to an image. We do this to maintain the spatial representation of the card type on the SERP. This approach also allows us to visualise the SERPs at a higher level and further cluster similar-looking SERPs. We assign colours to the different card types from our study, in this case: T is Blue, TIS WaPo is Yellow, TIS is green and TI Google is red. Page breaks are represented by purple. To construct the image we draw rectangles of equal width to represent the card types of results and we adjust the height of the rectangles according to the relative row space occupied. Therefore a TIS card would be six times larger in height as compared to a T card. This way, we can also capture the space trade-off and we therefore construct an abstract representation of a SERP for a given query. From Figure 6.1 we can observe that the speed affects the optimisation between the RU and TU optimisations. The optimisation obtained from the TU shows that the fast user would benefit more from having more title cards displayed. Comparing the optimisations between speeds, we can observe that within the rate optimiser, we observe different optimisations based on the user's speed, whereas in the TU optimiser, the optimisation remains the same. Meaning that the speed of interactions does not affect the TU optimisation and it is driven more by the interaction probabilities. In the figure, we do not dive into showing the optimisations from the user behaviours where users preferred only one type of card since all the optimisations converged to show only that card type to the user.

6.3 The User Study

This user study employed a mixed-design user study methodology to investigate the differences between the two optimisation algorithms: (a) maximising the rate of utility (RU) and (b) maximising the total utility. Our experimental conditions are organised as follows:

Between-Subjects Factor: Participants were assigned to one of two groups, each exposed to a different optimisation algorithm. Group **RU** experienced the algorithm focused on maximising the rate of utility, while Group **TU** interacted with the algorithm aimed at maximising the total utility. By adopting a between-subjects design, we could observe and compare how the choice of optimisation algorithm influences participants'

overall satisfaction and search behaviours.

Within-Subjects Factor: Within each group, participants were compared across three topics, where at each topic the interface was progressively optimised over the random baseline (results are presented in random result cards). This allowed us to investigate how participants' satisfaction and search behaviour changed when they encountered different levels of optimisation during their search tasks. Participant allocation to experimental conditions was conducted through a single Prolific page, employing Latin square rotations to assign topics and the topic order to participants. This approach was chosen to guarantee the uniform distribution of participants across conditions and to mitigate potential order effects.

6.3.1 Collection and System

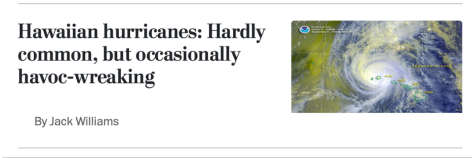
Similar to the previous experiment, we used the TREC Washington Post corpus (WaPo) that we had indexed in our system. We presented results on SERPs using these same result card types. We used the following result cards in our experiment:

1. Title + Image + Summary [TIS]
2. The Washington Post Style [TIS WaPo]
3. Google News Style, Title + Image [TI Google]
4. Title only [T]

We set our web page to display a maximum of 14 rows so that each result card can be displayed fully at least once in the random layout. The exact constraints for each card type will be shown in Figure 6.2. The total number of result cards shown on the page depends on the type of card and the number of rows it occupies according to the constraints highlighted above. For example, on a single result page, there could either be 14 title cards, 2 TIS cards, 7 TI Google Cards or 4 TIS WaPo cards.

(a) Title+Image+Summary (TIS): ∼6 rows

(b) Title+Image+Summary (TIS WaPo): ∼3 rows



(c) Title+Image (TI Google): ∼2 rows

(d) Title Only (T): ∼1 row

Figure 6.2: Example of the different result card types, with an approximation of the number of rows each card type occupies.

6.3.2 Search Topics and Tasks

In this iteration of the experiment, we focused exclusively on three topics, excluding the topic of Piracy at Sea. Consequently, the topics that we used were as follows:

1. Topic 341: Airport Security,
2. Topic 363: Transportation Tunnel Disasters,
3. Topic 408: Tropical Storms.

The exclusion of the topic of Piracy at Sea was predicated on several considerations detailed within our experimental procedure (§ 6.3.3). Primarily, the objective was to curtail participant fatigue by limiting the scope to three topics. Additionally, this particular topic was omitted due to its comparatively limited assortment of documents featuring images. Feedback from participants in the previous study also indicated that this topic presented substantial challenges, leading to its removal to maintain cognitive

consistency across the studied topics. In alignment with the previous methodologies (§5.2.2), the same generated queries were used for the selected topics.

6.3.3 Procedure

While the overarching structure of this experiment remains consistent with the previous study, several modifications were implemented to enhance its design. Participants were recruited from Prolific, similar to the previous methods. Upon accepting the terms and viewing the tutorial, participants were randomly assigned topics via a Latin square rotation, which determined the order of topic assignment and the specific optimiser allocated to each participant. Unlike the prior experiment (§5.2.4), participants were tasked with identifying and marking documents as relevant or non-relevant for the assigned topic without the need to extract specific document sections. This adjustment aimed to mitigate participant fatigue, particularly as the experiment encompassed three topics. Initially, in the first topic participants explored documents across all queries for the first topic to maximise interaction data collection for each result card type. In addition to collecting data on query and interface satisfaction, this iteration also captured interaction data (timing and click probabilities) for estimating the EPU post topic. This data informed the re-ranking of the retrieved list for subsequent topics based on calculated probabilities and timing values, enhancing the interface optimisation progressively, as shown in Figure. No minimum requirement for relevant document identification was imposed for the second and third topics, allowing participants to explore at their discretion. Following the completion of the third topic, satisfaction metrics were collected, akin to earlier stages, concluding the experiment.

6.3.4 Participants & Demographics

Participants were pre-screened for native English proficiency to ensure task accuracy, aiming to reduce language and cultural misunderstandings. This approach was taken to keep the experiment's findings consistent, though it may limit broader applicability. Participants were paid according to minimum wage laws in the UK and the US. We employed a total of 144 participants of which there were 91 male and 53 female. Their

ages ranged from 19 to 66 years. The participants reported various employment levels, the majority of which a majority were employed (90). The participants were also comprised mostly of non-students. Before conducting the study we obtained ethics approval from the department ethics committee (no. 2333) at the University of Strathclyde.

6.4 Results

6.4.1 Summary of Search Behaviours

Since the collection system for this experiment was similar to the previous experiment. We describe below a comprehensive summary of the search behaviours exhibited by users in this experiment. We examined differences in task completion rates, interaction times, and other user metrics, such as the number of queries, clicks, and time spent across various SERPs. Given the mixed-design nature of our analysis, we considered data from both between-groups and within-subjects perspectives. ANOVA tests were used to assess whether significant differences existed between the conditions and the measures under investigation. The primary effects were analysed at a significance level of $\alpha = 0.05$. Tukey's HSD tests were utilised for post-hoc analyses. For the reported tests, the F-score, p-value, and effect size η_p^2 are presented to two decimal places. The ranges of η_p^2 values correspond to small (< 0.06), medium (0.06 - 0.14), and large (> 0.14) effect size [85]. The $\pm$ values reported in the tables denote the mean and standard deviation. Within-subject comparisons were conducted using paired t-tests, and Cohen's d was utilised to report effect sizes [91].

From our analysis we found that, on average, participants inspected 5 queries, interacted with 7 SERPs per topic, and saved a total of 12 documents. The average time to initiate a click on a query was approximately 5 seconds, with an average of 19 seconds spent reviewing each SERP. The average document reading time was 13 seconds. Participants demonstrated a 66% accuracy rate in identifying relevant documents, where accuracy is defined as the proportion of documents participants deemed relevant that corresponded with the TREC qrels file. On average, participants required 37 minutes to complete all three topics. These findings are summarised in Table 6.2.

Table 6.2: Average user behaviour during the experiment. Timing data is reported in seconds, and the experiment time is reported in minutes (for easier interpretation).

(a) Count Based Behaviour

# of...			Accuracy
Q	**SERPs**	**Docs Saved**	
4.79 ± 1.38	7.13 ± 2.44	12.31 ± 5.95	66.38 ± 9.73

(b) Timing Based Behaviour

Time for ...			
Q Selection	**Browsing SERP**	**Doc Reading**	**Experiment**
5.32 ± 4.75	19.32 ± 15.56	8.36 ± 8.22	37.60 ± 15.80

Building on this foundational analysis, this section looks into the variation of performance metrics across different optimisation strategies and explores individual differences in response to progressively optimised SERPs.

Between Groups

Our analysis revealed no statistically significant differences in the number of queries issued, the number of documents inspected, the accuracy of marked documents, or the timing of interactions between both optimisers as seen from Table 6.3. However, statistically significant differences were observed in the number of SERPs viewed, with participants in the RU group viewing more SERPs than those in the TU group ($F(1,2083) = 14.98$, $p < 0.05$, $\eta_p^2 = 0.007$) and the number of documents marked as relevant ($F(1,1667)=4.001, p < 0.05$, $\eta_p^2 = 0.002$), with more documents marked as relevant by users of the RU optimiser. This behaviour indicates that while the TU optimiser led to fewer interactions in terms of SERP views and document markings, it did not compromise the accuracy, efficiency (time-wise), or timing of task completion compared to the RU optimiser. These findings suggest that increased interactions, facilitated by preferred presentation formats, do not inherently result in better identification of relevant documents.

Table 6.3: Descriptive statistics of user behaviour between the two optimisers RU and TU.

(a) Search behaviours, with the mean number of actions performed per user, per topic, per query. Here, $\mathcal{Q}$ denotes Queries, *Docs* denotes documents. The superscripts denote significant differences between groups.

Group	**#Q Issued**	**#SERPs Viewed**	**#Docs...**		**Accuracy**
			Inspected	**Marked**	
RU	4.04±1.74	$2.34 \pm 1.97^{\mathbf{TU}}$	4.64±6.03	$3.34 \pm 4.30^{\mathbf{TU}}$	0.66±0.32
TU	4.13±1.88	2.05±1.53	3.99±3.64	2.99±2.82	0.66±0.32

(b) Average timings for search behaviours per user, per query in seconds.

Group	**Q**	**R Doc**	**$\bar{R}$ Doc**
RU	5.36±4.76	9.73±08.04	10.90±08.44
TU	5.22±4.64	9.63±07.77	10.69±07.30

(c) Average time spent per topic, per user in minutes.

Group	**Topic 1**	**Topic 2**	**Topic 3**
RU	19.68±11.71	8.84±5.48	6.71±4.07
TU	18.32±10.56	9.47±7.87	6.34±3.69

Within Group

In the analysis of descriptive statistics within each optimiser group, several noteworthy findings emerged, elucidating the behavioural patterns of participants as they navigated through various topics in the task. A clear pattern of behaviour change was observed, particularly in the manner in which participants interacted with search queries and SERPs across different topics.

Initially, a significant reduction in the number of queries inspected by participants was noted when transitioning from the first to the second topic. This outcome aligns with expectations, given that participants were instructed to examine all queries within the first topic. The inspection of queries remained consistent between the second and third topics, indicating no significant variation in user behaviour in this regard.

Moreover, participants exhibited a noteworthy trend of inspecting fewer SERPs as they advanced through the topics, regardless of the optimisation strategy employed. During the initial topic, participants were observed to browse approximately three pages

Table 6.4: Search behaviours within each optimiser, with the mean number of actions performed per user, per topic, per query. Here, Q denotes Queries, Docs denotes documents. Superscripts denote significant differences within the topics in the group. For example, the number of queries issued significantly differ from topic 1 to topic 2 and topic 1 to topic 3.

Group	**Topic**	**#Q Issued**	**#SERPs Viewed**	**#Docs**		**Accuracy**
				Inspected	**Marked**	
RU	1	$6.29 \pm 1.29^{2,3}$	$2.73 \pm 1.98^{2,3}$	5.09 ± 6.61^{3}	3.58 ± 4.66	0.69 ± 0.31
	2	4.19 ± 1.73	2.24 ± 2.10	4.15 ± 5.87	3.42 ± 4.78	0.66 ± 0.34
	3	3.89 ± 1.98	1.96 ± 1.81	3.71 ± 4.87	2.89 ± 2.94	0.63 ± 0.34
TU	1	$6.15 \pm 0.40^{2,3}$	$2.52 \pm 1.88^{2,3}$	4.21 ± 4.27	3.05 ± 3.05	0.66 ± 0.33
	2	4.38 ± 1.82	1.73 ± 1.25	3.63 ± 3.30	2.90 ± 2.82	0.69 ± 0.32
	3	3.88 ± 2.12	1.75 ± 1.15	3.45 ± 2.86	2.99 ± 2.40	0.65 ± 0.33

Table 6.5: The table presents the mean and standard deviation of the number of cards per SERP, per user, topic and query. Superscripts indicate significant differences between topics within the group.

Group	**Topic**	**#Cards / SERP / Q**	**RBO**
RU	1	$7.24 \pm 1.97^{2,3}$	$1.00 \pm 0.00^{2,3}$
	2	6.30 ± 1.42	0.93 ± 0.11
	3	6.42 ± 1.43	0.97 ± 0.05
TU	1	$7.52 \pm 2.48^{2,3}$	$1.00 \pm 0.00^{2,3}$
	2	8.87 ± 4.26	0.96 ± 0.09
	3	8.56 ± 4.07	0.97 ± 0.08

before clicking on documents. This behaviour gradually shifted, with fewer pages being inspected in subsequent topics. This trend can be attributed to the randomisation of result presentation in the first topic, followed by optimised presentation in later topics, which presumably allowed users to find preferred content more efficiently on a single SERP, thereby reducing the need to inspect additional SERPs. We observe the following significant differences using paired t-tests as seen from Table 6.7,

Regarding the inspection of documents, no statistically significant changes were observed across the topics, except for the first optimisation strategy (RU), where a notable decrease in the number of items inspected was recorded in the first topic compared to the third. Group RU Topic 1 ($M1 = 5.15$, $SD1 = 5.05$) to Topic 3 ($M2 = 3.66$, $SD2 = 3.79$, $t = 2.59$, $p < 0.05$, $d = 0.33$). Conversely, the second optimisation strat-

Table 6.6: Average timings for various search behaviours during the study, within groups by each topic, per user, per topic, per query. Superscripts indicate significant differences between topics within the group.

Group	Topic	T(Q)	T(R Doc)	T($\bar{R}$ Doc)
RU	1	5.63 ± 5.01	$12.39 \pm 08.21^{2,3}$	$12.37 \pm 08.64^{2,3}$
	2	5.49 ± 5.12^{3}	08.72 ± 06.83	09.60 ± 07.58
	3	4.88 ± 3.99	08.30 ± 07.48	09.60 ± 08.55
TU	1	5.49 ± 4.83^{2}	$11.72 \pm 07.41^{2,3}$	11.66 ± 07.32
	2	5.61 ± 4.96^{3}	08.44 ± 06.45	10.12 ± 07.28
	3	4.48 ± 4.11	08.32 ± 07.31	09.55 ± 07.09

Table 6.7: Summary of Paired t-test Comparisons for SERP Inspection

Comparison	Mean (M)	SD	t-value	p-value	Cohen's d
RU T1 vs. T2	M1=2.72	SD1=1.61	t=2.74	$p < 0.05$	d=0.25
	M2=2.26	SD2=1.97			
RU T1 vs. T3	M1=2.72	SD1=1.61	t=4.45	$p < .001$	d=0.46
	M3=1.99	SD3=1.56			
TU T1 vs. T2	M1=2.51	SD1=1.52	t=6.75	$p < .001$	d=0.51
	M2=1.82	SD2=1.13			
TU T1 vs. T3	M1=2.51	SD1=1.52	t=4.24	$p < .001$	d=0.44
	M3=1.90	SD3=1.18			

egy (TU) showed consistent behaviour among users, with no significant difference in the number of documents inspected across topics, irrespective of interface optimisation.

