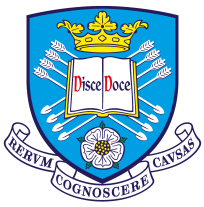


The University Of Sheffield.

Understanding and Predicting User Perception of Engagement through User Behaviour in Information Retrieval

Mengdie Zhuang

A thesis submitted in part fulfilment of the requirements
for the degree of Doctor of Philosophy

Management School
University of Sheffield
Submitted April 2019

Abstract

User engagement is an assessment of the quality of user experience while interacting with information systems. Typically, two types of measures are used to assess engagement, one based on user perception, and one on user behaviour. They are either obtrusive to user interaction, or respectively, implicit. Moreover, they have been treated as discrete types of measures and the current body of literature does not suffice to verify their connections.

The purpose of this research is to analyse the relationship between user behaviour and user perception in the assessment of engagement in two information retrieval contexts, searching and browsing. In Phase 1, we investigate the role of discrete behavioural features in predicting user perception of engagement through correlation analysis. The resulting predictive model then serves as a baseline and basic framework for phases presented subsequently. In Phase 2, we investigate the added benefits of behaviour sequences through the chi-square test. In Phase 3, based on the findings from our previous phases, we developed and evaluated context-based measures of engagement. Our measures perform as well as the state-of-the-art approach, without the need for finely-grained data and are interpretable and transferable between different contexts with ease.

Findings confirm that a relationship exists between user behaviour and user perception of engagement. Three behaviour patterns that are indicative of engaged/disengaged users were identified. Along the way, we contrast the two information retrieval contexts, drawing parallels and shedding light on particular behaviour patterns only apparent in one or the other.

This thesis contributes to a greater understanding of measuring engagement in information retrieval. For the first time, we demonstrate how user behaviour is correlated with user perceived engagement in the context of searching and browsing. Furthermore, we extract easily computable and interpretable measures of engagement, which contribute to the bottom-up methodological approach for measure design.

Acknowledgements

This work would not exist without the continued guidance and support of my supervisor, Professor Elaine G. Toms. Her commitment in providing me with achievable goals and encouragement are certainly the foundation of all my achievements over these many years. My second supervisor, Professor Andrew Simpson, has invested an enormous amount of effort into my research theory. Without his generous comments, I would have spent more time locating an appropriate one. In equal measure, I must express my deepest thanks to Dr. Gianluca Demartini, who supervised me in my first year of PhD. I am grateful that we kept our collaboration live after that. The many hours he spent teaching and explaining to me the various facets of research I was unfamiliar with I have found invaluable. My one regret is that I was unable to bring to fruition all of the promising research topics he suggested. I hope our connection will continue after this project.

Many others made the endeavour of producing and writing up this text possible. Notably, without these two that this research idea would not even have started, Dr. Heather O'Brien, who started the user engagement in information system discussion, and Dr. Mark Hall who kindly collected and offered the datasets to me. I hope I can have the opportunities to meet and thank them in person in the near future. I also want to thank the PGR Administrators, Mandy Robertson, Rebecca Roberts, Josie Smith and Matt Jones. A PhD is an adventure marathon, and they are the guardians.

Additionally, I am grateful to Professor Emine Yilmaz, who hosted me in the Alan Turing Institute, and provided me the opportunity to work in an exciting project with amazing people; Professor Avishek Anand, and Dr. Ujwal Gadiraju who kindly hosted me in Germany, and have shown a humbling interest in my research; my colleagues and friends who supportted me during the PhD time, and in particular, Sophie Rutter, Paula Goodale, Simon Wakeling, Christina Founti, Alessandro Checco, Ying Lan Ang, Michele Schirru, and Laura Vergoz, who I had countless conversations which enriched my experience as a student; the University of Sheffield has been my home over the past years. They have all made this endeavour possible through their support, both financial and moral.

Mostly, I want to thank my family, my parents who taught me to be the brave and curious woman I am, and Alex, who offered the woman support, love and confidence.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

The myriad of online services that have become available to us over the rise of the World Wide Web as a commonplace presence in our daily lives represents the most affluent source of information in the modern age. Decades of research have been carried out with the sole purpose of improving information retrieval (IR) systems in terms of relevance, thus recognising how online retrieval of information has become a ubiquitous requirement for the contemporary person. The development of Human - Computer Interaction (HCI) as a field of research has been undertaken in tandem with the aforementioned research scopes and, through its applications, which benefit from industrial appeal and constant user feedback, developed a new centralized set of goals. Among these, is the principle that *a positive user experience* contributes greatly to the system's viability, usefulness and staying power. However, we still lack robust methods of quantifying this phenomenon or even imparting a weakly-deterministic measure as a function of the system properties alone (Kelly, 2009; O'Brien, 2016; Lalmas et al., 2014). This thesis aims to address this gap by approaching it from the perspective of *user engagement*.

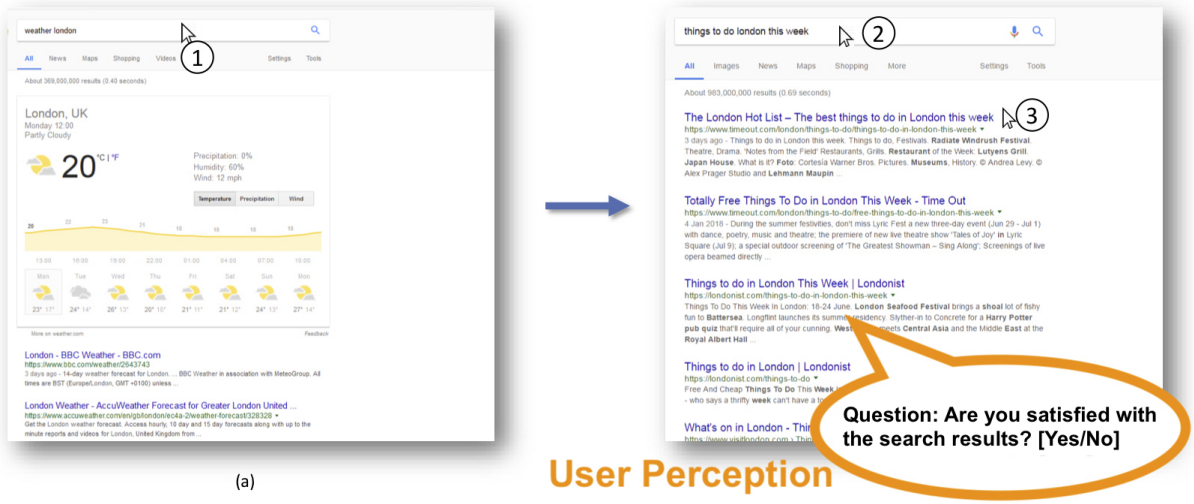
User engagement is an assessment of the quality of user experience in interacting with information resources. It focuses on the positive aspects of user experience, manifested in different forms, and connects the user and resources on an emotional, cognitive, and behavioural level

(O'Brien and Toms, 2008). A priori, one would require a precise definition of the characteristics of this type of positive user experience in relation to information retrieval systems. We shall delve into the particulars of the associated concepts of *user experience*, *user engagement* (section 2.4) and finally *user perception of engagement* (section 2.5.1) in the review of prior research chapter (chapter 2). The common agreement of all forms of definition of user engagement revolve around the positivity of user experience, namely the acquisition of a *pleasant*, *stimulating* or otherwise *rewarding* feeling through the interaction. A feeling of addiction to the interaction has been discussed in certain contexts (e.g. video games) and many other interpretations abound. Further in our study, we implicitly discuss these components in section 2.5.1 when selecting *engagement measures* by extrapolating these many paradigms. For now let us content ourselves with O'Brien and Toms (2008)'s definition that user engagement is "characterised by challenge, aesthetic and sensory appeal, feedback, novelty, interactivity, perceived control and time, awareness, motivation, and interest and affect" (p.949), reflecting its multi-dimensionality. To date, user engagement has been studied in multiple contexts such as multimedia, reading, and search systems (Quesenbery, 2003; Jacques, 1996; Webster and Ho, 1997; O'Brien and Toms, 2013).

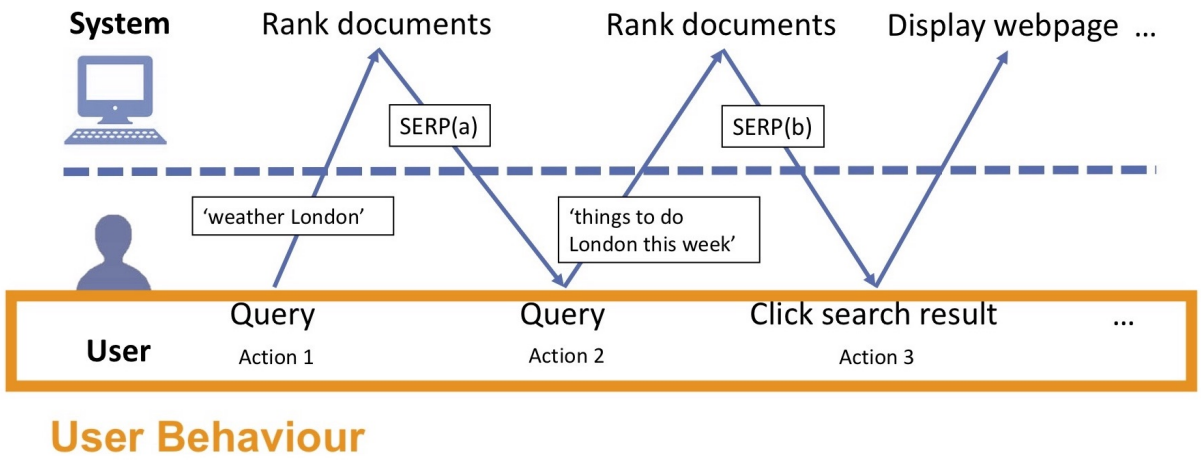
Defining user engagement is merely one part of the problem; one still needs to measure it. Typically, two main types of measures were used to assess engagement: the user's perception of the interaction, and user behaviour during the interaction. To ground this preliminary analysis in an attempt to see how some of these difficulties impact research, let us consider a very simple example:

A user is attempting to find out the weather in London on a certain day (figure 1.1) by querying 'weather London' online. In response to the query request, the system returns a Search Result Page (SERP), which is a ranked list of documents, to the user. Next, she searches for things to do this week in London and clicks the first link on the returned SERP. The system displays the linked webpage. In this case, the user issued three actions (figure 1.1 (b)). In the meantime, the system responds three times 1, return SERP for query 'weather London'; 2, return SERP for query 'things to do London this week'; and 3, display the link webpage. The actions or responses from both sides, are ordered by time, and thus describe the interaction between the

system and user. However, this data alone is insufficient to explain whether or how engaged the user is in the process.



(a)



(b)

Scenario: The user searches the weather in London and checks the results; then searches things to do this week in London, and clicks the link for best things to do in London.
Action 1: The user searches 'weather London' in the search box;
Action 2: The user searches 'things to do London this week' in the search box;
Action 3: The user clicks the first search result.

Figure 1.1: Example of a search process and the ways to collect (a) user perception of the interaction and (b) user behaviour during the interaction.

Understanding the user perception of that interaction requires a different type of data. User perception of that experience is collected through observation and/or self-reported methods (e.g., interviews, questionnaires). In figure 1.1 (a), such data can be collected from users response to a pop-up question "Are you satisfied with the search results?" during the process,

or immediately after the process. The major benefits of using user perception of engagement data is that this is feedback obtained directly from the users and is thus reflective of user's experience. But the collection is obtrusive to a user's natural interaction, interrupting the flow of that experience. In addition, these methods rely on users volunteering their responses.

User behaviour can be extracted from the interactions during the process (e.g., highlighted in figure 1.1 (b)). Low-level user behaviour, such as eye gaze (Buscher et al., 2010; Arapakis et al., 2014b; Lagun et al., 2014), cursor movements (Arapakis and Leiva, 2016; Lagun et al., 2014) and clicks (Dave et al., 2013; Dupret and Lalmas, 2013; Drutsa et al., 2015a; Lehmann et al., 2012), are automatically and inconspicuously collected, scale well, and are not restricted by time or place. Unlike self-reported data, data acquisition from log files provide sufficient amount of information for robust mathematical modeling, do not interrupt the user, or interfere with the user experience. However, user behaviour is not explicit as the same behavioural signal can relate to different user intentions (e.g., dwell time with satisfaction (Kim et al., 2014), or frustration (Odijk et al., 2015)). It is therefore essential to elucidate the link between a user's perception of experience and user behaviour during the experience.

The challenge, however, is whether there is a relationship between these interactions that encapsulate user behaviour and the user perception of engagement. Assuming that a uniform behaviour variable set can serve to measure or quantify engagement across all contexts is very optimistic (Lehmann et al., 2012). Several research studies outside the Information Retrieval (IR) field (Arapakis et al., 2014c; Arapakis and Leiva, 2016) have provided evidence that gaze behaviour and cursor behaviour can certainly correlate positively to engagement in online news reading. However, the current body of work conducted in searching and browsing fails to fully capture these relationships: previous studies (e.g.,(Su, 2003; Dupret and Lalmas, 2013)) have either used sets of features based on user behaviour to represent different dimensions of engagement but without sufficiently validating with user perception of engagement data as outlined in (Lalmas et al., 2014) or only single components of engagement (e.g., satisfaction (Kim et al., 2014; Al-Maskari and Sanderson, 2010), attention (Arapakis and Leiva, 2016)) have been examined individually against behavioural features, but this is a very limited analysis and does not provide a comprehensive, holistic picture of the user experience, as O'Brien and

Toms (2010a) suggests. This lack of a comprehensive system which captures engagement in its multifaceted forms represented the very motivation behind the development of a special user engagement evaluation scale which we shall discuss at length in this study.

In this thesis, the engagement evaluation problem is approached from both the user-centred perspective (which focuses on first-party evaluations collected from the users themselves) and the modelling perspective (in which we attempt to infer user's engagement from other variables and measures collected and processed online), by identifying and exploiting variables in the user behaviour during the process of looking for information online that characterize the extent to which people engage with the experience. Extracting from the user behaviour data, these variables represent indicators of the quality of the seeking process: the extent to which the matching of system returned results to user information needs occurs.

We also look into two types of information retrieval contexts, namely searching and browsing, and investigate how the links between user behaviour and user perception of engagement varies in the two contexts. The following section of this chapter introduces the research objectives that guided our study.

1.2 Research Objectives

The general purpose of this study is to analyse and model the relationship between user behaviour and user perception of engagement in information retrieval. From an empirical perspective, this work is motivated by the need for more pragmatic engagement measures that do not interrupt the user information retrieval process. From a theoretical perspective, it is motivated by the gap between the models of HCI in which user perception plays a major role, and current models of IR in which user perception is under-emphasized, and behavioural signals dominate. Bridging these gaps can increase the understanding of engagement in a multi-dimensional manner, and to develop an effective measure of engagement to measure engagement intra-session without interfering with the user.

We postulate that user behaviour contains information on the user perception of engagement.

Starting with identifying a detailed set of behavioural features used in previous studies, this work first focused on the importance of each behavioural feature in terms of describing four selected engagement dimensions, namely perceived usability, felt involvement, novelty, and endurability (detailed description in section 3.4.1). Then, the associations between behaviour sequences and user perception of engagement were explored as a means of identifying the key subsequences that represent engagement in the process. The differences in searching and browsing, were also compared. A set of proposed measures of engagement was evaluated in both searching and browsing. The study progressed through a series of research phases to identify and test the relationships between a rich spectrum of user behaviour and user perception of engagement. We outline below the broad research objectives that guided the study. More specific research questions and hypotheses were formulated for the various phases of the research and are described in the research design (Section 3) and relevant chapters to follow. The first three general objectives *Obj.1*, *Obj.2*, and *Obj.3* are devoted to one research phase each. The latter two general objectives *Obj.4*, and *Obj.5* are overarching ones that guide the discussion throughout the project.

Objectives:

Obj.1 To identify and validate the role of behavioural features in inferring user perception of engagement.

What are the recurring actions the user emits to the system when searching or browsing? Are any of these easily quantifiable, and if so, how much can we summarize through logging these interactions? And, finally, are any of these an indicator for higher or lower user perception of engagement? Behavioural features were suggested as engagement proxies (e.g., online news reading (Arapakis and Leiva, 2016), general web applications (Lehmann et al., 2012), general search (Drutsa et al., 2015a)), but little work has been conducted in verifying behavioural features with user perception of engagement in the information retrieval context. Therefore, it was necessary to approach this objective from the start by selecting the behavioural features that are suggested as engagement proxies and validating how they can describe user perception of engagement. The outcome of resolving

related research questions was a set of ranked behavioural features ordered by their ability of explaining user perception of engagement.

Obj.2 To identify, and characterize behaviour sequences that are indicative of user perception of engagement.

The aforementioned user actions do not occur independently, but rather, some of these actions immediately motivate subsequent ones in the user action sequence. These patterns may be trivial such as submitting a query and then clicking on the first result, or far more intricate such as looping through a set of documents and making comparisons. How can one best capture these patterns or sequences, and differentiate them from random subsequences of actions? If one does succeed, how much more information is contained therein than the simple collections of user actions? Is this extra information useful in predicting user engagement? What other properties and insights can we extract that are specific of user behaviour in these contexts?

Studies have found that certain patterns in behaviour are informative in indicating positive or negative user experience, evidenced by (Mehrotra et al., 2017). Extensive signals can be engineered based on the behavioural features selected in the first objective, but they could not be tested all together without confounding the results or reducing the interpretability. Therefore, it is necessary to extract a set of key actions and form *behaviour sequences*. Thus, our second objective focuses on leveraging the patterns in the *sequentialized action space* that occurs as a function of user engagement. This phase of the research produced evidence of a significant association between search behaviour sequences and user perception of engagement. Furthermore, it provided evidence to suggest properties of an ideal measure of engagement, which takes into account the sequential nature of user behaviour.

Obj.3 To develop and evaluate measures of engagement using behaviour - perception relationships.

How can one extract simple numerical measures of engagement from the above insights that are interpretable and easily computable from a single user's action sequence?

Metrics and models have been proposed for user perception during information seeking, such as search satisfaction (Hassan et al., 2011), and utility (Machmouchi et al., 2017). However, user perception of engagement is collected directly from users, which is different from the metrics used in previous studies (e.g. satisfaction), in which the authors focus primarily on the searching scenario and assume successfully locating the correct documents is highly correlated to positive perception. They collected either third-party judgements or labelled user session by the occurrence of search success. Our work is closest to Machmouchi et al. (2017)'s; however, the latter does not take user behaviour order in to account. This phase of the research proposed a set of measures of engagement based on the behaviour - perception relationships discovered in this research, and produced evidence of improved prediction performance of developed measure compared to the state-of-art measure.

Obj.4 To compare behaviour - perception relationship among four different engagement dimensions.

How does one measure user reported engagement? Once these measures have been established, how do they relate to user behaviour? Is the correlation equally strong amongst all these measures? How about across the two information seeking contexts?

Previous studies have identified many potential proxies for engagement. However, in theory, these capture significantly different aspects of the user experience, and it is therefore more appropriate to refer to them as engagement dimensions. The concept of engagement benefits from this multi-dimensional nature, as we wish to capture the user experience in the most comprehensive way, including novelty, usability, felt involvement and endurability, as they describe user experience in a more comprehensive way. Across the first two research phases, we compared the links between different engagement dimensions.

Obj.5 To compare behaviour - perception relationship in browsing and searching.

How do our intuitions hold up comparatively in the two information seeking contexts? What are the inherent differences?

Searching and browsing are two types of online information seeking, and researchers have

discussed them from different perspectives. Across the three research phases, we compare the results using two selected datasets.

1.3 Scope

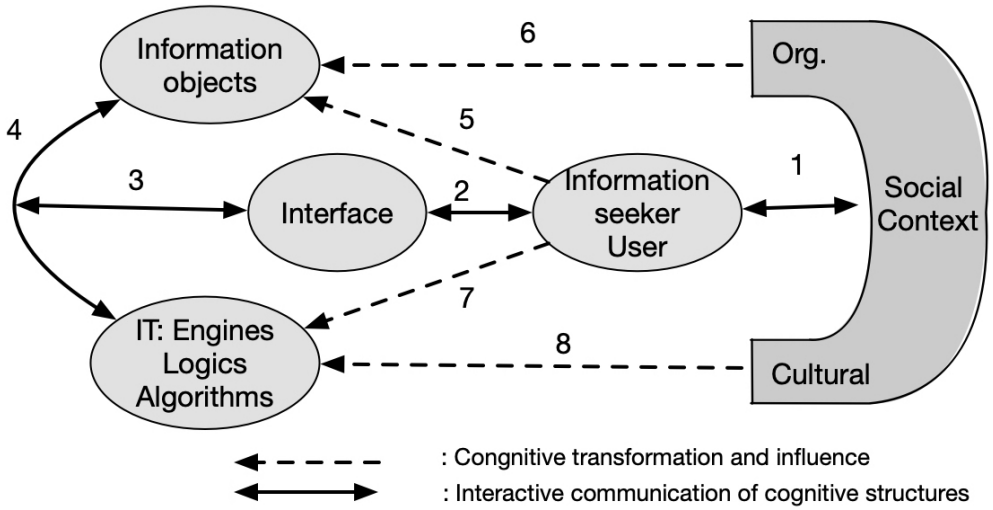


Figure 1.2: Generalized model of the Information Seeking & Retrieval processes with participating cognitive actor(s) in context (Recreated from Ingwersen and Järvelin (2006), p. 261).

Ingwersen and Järvelin (2006)'s model (figure 1.2) emphasizes the information processes that are executed during Information Seeking & Retrieval. The two types of links are illustrated, first the process of interaction (link number 1-4 in figure 1.2) and generation and transformation of cognition or cognitive influence (link number 5-8 in figure 1.2). In terms of reflecting user perception of the process or experience, these links become interwoven.

This thesis is a study of the links between two levels of such a cognitive framework, namely the cognitive-emotional level and social & physical level (figure 1.3), which have received separate attention in prior work. More specifically, we focus on the relationships (link number 2, 5, and 7 in figure 1.2) between user perception of the process, which are the objects at the cognitive-emotional level and user interaction with the IR system, which we refer to as user behaviour, while performing browsing and searching tasks in a single session. It is at the same time intra-session and user-centred IR evaluation. Our focus is on user common behaviour patterns and their proper interpretation in terms of reflecting predefined user engagement

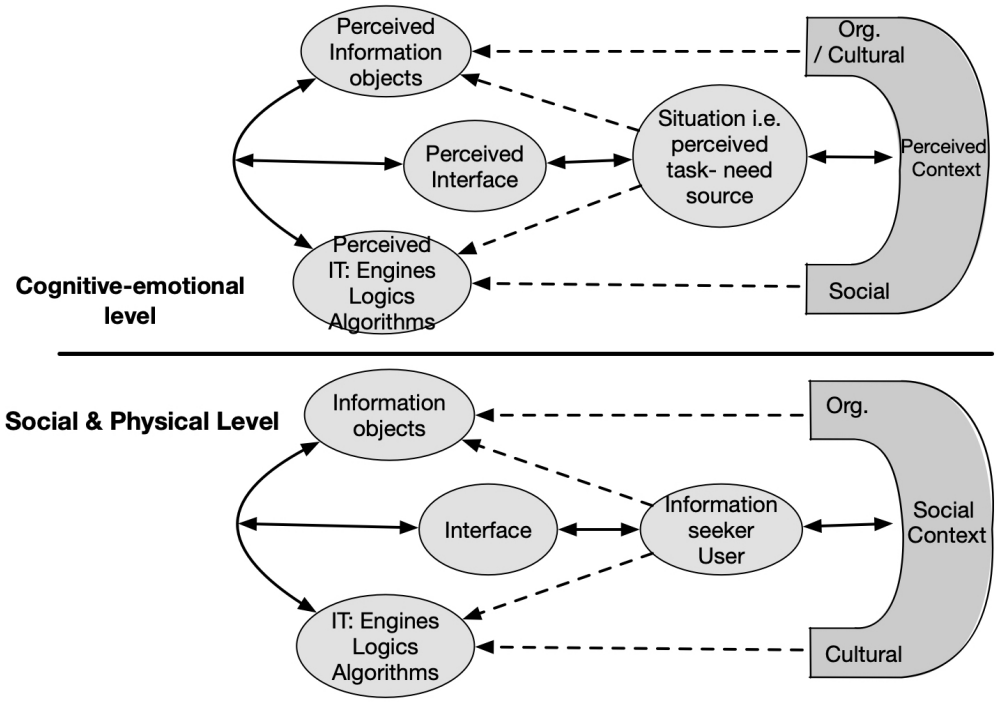


Figure 1.3: The cognitive-emotional level and social& physical level of the cognitive framework of Information Seeking & Retrieval processes (Recreated from Ingwersen and Järvelin (2006), p. 278). Arrows correspond to arrows in figure 1.2.

dimensions (O'Brien and Toms, 2010a) within one session. We do not extend our discussion to any particular group of people, nor argue for the definitions of dimensions of engagement.

1.4 Ethics

The two datasets used in thesis are secondary data that has already been gathered, by persons other than the student, are completely anonymised and there is no possibility of their being de-anonymised. The two datasets were collected at two different times (see details in (Toms et al., 2013; Hall et al., 2013)) and so two different ethics approval applications were made before the collection. As part of the informed consent process, participants gave permission for their data to be re-used in future research. The approval letter of the self-declaration of this study is attached in appendix.

1.5 Thesis structure

Following this introduction, in chapter 2, the literature is reviewed to situate this study in relation to the scope, theoretical frameworks, and the state-of-art empirical studies. In chapter 3, the overall research design is presented with the datasets and measures. In chapter 4, the relationships of behavioural features and four engagement dimensions are identified with the phase 1 study. In chapter 5, the relationships between behaviour sequences and four engagement dimensions are identified with the phase 2 study. The findings of phase 1 and 2 are reviewed in the beginning of chapter 6 to develop a set of measures of engagement, and these measures are further validated in chapter 6 with the phase 3. Chapter 7 presents the summary of findings across three research phases. The contributions, limitations and ideas for future work are stated in chapter 8.

Chapter 2

Prior Research

2.1 Overview

The research literature presented here is used to provide background information, identify gaps in understanding, and to situate this study in relation to what is already known about user engagement during the information retrieval process. We start by looking at the literature that sets the stage for this study, exploring important background concepts such as information retrieval (IR), information seeking (IS - an overarching concept of IR) as well as IS as a process, and how the process is evaluated. Then we gradually move on to user experience and engagement, and explain its connections with information retrieval. We further move on to discussing behaviour and user perception in the large and progress through studies measuring user perception and behaviour of user engagement in IR and their relationships.

The following sections of this chapter are divided into four main parts. We start with the literature review methodology. The following section describes research relating to the information seeking and retrieval theory and related concepts (section 2.3) that are relevant to information retrieval from the user-centred IR evaluation perspective. More specifically, this section introduces theoretical models for the information seeking process and provides a justification for the selection of the Information Seeking Process model (Marchionini, 1995) for use in this thesis. We also introduce evaluating IR process via user perception and user behaviour. After

this section, two concepts, user experience and user engagement (section 2.4) details how these two are related to IR.

We discuss how measuring user perception and user behaviour in IR was done in section 2.5.2, which includes the development of measuring user perception and user behaviour in IR, and summarizes studies specific to user engagement. The following section (section 2.6) provides an overview of verifying the intersection of these two types of measures in the past studies, especially in IR or in user engagement. The last section provides a summary of the literature review.

2.2 Literature review methodology

The literature review is conducted following the steps suggested in Bryman (Bryman, 2016) (p.99):

Step.1 "Define the purpose and scope of the review."

Step.2 "Seek out studies relevant to the scope and purpose of the review."

Step.3 "Assess the relevance of each study for the research question(s)."

Step.4 "Appraise the quality of studies from Step 3."

Step.5 "Extract the results of each study and synthesise the results."

It should be noted that the order of steps was not entirely linear and earlier steps were returned to and revised as the review progressed. We now state what we did in each step:

Step 1: "Define the purpose and scope of the review.": the purpose of reviewing the user engagement measures in the IR literature and related concepts was:

- Establishing the context of and providing theoretical frameworks for the topic of IR evaluation and user engagement.

- Situating this study in relation to what is already known about measuring user engagement in IR.
- Identifying the main methodologies, data collection tools, and measures that have been employed in measuring user engagement in IR.

The key criteria for considering whether to include a study were that it must be about user-centred evaluations of information systems with a focus on user engagement. These criteria are broad and there were potentially many studies that could be included. The scope was narrowed as follows:

- Information systems: The focus was on use of internet search engines, but also included studies of information systems such as news reading platforms, and multimedia applications.
- User engagement: The focus was on using engagement as the assessment of user experience. If studies focused on a related concept of engagement (e.g., satisfaction) these were also included.
- Language of published studies: Only studies published in English were included. This is a limitation of this review.
- Date of published studies: There is no time criteria for reviewing the related theoretical concepts. In the early 1990's, we see a rise in the number of empirical studies on the newly emerging branch of interactive IR (Borlund, 2013); hence this point in time serves as our criterion for selecting empirical studies for review. In terms of methodology we focus on studies conducted post 2010, which present the current state of the art in experimental methodology, and measures of user perception and behaviour.