The analysis also revealed consistent marking of documents as relevant by users across all topics and optimisation strategies, indicating no significant difference in the evaluation of document relevance. Except in the RU optimiser, we observe that a higher proportion of documents were marked as relevant after the third topic

A distinct behavioural pattern was observed in the timing aspects of user interactions. Specifically, within the RU optimisation strategy, participants took significantly less time to issue a query as they progressed through the topics. This trend was also evident in the TU optimisation strategy, albeit with a gradual reduction in query issuance time. This observation suggests that users possibly became more accustomed to the interface, enabling faster interaction.

Furthermore, a significant reduction in the time taken to read and mark a document as relevant was noted as participants progressed through the topics. This efficiency can

be attributed to the influence of card presentation on user decisions to click or skip, as well as the optimised presentation of content according to user preferences, facilitating quicker document inspection as seen from Table 6.8.

Table 6.8: Summary of Paired t-test Comparisons for Time Taken to Mark Documents

Comparison	**Mean (M)**	**SD**	**t-value**	**p-value**	**Cohen's d**
RU T1 vs. T2	M1=12.77	SD1=7.61	t=5.03	$p < .001$	d=0.48
	M2=9.45	SD2=6.18			
RU T1 vs. T3	M1=12.77	SD1=7.61	t=5.83	$p < .001$	d=0.55
	M3=8.84	SD3=6.72			
TU T1 vs. T2	M1=12.87	SD1=6.47	t=6.12	$p < .001$	d=0.71
	M2=8.72	SD2=5.17			
TU T1 vs. T3	M1=12.87	SD1=6.47	t=5.90	$p < .001$	d=0.70
	M3=8.32	SD3=6.54			

Conversely, the time taken to identify non-relevant documents also decreased, particularly between the first and last topic only in the RU group, underscoring an enhanced ability to discern non-relevant information quickly, Table 6.9 shows the paired t-tests with significant differences.

Table 6.9: Summary of Paired t-test Comparisons for Time Taken to Identify Non-Relevant Documents in RU Group

Comparison	**Mean (M)**	**SD**	**t-value**	**p-value**	**Cohen's d**
RU T1 vs. T2	M1=13.19	SD1=6.86	t=3.03	$p < 0.05$	d=0.35
	M2=10.80	SD2=6.70			
RU T1 vs. T3	M1=13.19	SD1=6.86	t=4.32	$p < .001$	d=0.49
	M3=9.79	SD3=6.87			

The analysis underscores minimal differences between optimisation strategies concerning the number of documents marked and SERPs viewed. However, within each optimisation strategy, variations were observed in the number of SERPs viewed, documents marked, and timing aspects, such as the time taken to read relevant documents and issue queries. These findings highlight the impact of result list optimisation on user behaviour, as reflected in the metrics analysed.

A good question to ask at this point is whether these behaviours are a consequence of the optimisation strategy or simply an artefact in the experimental methodology that caused users to change their behaviour. We assert that the changes in behaviour

are caused by the optimisation and we provide the following pieces of evidence from our data to support our hypothesis.

Change in the number of cards shown

Firstly, as we can observe from Table 6.5, in the Rate Optimiser, there is a discernible decrease in the average number of cards viewed per page when compared to a random layout Topic 1 ($M1 = 7.23$, $SD1 = 1.20$) to Topic 2 ($M2 = 6.27$, $SD2 = 0.97$, $t = 5.13$, $p < 0.001$, $d = 0.89$) and Topic 1 ($M1 = 7.23$, $SD1 = 1.20$) to Topic 3 ($M2 = 6.44$, $SD2 = 0.98$, $t = 4.17$, $p < 0.001$, $d = 0.72$). This suggests that the optimiser effectively streamlined the presentation of information, allowing participants to encounter fewer cards but still maintain accuracy in identifying relevant information. Notably, despite the reduction in the number of documents presented, participants in the Rate Optimiser marked a greater proportion of documents as relevant, indicating a more targeted and efficient search experience.

Conversely, within the Total Utility Optimiser, participants were exposed to a higher number of cards on average. Topic 1 ($M1 = 7.48$, $SD1 = 1.40$) to Topic 2 ($M2 = 9.03$, $SD2 = 3.92$, $t = -3.31$, $p < 0.05$, $d = -0.53$) and, Topic 1 ($M1 = 7.48$, $SD = 1.40$) to Topic 3 ($M2 = 8.59$, $SD = 3.80$, $t = -2.35$, $p < 0.05$, $d = -0.39$). However, this increase did not correlate with a higher number of documents being inspected, suggesting that the additional information presented did not overwhelm the participants or significantly alter their inspection behaviour. This could imply that although participants were presented with more options, they remained selective, focusing on relevance rather than quantity.

No change in accuracy of marked documents

Given that the accuracy of the marked documents did not significantly change between topics as the interface became more optimised, these observations support the hypothesis that the layout optimisation was effective and participants indeed experienced a variation in SERPs. The Rate Optimiser, by presenting fewer (and possibly more relevant) results, appears to have facilitated a more efficient search process, reducing the need for participants to view multiple SERPs. This efficiency is highlighted by the

participants' ability to discern and mark a higher proportion of the reduced number of presented documents as relevant, optimising their search strategy.

Change in RBO

We also observe from Table 6.5, that there is a significant shift in RBO – meaning that the orderings of the documents significantly shifted from the original baseline BM25 ranking. We know that from the first topic, poorly presented documents caused users to needlessly view more SERPs. Users in this case needed to look at fewer SERPs to find relevant information more quickly.

In summary, the application of these optimisation strategies demonstrates a clear influence on user behaviour. The Rate Optimiser aligns closely to enhance the relevance of information encountered by the user, thereby minimising the need for extensive SERP inspections. The Total Utility Optimiser maintains inspection behaviours while increasing the breadth of information presented, suggesting a balance between information quantity and user selectivity.

Given these insights and the inclusion of a self-report questionnaire, the subsequent section of this thesis will explore whether users perceived these behavioural changes sufficiently to report them, addressing the first research question. This examination aims to bridge the gap between observed behavioural data and user perceptions, offering a comprehensive understanding of the implications of result list optimisation on user interaction patterns.

6.4.2 RQ2: How do user satisfaction and cognitive load metrics evolve as the user interface is iteratively optimised across multiple topics within a search session?

Our second research concerns the evolution of cognitive load and other user satisfaction metrics as the ranked list and interface was progressively optimised. The analysis was grounded on the comparative assessment of mean metrics across various stages of user interaction with the topics, as seen in Table 6.10.

Upon initial examination of the data, it might appear that certain metrics, such as overall satisfaction, exhibit fluctuations with the optimisation of the interface. How-

Table 6.10: Comparative Analysis of Participant Responses by Group and Topic

(a) Topic and Overall Satisfaction

Group	Topic	Overall Satisfaction
RU	1	4.21 ± 1.14
	2	4.03 ± 1.34
	3	3.97 ± 1.35
TU	1	4.24 ± 1.07
	2	4.23 ± 1.06
	3	4.17 ± 1.26

(b) Other Metrics

Group	Topic	Distracting	Liked Engine	Felt Productive	Mentally Taxing
RU	1	2.75 ± 1.27	3.93 ± 1.12	4.07 ± 1.13	3.36 ± 1.38
	2	2.99 ± 1.44	3.85 ± 1.31	3.93 ± 1.37	3.46 ± 1.26
	3	3.00 ± 1.45	3.90 ± 1.31	3.99 ± 1.41	3.45 ± 1.38
TU	1	2.78 ± 1.25	4.01 ± 1.08	4.06 ± 1.17	3.24 ± 1.27
	2	2.71 ± 1.14	4.03 ± 1.05	4.03 ± 1.17	3.19 ± 1.27
	3	2.79 ± 1.08	4.11 ± 1.10	4.08 ± 1.32	3.23 ± 1.36

ever, a deeper statistical analysis reveals that these changes do not reach statistical significance. We performed ANOVA tests to ascertain if any statistically significant differences existed between groups, alongside paired t-tests to investigate within-group differences. The outcomes of these analyses uniformly indicate the absence of statistically significant variances, suggesting that the users did not perceive the modifications in the interface strongly enough to report a significant shift in the measured metrics.

Interestingly, despite observable alterations in user behaviour patterns attributable to the interface optimisations, such changes were not mirrored in the satisfaction metrics. This dichotomy suggests that while the optimisation indeed influences user behaviour, the degree of change does not surpass the threshold needed for users to report it. This implies that if there is any change, it is minimal, and the study's design is only equipped to detect changes of a medium scale with reliability. Consequently, to ascertain the presence of subtle changes, a study with greater detection power would be necessary.

Given these findings, it becomes important to explore how user preferences either converged or diverged in light of the observed optimisations. The subsequent research

question aims to further dissect the evolution of the interface as the topics progressed, seeking to categorise user preferences and understand the implications of these behavioural patterns on the optimisation process.

6.4.3 RQ3: To what extent do user preferences converge towards a unified SERP configuration, and what are the cognitive load variations associated with different SERP optimisation strategies across tasks?

To address the second research question on how the optimisations converged or diverged, it was first essential to ascertain whether optimisations in search result presentation yield noticeable satisfaction changes among users. Despite preliminary observations suggesting minimal discernible differences in user satisfaction, irrespective of whether the optimisations increased or decreased satisfaction levels, a thorough examination is warranted to determine if these optimisations converge at any point. This will allow us to ascertain the level of differences in the presentation of the SERPs to determine how the optimisations evolved.

To facilitate this investigation, we devised a methodological approach, focusing on the visualisation of the shown SERPs encountered by users. Given that we had recorded the card types for each result on a SERP, we first constructed an abstract representation of the shown SERP for a query for a user. We assign a colour to each result card in our study and construct a square image to represent the SERP. In our image, we also considered the spatial constraints of different card types, as observed in simulations. The resultant pictorial abstract representations of SERPs serve as a foundation for subsequent analysis.

To evaluate the similarity of these visual representations, t-Distributed Stochastic Neighbour Embedding (t-SNE), a dimensionality reduction algorithm, is employed. This technique allows for the high-dimensional data associated with SERPs to be projected onto a two-dimensional plane, facilitating an analysis of clustering patterns. Each data point represents a SERP corresponding to a user query, enabling the identification of patterns across multiple queries per user and allowing us to observe whether clusters

of similar SERP layouts emerge. We can observe how these clusters have formed from Figures 6.3-6.8.

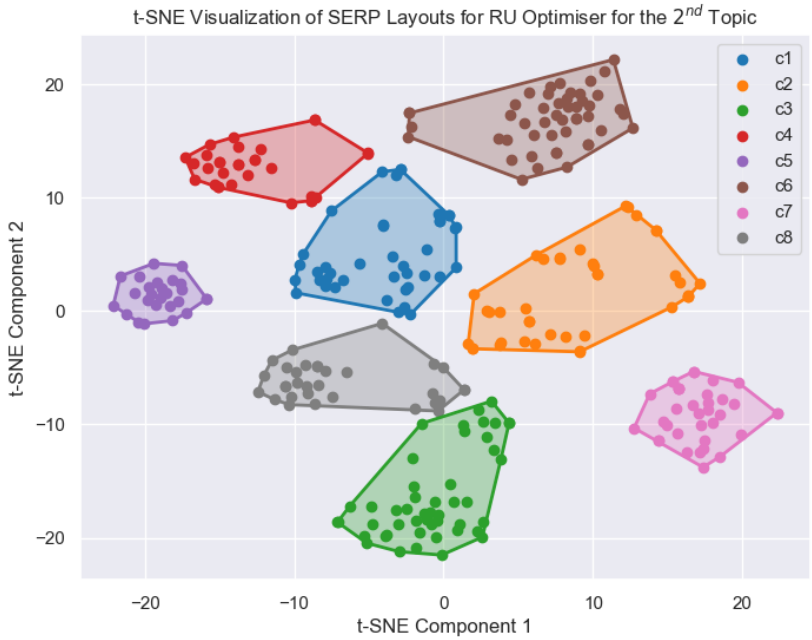


Figure 6.3: t-SNE Visualisation of SERP Layouts for RU Optimiser when Topic Count is 2

Cluster Analysis

Our analysis offers significant insights into the optimisation of user interface layouts. When the focus of the optimisations was on enhancing the rate of utility, the resulting data clusters demonstrated a considerable degree of overlap. This overlap suggests a uniformity in layout preferences among users, as depicted in Figures 6.3-6.5. In contrast, when the optimisation endeavoured to maximise total utility, a clearer distinction in the clustering emerged. This distinction indicates a trend towards specific layout preferences among particular groups of users, as evidenced by the distinct clusters formed around different card types, visible in Figures 6.6-6.8. We also observe from Figure 6.6-6.8 that the points within the cluster becoming more dense and thus exhibiting lesser variance, suggesting that if we were to optimise the interface even further we might have observed very little difference in user preferences overall.

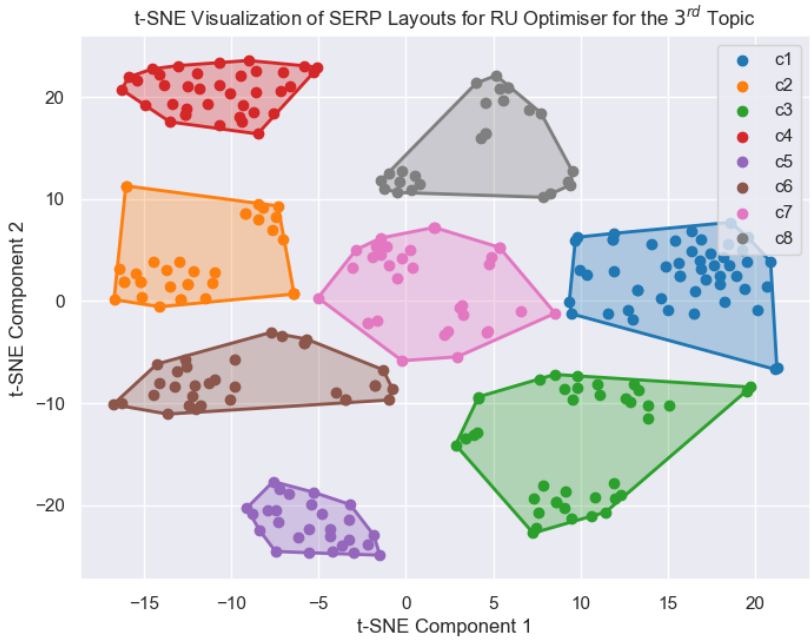


Figure 6.4: t-SNE Visualisation of SERP Layouts for RU Optimiser for Topic 3

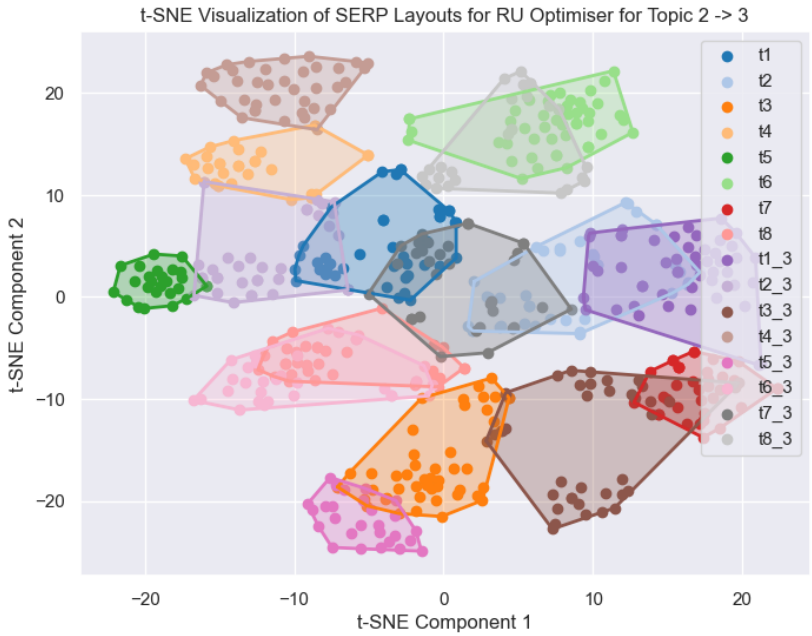


Figure 6.5: t-SNE Visualisation of SERP Layouts for RU Optimiser from Topic 2 to 3

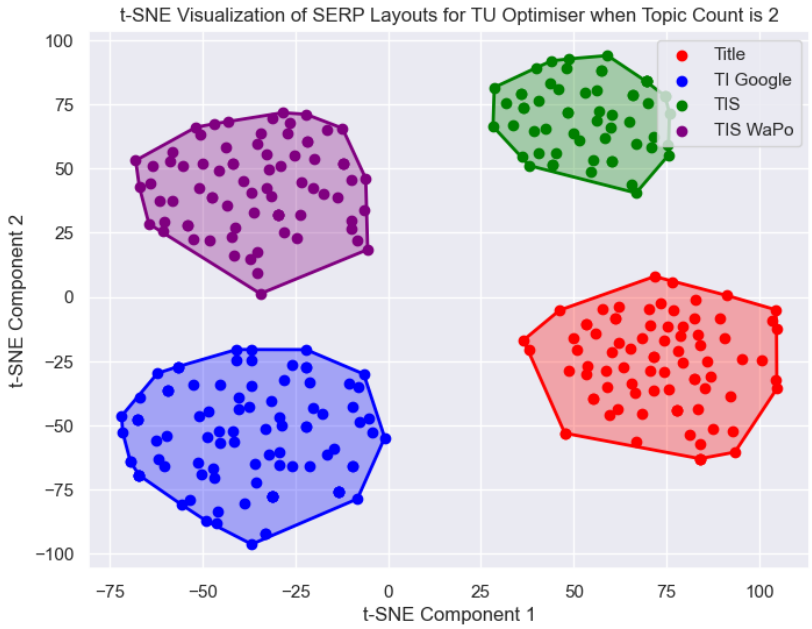


Figure 6.6: t-SNE Visualisation of SERP Layouts for TU Optimiser when Topic Count is 2

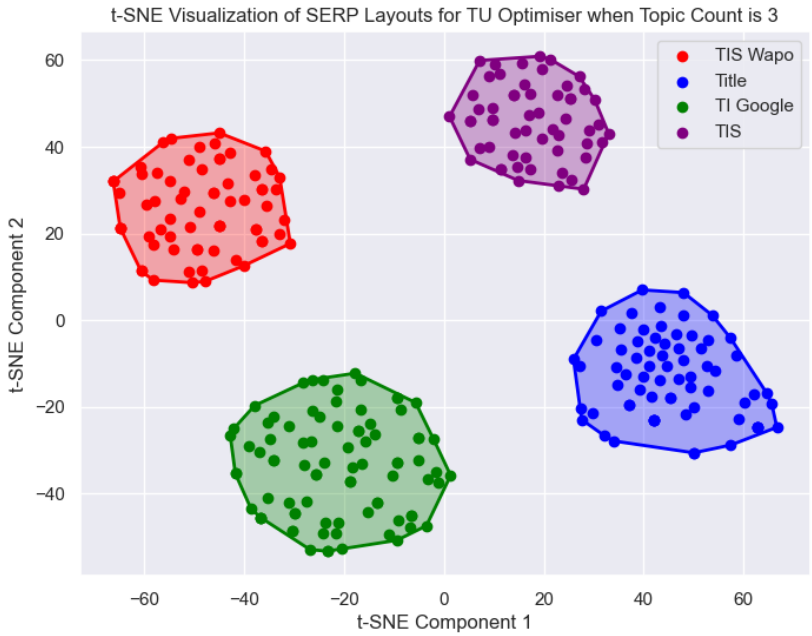


Figure 6.7: t-SNE Visualisation of SERP Layouts for TU Optimiser for Topic 3

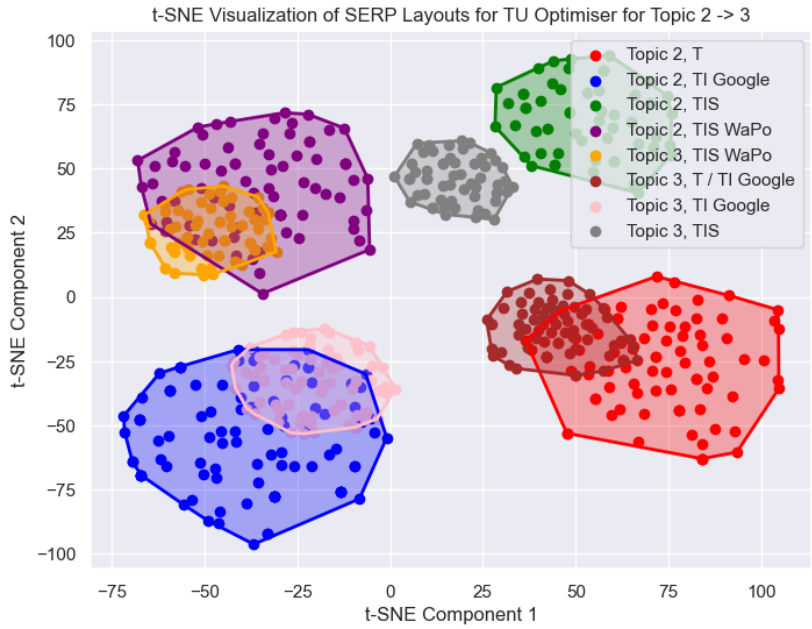


Figure 6.8: t-SNE Visualisation of SERP Layouts for TU Optimiser from Topic 2 to 3

Figure 6.9 and Figure 6.10 highlight the distribution of card types within each cluster. In the TU optimiser, we can clearly see a dominant distribution of card types within each cluster. For example, in the TIS cluster, the TIS card is dominant, with the participant seeing the other card types as well, but only at later ranks. Whereas, in the RU optimiser, such a clear differential is not observed. Instead, we observe varying distributions of different card types in each cluster. We also interestingly observe that the TI Google card type is more dominant in each distribution. This suggests that the cost associated in displaying multiple such cards (even though their individual utility is lower) is outweighed by the overall benefit gained from them.

The distribution of users across the clusters is further elaborated in Figure 6.11. For instance, within the context of the rate optimiser for the second topic, it was observed that 67 out of 72 users were categorised into multiple clusters. Similarly, for the third topic using the rate optimiser, 60 out of 72 users were found to belong to more than one cluster, with the most frequent cluster membership being Cluster Three. Notably, Cluster 3 predominantly comprised TIS and TIS WaPo cards.

Conversely, the total utility optimiser presented a different pattern. For the second

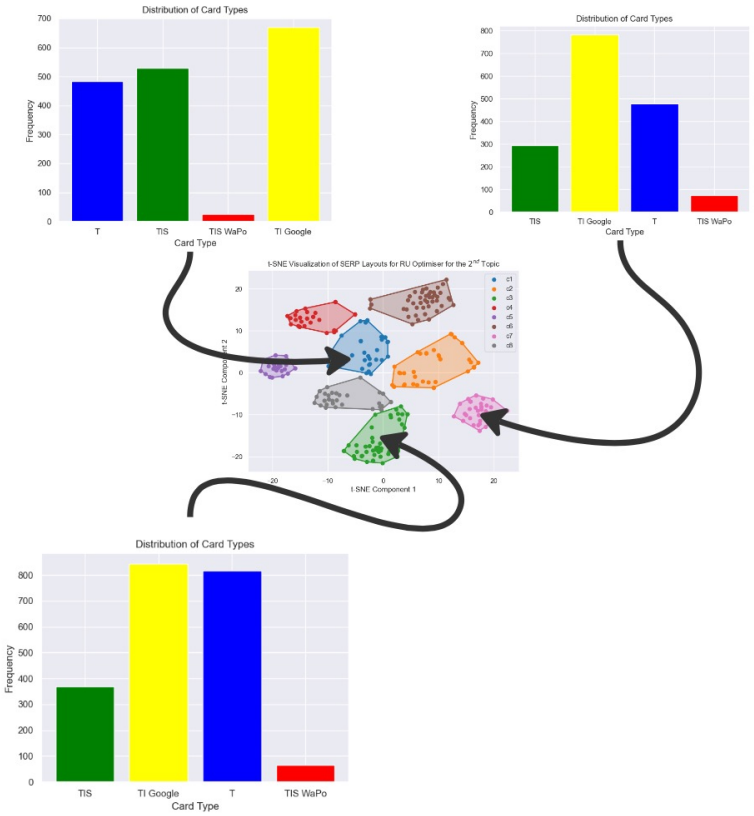


Figure 6.9: Distribution of card types inside the cluster blobs for the RU optimiser

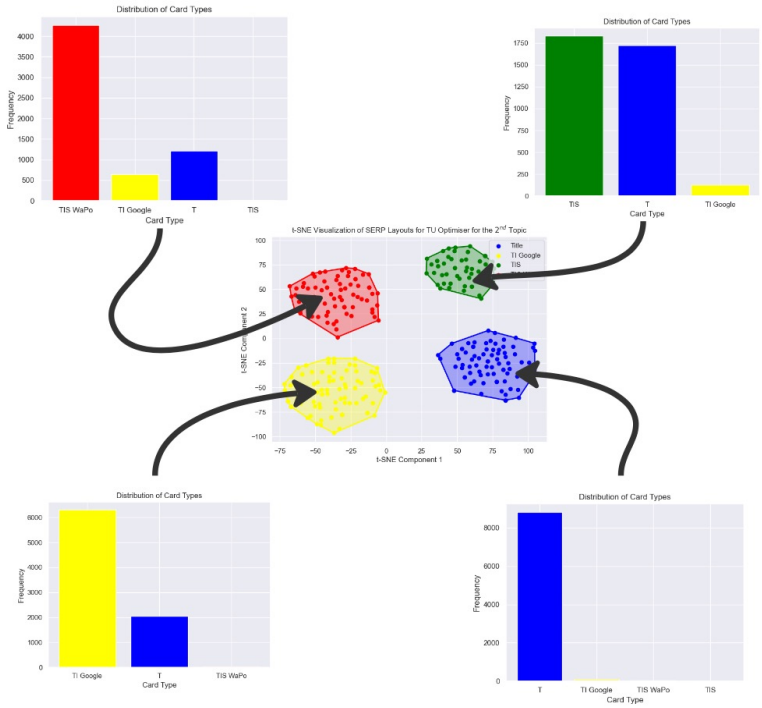
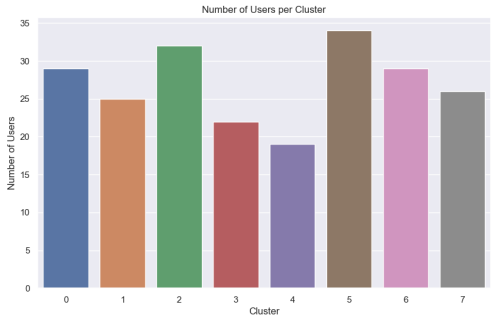


Figure 6.10: Distribution of card types inside the cluster blobs for the TU optimiser

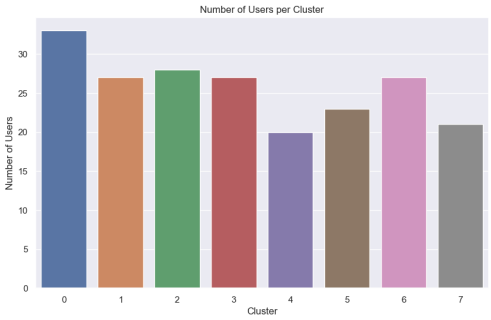
topic, only 19 out of 72 users were identified in multiple clusters. For the third topic, the number decreased to 11 out of 72 users, with nobody being part of all clusters. The majority of users were found to be associated primarily with TIS cards, highlighting a more selective clustering phenomenon compared to the rate optimiser scenario. This analysis underscores the impact of optimisation strategies on user grouping and layout preference identification.

Further analysis of cluster dynamics revealed that a significant proportion of users ($>$ half) transitioned to a new cluster when the optimisation criterion shifted from the second to the third topic under the rate optimisation. This suggests that these users' preferences evolved, aligning with distinct layout types. In contrast, users subjected to the total utility optimisation exhibited minimal movement between clusters, indicating a persistence of layout preferences.

These observations underscore the influence of optimisation strategies on user interaction patterns with SERPs, albeit without a significant perceptual difference or impact



(a) Rate Optimiser Topic 2



(b) Rate Optimiser Topic 3

(c) Total Optimiser Topic 2

(d) Total Optimiser Topic 3

Figure 6.11: Number of Users per Cluster

on satisfaction levels. This is further corroborated by the Rank-Biased Overlap (RBO) metric, which demonstrated only a weak correlation between improvements in RBO and user satisfaction. This suggests that while optimisations tailored to user preferences may slightly enhance satisfaction, the overall perceived impact remains marginal and its significance is yet to be fully understood.

This underscores the critical role of result presentation in search ranking processes, advocating for an approach that emphasises the explicit consideration of presentation aspects without resorting to deep learning or computationally intensive methods. This strategy aims to facilitate explainable rankings based on user preferences without implicitly encoding extensive information.