Step 2: "Seek out studies relevant to the scope and purpose of the review.":

First, the University of Sheffields library catalogue and Google search engine were used to locate relevant books and articles. In addition, key journals in the various fields were consulted

(e.g. Transaction on Information Systems [1], JASIST [2]), along with the proceedings of relevant conferences (e.g., SIGIR [3], CHI [4]). The Google Scholar search engine was also employed. At the same time, we adapted the forward and backward snowballing approach (Wohlin, 2014) in order to follow the references of identified literature and the works cited them. Monitoring new literature which emerged during the study is deployed by setting up alerts in Google scholar based on the key search terms and new content alerts for key journals.

Seven key general search terms were used ("information retrieval"; "information seeking"; "information seeking process"; "information retrieval evaluation"; "user experience"; "engagement"; "user behaviour"), which resulted in the discovery of immediately relevant resources from which useful citations could be garnered to widen the search. To ensure completeness increasingly specific search terms were employed (e.g. "engagement" and "searching" and "behaviour"; "engagement" and "measurements").

Step 3: "Assess the relevance of each study for the research question(s)."

Studies were considered relevant if information retrieval process models or user engagement measurements were described.

Step 4: "Appraise the quality of studies from Step 3."

Influential studies in the field were primarily selected. Studies should be in published in journals or top conferences or be cited by others.

Step 5: "Extract the results of each study and synthesise the results."

These studies are summarised in the following sections of this chapter.We start with the theoretical concepts that are related to information seeking and information retrieval evaluation. Subsequently, we outline the state of current understanding on the topics of *user experience*, *user engagement*, and the links we stipulate in between *user perception* and *user behaviour*.

[1] https://tois.acm.org/index.cfm
[2] https://onlinelibrary.wiley.com/journal/23301643
[3] http://sigir.org
[4] https://sigchi.org

2.3 Information seeking and retrieval and related concepts

2.3.1 Information seeking and retrieval

Information Seeking (IS) is aptly characterized by Case and Given (2016) as "a conscious effort to acquire information in response to a need or gap in your knowledge" (p. 6). While this definition encompasses many types of processes and interactions, we are primarily concerned with information seeking in the space of online information acquisition, via information retrieval systems designed to facilitate this type of access to users. In this study, we focus on Information Retrieval (IR) evaluation, and IS serves as the boarder context as studies in information seeking and retrieval (IS&R)(Byström and Järvelin, 1995; Ingwersen and Järvelin, 2006; Kuhlthau, 1993; Wilson, 1999) suggested. "[Information retrieval] is but one means of information seeking which takes place in a context determined by, e.g., a persons task, its phase, and situation." (p. 1)(Ingwersen and Järvelin, 2006). Therefore, we rooted our theoretical ground in IS process studies, and draw insights from the IR studies.

An information retrieval system is "an information system which is constituted by interactive processes between its information space, IT setting, interface functionalities and its environment, and capable of searching and finding information of potential value to seeker(s) of information." (p. 387) (Ingwersen and Järvelin, 2006). A hallmark of a good information retrieval system in the modern age is the ability to captivate users, causing users to invest time, effort and emotion (Lehmann et al., 2012) into the interaction, which represents much more than just satisfaction (O'Brien and Toms, 2008). Thus, our main concern is not elucidating the nature of information retrieval, but in evaluating its quality with respect to measures of user interaction and experience, which we shall define below. Some of the literature aiming at enlarging such descriptions will be omitted. However, in order to begin speaking about *process evaluation*, a certain foray into the theoretical models which have guided past academic discussion on this topic and grounded previous attempts at quantitative analysis of this interaction is necessary. In the following sections we collect a summary of models which describe interaction (in the

sense of Human Computer Interaction) with a focus on process modelling.

2.3.2 Theoretical models

We start by describing an information seeking process model (Kuhlthau, 1991, 1993) which focuses on users emotional changes and the subsequent effects on their behaviour and is based upon observations from students working on demanding tasks such as school assignments. Kuhlthau (1991)'s model is a universal one that is applicable to any domain (Case and Given, 2016). For us, it represents an example of a process model devoted to information seeking, and despite its context lacking the required focus for our current needs, it features the generalized cognitive and affective processes specific to this type of interaction. In particular, it treats information seeking as a process of the gradual refinement of the problem area (Wilson, 1999) rather than mental iterations in cognitive models. It devotes a special focus to feelings, actions and thoughts, and was developed in the same manner as the Affect, Behaviour and Cognition model (the ABC model of attitude) (Breckler, 1984; Eagly and Chaiken, 1998). Figure 2.1 illustrates the seven states associated to the user interaction: initiation, selection, exploration, focus formulation, collection, presentation, and assessment. There are two properties worth highlighting at this point. The first is the sequential development of the states of user interaction, which is characteristic of any process model. The second is the interplay between these states and the affective, cognitive and physical activities of the user which is definitive of affective models in general.

Apart from the process property of IS, it is also treated as an interaction between human and the external system. In the following section, we first introduce the interaction model and information seeking as a process in detail, then we move on to the theoretical models of information seeking process.

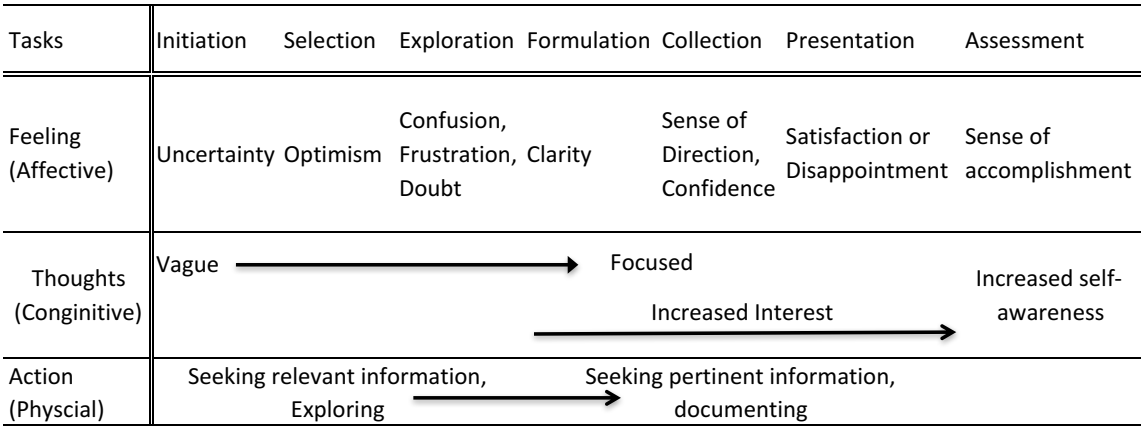


Figure 2.1: Kuhlthau model of information seeking process (Recreated from Kuhlthau (1993)).

Interaction Model

Donald Norman, one of the foremost researchers in user-centred design, describes how users interact with systems as an iterative process. This widely referenced early model of user interaction (Norman and Draper, 1986) employs seven stages to present an essential and summative process model of a general user's interactions with the system (figure 2.2). It propounds a highly generalized framework for any type of interaction on the user's end, with unconstrained goals, expectations, actions and reactions in terms of perception. Despite its generality, two features of its design are worth extracting, as they will constitute the pivotal theoretical underpinning of further arguments in this text.

1. It showcases not only the distinction between mental and physical activity within the process of interaction with the system, but also draws the line that separates them. It emphasizes how the user's goal is a precursor to the activity cycle which bridges into the physical interaction with the system, the observable behaviour of the user, which in turn brings us to our second point:

2. There are two types activities (or *states* in the interaction cycle) that bridge the gap between mental activity and the system. The first is the ability and process through the user goals, decide and executes their actions. The second, as a response to the user's physical interaction is an interpretation step that takes the user from their perception of the results of the interaction to a personal evaluation of their current state.

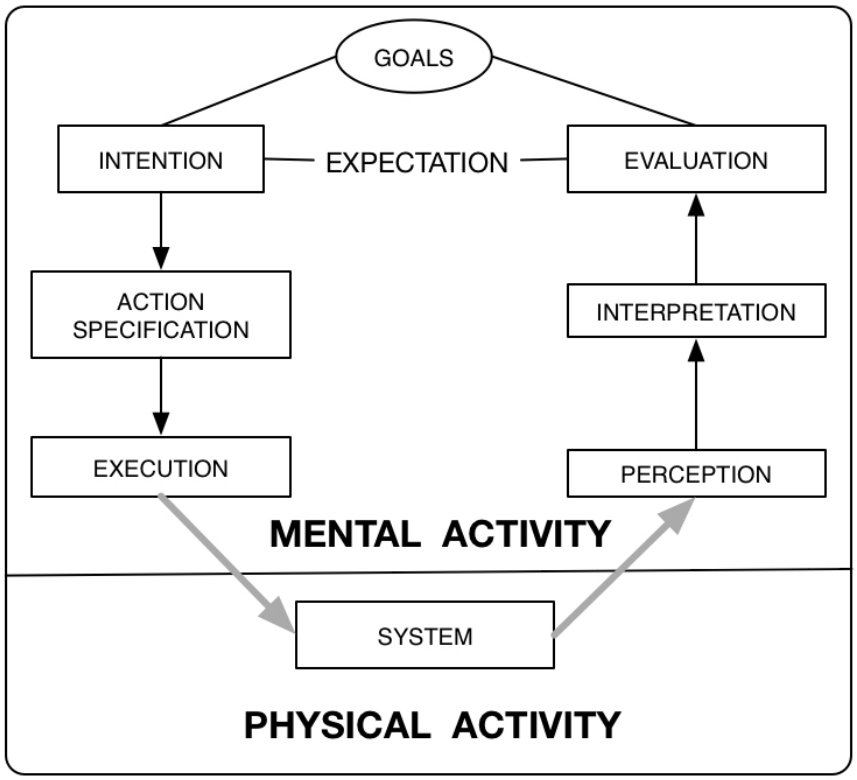


Figure 2.2: Norman's model of interaction (Recreated from Norman and Draper (1986))

This active-reactive process through which the user establishes a cognitive feedback loop with the system is the entire linchpin of the cognitive framework that brings credibility and draws interest upon all models that seek to evaluate the user's experience in terms of their sequential interactions with the system, including our own in the following chapters.

Also due to its generality, Norman's model holds tight in the face of numerous mappings to partial interactions, making it relevant both at a micro and at a macro-scale (Case and Given, 2016). In its most encompassing form, it describes the interaction of the user at the session or even multi-session level, where sprawling, over-arching themes give rise to goals that the user must carry through arbitrary complexity and abstraction. At the most granular level, the minute tasks that the user issues to himself, the subconscious imperatives that detach themselves from the higher goals execute in the same action-feedback loops that Norman's model describes going through all seven stages before completing and tying into the next micro-task.

Information seeking process

Information seeking is one type of the more general human computer interactions described in Norman's model (figure 2.2). A number of theoretical models were published to describe information seeking over the last six decades, and they "vary as to their assumptions, structure, purposes, scope, and intended uses" ((Case and Given, 2016), p. 144). In order to introduce the conceptual models of the information seeking process, we first introduce the properties of such process.

Looking back at the original definitions of the word 'process', many areas ranging from natural phenomena, legal actions to manufacture operations were involved. The definition of process from Merriam-Webster is a series of actions or operations conducing to an end (Webster, 2016). This definition, in a rather more general scope that is of interest here, captures the fundamental fact that a process contains several stages that link in a temporal scaffold containing a clear start and end. In this text we shall see definitions of this form abound, whereas here it is necessary to mention that for the generalized information seeking processes under consideration, the user is constrained by time and their ability to interact with the system in such a way that their interactions become *sequential*.

From early studies through to the more recent ones, taking into account components of the information seeking process such as starting points, ending criteria, influential factors, stages and associated outcomes, various definitions have been proposed. Researchers in IS generalised it in a descriptive and conceptual way, focusing on the high level mental activities of the users (see reviews of the models in (Case and Given, 2016; Ingwersen and Järvelin, 2006)).

Research on the starting point of the process was focused on investigating the motivations behind the act itself. For example, the Leckie model (Leckie et al., 1996) suggests "work roles" and "tasks" as a motivator of information needs, as their model applies to professionals. Some other researchers agree that the awareness of information needs, or sometimes described as a gap in knowledge, is the first step in the search process (Belkin et al., 1997; Byström and Järvelin, 1995; Kuhlthau, 1991; Wilson, 1999).

Unlike the start of a process, there are various discussions devoted to describing the antecedents of an ending, or the criteria of the termination. "If researchers agreed that information need is the natural start point, then when the need is met or abundant should be the natural stopping point" (p. 24) (Zach, 2005); however, real search scenarios are far more complex (Case and Given, 2016; Ingwersen and Järvelin, 2006; Zach, 2005) due to several influential factors such as time pressure and task types. Acquiring too much or too little information than users really need will lead to information avoidance (Case et al., 2005) and information overload (Rogers, 1986), or information anxiety (Wurman, 1989), all of which may lead to outcomes that are not optimal (Wilson, 1981). On the other hand, regarding the stopping criteria, theories from different perspectives and disciplines were introduced, such as stopping rules (e.g. the satiation rule, the disgust rule, and the combination rule (Kraft and Lee, 1979)) from a psychological aspect with a focus on estimation of the relevance of new information; utility theory (Cooper, 1976) and production theory (Varian et al., 1996) from economics; and *satisficing* strategy (Agosto, 2002; Prabha et al., 2007) from decision making studies (Simon, 1955; Zach, 2005). These theories and rules lead to the establishment of several formal models, simulating the information seeking process towards the stopping point, such as the notable Probability Ranking Principle (Robertson, 1977), which is mainly applied on the decision theory, and more recently the Search Economic Theory (Azzopardi, 2011) is based on Production theory. Despite research that focuses directly on information needs, Kuhlthau (1991) evaluates the factors that lead to the changes over time in user response to their information needs. This model was mainly based on observations of students, and suggested that the termination of search is connected to the users feeling of relief/satisfaction or disappointment (Kuhlthau, 1991). However, more models are illustrated in an iterative manner (e.g. Wilson (1981, 1999); Marchionini (1995)), in which it is difficult to define the stopping point systematically.

The properties listed above encompass some of the early attempts in understanding the information seeking process and mitigating the shortcomings of models such as Norman's, especially around evaluating the user's interaction patterns and their perception of the interaction. The following models differ in design and scope, yet they have all nevertheless been engineered in order to capture different facets of the intricate information seeking process.