In conclusion, while the current investigation sheds light on the nuanced dynamics between optimisation strategies and user perceptions, it also highlights the need for further research. The subsequent chapter will delve into areas requiring improvement and potential avenues for future inquiry, thereby encapsulating the thesis's overarching contributions and delineating a road-map for continued exploration in this domain.

6.5 Summary

In this chapter, we continued our investigation into the impact of presentation on ranking effectiveness and user satisfaction. Drawing from the insights of the preceding two experiments (which answered **HL-RQ1** and **HL-RQ2**), it was clear that the presentation of different result cards on SERPs influences key metrics such as query satisfaction, nDCG and total gain on a page. These initial findings led to the hypothesis that optimising ranked lists by considering factors such as presentation and EPU, as introduced in the first experiment (§4), could enhance user satisfaction and their ability to locate relevant information.

To address the third overarching research question (**HL-RQ3**) concerning the effects of interface optimisations on user satisfaction, we answered three crucial subsidiary questions. The first aimed to determine whether distinct optimisation strategies exist that could maximise utility within a given space. We identified that optimisation

could be approached by either maximising the rate of utility gain per page or the total utility gain per page. This differentiation led to the second question of whether these strategies yield divergent optimisations based on user behaviour variations, especially considering the pace of user interaction. Through simulations of different user behaviours (answering the first RQ), we discovered that optimisation strategies indeed result in varied page configurations contingent on user speed. This finding underscored the time-sensitive nature of EPU and suggests that different user behaviours necessitate distinct page formations.

With this understanding, we examined whether users could perceive these optimisations, as measured by user satisfaction metrics. We found that the progressive optimisation did not translate into statistically significant changes in user perception. This observation raised further questions about whether users exhibiting similar behaviours encountered comparable SERPs (i.e., did SERPs converge or diverge?). To investigate this, SERP layouts were converted into images, and an unsupervised machine learning technique, t-SNE, was employed to ascertain if users clustered together based on the optimisations they experienced.

Our analysis revealed distinct patterns of optimisation. Specifically, for the rate optimiser, the type of optimisation a user encountered was highly dependent on the query, resulting in disparate optimised SERPs with minimal overlap between users. Conversely, for the total utility optimisation, users could be categorised into four distinct clusters based on their card type preferences. Further examination within each cluster did not reveal significant shifts in reported user behaviour across topics, further confirming our findings from RQ2. Therefore, despite the optimisation of users' ranked lists and the resulting behavioural adaptations, such as needing to view fewer pages and spending less time on relevant items, these changes were not perceived strongly enough by users to be reported. This finding concludes the investigation of the third high-level research question (**HL-RQ3**), marking the end of this thesis. The subsequent chapter will discuss the implications of these findings and suggest directions for future research.

Chapter 7

Conclusion

In this thesis, the objective was to contribute to the understanding and improvement of document rankings in a result list via consideration of result presentation. Initially, through our background research we established that the information retrieval process involves multiple steps, including query issuance, document retrieval based on matching functions, and the presentation of results to users. These various steps can be looked at from two perspectives: the system-side and the user-side. We explored work through both these perspectives in considerable detail, revealing that there are costs associated with both the user and system sides. Numerous studies have aimed at optimising these costs through various models. The evolution from simple retrieval systems to complex interactive information retrieval (IIR) systems highlighted the significant impact of user interaction in the retrieval process. This led to a spectrum of models focusing on either user or system aspects, with pioneering work such as the Probability Ranking Principle (PRP) suggesting the ranking of results by decreasing relevance. This principle underpinned many subsequent models aimed at optimising document ranking, which we covered in detail in §2.

In our background, we also examined the influence of result presentation on user perception, noting how factors like layout and presentation styles affect satisfaction. We posited that the optimal ranked list combines relevant documents with engaging formats, which we call result cards. The Interactive Probability Ranking Principle (iPRP) was identified as a suitable basis for document ranking, acknowledging varying

user interactions with different items, which was not addressed by other models. We grounded our analysis by implementing the card model, an extension of the iPRP, which additionally highlights the trade-off between utility and screen space with different presentation formats.

We posited and addressed three main research questions (**HL-RQ1**,**HL-RQ2** and **HL-RQ3**) for our goal to optimise document ranking and presentation. **HL-RQ1** explored how different presentation formats are perceived by users, introducing the concept of expected perceived utility (EPU) to account for varied user preferences and perceptions. Experiments revealed distinct presentation preferences, although they focused on individual utility measurement without considering the overall Search Engine Results Page (SERP) impact. This led to **HL-RQ2** about the interplay between system and user aspects, particularly whether presentation could compensate for lower query quality to maintain user satisfaction. We conducted a more sophisticated, but still controlled experiment which confirmed that interface design influences satisfaction and corroborated system-side metrics' such as nDCG@10 and total gain on the page, showing the influence of presentation and performance on user satisfaction.

The third, and final research question (**HL-RQ3**) aimed at identifying strategies to optimise SERPs based on the space-utility trade-off. Through simulation, distinct optimisation strategies were developed, along with an efficient algorithm. This optimisation appeared to improve user interaction efficiency, though users did not report significant differences in satisfaction or cognitive load. However, users did notice optimised SERPs, indicating subtle but unreported perceptions of improvement.

7.1 Limitations and Future Outlook

Acknowledging the limitations of this work is crucial. The use of time to estimate user benefits and costs, while logical, oversimplifies the utility estimation, especially for tasks beyond ad-hoc searches. For example, in future work we could estimate the benefits depending on the task at hand. Suppose we know that the information need is fixed, we can assign a fixed total benefit to a given query, and the user collects these fixed benefits as the progress through exploring relevant documents. This way, the

benefits acquired by the users would not depend on time and remain dimensionless, allowing us to also estimate the other costs differently.

In our study we approximated mouse movements to eye-gaze, and while lots of research has shown this to be reasonable measure of attention, it is still imperfect. This complicates the accurate assessment of time spent per item. As eye tracking technology becomes increasingly cheap and more sophisticated (see Apple Vision Pro, 2024 or Meta Quest 3), we can completely make the shift to using user gaze to estimate attention.

In this thesis, we proposed an optimisation algorithm based on DP, that utilises the estimated EPU along with a dimensional estimation of the card size to give the optimal presentation strategy. However, in reality, elements that are rendered on SERPs are not going to be of similar dimensions. For example, future SERPs may choose to render custom elements based on a user need. Users may also use completely different services that do not require ranking of items, but instead have information combined from multiple sources and presented to them. Future directions can look into how to best optimally combine several documents though latest innovations such as Retrieval Augmented Generation (RAG) to better present retrieved results. Also, our optimisation algorithm, though effective, does not allow fine-tuned control over the degree of optimisation over the ranked list. There could be a finely balanced optimisation which is an in-between of the user preference and the system re-rank. Finally, the experimental setup's controlled nature, including pre-selected queries and pre-defined test collections, may limit the findings' applicability to real-world ad-hoc search scenarios. Future work could look at other types of search tasks with more open test collections and realistic user behaviour.

In conclusion, the research I carried out for the investigations in this thesis has been both a rewarding and challenging endeavour, significantly contributing to the understanding of how presentation can affect the ranking of items. Future research should aim to refine EPU calculations and explore ways to translate these findings into practical applications, enhancing both the fairness and transparency of ranked lists. The appendices provide additional insights into information retrieval and supplementary

analyses not directly related to the research questions.

Appendix A

Some Additional Background

A.1 The IR Process

A.1.1 Indexing

Indexing is a key component in search systems, as it organises documents in a way that helps in matching queries with relevant documents. This process requires extra storage space, but this is necessary for quick information retrieval.

There are two main types of indexes: direct and inverted. An inverted index, also known as a postings list, stores a mapping from content, like words or numbers, to where they appear in a document or set of documents. This is different from a forward index, which maps documents to their content. The main purpose of an inverted index is to allow quick searches of full text, but it requires more processing when adding a document to the database. The inverted index is often used as the main data structure in document retrieval systems [92–94]

For this thesis, an inverted index will be used because we need to match full-text documents quickly. To build an inverted index, which we will call a 'document index' in this thesis, two things are needed: the document corpus and the indexing process. The document corpus is a collection of documents we want to search in our Information Retrieval (IR) system. The indexing process has three main steps: tokenization, stop word removal, and stemming.

Tokenization: Tokenization is the process of breaking down the text in a document

into smaller units, known as tokens. A lexical token is a string that has a specific meaning, different from the probabilistic tokens used in large language models. In our context, we focus on lexical tokens, which consist of a token name and sometimes a token value. The token name categorises a token based on rule-based lexical units [95].

Stop Word Removal: Stop words are common words that are typically filtered out in the processing of text for natural language processing and information retrieval due to their perceived lack of significance. These words often include articles, prepositions, and conjunctions, such as "the", "and", "but", "or", "in", and "on". There is no universal list of stop words used across all-natural language processing tools, and the selection of stop words can be tailored based on specific requirements. For example, in a search query, words such as "is", "at", "which", and "on" might be ignored to focus on more meaningful words. The trend in information retrieval systems has evolved from using extensive stop lists, comprising 200–300 terms, to much smaller lists, or even opting to not use a stop list at all. The choice depends on the specific context and goals of the information retrieval system [96, 97].

Stemming: Stemming in linguistic morphology and information retrieval refers to the process of reducing words to their basic form, which might not be the actual root of the word. This helps in treating related words as synonyms, a process known as conflation. Stemming algorithms have been a topic of study since the 1960s, and they play a crucial role in information retrieval systems by expanding queries to include various word forms [98, 99].

A.2 Retrieval Models

A.2.1 Boolean Model

An age-old model for IR based on set theory is the Boolean model, first introduced in 1950 and widely used at that time. The Boolean model of information retrieval is an example of a set-theoretic approach within the broader classification of retrieval models [100, 101]

In the Boolean model, the documents $\boldsymbol{d_i}$ and queries $\boldsymbol{q_i}$ are represented using sets

of terms. A query in this model is formulated using Boolean operators such as AND, OR, and NOT. These operators are used to specify the logical relationships between terms in the query. For instance, a query represented as $\boldsymbol{q_i}$ = "term1" AND "term2" would retrieve documents that contain both 'term1' and 'term2'. Similarly, using OR would retrieve documents that contain either of the terms.

Mathematically, the similarity function $\boldsymbol{sim}(d_i \mid q_i)$ in the Boolean model evaluates to a binary outcome – either a document is relevant (true) or not relevant (false) to a given query. This model does not rank the results based on relevance but merely selects documents that exactly match the query criteria.

While the Boolean model provides a clear and straightforward mechanism for matching queries to documents, its limitations are notable. It lacks the nuance of ranking documents by relevance and does not account for partial matches or the frequency of terms within the documents. This often leads to either too many or too few results, depending on the specificity of the query. Therefore, the order in which a user may evaluate returned results is not deterministic.

In the context of the broader mathematical constructs of information retrieval, the Boolean model serves as a foundational approach, demonstrating the application of set theory in document retrieval. However, the evolution towards more sophisticated models like probabilistic and vector space models reflects the need for more refined and user-centric approaches in information retrieval systems.

A.2.2 Vector Space Model

The Vector Space Model (VSM) represents documents and queries as vectors of identifiers, where each vector

$$d_i = (w_{1,i}, w_{2,i}, \ldots, w_{n,i})$$

where w represents a term within a document. And the queries can be represented by

$$q_i = (w_{1,i}, w_{2,i}, \ldots, w_{n,i})$$

In VSM, ranking is achieved by comparing the similarity between these vectors. If a term occurs in a document then its value is non-zero. Typically terms are single words, keywords, or longer phrases. Each dimension corresponds to a separate term in the document, and if the terms are chosen to be words, the dimensionality of the vector is the number of words in the vocabulary of the corpus. Documents and queries can then be compared for similarity using standard vector operations. The cosine similarity is a well-known method to achieve this. In the cosine similarity we measure the deviation of the angle between the document and query vector, a value closer to 1 suggests a higher match, enabling the ranking of documents in relation to the query as shown in Equation A.1:

$$sim(d_i \mid q_i) = \cos(d_i, q_i) = \frac{\mathbf{d_i} \cdot \mathbf{q_i}}{\|\mathbf{d_i}\| \, \|\mathbf{q_i}\|} = \frac{\sum_{i=1}^{N} w_{i,j} w_{i,q}}{\sqrt{\sum_{i=1}^{N} w_{i,j}^2} \sqrt{\sum_{i=1}^{N} w_{i,q}^2}} \tag{A.1}$$

The precision of the retrieval process in information retrieval systems can be significantly enhanced by implementing a weighting scheme. Such schemes are designed to assign varying weights to words, thereby prioritising terms that contribute more substantially to the discriminative power of the retrieval process.

One commonly used weighting is Term Frequency (TF), which considers the frequency of a term's occurrence within a document. However, to refine the retrieval process further, the concept of Inverse Document Frequency (IDF) has been introduced. Proposed by Karen Sparck Jones, IDF is predicated on the principle that the importance of a term is inversely proportional to its frequency across all documents in the index. This approach effectively reduces the weight of common, non-discriminative terms, while amplifying the significance of rarer, more discriminative terms.

The amalgamation of TF and IDF, known as TF-IDF, offers a comprehensive metric that encapsulates both the frequency and the importance of terms. Calculated as the product of the Term Frequency (TF) and the Inverse Document Frequency (IDF), the TF-IDF value for a term in a document is given by

$$\text{TF-IDF}(w, d_i, D) = \text{TF}(w, d_i) \times \text{IDF}(w, D) \tag{A.2}$$

where, TF is the number of times (denoted by f or frequency) the term w appears in a document d_i (f_{w,d_i}) divided by the total number of terms in d_i

$$\mathrm{TF}(w, d_i) = \frac{f_{w,d_i}}{\sum_{w' \in d_i} f_{w',d_i}} \tag{A.3}$$

and IDF(w,D) is the logarithm of the number of documents (N_D) divided by the number of documents containing term w (N_{D_w}).

$$\mathrm{IDF}(w, D) = \log \frac{N_D}{N_{D_w}} \tag{A.4}$$

Appendix B

A better Intuition of EPU

In this section, we aim to deepen our understanding of how certain key components, particularly the probability of relevance and the benefits of selecting a relevant item, shape the Expected Perceived Utility (EPU) function. We will explore how adjustments to the probability of relevance can impact the EPU and investigate the effects of altering the benefits associated with clicking on a relevant item on the EPU, as well as on Discounted Cumulative Gain (DCG), Rank-Biased Overlap (RBO), and Time-Biased Gain (TBG). This analysis will be grounded in the data derived from our initial user study, specifically focusing on the variables presented in Table 4.1a. Moreover, by applying these considerations across all 50 TREC topics within the Washington Post (WaPo) collection, we aim to quantify the extent to which topic variability influences these metrics.