Information seeking process model

We turn now to set of models which attempt to address these issues. As we shall see, they each have their own strengths and provide differentiated insights into the more general problem. To begin, Marchionini constructed an integrated model of the information seeking process (Marchionini, 1995). His focus is on information seeking itself, and describes information seeking as an 8-step process (figure 2.3) comprised of: recognising and accepting an information problem, defining and understanding the problem, choosing a system, formulating a query, executing (an atomic) series of commands, examining results and extracting information, and reflecting/iterating/stopping. A process begins with the realisation that a problem/information need exists, upon which complex ending criteria that depends on context such as system function and information goal are formulated either consciously or subconsciously (Marchionini, 1995). Although the transition from one step to another provides a basis to describe information seeking as a sequence, each step in the process does not carry with it instructions about measurable outcomes of the search. On the other hand, the transitions between the steps are clearly categorised according to their likelihood within the user's process. The model categorises transitions into three types: the default, highest probability transition and two non-standard types dubbed high probability and low probability, without a measurable condition when a certain type of transition takes place. On the other hand, this descriptive feature of the model already confers upon it the flexibility to capture interactions that the previous did not, and evaluate the relative likelihood of these sequences of interactions. A priori, many possible sequences are allowed, but some are more *natural* than others to the user's interaction and thus occur with higher probability. This allows complex and novel information to be captured about the interaction, which differentiates patterns at a sequence level. We shall make use of this philosophy heavily in this text. However, the seemingly random iterations of transitions make it difficult to analyse the nature of complex and dynamic information seeking process as a whole, and we are no closer to providing a clear measurement of its outcome.

In a structuring attempt, Marchionini (Marchionini, 1995) grouped these stages into a 3-subprocess framework for the information seeking process (figure 2.4): understanding, planning

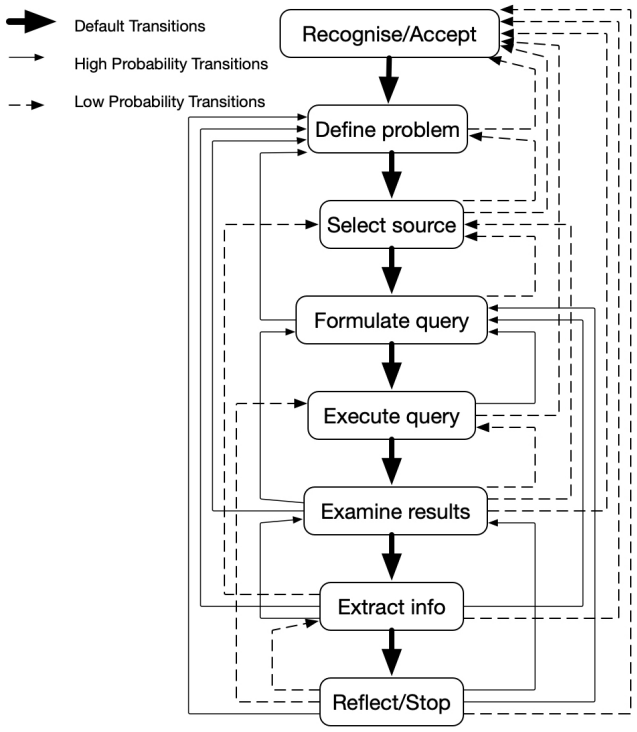


Figure 2.3: Marchionini's information seeking process model (Recreated from Marchionini (1995)).

and execution, and evaluation and use. These categories shed light on the higher order processes involved in the search, organize the previous eight stages and align them in a parallel structure ordered by the time axis.

Recognising and accepting an information problem, and defining and understanding a problem, are grouped under the broader *understand*; choosing a system, formulating a query, executing the search, and examining results belong to a the general category dubbed *plan and execution*; lastly, examining results and extracting information as well as reflecting/iterating/stopping belong to *evaluation and use*. Usually capturing a process at a more granular level provides more flexibility and finer information when tackling applications, but when the higher level categories suffice to explain much of the variance in our process (or to capture the patterns which are characteristic of the process), such an exercise can be regarded as a worthwhile effort in summarization. In our own studies further we will apply this philosophy, moving from *fine* to *coarse* in order to capture types of interactions at each level of granularity. We shall see that even though subprocesses get aggregated at the level of the higher categories, evaluation becomes easier, and we sometimes even manage to eliminate much of the noise in our data

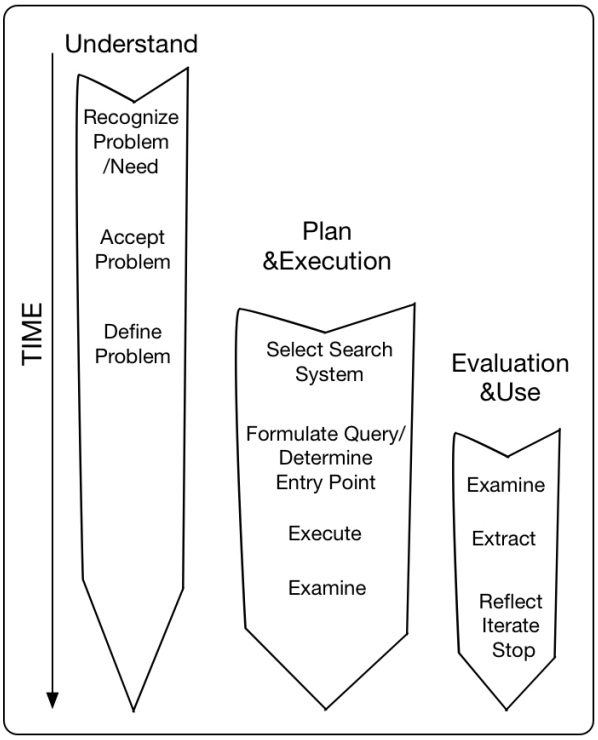


Figure 2.4: Parallel information seeking subprocess (Recreated from Marchionini (1995)).

through such techniques. However, while these two process models (Marchionini, 1995) in figure 2.3 and figure 2.4 share the property that they are describing high-level mental activities and are intrinsically conceptual, lacking specific, empirically confirmed relations with real time tasks, we shall see how to capitalize on the insights they provide while incorporating dimensions of the user interaction that lend themselves well to quantitative study.

One further model stands as a critical precursor to the research ideas which motivate our study. Unlike the two models presented above, Bates (Bates, 1989) developed a dynamic model, the berry-picking model, which " entails that each new piece of information that searchers encounter provides them potentially with new ideas and direction to follow" (p. 218) (Ingwersen and Järvelin, 2006). The three models (Norman and Draper, 1986; Marchionini, 1995) mentioned above suggest that users recognise a specific goal at the beginning of the process, which remains static throughout the interaction, which, in turn is driven by their struggle to fulfil their original goal. The berry-picking model showcases how the information needs of users change throughout the information seeking process, leading to dynamic outcomes (figure 2.5). Instead of suggesting that users hold an un-changed information need throughout the entire information seeking

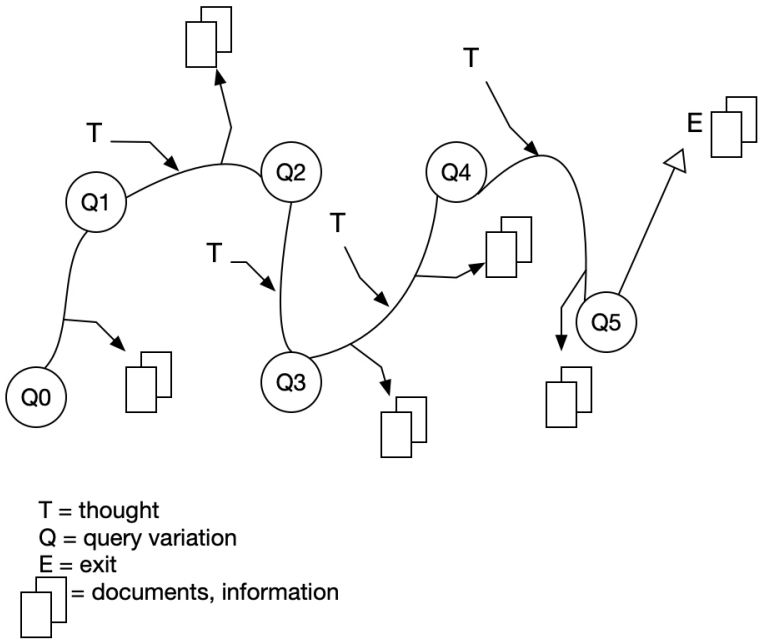


Figure 2.5: Bates berrypicking model (Recreated from Bates (1989)).

process, Bates observed in real world cases that users actually encounter new information during the process and modify their thoughts and perspectives, which further affects their information needs. That is to say, the process conducted by a user does not necessarily point to only one static information need. As information encountering happens at the mental level, we do not have a set of clear and conveniently measurable outcomes of how users reflect on the documents, and how users use the documents to redirect their search. Therefore, this model invokes the need to specifically and accurately assess the user's goal suggesting a more organic approach to task completion. The theory of *information foraging* perhaps best applies to describe the mental patterns which govern the behaviours underlying this model. Developed by Pirolli and Card in the early 90s (Pirolli and Card, 1995), this theory briefly draws analogies between information seekers and organisms in search of food. Such organisms consume a particular *diet* - or type of information - in a certain *patch* of their habitat - a particular information seeking resource. When the organism reaches saturation in a current patch the process ends. However if the patch is depleted before saturation, a transition occurs. The dynamics of the system is governed by the organism's immediate needs -maximizing their intake - and their resources in the form of time and effort. Furthermore, both the organism's patches and diet can change dynamically, and this complexity stands at the heart of what makes this model so

comprehensive. We note here that devising measures to evaluate the users interaction under the assumptions of this model intrinsically entail being able to capture transitions, and assessing the costs and gains of each transition to each patch, as well as the *perceived* costs and gains on the part of the user, in order to understand the triggers behind such transitions. Evidently, for an arbitrary search, it is not necessarily the case that the interaction happens in such a nonlinear fashion. Nevertheless, the allowance for this flexibility in the model represents a crucial step forward in accurately describing the user interaction. The way researchers investigated the information-encountering phenomena is observing the changes in user behaviour, such as query inputs, and have applications in query suggestion (Kruschwitz et al., 2013) and user search behaviour (Downey et al., 2008).

Although these theoretical works mentioned above give guidance on the information seeking process, all of them lack clear, specific outcome sets and transition criteria between stages, bringing difficulty to observation, and therefore limit large-scale replication experiments compared to the scale of search algorithm tests. Still, these models have their merits. Concluding from these models, the interaction between user and system is iterative (Marchionini, 1995), and contains two main components (Norman and Draper, 1986; Kuhlthau, 1993): a physical element with outcomes such as actions and behaviours, and a mental component with the outcomes of user feelings and perceptions. In addition, these two components may affect one another, or at the very least the changes in user behaviour reflect users mental changes (Bates, 1989). Overall, these models guided a plethora of user studies of the information seeking process with many different implications. With these studies came an emphasis on users and their needs, shifting focus to assessing the quality of user experience during the information seeking process. As the process is comprised of a mental and a physical component, researchers may examine user experience using the user perception (mental) and user behaviour (physical) outcomes from the process.

2.3.3 Evaluating IR through user perception and behaviour

In evaluating user interaction with the information retrieval systems, typically measures are based on two approaches: one based on user perception of the system, and one on how the user behaves while engaging with the system.

Perception refers in general to an agent's subjective response to an objective stimulus. As such, when speaking of perception, it is of central importance to be able to isolate unique user experiences and differentiate this individual impact from the objective nature of the result of the interaction outputted by the system. The dependence of perception on individual experience (Borgman, 1989; Fenichel, 1981; Dillon and Watson, 1996), search context (Ingwersen and Järvelin, 2006), and information retrieval tasks (Leckie et al., 1996; Byström and Järvelin, 1995) has been examined throughout the past decades. Although the variety and vast influential factors mean that user perception is ultimately a somewhat abstract measure, perception is no doubt a crucial element of evaluating IR that attracts research attention.

Outside of the IR discipline, perception is studied in various ways in different fields, ranging from cognitive psychology to neuroscience. Even as a concept rooted in the field of cognitive science, perception is somewhat vaguely defined, as the distinction between perception, cognition, representation, and understanding is rather difficult (Ingwersen, 1992). According to two cognitive psychologists (Gibson, 1966; Gregory, 1970), perception is a process of receiving information from the environment which is then transmitted to the brain. However, it can be distinguished into two types of processes; bottom-up processing (Gibson, 1966) and top-down processing (Gregory, 1970); Gibson argues that perception begins with the stimulus itself whereby Gregory refers to the use of contextual information in pattern recognition. Despite the long history of discussion, there is no consensus on how perception is developed (e.g. bottom-up processing (Gibson, 1966) versus top-down process (Gregory, 1970)), and the influential factors (e.g. context (Bruner and Minturn, 1955), emotion (Allport, 1955)). What is certain, however, is that perception is based on both freshly-received and previously stored information, and involves signals in the nervous system, which in turn result in physical behaviour signals (Allport, 1955). Indeed, the relationship between perception and information, and between

perception and behaviour, is investigated in various domains including information retrieval (e.g., satisfaction (Al-Maskari and Sanderson, 2010), usefulness (Cole et al., 2009)).

User behaviour in the IR process refers to the actions and selections made by the user while interacting with a system (e.g, clicks, keystrokes, and eye movements). An overarching concept of it is "Information Behaviour", which as defined by Case and Given (2016): "encompasses information seeking as well as the totality of other unintentional or serendipitous behaviors (such as glimpsing or encountering information), as well as purposive behaviors that do not involve seeking, such as actively avoiding information." (p.6). This is meant to be an all encompassing definition, capturing user behaviour irrespective of intent or purpose, and referring to all information encountered regardless of its significance. Wilson (1999) introduce a nested model of research areas within the general field of Information Behaviour, in which Information Behaviour is the macro field where information seeking behaviour happens. Information seeking behaviour is the most general type of behaviour invested in order to acquire information, while information searching behaviour is more specific, and relates to the specific interactions between user and system that take place within any information seeking endeavour. Jansen and Rieh (2010) adapted Wilson (1999)'s nested model and link the structures with information systems (figure 2.6). The model suggests that the nested research field of information behaviour can be mapped to a nested framework of information systems.

The principal research objective of this study is to analyse the relationship between user behaviour and user perception of engagement using information retrieval systems and implement the unfolded relationships to measure engagement. Out of the three levels of human-system interaction described in figure 2.6, this objective applies most naturally to the lowest of the three levels, the one which captures the relation between information searching behaviour and information retrieval systems. The user behaviour here refers to the participants' behaviour during the interaction with information retrieval systems. It covers both active and passive behaviours, such as clicking new results vs. merely staying on the results page without moving the cursor, and may involve conscious effort to acquire information. The motivation behind one behaviour may differ. For instance, when inputting queries, users may be motivated by a knowledge gap, or just casually exploring contents online out of leisure.

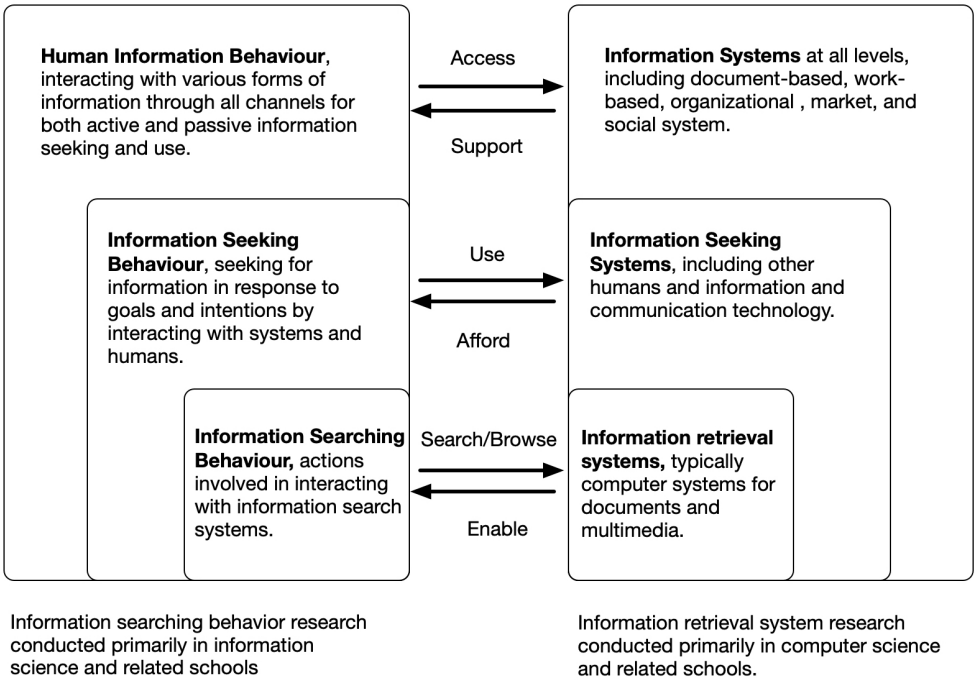


Figure 2.6: Framework of human information behaviour and information systems (Recreated from Jansen and Rieh (2010))

The intention in this section therefore is not to critically evaluate the multitude of models and concepts relating information seeking/searching interaction. Instead, most attention will be paid to research most closely relating to perception and behaviour measures to evaluate the IR process. However it is important to recognise, as Wilson (1999) has shown, that information searching activities occur within a broader theoretical context. The rest of this chapter (section 2.4, 2.5, 2.6) focuses on two concepts, experience and engagement exploring previous research that informs our understanding of modelling user behaviour and user perception in order to measure user engagement in the context of their interaction with information retrieval systems.

2.4 Experience and engagement

Traditionally, the success of information retrieval systems has been evaluated according to search system effectiveness (e.g., relevance, precision and recall)(Ingwersen and Järvelin, 2006), but this is only part of the story. Modern research interests have been elevated beyond assessing the quality of retrieved results. For instance, judging only effectiveness does not identify

whether users consider or value highly ranked documents in the same way as the system, or indicate whether the user perceives the search experience as a positive one (Turpin and Scholer, 2006). We know from Human Computer Interaction (HCI) that an unpleasant experience will influence whether a user continues to interact with a system or moves on to another (Lehmann et al., 2012) (cf. the notion of cost or effort in the act of searching as described by the information foraging theory (Pirolli and Card, 1995, 1999), whereby an unpleasant experience can be seen as a positive increase in effort on the part of the user). The third wave of HCI (Bødker, 2006) started earlier this century, shifting focus from technology for work to user experience, and the trend continues (Bødker, 2015) with even more focus on its interactive nature. The ISO defined the term "user experience" as (ISO, 2010):

"Person's perceptions and responses resulting from the use and/or anticipated use of a product, system or service." (p.7)

Although the definition above sounds like a common understanding of user experience, user experience is surrounded by a wide variety of meanings in the existing literature (Forlizzi and Battarbee, 2004; Law et al., 2009). Hassenzahl and Tractinsky (2006) reviewed a mixture of sources and describes user experience as :

"User experience is a consequence of a user's internal state (predispositions, expectations, needs, motivation, mood, etc.), the characteristics of the designed system (e.g. complexity, purpose, usability, functionality, etc.) and the context (or the environment) within which the interaction occurs (e.g. organisational/social setting, meaningfulness of the activity, voluntariness of use, etc.)" (p.5)

What we may extract from the above definitions and interpretations is that user experience represents a unique product of the user's perception, affective capability, mood, emotions and goals on the one hand and the result of their interaction with the system, as illustrated by the action-reaction cycles described by Norman's model (Norman and Draper, 1986) - in effect the entirety of the user's behaviour factors into the cognitive representation they construct of their interaction with the system.

Forming user experience consists of a number of steps, which can be considered as a process of decision-making and user reaction to interactive products (Sutcliffe, 2016; Hartmann et al., 2008). To assess the quality of user experience, measures such as usability (Norman and Draper, 1986) were considered over decades. The word "engagement" was also introduced to access the quality of user experience (O'Brien and Toms, 2008).

User engagement is a closely related concept to user experience, whose connection to the latter has been re-evaluated through the years. Let us explore some of the definitions and concepts pertaining to engagement, as this will be a central topic to our study. Starting off with *discrete engagement*, one may define it as "a feeling that one is directly manipulating the objects of interest" (p. 317) (Hutchins et al., 1985). Moreover, Laurel (Laurel, 1993) emphasised the enjoyable component by describing user engagement as "the state of mind that we must maintain in order to enjoy a representation of an action"(p.112-113). Recently, the definition of engagement has focused on numerous other perspectives, including emphasising engagement as a dimension of usability that encourages interaction (Quesenbery, 2003), the ability of a system to capture users attention and interest (Jacques, 1996), or simply a feeling of control from user (Bødker, 2006) or of playfulness (Webster and Ho, 1997). O'Brien and Toms (2008) summarised the studies above, constructed their definition of engagement around cognitive (Laurel, 1993), affective (specifically intrinsic motivation)(Jacques, 1996), and behavioral (Hutchins et al., 1985) states of interaction with a computer application, and further concluded (O'Brien, 2016) that :

"UE [User engagement] consists of affective (e.g. motivation, positive and negative feelings), cognitive (e.g. challenge, interest), and behavioural (e.g. interactivity) components." (p.5)

What we immediately notice is how this aligns with the structure of user experience, as both user perception and user behaviour are a part of user engagement. User engagement focuses on the *positive aspects of user experience* (O'Brien and Toms, 2008), especially how users are attracted by the system they are interacting with, and hence how users are driven to use and engage with the system. A hallmark of a good system in the modern age is the ability to captivate users, causing users to invest time, effort and emotion (Lehmann et al., 2012) into the interaction, which represents much more than just satisfaction (O'Brien and Toms, 2008).

The major difference between two recent definitions of user engagement ((O'Brien and Toms, 2008; Sutcliffe, 2009)), is that in the previous (O'Brien and Toms, 2008) the users' intention of re-use is considered as long-term engagement, while Sutcliffe (2009) considers the concept of user engagement is defined "primarily to explain how and why applications attract people to use them within a session while making the experience exciting and fun" (p.3) and thus is different from the long term effect ("many sessions and even years" (p.4)) captured by user experience. In this study we adapted the definition of user engagement by O'Brien and Toms (2008), in which they connect different perspectives (Webster and Ho, 1997; Jacques, 1996), as, "a quality of user experiences with technology that is characterized by challenge, aesthetic and sensory appeal, feedback, novelty, interactivity, perceived control and time, awareness, motivation, and interest and affect" (p.949). In other words, engagement is complex, multi-dimensional, and captures many features of user experience.

In this study, we analyse user engagement, an indicator of the quality of a users interaction with information retrieval systems, which is a multidimensional construct that integrates multiple variables that are indicative of human perception of the one's experience. We examine how user engagement as defined in O'Brien and Toms (2008) reflects user behaviour. We have seen that both perception and behaviour on the user's part play a pivotal role in all definitions of engagement. Hence it should not be surprising that currently, the main research avenues for assessing user engagement focus on two main types of measures: user perception of being engaged and user behaviour while interacting.

2.5 Measuring Behaviour and Perception of user engagement

To measure user experience in IR, and more specifically what we are going to examine in this study, user engagement, we choose to investigate and assess the outcomes of the process. Revisiting Norman's model of interaction (Norman and Draper, 1986) (figure 2.2), two questions can be asked based on the mental and physical activities respectively: what do the users think

of their own engagement - how do they perceive it for themselves? and what did the users do - what is the actual set of actions they undertook and how did they go about it? We refer the answer to these two questions as measures of user perception of engagement and user behaviour.

2.5.1 Measuring perception

This section first focuses on the user's perception of the information retrieval experience, with approaches used in collecting perception data in Information Retrieval and Human Computer Interaction studies. Then, we focus on validated measurements for user perception of engagement.

Measuring user perception of IR

Multiple concepts of user perception are used to assess the quality of user experience during IR, as reviewed by Kelly and Sugimoto (2013). Perceived usability (Nielsen, 1994), for example, measured via questionnaires such as the System Usability Scale (SUS) (Brooke, 1996) and the Computer System Usability Questionnaire (CSUQ) (Lewis, 1995), tends to be consistently deployed in IR studies. Other concepts include aesthetics of the system interface (Lavie and Tractinsky, 2004), satisfaction (Su, 1992), and user-perceived usefulness (Cole et al., 2009). Although it may seem that these concepts are highly correlated with one another (e.g. user perceived satisfaction, aesthetics and usability (Tractinsky et al., 2000) perceived usefulness and satisfaction (Calisir and Calisir, 2004)), as most of them were developed from usability research, the results are vast and controversial. Past studies (Tractinsky et al., 2000; Calisir and Calisir, 2004) indicate that there is no uniformity in the correlation analysis of these measures and it has become commonplace for each concept to have its specific emphasis and require separate measurement. As such, adopting multidimensional measures associated with these concepts may be required, especially in as much as our studies are concerned. We recall our discussion of the plurivalent nature of engagement here in the hope of instilling in the reader the parallelism between these methods and the nature of the object to be measured.

Traditionally, most user perception measures are collected using observations and self-reported methods, such as interviews, diaries, questionnaires, and stimulated recall (see reviews about collection methods in Interactive IR studies (Kelly, 2009), emotion studies (Lopatovska and Arapakis, 2011), and user engagement (Lalmas et al., 2014)), and thus require user responses to a set of questions or items. Directly measuring user perception relies on user's volunteering their responses (Lalmas et al., 2014; Kelly, 2009). Such dynamic events are obtrusive to a users natural interaction, interrupting the flow of the user experience, and making the collection impractical with large instruments (e.g., large number of questions or long guidelines). In addition, designing user perception experiments and analysing the data requires substantial background knowledge on the part of the researchers to account for the various influential factors mentioned above. Therefore, dynamically assessing user perception (e.g., at any point in the middle of the session) during the information seeking process has not yet been conducted fully and extensively. Moreover, Lalmas et al. (2014) concluded three main types of considerations while commenting on user perceived measures of engagement, namely communication, method bias, and reliability and validity, that obstruct the development and implementation of such a dynamic assessment. A more detailed summarization of the limitations of using questionnaires in Interactive IR studies can be found in Kelly (2009). Communication issues are mainly caused by the dependence on users and researchers' interpretation. Inflation is the phenomenon characterized by the tendency of using the high-end of the scale when evaluating systems (Kelly, 2009). This phenomenon was confirmed empirically in several studies such as Kobayashi and Boase (2012) and Junco (2013), in which discrepancies were found between participant-reported data and their actual behaviour. User-subjectivity (Sauer and Sonderegger, 2009) and social desirability bias (Donker and Markopoulos, 2002) were also observed. Method bias is also heavily discussed, especially in social science studies (e.g., (Podsakoff et al., 2012)) using self-reported methods. Suggestions for improving such methods include adding alternative measures, implicit measures, and objective behavioural measures (Fazio and Olson, 2003), considering from the perspective of the participants, context and the procedure (Burton-Jones, 2009), and specifically for IR studies, developing standardised, robust measures (e.g., questionnaires) rather than in an "ad-hoc" fashion (Kelly, 2009). To avoid methods bias, it is

essential to select well-developed measures that have been tested for reliability and validity.

Although the need for testing measures for reliability and validity is outlined, Kelly (2009) expressed that not enough emphasis is put on assessing the reliability of individual perception measures in IR research. A valid and replicable perception scale needs to go through a structured developmental process (DeVellis, 2003). Also, the wide usage of Likert-type scales limits the statistical interpretation of answers (Norman, 2010), as the scale is ordinal rather than interval.

Despite these concerns, perception measures can provide the insights that describe a user's experience, as perception outcomes obtain feedback directly from users. This represents an in-depth, personalized quantum of information from which we can extract accurate depictions of the user experience.

Measuring perception of user engagement

More recently, measuring user engagement has emerged as a more inclusive way of assessing user experience on digital systems, and associated measures based on perception in respect to certain domains are slightly different due to dependence on context and environment of the IR tasks. Due to the distributable and easy-to-replicate nature of questionnaires, efforts on developing reliable context-based psychometric scales and validating them have been invested for decades. Existing psychometric scales cover various attributes and have been applied in various ways. For instance, in an online shopping scenario, intentions of return and recommendation to others are considered (O'Brien and Toms, 2010a). On the other hand, in the eHealth domain, Kostkova (2016) concluded that engagement reflects on knowledge and attitude changes, and Lefebvre et al. (2010) suggests the credibility of the health information always plays a role. While for online news reading systems (Arapakis et al., 2014c), the focused attention sub-scale from the User Engagement Scale (UES) (O'Brien and Toms, 2010a) was selected together with the Positive and Negative Affect Schedule (PANAS) (Watson et al., 1988), which measures the affect before and after the experience, to capture user engagement. Four of the questionnaires, originally from different domains, have been replicated or discussed in various environments:

the Survey to Evaluate User Engagement (Jacques, 1996), Engagement and Influences on Engagement Questionnaires (Webster and Ho, 1997), and two versions of the User Engagement Scale (O'Brien and Toms, 2010a; O'Brien et al., 2018). Comparison of these four questionnaires and associated engagement dimensions are listed in table 2.1.