Figure B.1 elucidates the relationship between the probability of relevance, denoted as P(R), and the EPU for different card types. The graph demonstrates a significant trend: the EPU transitions into a positive domain once P(R) exceeds approximately 20%. This suggests that, beyond this threshold, it becomes increasingly advantageous to present results in either the TS or TIS formats. The '+' symbols mark the crossover points on the graph, highlighting, for instance, that at a 38% probability of relevance, the TS card type becomes preferable over the T card. Moreover, the graph illustrates that the TI card type consistently underperforms relative to other formats, even when the displayed item is of high relevance. This indicates that merely improving the

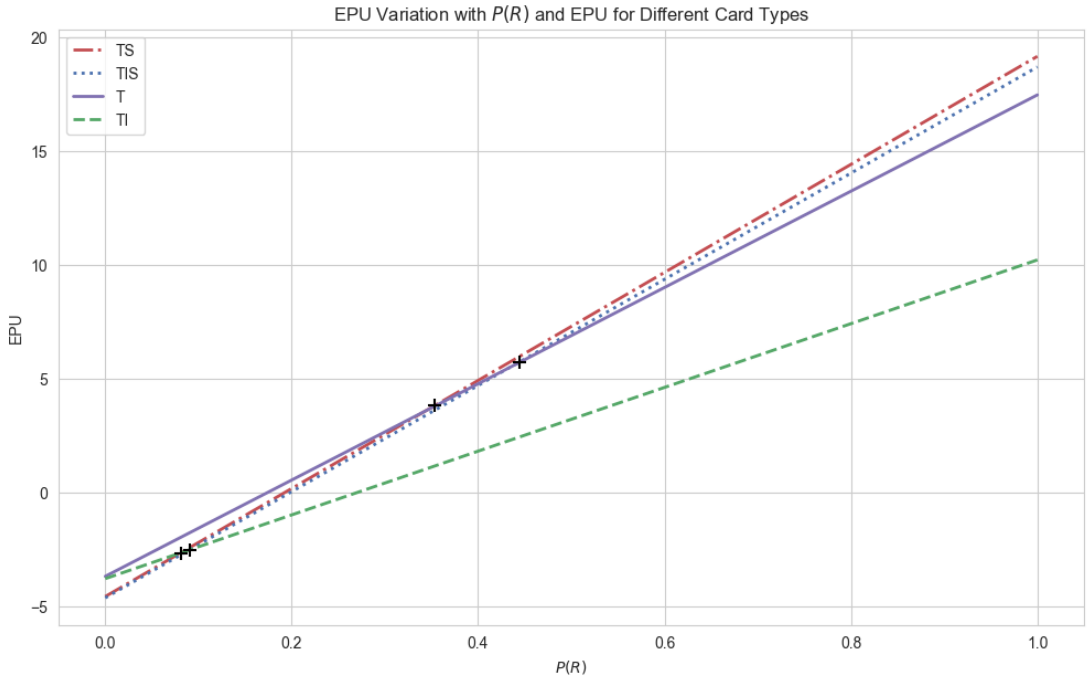


Figure B.1: $P(R)$ vs EPU_{card}

presentation format of a significantly less relevant item is unlikely to substantially alter its ranking if sorted by EPU. However, when the relevance is marginal, optimising the presentation could potentially enhance user satisfaction without the need to promote inferior items.

The graph in Figure B.2 offers a clear visual on how the Expected Utility (EPU) for different card types responds to the benefit $B(c|R)$.It highlights that while an uptick in benefit typically corresponds to a rise in EPU, it's only with notable changes in the time spent reading that this difference becomes pronounced. The similarity in EPU across card types, when benefits are equal, tells us that user engagement time with the content, rather than just the content's relevance, shapes the EPU. Our approach to cap the benefit reflects a real-world scenario where spending more time on an item does not necessarily increase its utility. Our data also shows that users' reading time does not fluctuate much between relevant and non-relevant items, underscoring that it's the likelihood of relevance that's a key driver for EPU. However, when we factor in the benefits, it creates subtle variations in EPU between different card types, allowing for

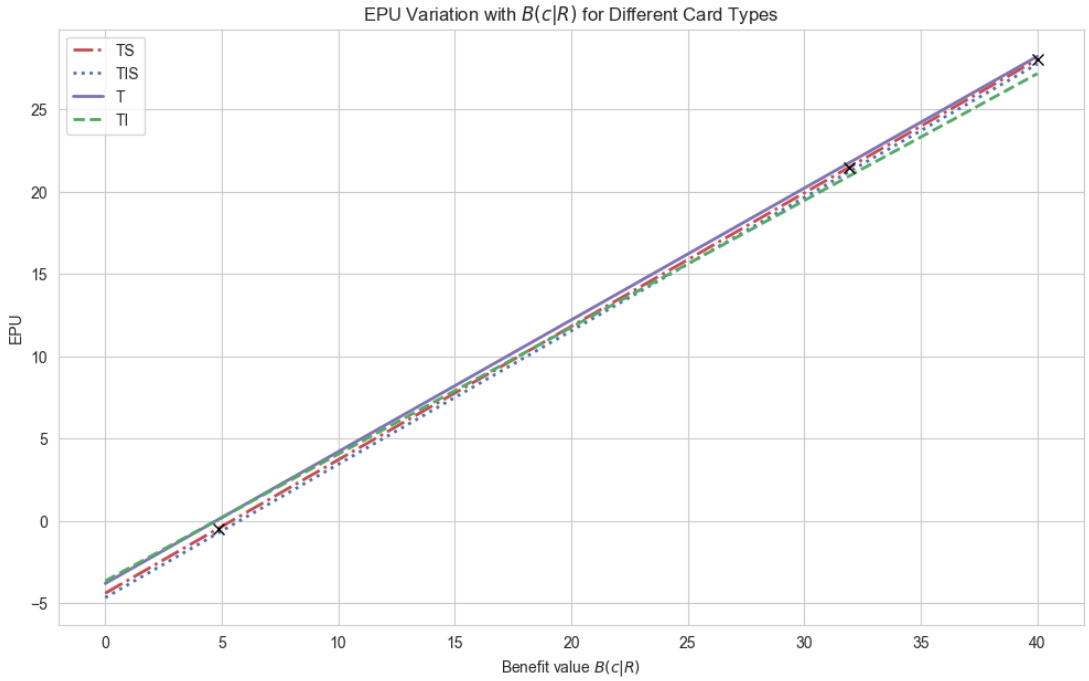


Figure B.2: $B(c|R)$ vs EPU_{card}

a way to optimise ranked results to consider presentation. Moving forward, we'll need to understand how tweaking the benefit for relevant items affects our key metrics like DCG, RBO, and TBG, to ensure that adjustments in benefits don't unduly influence these rankings.

Examining Figure B.3, it becomes apparent that DCG values remain consistent regardless of the benefit changes. In contrast, Figure B.4 reveals a notable initial fluctuation in RBO for title cards; this implies that if a user spends a brief period, such as two seconds, to assess relevance, it could significantly alter the ranking. However, typical user behaviour involves a more extended evaluation time, around ten seconds. As for the Time Biased Gain, as illustrated in Figure B.5, it is clear that the initial seconds of reading time have a pronounced effect on TBG values. This aligns with the TBG's intent to reflect the time spent on an item, affirming that while benefit influences TBG, it doesn't lead to major shifts in DCG and rankings.

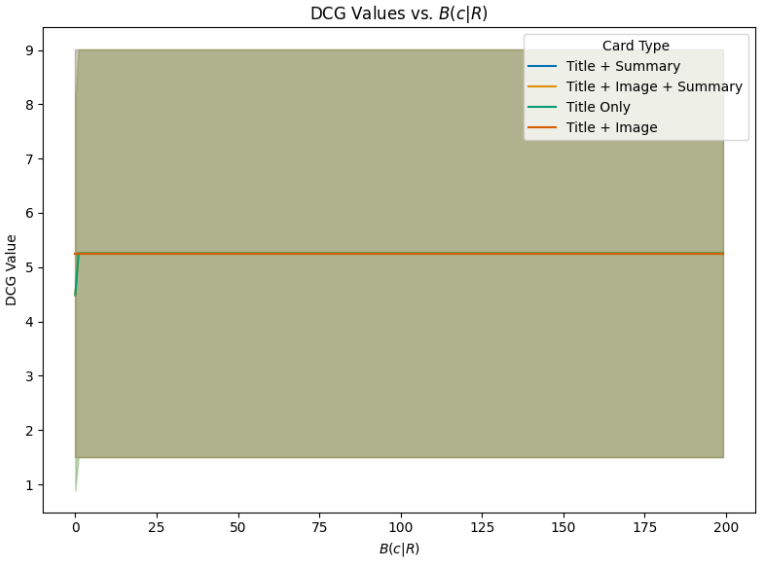


Figure B.3: $B(c|R)$ vs DCG

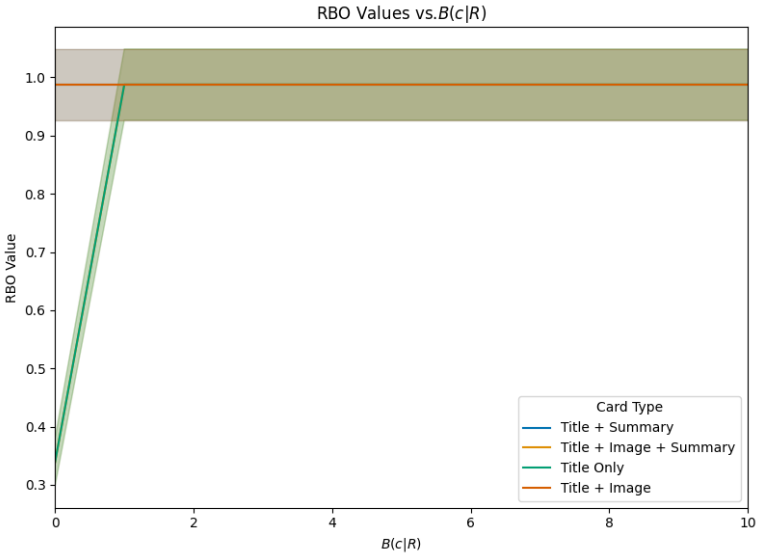


Figure B.4: $B(c|R)$ vs RBO

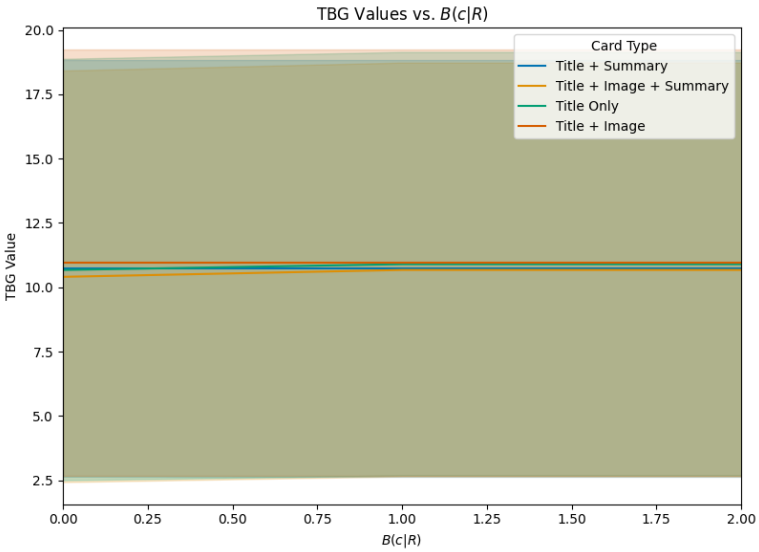


Figure B.5: $B(c|R)$ vs TBG

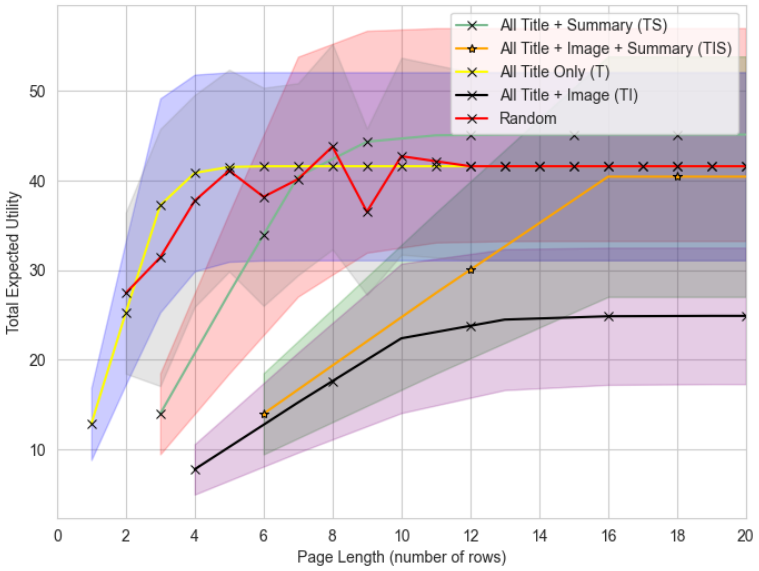


Figure B.6: Space Utility Trade-off, as a function of increase in page space to total utility.

B.0.1 The Space Utility Trade-off

To gain a deeper comprehension of the EPU function's responsiveness to modifications in its principal components, it is essential to evaluate the practical implications of these utility trade-offs. This necessitates an examination of the impact that varying the display of SERPs in distinct card types has on the page's overall utility.

An investigation was conducted into the fluctuation of the EPU when SERPs incorporated different combinations of result card types. The objective was to discern the dynamic between the variety of card types and the EPU of the SERPs. The EPU for the top 20 documents retrieved for every subject, using the topic name as the search query, was computed. Subsequently, "virtual" SERPs were constructed under the hypothetical scenario where each SERP exclusively presented one type of result card. For instance, a SERP composed solely of title cards would encompass 20 documents, provided that a single page could accommodate 20 rows. Conversely, a SERP featuring Title, Image, and Summary (TIS) cards would display merely three documents. This methodology facilitated the calculation of the total utility for each distinct virtual SERP, thereby allowing a comparative assessment of various card layouts' effectiveness.

The results, depicted in Figure B.6, elucidate that the selection of result card type exerts a significant influence on the effectiveness of the SERPs, particularly when screen real estate is at a premium. It is evident from the figure how the total expected utility of the list varies with the expansion of screen space. The inclusion of more detailed cards, such as TIS or Title and Summary (TS), markedly elevates the cumulative utility as screen space proliferates. Nonetheless, there emerges a balance to be struck between the number of result cards exhibited and their individual EPU. To illustrate, showcasing a mere quartet of TS cards yielded a more substantial total EPU than a dozen T cards.

This suggests that optimising the utility for a given space necessitates a trade-off between the number of result cards displayed and their respective EPUs.

Moreover, the analysis demonstrated that the overall utility of a page reached a plateau at divergent junctures for different result card configurations. For instance, the aggregate utility of the page hit a plateau beyond the display of six T cards, whereas for TS cards, the plateau manifested after five cards were shown. This accentuates the

correlation between the spatial occupation of a result card and its aggregate utility to the end-user, thereby underscoring the imperative to thoughtfully weigh the trade-off between these two variables.

Appendix C

Additional Graphs

C.1 Comparison Graphs

In our second experiment, we amassed a significant dataset from user interactions and described various user behaviours. Presented here are some illustrative graphs that provide a more granular view of how these behaviours vary. While these visualisations do not fundamentally alter our understanding of the problem, they do offer a more complete picture of the different behaviours observed. This detailed view is particularly pertinent for the second experiment, which included a multitude of layouts for comparison. The third experiment diverged in its focus, with a comprehensive analysis and clustering of graphs provided in Chapter 6.

C.1.1 User Interaction with Different Card Types

This graph in Figure C.1 showcases the number of user clicks for different card types in the search results. We observe here that the T cards receive a significantly higher number of clicks compared to the rest, possibly because users needed to click every single T item to properly assess it. The lower TIS clicks possibly suggest that they are more informative, needing fewer clicks to inspect the relevance.

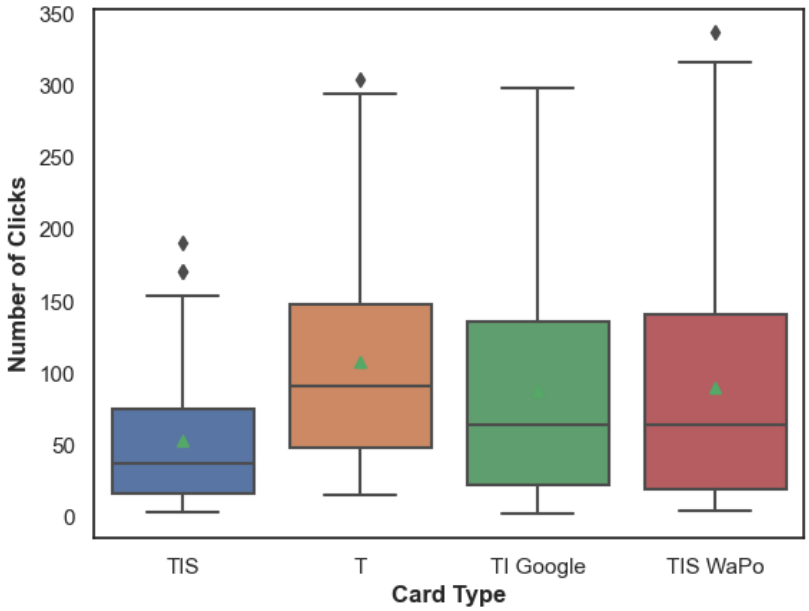


Figure C.1: The number of clicks across various SERP card types.

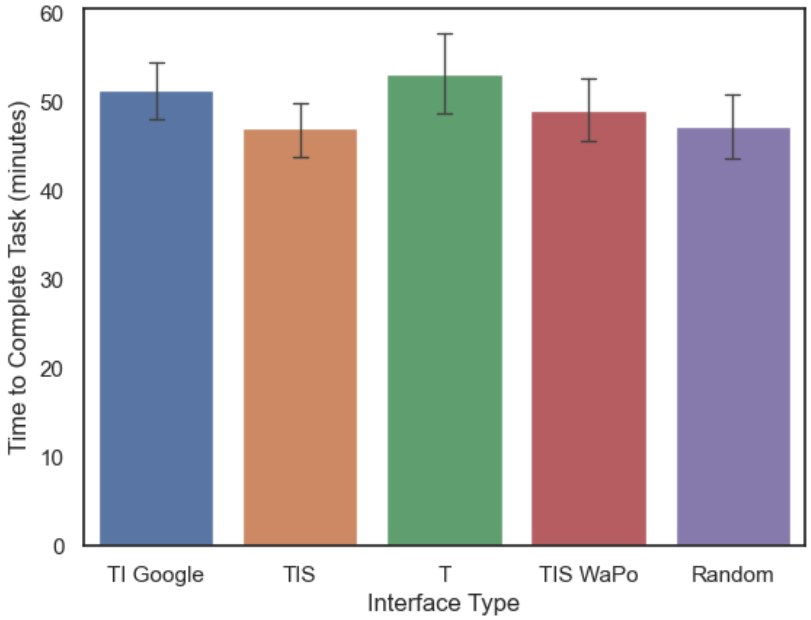


Figure C.2: Comparison of task completion times across different interface types.

C.1.2 Task Completion Time Across Interfaces

Figure C.2 shows the task completion time across various interfaces. We observe no differences here, meaning that the layout itself did not influence the speed with which users completed the search task.

C.1.3 User Satisfaction by Interface Layout

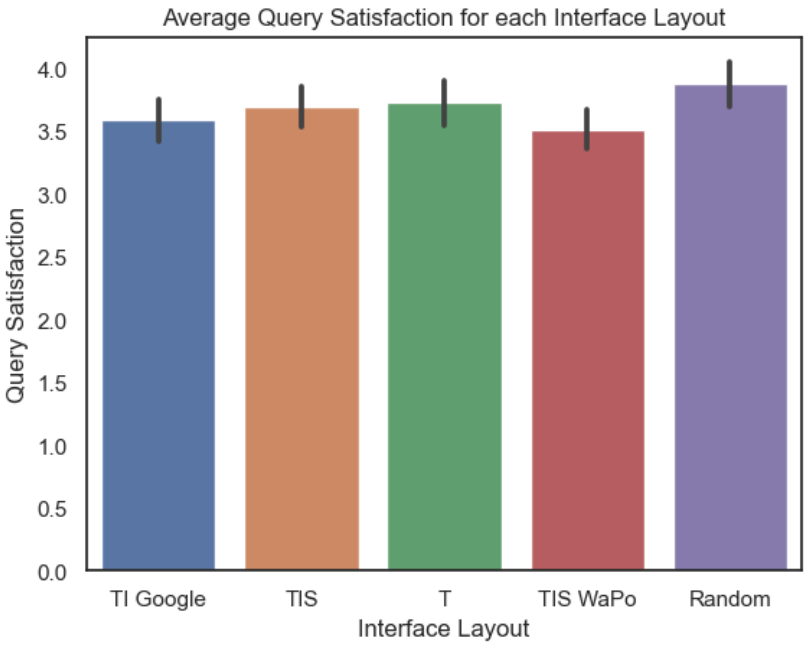


Figure C.3: Average query satisfaction for each interface layout.

Figure C.3 presents the average user satisfaction with different interface layouts for queries. We observe that the interface layout has no direct impact on query satisfaction, but as we delineated in §5, there is a deeper relationship that relates to the nDCG and total gain on the page.

C.1.4 Time Spent on Page by Interface Type

The mean time users spent on each page, shown in Figure C.4, shows the proportion of time spent on SERPs by the interface layout. We can see that in the T layout spent more time on the SERPs possibly due to having more cards to look at overall on the

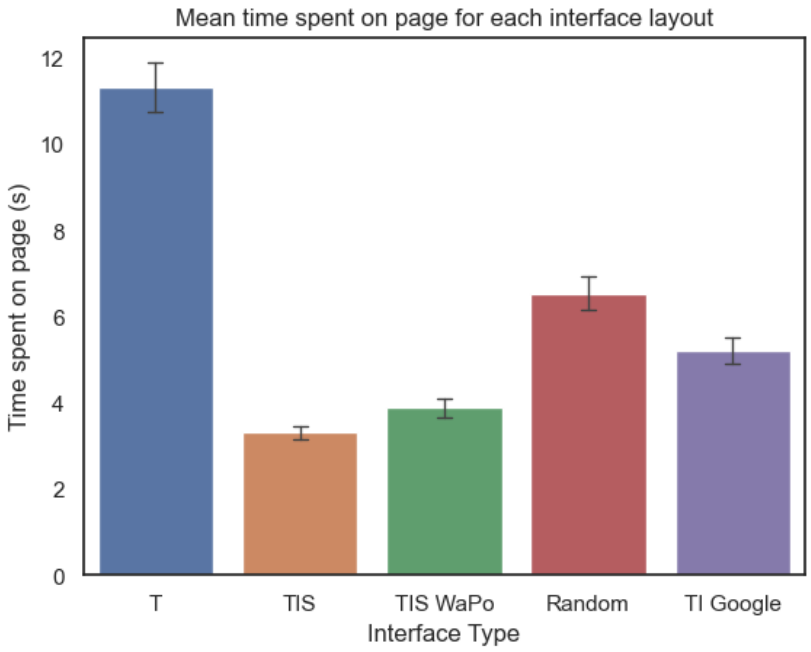


Figure C.4: Mean time spent on a page for each interface layout.

SERP.

C.1.5 Query Satisfaction by Topic Order

Figure C.5 explores how the order in which topics are presented affects user satisfaction. Our graphs suggest that the topic order does not affect query satisfaction. That is to say that it does not matter whether the topic appeared first or second, the satisfaction with the query remained the same

C.1.6 Feedback on Query Satisfaction by Topic

Lastly, Figure C.6 assesses user feedback on query satisfaction across different topics. We observe that the topic itself has little effect on query satisfaction, meaning that the possibility of order effects on query satisfaction does not exist.

We now explore how the different metrics related to interface satisfaction varied by the layout.

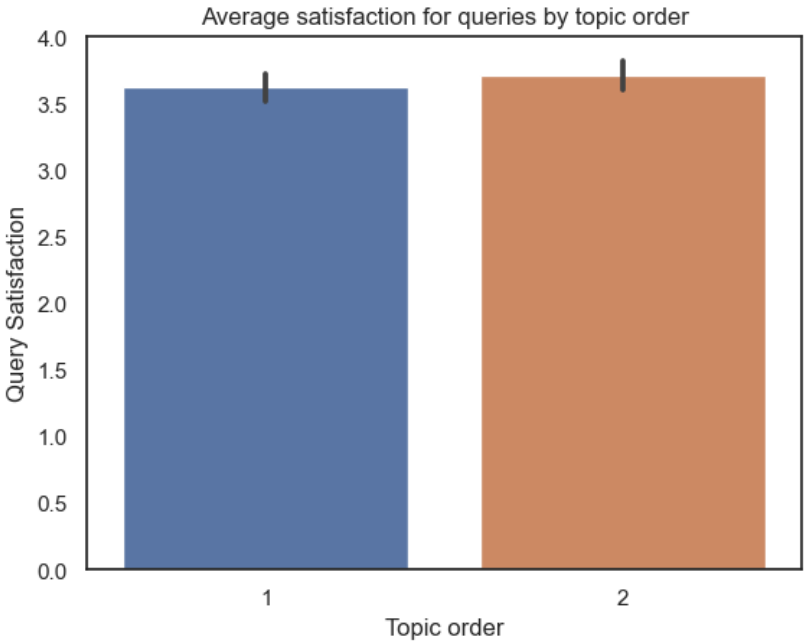


Figure C.5: Average satisfaction for queries by topic order.

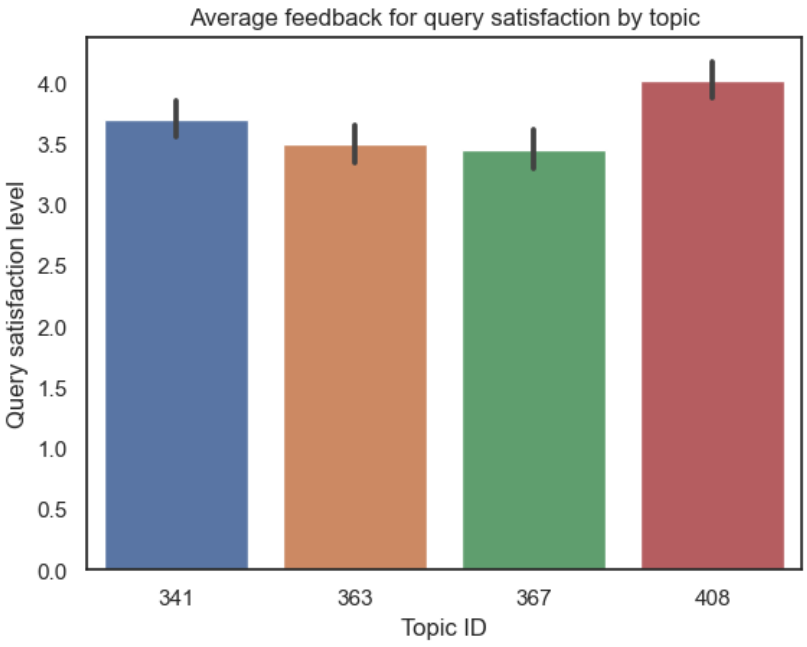


Figure C.6: Average feedback for query satisfaction by topic.

C.1.7 Cognitive Load Across Interface Layouts

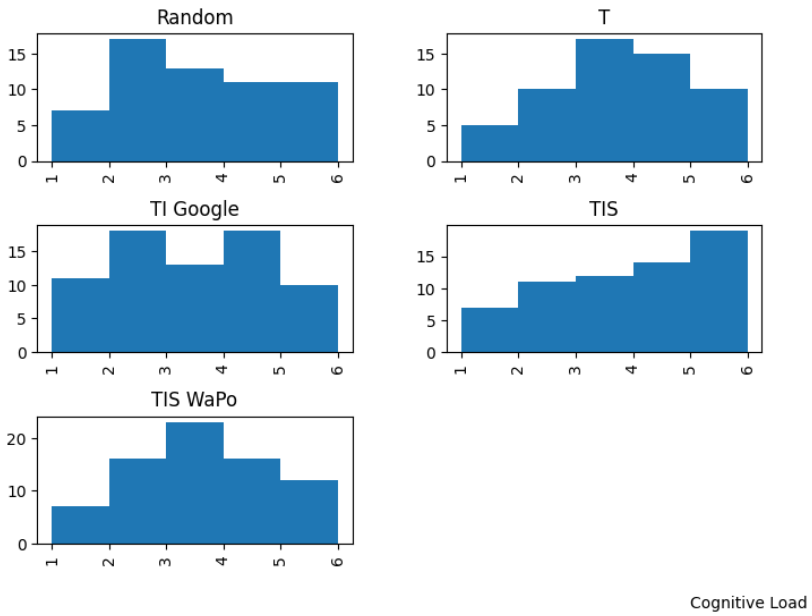


Figure C.7: Distribution of cognitive load across different interface layouts.

Figure C.7 illustrates a clear distribution of cognitive load across various interface layouts. The “TIS WaPo” layout results in a relatively even distribution, indicating that this interface likely balances information density with user cognitive capacity. In contrast, the ’Random’ layout shows a skewed distribution towards higher cognitive load, which is expected as the results should have been more difficult to parse.

C.1.8 Distraction Level by Interface Type

In Figure C.8, the ”TI Google” layout demonstrates a bimodal distribution, suggesting that users are either not distracted at all or quite significantly distracted. The “TIS” layout, with its peak in the middle, suggests a moderate level of distraction.