Table 2.1: Comparison of engagement questionnaires.

Measure (citation)	#Items	Dimensions	Original domain	Environment and studies tested this measure
SEE (Jacques, 1996)	14	Attention Motivation Controls Needs Time Perception Attitude	Multimedia	N/A
Engagement measure (Webster and Ho, 1997)	7	Attention Focus Curiosity Intrinsic Interest Overall.	Multimedia	Multimedia system for presentation (Webster and Ho, 1997); Multimedia system for training (Chapman et al., 1999)
UES (O'Brien and Toms, 2010a)	31	Aesthetic Appeal Focused Attention Novelty Felt Involvement Endurability Perceived Usability	e-Shopping	Multimedia webcast (O'Brien and Toms, 2010b); Facebook (Banhawi and Ali, 2011); Video game (Wiebe et al., 2014); Online news (O'Brien, 2011); Search (O'Brien and Toms, 2013)
UES revision (O'Brien et al., 2018)	31	Aesthetic Appeal Focused Attention Perceived Usability Reward	e-Shopping and book search	Search (Capra et al., 2018)

Systematically constructing associated survey instruments started with the Survey to Evaluate User Engagement (SEE) by Jacques (1996) in the last century. The SEE contains six attributes with 14 questions in total: attention, motivation, controls, needs, time perception, and attitude. As the first of its kind, it provides a starting point for designing multidimensional questionnaire for engagement.

Later, Webster and Ho (1997) introduced a 15-item questionnaire, with 7 items for engagement and 8 items for influence on engagement, to measure the influence of using multimedia systems for presentation on engagement. The influence on engagement part contains four factors, which are challenge, feedback, presenter control, and variety (Webster and Ho, 1997). Engagement was roughly divided into attention focus, curiosity and intrinsic interest, and overall (Webster

and Ho, 1997). Compared with the other questionnaire, Webster and Ho (1997) examined listeners more specifically, who are exposed to presentation systems. The engagement portion of Webster and Ho (1997)'s questionnaire (7 items) was further adapted by Chapman et al. (1999), where participants interact with a computer-based system comparing video, audio and text, and compared the effect of media environment on engagement. These 7 items were used in Webster and Ahuja (2006) to explore the relationship between engagement and perceived disorientation with the multimedia interaction context. Apart from being directly applied as a whole, subsets of items in Webster and Ho (1997) has been used to form measures for other psychological concepts such as cognitive absorption (Agarwal and Karahanna, 2000). This questionnaire has proven relevant in a multitude of scenarios, but it has not been generalized to domains beyond multimedia, which is where it was created and first administered.

The User Engagement Scale (UES) developed by O'Brien and Toms (2010a) builds upon Webster and Ho (1997) and Jacques (1996) and combines more attributes identified in O'Brien and Toms (2008) such as user's future intention of using the service, or the visual appearance of the interface. The UES (O'Brien and Toms, 2010a) is evaluated through Exploratory Factor Analysis (EFA) and Structural Equation Modeling (SEM). The final version of the UES (O'Brien and Toms, 2010a) contained 31 items and six dimensions (O'Brien and Toms, 2010a): Aesthetic Appeal, which is perception of the visual appearance of interface; Felt Involvement, the feelings of being drawn in and entertained in interaction; Focused Attention, the concentration of mental activity, flow and absorption; Novelty, curiosity evoked by content; Perceived Usability, affective and cognitive response to interface/content; and Endurability, which is the overall evaluation of the experience and future intentions. Unlike other engagement measures, the model underpinning the UES shows how Endurability is explained, either directly or indirectly, by the other five dimensions (O'Brien and Toms, 2010a). The UES has been used to evaluate multiple system types, and this original six-dimensional structure that emerged from an online shopping environment (O'Brien and Toms, 2010a) does not always hold. Some studies used a subset of the six dimensions, suggesting that certain items may be more salient in each type of context. For example, in the webcast (O'Brien and Toms, 2010b) study, the entire Felt Involvement dimension was removed, but Endurability and Novelty were still distinct. In another

study of Facebook (Banhawi and Ali, 2011), which also removed Felt Involvement, Novelty and Enduribility merged into one dimension, and resulted in a four-dimensional structure. For studies using all six-factors in gaming (Wiebe et al., 2014), online news (O'Brien, 2011) and online web search (O'Brien and Toms, 2013), Novelty, Felt involvement and Endurability merged into one factor in factor analysis while all other three dimensions remained distinct.

Recently, O'Brien et al. (2018) revised the UES scale using Multidimensional Item Response Theory, and further suggested a four dimensional structure by merging Novelty, Felt involvement and Endurability into one dimension, Reward. Apart from re-examining the two datasets used in the first study (O'Brien and Toms, 2010a), they included a new dataset collected during studies in which people perform a mixture of book search tasks. A short form, USE-SF, is also introduced with 12 items, which requires less response from the user thus is designed to avoid user fatigue, with none from the original Felt Involvement dimension. Capra et al. (2018) reported a four factor structure for the UES-SF using search tasks with various task determinability. O'Brien et al. (2018) suggests that the revised UES with a four-factor structure aims at wider human computer interaction applications rather than only information systems, and it is valid to select a subset of the four dimensions in context-based studies. Although differences have emerged in the various applications (table 2.1), the UES series questionnaire is the most thoroughly tested measure of user engagement on an information system.

2.5.2 Measuring user behaviour

This section first reviews research focused on user behaviour to assess the user experience during information retrieval. We also discuss the approaches used in collecting behaviour data based on Lalmas et al. (2014), and demonstrate that web analytics is a proper method for this study according to its scalability, the task-adapted method of data collection, which allows for our results to be easily replicable and our experiments to be conducted in a lightweight fashion minimizing the user input needed beyond that of their natural interaction. Then, we focus on behavioural features that have been suggested as proxies for user engagement.

Measuring user behaviour during IR

How a user interacts with an IR system is measured typically by a set of low-level user actions during the interaction and selections at the interface. These include search queries (Kotov et al., 2011), clicks (Chuklin and de Rijke, 2016), movement or focus of cursor or eye (Arapakis et al., 2014b), facial gestures (Arapakis et al., 2009), dwell time (Kim et al., 2014), and sequence of user actions (Shah et al., 2015; Ageev et al., 2011). These measures are categorised into groups by criteria such as the type of information they capture (e.g. query-related features, and session-related features (Kotov et al., 2011)), techniques used (e.g. functional magnetic resonance imaging (fMRI) (Moshfeghi et al., 2013), electromyogram (EMG)(Barreda-Àngeles et al., 2015)), and active or passive behaviour (e.g. (Huurdeman et al., 2016)). Lopatovska and Arapakis (2011) summarized and discussed the methods to study human emotions through information behaviour in library and Information Science, IR and HCI, and pointed out that more work is needed to verify prior findings in terms of the correlation between the two. Such studies, geared towards linking behaviour to a certain state of mind are ubiquitous within IR research and our current endeavour is no exception. In our case, user perception of engagement plays the role of the true psychological state under evaluation.

Lalmas et al. (2014) gave a very thorough tour of measuring engagement, in which they group behaviour into two groups according to the methods used to capture or extract them: physiological techniques, and web analytics. Each type of method possesses specific advantages and disadvantages, and thus none of them are optimal and applicable for every scenario. In the following, we discuss these two methods and elaborate on what each of them brings to the analysis (table 2.2).

Physiological techniques, which are widely used in neuroscience and psychological science (Cacioppo et al., 2007), bridge psychological states and physiological responses, and therefore can provide data and information related to psychological processes such as motivation, cognition, emotion, learning and the interaction among these processes. Several advantages, including being more directly connected to users, more objective, and the ability to capture changes over time (Lopatovska and Arapakis, 2011; Lalmas et al., 2014), resulted in their increased popular-

Table 2.2: Comparing methods in collecting behaviour data during IR tasks.

Method	Examples	Advantages	Disadvantages	Example of Studies
Physiological techniques	fMRI, EDA, EMG; Facial expressions; Body movements; Eye movements; Cursor movements.	Continuous measurement; universal interpretation of gestures; potential automatic extraction	Lowers participant mobility and may cause distractions; Assumption of context independence, which is not always the case.	Moshfeghi and Pollick (2018); Barreda-Àngeles et al. (2015); Arapakis et al. (2009); Mauri et al. (2011); Arapakis and Leiva (2016); Buscher et al. (2010); Cole et al. (2013); Lagun et al. (2014)Guo et al. (2012); Diaz et al. (2013); Huang et al. (2012); Buscher et al. (2012); Lagun et al. (2014)
Web analytics	Clicks; Queries; Time.	Covers entire user population; Ideal for large-scale modelling and cross validation due to cost efficiency	Superficial and requires researcher's interpretation	Savenkov et al. (2013); White et al. (2013); Agichtein et al. (2012); Diriye et al. (2012); Teevan et al. (2011); Bennett et al. (2012); White et al. (2013); Odijk et al. (2015); Kim et al. (2014); Mao et al. (2016).

ity in the field of information retrieval recently (e.g., psychophysiological measures (Moshfeghi et al., 2016; Barreda-Àngeles et al., 2015), eye tracking and cursor tracking (Arapakis et al., 2014b; Arapakis and Leiva, 2016))).

Employing collection of psychophysiological measures, including fMRI, electrodermal activity (EDA), as primary example of the techniques outlined above, has garnered a lot of interest in the academic community recently. Moshfeghi et al. (2013) identified three brain regions that register a positive response when presented with the task of processing relevant and non-relevant images of given queries. Moshfeghi et al. (2016) adapted the same approach on testing Question Answering tasks and Question Answering tasks with search functions. Again, differences of the distribution of brain activities appear between scenarios in which the user does not know the answer, which they described as having an information need, and the user actually know the answer, which is having no information need. Moshfeghi and Pollick (2018) continued this work with more finely grained search time periods (e.g., query formulation, relevance judgement), and found that the differences of brain activities appear on the transition rather

than the state. Barreda-Àngeles et al. (2015) found significantly different users' reaction to search engine latency with electrodermal activity (EDA), and electromyography (EMG). The approach is also applied to related concepts of user engagement. In one physiological study, Skin Conductance Level (SC), EEC and EMG were applied on Social Network Systems (Mauri et al., 2011) to compare the experience using Facebook with the relaxed and stressed conditions. It assigns a special focus on the engagement state using the Lang model of emotions (Lang, 1995), showing that using Facebook is a positive, affective state, which was distinct from the two other states (relaxed and stressed). However, it is worth noting that the emotional and sensory processing during search is a complex network of interaction between higher-level thoughts, and therefore cannot be fully captured by one single physiological measurement. Also, due to the effort required from participants (e.g. physical attendance and fitting with equipment), current studies using psychophysiological measures have a limited number of samples. The psychological processes captured depend largely on the research context and interpretation, and thus more research is required in the future.

In addition to psychophysiological measures, several studies confirmed that behaviour signals like eye movements (Arapakis et al., 2014b; Buscher et al., 2010; Cole et al., 2013), mouse movements (Arapakis and Leiva, 2016; Lagun et al., 2014) and facial expressions (Arapakis et al., 2009) are associated with user affective levels. Physiology studies on these signals have built the foundation for their interpretation, and resulted in coding standards for automatic extraction, such as the Facial Action Coding System (FACS) (Ekman and Friesen, 1976; Ekman and Rosenberg, 2005). Arapakis et al. (2009) shows that by adding features extracted from facial expressions, prediction of topic relevance was improved. However, it is worth noting that although such behaviour signals, especially tone and facial expression, are cross-culturally universal, it is assumed that data collected by automatic extraction is noise-free and independent of context (Lopatovska and Arapakis, 2011), which is not always the case. Eye movement has been studied with a special focus on attention, which is an important component of engagement (O'Brien and Toms, 2010a). The relationship between gaze behaviour and attention (Fischer, 1999; Rayner, 2012) has been revealed in various fields outside IR. The IR studies suggest that visual attention devoted to page results depends not only on the quality of page results (e.g.

quality of the ranking list (Buscher et al., 2010), sentimentality and popularity of the news article (Arapakis et al., 2014b)), but also on the quality of other components appearing in the same page and in the previous search experience (Buscher et al., 2010). The insight that eye movements can provide information about how users engage with information systems should be emphasised.

Furthermore, the alignment between eye movement and cursor movement in search became especially topical (e.g., Arapakis et al. (2014b); Hutchins et al. (1985)), due to the former requiring equipment such as an eye tracker or camera, whereas the latter can be collected through technique such as Java Script, which enables a more natural environment for the users, thus providing more possibilities for large scale modelling. The cursor movement here is the number of coordinates, total moving distance, and speed, from which some patterns can be extracted (e.g., scrolling). Although, some other user behaviours, such as clicks, are also cursor related, they can usually be collected through system log files as system responses to certain user actions, and are thus classified into the web analytics type. The gaze - cursor alignment is always discussed as situational. Some works claimed that the link is stronger in the y-axis of the display than x-axis (Rodden et al., 2008). Later, Huang et al. (2012) examined more factors and confirmed the alignment between gaze and cursor movement changes based on user, time, and search task, and grouped five types of cursor interactions: inactive, examining, reading, and action. In the case of engagement, studies were performed in an online news domain (Arapakis et al., 2014c) with cursor features such as speed and acceleration, and found that cursor features had more profound correlations with engagement than does gaze data. The meta-features extracted from cursor movements is further applied to successfully predict users' attention on knowledge module display (Arapakis and Leiva, 2016).

Cursor movement is also complementary to measures collected through web analytics such as clicks, queries, and dwell time, in predicting session success (Guo et al., 2012), estimated user attention on interface components (Diaz et al., 2013), in characterizing SERP examining strategy (Buscher et al., 2012), in predicting document relevance and ranking results (Lagun et al., 2014). But just like other physiological measures, it is vague. Privacy also represents an issue that requires care when applying this method at a large scale, since the implementation

of cursor trackers requires user consent.

The most common method is using web analytics, which is to extract signals from the "digital traces left by users" during the interaction of systems, "often referred to as web logs" (Lalmas et al., 2014) (p.48). The key drawback of using web analytics is that the captured raw data are less in-depth compared to physiological measures such as cursor movement, and therefore relies on researchers to identify and extract measures that may address specific research questions. Regarding its advantages, web analytics is ideal for large-scale modelling and cross validation, as it generally captures information from the entire population of users during the experimental period. Thus, the applications based on behavioural data collected through web analytics range from formal models to prediction studies. Some of the earliest studies into user behaviour fall into this category, and much attention has been paid to the subject in the past. Here, we will only recall only the most relevant concepts.