C.1.9 Engine Likability Among Different Interfaces

The likability ratings, depicted in Figure C.9, show that the ’T’ layout has a concentrated distribution around a middle rating, indicating consistent but not high satisfaction. The “TIS WaPo” graph suggests a polarised user base, where the interface is

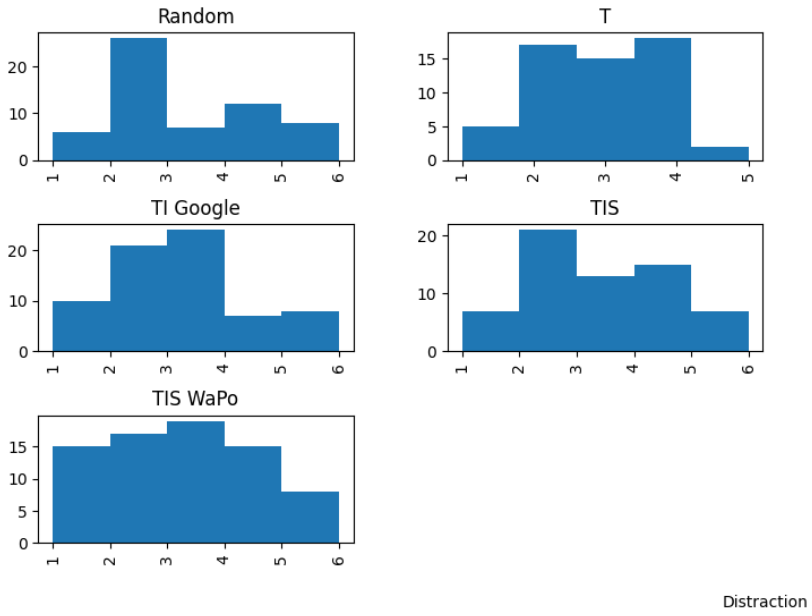


Figure C.8: Levels of distraction experienced by users on different interface layouts.

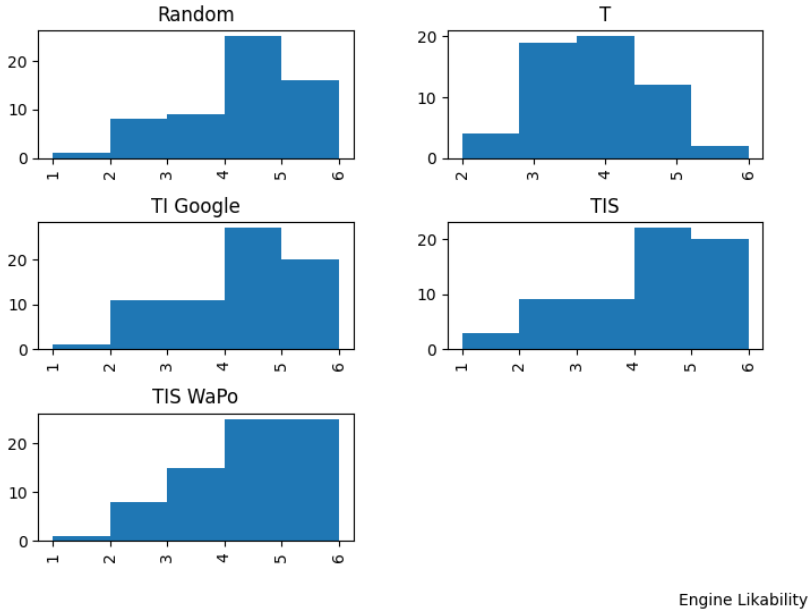


Figure C.9: User likability ratings for the search engine across different interfaces.

highly liked by some users but not as much by others.

C.1.10 Overall Satisfaction with Search Interfaces

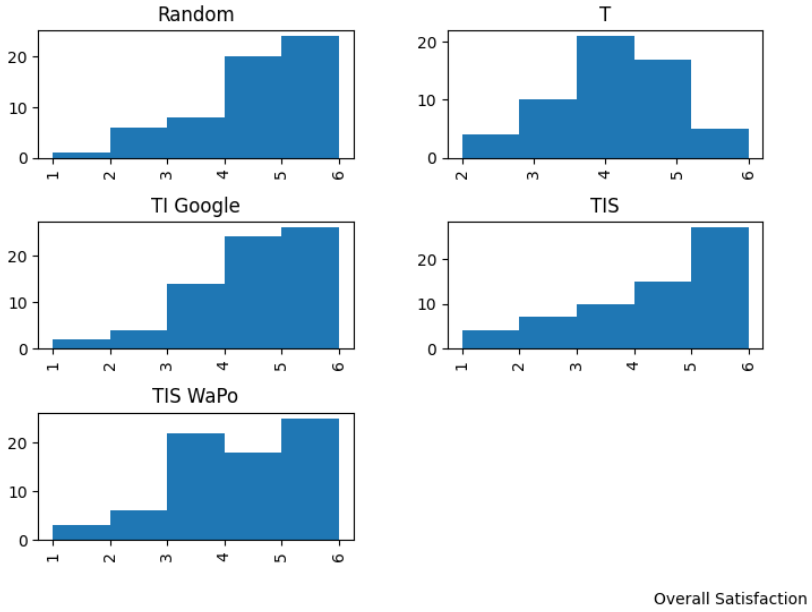


Figure C.10: Overall user satisfaction across various interface layouts.

Figure C.10 conveys that satisfaction is uniformly high across the "TIS WaPo" interface. The 'Random' interface shows a decline in satisfaction, indicating that a lack of structure in information presentation detracts from user satisfaction.

C.1.11 Productivity Scores for Different Interface Layouts

The productivity graph in Figure C.11 for the 'TIS' layout shows an ascending trend, which can be interpreted as an interface that supports increased productivity as users adapt or learn to navigate it. On the other hand, the 'Random' layout shows a peak at the lower end of the scale, suggesting that the lack of a coherent layout hinders user productivity.

Although these trends support existing research which suggests that the inclusion of a summary greatly improves user satisfaction, we found no statistically significant differences to provide any empirical evidence of this.

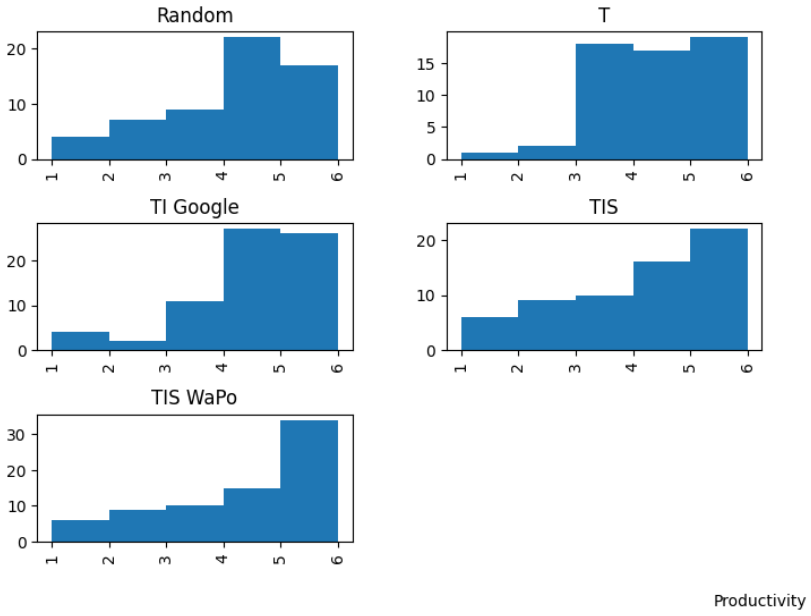


Figure C.11: Productivity distribution as influenced by different interface layouts.

Bibliography

[1] D. Kelly and D. Kelly, "Methods for Evaluating Interactive Information Retrieval Systems with Users," *Foundations and Trends R in Information Retrieval*, vol. 3, no. 2, pp. 1–224, 2009.

[2] "made with paint : r/physicsmemes." [Online]. Available: https://www.reddit.com/r/physicsmemes/comments/kncao1/made_with_paint/?share_id=McUcMrZxsF8KJT2qyyctB&utm_content=1&utm_medium=ios_app&utm_name=ioscss&utm_source=share&utm_term=1&rdt=43406

[3] T. J. Berners-Lee, "The world-wide web," *Computer Networks and ISDN Systems*, vol. 25, no. 4-5, pp. 454–459, 11 1992.

[4] N. J. Belkin, "Anomalous states of knowledge as a basis for information retrieval," *Canadian journal of information science*, vol. 5, no. 1, pp. 133–143, 1980.

[5] P. Ingwersen and K. Järvelin, *The turn: Integration of information seeking and retrieval in context*. Springer Science & Business Media, 2005, vol. 18.

[6] I. Ruthven, "Interactive information retrieval," pp. 43–91, 1 2008. [Online]. Available: https://asistdl.onlinelibrary.wiley.com/doi/full/10.1002/aris.2008.1440420109https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/aris.2008.1440420109https://asistdl.onlinelibrary.wiley.com/doi/10.1002/aris.2008.1440420109

[7] C. Cleverdon, J. Mills, and M. K. N. Title), "Factors determining the performance of indexing systems," *cir.nii.ac.jp*, vol. 1, no. 2, 1966. [Online]. Available: https://cir.nii.ac.jp/crid/1130000794058401664

[8] G. K. Zipf, *Human behavior and the principle of least effort: An introduction to human ecology.* Ravenio Books, 2016.

[9] S. E. Robertson, "The probability ranking principle in ir," pp. 294–304, 1977.

[10] N. Fuhr, "A probability ranking principle for interactive information retrieval," *Information Retrieval*, vol. 11, no. 3, pp. 251–265, 6 2008.

[11] J. Wang, "Mean-Variance analysis: A new document ranking theory in information retrieval," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 5478 LNCS, 2009.

[12] G. Zuccon, L. A. Azzopardi, and K. Van Rijsbergen, "The quantum probability ranking principle for information retrieval," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 5766 LNCS, 2009.

[13] X. Shen, B. Tan, and C. X. Zhai, "Implicit user modeling for personalized search," *International Conference on Information and Knowledge Management, Proceedings*, pp. 824–831, 2005.

[14] M. Sloan and J. Wang, "Dynamic information retrieval: Theoretical framework and application," in *ICTIR 2015 - Proceedings of the 2015 ACM SIGIR International Conference on the Theory of Information Retrieval*, 2015.

[15] Y. Zhang and C. Zhai, "Information retrieval as card playing: A formal model for optimizing interactive retrieval interface," in *SIGIR 2015 - Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2015.

[16] L. Azzopardi and G. Zuccon, "An analysis of the cost and benefit of search interactions," in *ICTIR 2016 - Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval*, 2016.

[17] D. Maxwell, L. Azzopardi, and Y. Moshfeghi, "A study of snippet length and informativeness behaviour, performance and user experience," *SIGIR 2017 - Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 135–144, 8 2017.

[18] L. Azzopardi, "The economics in interactive information retrieval," *SIGIR'11 - Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 15–24, 2011.

[19] H. Bota, K. Zhou, and J. M. Jose, "Playing your cards right: The effect of entity cards on search behaviour and workload," *CHIIR 2016 - Proceedings of the 2016 ACM Conference on Human Information Interaction and Retrieval*, pp. 131–140, 3 2016. [Online]. Available: https://dl.acm.org/doi/10.1145/2854946.2854967

[20] R. S. Rele and A. T. Duchowski, "Using eye tracking to evaluate alternative search results interfaces," in *Proceedings of the Human Factors and Ergonomics Society*, 2005.

[21] J. Teevan, E. Cutrell, D. Fisher, S. M. Drucker, G. Ramos, P. André, and C. Hu, "Visual snippets: Summarizing web pages for search and revisitation," in *Conference on Human Factors in Computing Systems - Proceedings*, 2009.

[22] S. Dziadosz and R. Chandrasekar, "Do thumbnail previews help users make better relevance decisions about web search results?" in *SIGIR Forum (ACM Special Interest Group on Information Retrieval)*, 2002.

[23] H. Joho and J. M. Jose, "A comparative study of the effectiveness of search result presentation on the Web," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 3936 LNCS, 2006.

[24] A. Tombros and M. Sanderson, "Advantages of query biased summaries in information retrieval," *SIGIR Forum (ACM Special Interest Group on Information Retrieval)*, 1998.

[25] F. Chierichetti, R. Kumar, and P. Raghavan, "Optimizing two-dimensional search results presentation," *Proceedings of the 4th ACM International Conference on Web Search and Data Mining, WSDM 2011*, pp. 257–266, 2011.

[26] T. Joachims, "Optimizing search engines using clickthrough data," *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 133–142, 2002. [Online]. Available: https://dl.acm.org/doi/10.1145/775047.775067

[27] C. Wang, Y. Liu, M. Zhang, S. Ma, M. Zheng, J. Qian, and K. Zhang, "Incorporating vertical results into search click models," *SIGIR 2013 - Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 503–512, 2013. [Online]. Available: https://dl.acm.org/doi/10.1145/2484028.2484036

[28] A. Chuklin and M. De Rijke, "Incorporating clicks, attention and satisfaction into a search engine result page evaluation model," *International Conference on Information and Knowledge Management, Proceedings*, vol. 24-28-October-2016, pp. 175–184, 10 2016. [Online]. Available: https://dl.acm.org/doi/10.1145/2983323.2983829

[29] D. Chen, W. Chen, H. Wang, Z. Chen, and Q. Yang, "Beyond ten blue links: Enabling user click modeling in federated Web search," *WSDM 2012 - Proceedings of the 5th ACM International Conference on Web Search and Data Mining*, pp. 463–472, 2012. [Online]. Available: https://dl.acm.org/doi/10.1145/2124295.2124351

[30] Y. Wang, D. Yin, L. Jie, P. Wang, M. Yamada, Y. Chang, and Q. Mei, "Optimizing whole-page presentation for web search," *ACM Transactions on the Web*, vol. 12, no. 3, 2018.

[31] H. Oosterhuis and M. De Rijke, "Ranking for relevance and display preferences in complex presentation layouts," *41st International ACM SIGIR Conference on*

Research and Development in Information Retrieval, SIGIR 2018, pp. 845–854, 6 2018. [Online]. Available: https://dl.acm.org/doi/10.1145/3209978.3209992

[32] L. Wu, M. Grbovic, and J. Li, "Toward User Engagement Optimization in 2D Presentation," *WSDM 2021 - Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pp. 1047–1055, 8 2021. [Online]. Available: https://dl.acm.org/doi/10.1145/3437963.3441749

[33] "Text REtrieval Conference (TREC) Home Page." [Online]. Available: https://trec.nist.gov/

[34] M. Dewey, *Decimal Classification and Relative Index for Libraries, Clippings, Notes, Etc*, 4th ed. Forest Press, Lake Placid Club, N.Y., 1891.

[35] G. Moore, "Cramming more components onto integrated circuits," *Electronics Magazine*, vol. 38, no. 8, p. 114 ff, 4 1965.

[36] Wikipedia Contributors, "JumpStation," https://en.wikipedia.org/wiki/JumpStation, 2023.