User behaviour data harvested through web analytics can help to build or shape formal models (Clarke and Wing, 1996; Wing, 1990) that are mathematically based techniques for describing system properties (Wing, 1990) and their role is delineated (Wing, 1990), which enables mathematical applications with a focus on the correlations between the most salient features. One example is the search economic theory (Azzopardi, 2011), an analogy of the Production Theory (Varian et al., 1996) from the field of economics. Azzopardi (2011) conceptualised users gain and cost during search to address how features like query cost and quality of the result list affect user behaviour, and how they lead to the search stop point. However, one imperfection of formal models is that the abstraction of parameters leads to negligence of certain factors, and thus, the formation of formal models typically requires gradual amendments and improvements (e.g.Azzopardi and Zuccon (2015)).

In addition, multiple efforts in IR process studies have attempted to look for patterns from the behaviour data, patterns that may have the capability to predict likelihood of a specific outcome during search such as switching search engines (Savenkov et al., 2013; White et al., 2013), search abandonment (Agichtein et al., 2012; Diriye et al., 2012)), navigation (Teevan et al., 2011), personalisation (Bennett et al., 2012; White et al., 2013), query reformulation

(Odijk et al., 2015), and user satisfaction (Kim et al., 2014) and user perceived usefulness (Mao et al., 2016).

Still, several challenges remain, with a major one being that measures based on behaviour are only descriptive, rather than interpretative, of the information seeking process and outcomes. That is to say, measures based on behaviour alone cannot fully explain the rationale behind user actions, and therefore require researchers background knowledge. Thus, researchers need to not only make assumptions about behaviours and the rationale behind them, which requires understanding of the underlying theories, and current context, but also verify the links (Lopatovska and Arapakis, 2011). That is also why behaviour measures are often used as additional measures to enhance the conclusions, rather than as stand-alone measurements (Charlton and Danforth, 2010). Despite these drawbacks, inference at large scale from behaviour data remains possible, as the data-driven approach lends generalizability to recurring patterns emerging from large scale behaviour analysis. However, not all types of behaviour data can be collected at large scales, due to the complexity of data protection schemes and the permissions required to collect it. In smaller settings, discerning patterns in behaviour while factoring out the many variants of behavioural bias induced by one's experimental framework and problem statement remains challenging, and extracting significant patterns without expert knowledge is difficult to ground theoretically. The following section discusses the behaviour measures that can be extracted from log files, which belong to the web analytic type, that are used in existing studies as engagement proxies.

Measuring user behaviour during IR to access engagement

To date, there is not yet a precise engagement measure based on user behaviour. We have mentioned current physiological measures used as engagement proxies (e.g., (Arapakis and Leiva, 2016)) with high recognition in the previous section (section 2.5.2). User behaviour measures collected through web analytics are used as engagement proxies as positive experience of interaction is expected to result in user actions such as positive feedback, returning users, and recommendation to other users (O'Brien and Toms, 2008). Lehmann et al. (2012) described

such assumptions as "the higher and the more frequent the usage, the more engaged the user" (p.165). The way that users engage with information systems depends on the type of systems (Lehmann et al., 2012), and therefore these behavioural measures captured in these ways cover a wide range of features from number of unique users, click rates, page views, to dwell time. For example, for an online news system, in-depth interactions such as reading until the end of a news article and leaving a comment are indicative of the fact that users are engaged, while for shopping websites, user return rate plays that role.

It is impossible, and as far as research is concerned, practically infeasible to list and address each context individually, developing specific behaviour measures for each. It is thus necessary to either bring the problem down to a tractable scale or categorize them in a more generalized manner. A simple way of grouping behaviour measures is by their measurement scope, be it intra-session, which refers to the user engagement within a single session, and inter-session, which refers to long term engagement. Another line of separation is the engagement dimension represented; for instance, Lehmann et al. (2012) examined the website visitors' behaviour through three perspectives, which are referred to as "engagement metrics": Popularity, Activity and Loyalty (Peterson and Carrabis, 2008), and looked into the behaviour patterns across 80 websites. Popularity refers to how frequently a website is visited; Activity of a website is how users browse or use the system; while Loyalty refers to the frequency of returning users. It is worth noting that these three metrics are not necessarily positively correlated. For instance, a website for a special one-time event will be high in popularity and activity in the short term, but may not have many returning users, hence low loyalty value. Therefore, depending on a website's nature, the engagement model may vary. The three engagement metrics mentioned by Lehmann et al. (2012) were also examined in other studies, focusing more on activity and loyalty metrics. Typically, number of distinct users, visits (Lehmann et al., 2012), and comments after a news article (Arapakis et al., 2014a) represent popularity; number of queries, page views (Donato et al., 2010), click ratio (Arapakis et al., 2014a), dwell time (Bateman et al., 2012), search engine abundance (Juan and Chang, 2005), hovers (Bota et al., 2016), and absence time (Dupret and Lalmas, 2013) stand for activity; and frequency of access per user represents Loyalty (Song et al., 2013). Recently, Drutsa et al. (2015a) applied the discrete

Fourier transform with four activity and loyalty measures: the number of sessions, presence time, number of queries, and number of clicks, in order to capture periodicity in user engagement and represent it as a frequency spectrum of its sequential metrics. The dimensions and examples of engagement measures based on behaviour are listed in table 2.3.

Table 2.3: Examples of user behaviour as engagement proxies.

Dimension	Description	Example of Measures	Example of Studies
Popularity	How much a website is used	Number of distinct users; Number of visits; Comments left after news articles	Lehmann et al. (2012); Arapakis et al. (2014a); Williams et al. (2016)
Activity	How a user use the system	Number of queries; Pages views; Click through rate; Dwell time	Donato et al. (2010); Arapakis et al. (2014a); Dupret and Lalmas (2013); Bateman et al. (2012); Juan and Chang (2005,?); Bota et al. (2016)
Loyalty	How often users return to the system	Frequency of usage per user; User return rate	Song et al. (2013); Drutsa et al. (2015a)

One big challenge of modelling engagement from the behaviour signals is the diversity of user engagement(Lalmas et al., 2014; Lehmann et al., 2012). That is to say, users engage with the website in different ways, rather than in a uniform manner. This inherent intricacy also supports the idea that engagement is hard to capture through formulaic measures or a single low-level feature extracted from the behaviour signal.

To reinforce the above, as an example with respect to dwell time, Kotov et al. (2011) refers to deep engagement as viewing search results with dwell time more than 30 seconds. 30 seconds has been used as a threshold for satisfaction (Fox et al., 2005), but still depends on context, as search tasks do not always require such a long time (White and Kelly, 2006). Thus, a single threshold value, or one measure alone, is not universally applicable. This can be mitigated by various techniques, such as adding more informative measures, either from the context or user, such as reading difficulty of clicked pages, search topics, and query types (Kim et al., 2014). Furthermore, different types of website have various engagement models, which are initially identified by Lehmann et al. (2012), a fact which again complicates the ability of single behavioural features to generalize. Therefore, discussion and interpretation of data based on

the information retrieval context is necessary, and more studies covering the variety of search environments are needed to augment our understanding of the precise relationship between user behaviour and engagement.

2.6 Verifying user behaviour with user perception

To assess user experience, measures based on either user perception or behaviour have certain drawbacks, challenges, and trade-offs. Briefly, approaches based on perception are collected directly from users, which can provide more insights that fully represent what users perceive of the search experience, but are obtrusive to users' natural interaction, and are difficult to collect over time. On the other hand, approaches based on user behaviour provide the potential to capture not only scalable but also dynamic data, which would enable powerful in-situ mathematical modelling that may lead to wide applications (e.g., real time search assistant function), but fall short in the understanding of user rationale. Thus, neither group of measures, solely, are sufficient for assessing user engagement during the process.

For a robust measurement, we need both, as outlined by Lalmas et al. (2014), the "optimal situation is when we can use subjective and objective measures in concert and with confidence that they corroborate our findings." (p.82). We have established in section 2.4 how user engagement lies at the intersection between the user's perception and their behaviour. To focus on the nature of engagement one must bridge this gap. Could we possibly uncover and describe these relationships, using cost-effective behavioural data to precisely represent perceived feelings, in order to assess user engagement and thus the quality of user experience? In fact, researchers have already started working towards this direction (e.g., (Al-Maskari and Sanderson, 2010; Arapakis et al., 2014c)). Although some studies have combined user perception and user behaviour variables into one measure to assess user experience, or verified the relationships between some of these variables, there were inadequate attempts to examine the interactions between the two in IR, or specificity about user engagement. Therefore progress in this area is considered preliminary. This section first describes the studies that mapped user behaviour

to user experience other than engagement, in the broader domain of IR (section 2.6.1), then specifically with respect to engagement in the wider field (section 2.6.2).

2.6.1 Verifying user behaviour with user perception in IR

In most studies based on behaviour measures (e.g. Kotov et al. (2011)), authors outline links between behaviour and perception (such as the longer users stay on a result, the more satisfied they feel). But do these links really exist? Does one behaviour represent just one perception? For example, when the user stays long on a result page, does he feel frustrated or interested in the result? Are these links systematic enough to explain the perception of interest? Limited studies in IR to date have tended to analyse the combined elements of user behaviour measures and perception measures in order to integrate their respective advantages and circumvent the disadvantages: to evaluate search in a cost-efficient, explicit and scalable manner (Lalmas et al., 2014).

Some studies collected user perception directly from the participants, which are called the *first party labels*. One early study that integrates user behaviour with user perception into one measure is Tague-Sutcliffes informativeness measure (Tague, 1987; Tague-Sutcliffe, 1995) that assesses the performance of the system simultaneously with the perception of the user. But this is atypical and, due to the effort (e.g., constant user feedback) required in implementation, is rarely used (Freund and Toms, 2007). Other attempts were mainly applied in satisfaction. Al-Maskari and Sanderson (2010) found significant associations between user satisfaction and user effectiveness as measured by the number of relevant documents found by user and the time they spent to locate the first relevant document, and user effort as measured by the number of queries that a user submitted. Other studies suggest that as time increases, satisfaction decreases (Su, 2003; Law et al., 2006), as the time spent on a search task is considered as user effort (Lancaster, 1981). However, this correlation was not always found significant (Hersh et al., 2000). Based on various correlation findings between satisfaction and behaviour measures such as dwell time (Liu et al., 2011; White and Kelly, 2006), last click in the query (Chapelle and Zhang, 2009), and query position and query formulation types (Odijk et al., 2015), Mao et al.

(2016) provide a predictive model to generate usefulness label for search results automatically using Gradient Boosting Regression Tree (GBRT) (Friedman, 2001). Such label represents click level satisfaction, and was found to be moderately correlated with feedbacks from users (Mao et al., 2016). Although these findings were only on the satisfaction of one search result, while engagement is built up throughout the overall search process, the results of Mao et al. (2016) provide an optimistic potential for modelling user engagement using user behaviour. In music discovery, Garcia-Gathright et al. (2018) showed that together with users goals, user behaviour measures such as average usage, and peak interactions on individual tracks, are informative for predicting user satisfaction.

Other studies (e.g., (Machmouchi et al., 2017; Mehrotra et al., 2017; Williams and Zitouni, 2017)) collected perception data using 3rd party judgements (e.g., crowd sourcing) rather than directly from the user, and such labels are referred to as the *third party labels*. Those studies usually contain a relatively larger number of participants (e.g., thousands of user sessions) and thus lend themselves to more complicated mathematical modelling. Machmouchi et al. (2017) mapped user behaviour into satisfaction and suggested one single measure, the "utility metric", learnt from a linear model, as the quality of session. A higher step in complexity brings us towards studies where either large sets of features or neural network learning structures were adapted. Williams and Zitouni (2017) construct a Recurrent Neural Network (RNN) to model the sequence of user interaction to predict positive/negative abandonment. Abandonment - the practice of issuing a query not followed by any clicks on the subsequent SERP - is traditionally interpreted as a negative signal of the user's disinterest or disengagement. However, facilitated by the functionality of modern search interfaces, abandonment can mean a result was displayed clearly through the SERP (via snippets, suggested answers etc.) to the extent that the user *abandons* the SERP positively. The fact that behaviour can be used to model this distinction gives us an insight both into user engagement in general and into the unique properties of this signal in particular. The caveat present in these techniques in general is that acquiring labels from third-party annotators is usually only a soft reference to user perception. In an ideal scenario, the reliable way to access this information is collecting the label directly from users.

2.6.2 Verifying user behaviour with user perception of engagement

User engagement studies have considered the data collected from user feedback and their behaviour. As mentioned before, Arapakis et al. (2014c) used three psychometric scales: Focused Attention sub scale from the UES (O'Brien and Toms, 2010a), the Positive and Negative Affect Scale (PANAS) (Watson et al., 1988) on a 5 point Likert-type scale, to assess within-content engagement in online news reading website on users perceived interest of the news article. They demonstrated mouse tracking as a scalable, cost-effective alternative to eye tracking in predicting engagement (Arapakis et al., 2014b), and also that certain cursor patterns indicated negative user experience, with a general yet profound correlation with engagement. They (Arapakis and Leiva, 2016) looked into user interaction with direct displays and simplified the self-reported measures of engagement into only three questions to cover attention, usefulness, and perceived task duration, and found that the prediction based on cursor features outperformed existing classifiers based on click and hoverover behaviour. In addition to this, most of the relationships between user behaviour and user perception of engagement were identified using correlation analysis (e.g., (Thomas et al., 2016)), which is insufficient to infer complex relationships and therefore cannot provide insight into the broader information retrieval scenario.

2.7 Summary

User perception of engagement and user behaviour are examined in order to assess the quality of user engagement in IR. However, neither of them individually is optimal for measuring user engagement. User perception measures can provide insights that describe a user's experience, as perception measures obtain feedback directly from users. Such in-depth information is not encoded in the behaviour signal alone.

However, collecting user perception of engagement is obtrusive to users' natural interaction. In addition, due to dependence on context and individual experience, substantial time and effort are required in designing user perception experiments and collecting and analysing the data. On the other hand, as objective, quantitative data, measures based on user behaviour have

the potential to capture scalable data over time without interrupting the users, and efficiently provide quantitative data for mathematical modelling and further applications. Nonetheless, with the major disadvantage of providing rather implicit information, such measures require substantial background knowledge and insights from researchers to formulate hypotheses and provide meaningful interpretations of the data. Although the importance of verifying user behaviour with user perception of engagement is evident, limited work was conducted in the context of information retrieval.

This disparity outlines an underlying gulf in our understanding of these variables and the implications of their interaction in practical contexts. Our objectives aim to capitalize on the following research goals, formulated as a result of this gap in understanding.

1. At present, the state of the art is unable to theoretically reconcile the conceptual models of HCI and IS in which user perception plays a major role and behaviour is merely treated as an abstract actuator of the interaction with the empirical studies in IR in which user perception is under-emphasized, and low-level behavioural signals dominate.