[37] H. P. Luhn, "A Statistical Approach to Mechanized Encoding and Searching of Literary Information," *IBM Journal of Research and Development*, vol. 1, no. 4, pp. 309–317, 1957. [Online]. Available: https://doi.org/10.1147/rd.14.0309

[38] S. E. Robertson and K. S. Jones, "Relevance weighting of search terms," *Journal of the American Society for Information Science*, vol. 27, no. 3, pp. 129–146, 5 1976. [Online]. Available: https://onlinelibrary.wiley.com/doi/full/10.1002/asi.4630270302https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.4630270302https://asistdl.onlinelibrary.wiley.com/doi/10.1002/asi.4630270302

[39] C. T. Yu and G. Salton, "Precision WeightingAn Effective Automatic Indexing Method," *Journal of the ACM (JACM)*, vol. 23, no. 1, pp. 76–88, 1 1976. [Online]. Available: https://dl.acm.org/doi/10.1145/321921.321930

[40] S. E. Robertson and S. Walker, "Some Simple Effective Approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval." [Online]. Available: https://dl.acm.org/doi/10.5555/188490.188561

[41] "Okapi BM25: a non-binary model," 2023. [Online]. Available: https://nlp.stanford.edu/IR-book/html/htmledition/okapi-bm25-a-non-binary-model-1.html

[42] M. Sanderson, "Test Collection Based Evaluation of Information Retrieval Systems," *Foundations and Trends® in Information Retrieval*, vol. 4, no. 4, pp. 247–375, 2010. [Online]. Available: http://dx.doi.org/10.1561/1500000009

[43] K. Järvelin and J. Kekäläinen, "Cumulated gain-based evaluation of IR techniques," *ACM Transactions on Information Systems (TOIS)*, vol. 20, no. 4, pp. 422–446, 10 2002. [Online]. Available: https://dl.acm.org/doi/10.1145/582415.582418

[44] M. D. Smucker and C. L. A. Clarke, "Time-based calibration of effectiveness measures," *SIGIR'12 - Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 95–104, 2012.

[45] W. Webber, A. Moffat, and J. Zobel, "A Similarity Measure for Indefinite Rankings," 2010. [Online]. Available: http://dx.doi.org/10.1145/1852102.1852106.

[46] C. W. Cleverdon, "The aslib cranfield research project on the comparative efficiency of indexing systems," *Aslib Proceedings*, vol. 12, no. 12, pp. 421–431, 12 1960.

[47] "TREC Washington Post Corpus." [Online]. Available: https://trec.nist.gov/data/wapost/

[48] B. J. Jansen, D. L. Booth, and A. Spink, "Determining the informational, navigational, and transactional intent of Web queries," *Information Processing*

and Management: an International Journal, vol. 44, no. 3, pp. 1251–1266, 5 2008. [Online]. Available: https://dl.acm.org/doi/10.1016/j.ipm.2007.07.015

[49] M. D. Gordon and P. Lenk, "When is the probability ranking principle suboptimal?" *Journal of the American Society for Information Science*, vol. 43, no. 1, pp. 1–14, 1992.

[50] L. Azzopardi and G. Zuccon, "An analysis of theories of search and search behavior," in *ICTIR 2015 - Proceedings of the 2015 ACM SIGIR International Conference on the Theory of Information Retrieval*, 2015.

[51] G. Zuccon, L. A. Azzopardi, and K. Van Rijsbergen, "The quantum probability ranking principle for information retrieval," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 5766 LNCS, pp. 232–240, 2009.

[52] H. Markowitz, "Portfolio selection," *Harry Markowitz: Selected Works*, pp. 15–30, 1 2009.

[53] Y. Kammerer and P. Gerjets, "How the interface design influences users' spontaneous trustworthiness evaluations of web search results: Comparing a list and a grid interface," in *Eye Tracking Research and Applications Symposium (ETRA)*, 2010.

[54] S. Dumais, E. Cutrell, and H. Chen, "Optimizing search by showing results in context," *Conference on Human Factors in Computing Systems - Proceedings*, pp. 277–284, 2001. [Online]. Available: https://dl.acm.org/doi/10.1145/365024.365116

[55] C. Braganza, K. Marriott, P. Moulder, M. Wybrow, and T. Dwyer, "Scrolling behaviour with single-and multi-column layout," *WWW'09 - Proceedings of the 18th International World Wide Web Conference*, pp. 831–840, 2009. [Online]. Available: https://dl.acm.org/doi/10.1145/1526709.1526821

[56] M. Gambhir and V. Gupta, "Recent automatic text summarization techniques: a survey," *Artificial Intelligence Review*, vol. 47, no. 1, pp. 1–66, 1 2017. [Online]. Available: https://link.springer.com/article/10.1007/s10462-016-9475-9

[57] H. Luhn, "The automatic creation of literature abstracts," *IBM J Res Dev*, vol. 2, 1958. [Online]. Available: https://doi.org/10.1147/rd.22.0159

[58] H. P. Edmundson, "New Methods in Automatic Extracting," *Journal of the ACM (JACM)*, vol. 16, no. 2, pp. 264–285, 4 1969. [Online]. Available: https://dl.acm.org/doi/10.1145/321510.321519

[59] E. Hovy and C.-Y. Lin, "Automated text summarization and the SUMMARIST system," p. 197, 1996. [Online]. Available: https://dl.acm.org/doi/10.3115/1119089.1119121

[60] H. Daumé and D. Marcu, "Bayesian query-focused summarization," *COLING/ACL 2006 - 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, vol. 1, pp. 305–312, 2006. [Online]. Available: https://dl.acm.org/doi/10.3115/1220175.1220214

[61] Marshall Israel, Quinsulon L, H. Han, and I.-Y. Song, "Focused multi-document summarization: Human summarization activity vs. automated systems techniques," 2010. [Online]. Available: https://mds.marshall.edu/wdcs_faculty/1/

[62] L. L. Bando, F. Scholer, and A. Turpin, "Constructing query-biased summaries: A comparison of human and system generated snippets," *IIiX 2010 - Proceedings of the 2010 Information Interaction in Context Symposium*, pp. 195–204, 2010. [Online]. Available: https://dl.acm.org/doi/10.1145/1840784.1840813

[63] M. Verma and E. Yilmaz, "Search costs vs. User satisfaction on mobile," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10193 LNCS,

pp. 698–704, 2017. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-56608-5_68

[64] B. J. Jansen, A. Spink, and T. Saracevic, "Real life, real users, and real needs: A study and analysis of user queries on the Web," *Information Processing and Management*, vol. 36, no. 2, pp. 207–227, 3 2000.

[65] L. Azzopardi, D. Kelly, and K. Brennan, "How Query Cost Affects Search Behavior," *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 23–32, 7 2013. [Online]. Available: https://doi.org/10.1145/2484028.2484049

[66] D. Hienert, D. Kern, M. Mitsui, C. Shah, and N. J. Belkin, "Reading protocol: Understanding what has been read in interactive information retrieval tasks," *CHIIR 2019 - Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*, pp. 73–81, 3 2019. [Online]. Available: https://doi.org/10.1145/3295750.3298921

[67] M. Abualsaud, "The effect of queries and search result quality on the rate of query abandonment in interactive information retrieval," *CHIIR 2020 - Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*, pp. 523–526, 3 2020.

[68] X. Zhao, "Deep Reinforcement Learning for Search, Recommendation, and Online Advertising: A Survey."

[69] Q. Guo and E. Agichtein, "Beyond dwell time: Estimating document relevance from cursor movements and other post-click searcher behavior," *WWW'12 - Proceedings of the 21st Annual Conference on World Wide Web*, pp. 569–578, 2012. [Online]. Available: https://dl.acm.org/doi/10.1145/2187836.2187914

[70] Y. Kim, A. Hassan, R. W. White, and I. Zitouni, "Modeling dwell time to predict click-level satisfaction," *WSDM 2014 - Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, pp. 193–202, 2014. [Online]. Available: https://dl.acm.org/doi/10.1145/2556195.2556220

[71] P. Borlund and J. W. Schneider, "Reconsideration of the simulated work task situation: A context instrument for evaluation of information retrieval interaction," *IIiX 2010 - Proceedings of the 2010 Information Interaction in Context Symposium*, pp. 155–164, 2010. [Online]. Available: https://dl.acm.org/doi/10.1145/1840784.1840808

[72] "maxwelld90/wapo_indexer: Washington Post corpus indexer (for Whoosh)." [Online]. Available: https://github.com/maxwelld90/wapo_indexer

[73] "leifos/ifind: A Series of related Search Games, a Living Lab, and other related experiments." [Online]. Available: https://github.com/leifos/ifind

[74] M. C. Chen, J. R. Anderson, and M. H. Sohn, "What can a mouse cursor tell us more? Correlation of eye/mouse movements on web browsing," *Conference on Human Factors in Computing Systems - Proceedings*, pp. 281–282, 2001. [Online]. Available: https://dl.acm.org/doi/10.1145/634067.634234

[75] V. Navalpakkam and E. F. Churchill, "Mouse Tracking: Measuring and Predicting Users' Experience of Web-based Content," 2012.

[76] J. Huang, R. W. White, and S. Dumais, "No clicks, no problem: Using cursor movements to understand and improve search," *Conference on Human Factors in Computing Systems - Proceedings*, pp. 1225–1234, 2011. [Online]. Available: https://dl.acm.org/doi/10.1145/1978942.1979125

[77] Q. Guo and E. Agichtein, "Towards predicting web searcher gaze position from mouse movements," *Conference on Human Factors in Computing Systems - Proceedings*, pp. 3601–3606, 2010. [Online]. Available: https://dl.acm.org/doi/10.1145/1753846.1754025

[78] F. Mueller and A. Lockerd, "Cheese: Tracking mouse movement activity on websites, a tool for user modeling," *Conference on Human Factors in Computing Systems - Proceedings*, pp. 279–280, 2001. [Online]. Available: https://dl.acm.org/doi/10.1145/634067.634233

[79] N. Bhattacharya, "Record User Interactions on your Webpages: A tutorial," 2021. [Online]. Available: https://medium.com/@nilavra/60ccc19f0516

[80] W. Webber, A. Moffat, and J. Zobel, "A similarity measure for indefinite rankings," *ACM Transactions on Information Systems*, vol. 28, no. 4, 2010.

[81] A. Al-Maskari, M. Sanderson, and P. Clough, "The relationship between IR effectiveness measures and user satisfaction," *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'07*, pp. 773–774, 2007. [Online]. Available: https://dl.acm.org/doi/10.1145/1277741.1277902

[82] L. Azzopardi, "The economics in interactive information retrieval," *SIGIR'11 - Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 15–24, 2011.

[83] E. W. Morrison and J. B. Vancouver, "Within-person analysis of information seeking: The effects of perceived costs and benefits," *Journal of Management*, vol. 26, no. 1, pp. 119–137, 2000.

[84] P. Borlund and P. Ingwersen, "The development of a method for the evaluation of interactive information retrieval systems," *Journal of Documentation*, vol. 53, no. 3, pp. 225–250, 1997.

[85] J. Cohen, "Eta-squared and partial eta-squared in fixed factor anova designs," *Educational and Psychological Measurement*, vol. 33, no. 1, pp. 107–112, 4 1973. [Online]. Available: https://journals.sagepub.com/doi/abs/10.1177/001316447303300111?casa_token=DVox4WckkpgAAAAA%3AGQfTrqJUtdybsEp72fngh7dfA3WU9Bnbv-jvl0xiUi050l1w3nnev8Hn-ZqBEaJgy6m43NjM8cc

[86] A. Caprara, H. Kellerer, and U. Pferschy, "The Multiple Subset Sum Problem," *SIAM J. Optim.*, vol. 11, no. 2, pp. 308–319, 2000. [Online]. Available: https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.21.9826#id-name=CiteSeerXhttp://www.or.deis.unibo.it/alberto/mssp_siam.ps

[87] G. Gallo, P. L. Hammer, and B. Simeone, *Quadratic knapsack problems*, 1980.

[88] L. Caccetta and A. Kulanoot, "Computational Aspects of Hard Knapsack Problems," *Nonlinear Analysis*, vol. 47, no. 8, pp. 5547–5558, 8 2001.

[89] T. Dantzig and J. Mazur, "Number : the language of science," p. 396, 2007.

[90] Z. Y. Wu, Y. J. Yang, F. S. Bai, and M. Mammadov, "Global Optimality Conditions and Optimization Methods for Quadratic Knapsack Problems," *J Optim Theory Appl*, vol. 151, no. 2, pp. 241–259, 2011. [Online]. Available: https://api.semanticscholar.org/CorpusID:31208118#id-name=S2CID

[91] J. Cohen, "Statistical Power Analysis for the Behavioral Sciences," *Statistical Power Analysis for the Behavioral Sciences*, 5 2013. [Online]. Available: https://www.taylorfrancis.com/books/mono/10.4324/9780203771587/statistical-power-analysis-behavioral-sciences-jacob-cohen

[92] D. E. Knuth, *The Art of Computer Programming*, 3rd ed. Reading, Massachusetts: Addison-Wesley, 1997. [Online]. Available: http://www.projectetal.com/forums/index.php?threads/the-art-of-computer-programming-vols-1-3-by-donald-knuth.8415/

[93] G. Salton, E. A. Fox, and H. Wu, "Extended Boolean Information Retrieval," *Communications of the ACM*, vol. 26, no. 11, pp. 1022–1036, 11 1983.

[94] J. Zobel, A. Moffat, and K. Ramamohanarao, "Inverted files versus signature files for text indexing," *ACM Transactions on Database Systems*, vol. 23, no. 4, pp. 453–490, 12 1998. [Online]. Available: https://api.semanticscholar.org/CorpusID:7293918#id-name=S2CIDhttps://doi.org/10.1145%2F296854.277632

[95] "Anatomy of a Compiler and The Tokenizer," *www.cs.man.ac.uk*. [Online]. Available: http://www.cs.man.ac.uk/~pjj/farrell/comp3.html

[96] A. Rajaraman and J. D. Ullman, *Mining of Massive Datasets*. Cambridge University Press, 10 2011. [Online]. Available: http://i.stanford.edu/~ullman/mmds/ch1.pdf

[97] H. S. Christopher D. Manning, Prabhakar Raghavan, *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[98] J. B. Lovins, "Development of a Stemming Algorithm," *Mechanical Translation and Computational Linguistics*, vol. 11, pp. 22–31, 1968. [Online]. Available: http://chuvyr.ru/MT-1968-Lovins.pdf

[99] M. Porter, *Porter Stemming Algorithm*, 1980. [Online]. Available: http://tartarus.org/~martin/PorterStemmer/

[100] R. Baeza-Yates and B. Ribeiro-Neto, *Modern information retrieval*, 1999. [Online]. Available: https://people.ischool.berkeley.edu/~hearst/irbook/print/chap10.ps.gz

[101] F. W. Lancaster and E. G. Fayen, *Information Retrieval On-Line*. Los Angeles, California: Melville Publishing Co., 1973.

Bibliography