2. Additionally, not enough research is currently able to validate the models of users' perception of engagement as a multi-dimensional psychosomatic response to the interaction with the signals provided by users' behaviour, and to measure this relationship reliably, or in a way which captures the underlying complexity of engagement.

Our study aims to address these gaps and shore the shortcomings in our understanding of engagement as a whole. For this research, we posit that user behaviour is an indicator of user perception, and that evaluating user engagement in information retrieval should draw insights from both behaviour and perception measures. Therefore, in this study, we aim to bridge the understanding between user behaviour and user perception of engagement by examining both types of outcomes, and ultimately to create a cost-effective predictive model for user engagement using established relationships between user behaviour and user perception of engagement.

We pursue the resolution of this hypothesis based on the insights provided by the models of user interaction (Norman's Model in figure 2.2, Kuhlthau's ISP in figure 2.1, Marchionini's ISP in

figure 2.3, Bate's berrypicking model in figure 2.5), and employing the methodology detailed in 3. In particular, our focus will be on behaviour *as a process*, and on how this process translates into the user's perception of engagement. We wish to illustrate the advantages of this method over the static session-wide aggregates for interaction features, so we will explore both methods. We also want to shift some of the focus from assessing the user perception of engagement to being able to predict it from behaviour directly. The next chapter will be devoted to detailing the experimental design, methodology and data used in our studies.

In the introduction (chapter 1) we announced several objectives for the body of this research. Here, in retrospect of our review of the current state of the art in models and methodology for studying similar problems, we can connect these objectives to gaps in the current state of the IR domain knowledge.

Obj.1 refers to identifying and validating the role of behavioural features in inferring user perception of engagement. This sets the stage for the rest of our analysis by considering the primal source of information about the user's behaviour - the static behaviour features (which have previously been suggested as proxies of engagement), and addresses the issue of devising correlations between perception of engagement and behaviour in the most general setting available to us.

Obj.2 represents the materialization of our goal to incorporate the time-dependency encoded in a user's action sequence in order to better capture the above relationship.

Obj.3 aims to address the problem of engineering the behavioural diversity captured above into behavioural measures of user perceived engagement in an easily computable on-line fashion. We aim to further extend the state of the art in measure design by contributing new patterns and showing how the researcher's insights can be engineered into these measures.

Together, these three objectives break down the problem of addressing the relation between user perception and user behaviour in the setting of IR. A detailed analysis and formulation of these objectives is the object of the next chapter. There we also discuss our fourth objective, *Obj.4*, which is more technical in nature, but aims to provide statistical relevance to our previous

methods by diversifying and multiplying the perspectives through which this relation can be assessed.

Additionally, throughout our research we reap an extra benefit, which is the ability to contrast our theories and findings between the two major information seeking contexts, searching and browsing. *Obj.5* addresses the issue of understanding the relation between user perceived engagement and user behaviour in both contexts, assessing how the two differ with respect to this relationship and categorizing the differences according to the results of our analysis. Thus, this objective, together with *Obj.4* can be viewed as overarching goals of our analysis and we shall be referring back to them in all our subsequent chapters.

Chapter 3

Research Design

3.1 Introduction

The purpose of our research is to identify and statistically confirm relationships between user behaviour and user perception of engagement. This general objective is broken down into 5 smaller objectives to which individual quantitative studies can be allocated.

Obj.1: To identify and validate the role of behavioural features in inferring user perception of engagement.

Obj.2: To identify and model user behaviour sequences that have a significant association with user perception of engagement.

Obj.3: To implement and evaluate measures of engagement.

Obj.4: To compare behaviour - perception relationship among four different engagement dimensions.

Obj.5: To compare behaviour - perception relationship in browsing and searching.

The fourth and fifth objectives represent overarching objectives which guide most of our analysis throughout this thesis. To the first three objectives we devote three separate phases of our research as explained further in this chapter. Each objective is approached via a certain set of research questions. A detailed description of the architecture of our study can be found in figure 3.4. In total, three research phases are designed to address ten research questions. Two datasets, which were collected in previous studies and reflect two types of information retrieval contexts, browsing and searching, were chosen to be used in the research. In the remainder of

this chapter, we present the research philosophy (section 3.2), and the overview of the research design (section 3.3), then we describe the variables used in this study (section 3.4). Finally, we introduce the individual research designs (section 3.5) and the datasets used (section 3.6).

3.2 Research Philosophy

This study, as much of the quantitative research undertaken during the last century is supported by the epistemological principles of positivism (Bernard and Bernard, 2012). Positivist methodology, itself founded on empiricism, aims to explore and explain quantitative relationships in the large. Two major research frameworks have been derived from its principles, namely *correlational analysis* and the *experimental approach* (Bernard and Bernard, 2012). The latter pertains to scenarios in which the researcher (experimenter) has direct access to manipulating the dependent and independent variables in order to draw conclusions about their observed relationship - and usually such conclusions revolve around the causal nature of the interaction between variables. In this sense, the experimental approach is a holistic one, since it emphasizes the need to manipulate parameters and observe results globally. Correlational analysis, on the other hand, is appropriate when the researcher has access to a particular slice of the true parameter space of his variables, and is unable for natural reasons to manipulate these objects themselves. In general, when dealing with user interaction, we find ourselves more often in this paradigm rather than the former. Explaining causality relationships is not our aim in this text. Due to the nature of this study, we surmise that correlational analysis is the only appropriate framework in which to conduct our research. We cannot factor our individual preference and bias when collecting user perception of engagement, and we are less concerned with the cause-effect relationship between user behaviour and user perception of engagement, due to the complex network of psychological factors we have investigated in chapter 2. Our interest lies in the development of an explanation for the covariance of these two variables, and ultimately deducing what the knowledge of one can help infer about the other in a causality agnostic setting.

3.3 Overview of Research Design

In order to unfold the relationships between user behaviour and user perception of engagement, this study employs the correlational research design (Bernard and Bernard, 2012), a framework in which the statistical relationship between a set of two or more target variables is tested through analytical means.

This framework naturally carries some limitations - primarily its inability to distinguish causal relationships - however, we consider it an appropriate fit to the problem and would like to expound this argument currently. Primarily, we would like to draw attention to the nature of our variables of interest, namely user behaviour, in the large, but also in how much it relates to their interaction with the system, and, user perception of engagement, which we endeavour to fit into a rigorous framework for quantitative study. The challenge in performing any sort of statistical deduction or hypothesis testing is that both of our target variables are a function (at least partially) of many intangible extraneous variables, most relevant of which is the user's *true engagement* which we treat only as a heuristic, and avoid mentioning it past the current discussion. A number of other potential extraneous variables may be suggested here, all of which have been adequately tackled in the user psychology and HCI literature, namely *mood* and *individual differences*. It is challenging, or otherwise purely impractical to delve into the study of engagement by factoring in all the possible extraneous variables. Therefore, in this study, we hypothesize there is correlation between user behaviour and user perception of engagement without mention of a causal relationship between the two. In reality there is a potentially more complex conglomerate of factors at work behind the scenes, which accounts for the behaviour of these variables in what we shall see further.

Additionally, in the hope of cementing this argument, we point out that our data is, in essence, *static*, in that it cannot be altered once collected and there is no way for the experimenter to influence desired behaviours to consistently test certain patterns in the data. This will become apparent during the in-depth look at our datasets and how they are collected, later in this chapter. We point out here that, firstly, the data is subject to very little encoding, and is suitable for our analysis immediately after collection. During the body of our research,

we aggregate the data in various ways and extract many features to suit our testing needs, but the focus is always on interpretability, as we hope our conclusions will serve as actionable information for researchers to extend our methods and add their insights to ours. Secondly, in the effort or reducing or eliminating the observer-expectancy effect, which can prove detrimental in any user-centric study, our data is collected as unobtrusively as possible, and are thus resigned to the state of the world that this data describes unquestionably. We are careful to make no more assumptions than is possible given this information snapshot, and are conservative in drawing conclusions.

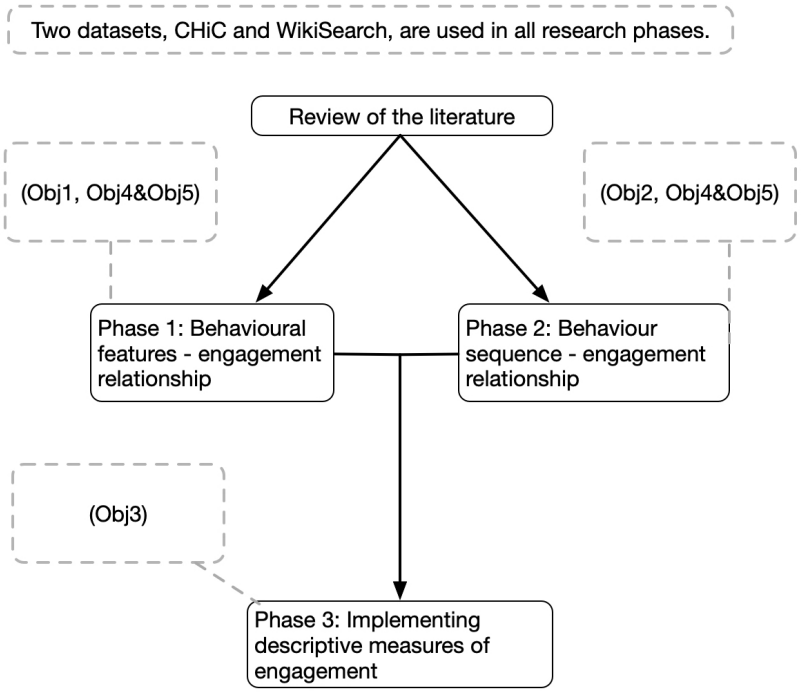


Figure 3.1: Overview of the research design.

This project contains three research phases (figure 3.1) guided by the current state of the literature, starting from identifying the main behavioural features that describe user perception of engagement in phase 1. In the second phase, we examine these relationships at the behaviour sequence level. Based on several properties suggested by the uncovered relationships from phases 1 and 2, the final phase focuses on implementing measures of engagement by proposing five new feature sets. Each phase includes two or three separate studies, replicated using two selected datasets, CHiC (browsing) and wikiSearch (searching). The two datasets are described in more detail in section 3.6.

In the following part of this chapter, we first describe the variables used in this study. This is followed by a description of each research phase, including the research questions addressed and the methods employed. More detailed descriptions of the methods used, and variables extracted in the different phases of the study are described in their respective chapters. We then describe the two selected datasets, including the systems, participants, tasks and procedure.

3.4 Variables

This section describes the variables used to assess user perception of engagement and how the variables used to capture user behaviour are extracted from behaviour logs in this study.

3.4.1 Variables of user perception of engagement

Following the literature review in chapter 2, we used a subset of the UES questionnaire (O'Brien and Toms, 2010a), which originally contained 31 items and six sub-scales, to assess user perception of engagement. Each item is presented as a statement using a 5 point scale or a 7 point scale ranging from "strong disagree" to "strong agree". The original six sub-scales represent different dimensions of engagement, namely, Aesthetics Appeal, Perceived Usability, Focused Attention, Endurability, Felt Involvement, and Novelty.

Figure 3.2 describes the data, and variables used to assess user perception of engagement. In this case, the data of user perception of engagement represent numerical values which describe a user's state of engagement through direct user response to the UES questionnaire after the task. In this text, the UES sub-scales are employed to represent the variables of user perception of engagement for which we keep the original denomination of *dimensions*, respecting the intuition that they capture different facets of the user's engagement.

We selected the UES questionnaire because it has been tested for validity and reliability (e.g., using Principal axis factor analysis) in multiple contexts as discussed in the prior research chapter (section 2.5.1), such as webcast (O'Brien and Toms, 2010b), Facebook (Banhawi and

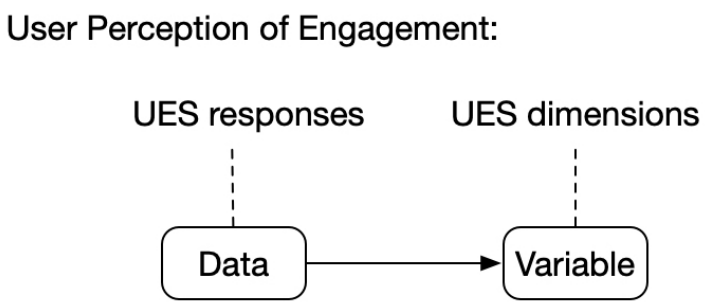


Figure 3.2: Data, and variables used to access user perception of engagement.

Ali, 2011), and more specifically in search (O'Brien and Toms, 2013). It is, to date, one of the most tested questionnaire that measures user engagement. As discussed in section 2.5.1, different structures emerge in past studies and the author also suggested certain dimension(s) will be more relevant than others considering the context (O'Brien and Toms, 2010a). We reflect on this point in the limitations section 8.2 of the conclusion.

Table 3.1: Description of selected four sub-scales in UES questionnaire (O'Brien and Toms, 2010a)

UES Sub-scale	**Description**
Novelty	Curiosity evoked by content.
Felt Involvement	Feelings of being drawn in and entertained in interaction.
Endurability	Overall evaluation of the experience and future intentions.
Perceived Usability	Affective and cognitive response to interface/content.

We extracted a subset of the original UES items (O'Brien and Toms, 2010a) which represent four sub-scales, Novelty, Felt Involvement, Endurability, and Perceived Usability. Descriptions of the four selected sub-scales are present in Table 3.1. The *Novelty (NO)* sub-scale contains 3 items, measuring curiosity evoked during the interaction, indicating that the system or experience contained surprising, unexpected, or new information at various points in time (O'Brien and Toms, 2010a). We select NO because it is known to be connected to curiosity which serves as a driving factor of browsing (Rice et al., 2001), which is one of the two major information retrieval context. *Felt Involvement (FI)* contains 3 items, measuring how well the experience with systems can satisfy users' needs, and thus reflecting feelings of being drawn in and entertained during the interaction. *Endurability (EU)* contains 5 items, measuring whether users perceived the interaction as successful, rewarding, or worthwhile (O'Brien and Toms, 2010a). Overall, this

Table 3.2: Items in the four engagement dimensions (O'Brien and Toms, 2010a).

Dimension	**Item**
Novelty (NO)	1, I continued to use this website out of curiosity. 2, The content of the website incited my curiosity. 3, I felt interested in this experience.
Felt Involvement (FI)	1, I was really drawn into this experience. 2, I felt involved in this experience. 3, This experience was fun.
Endurability (EN)	1, Using this website was worthwhile. 2, I consider my experience a success. 3, This experience did not work out the way I had planned. 4, My experience was rewarding. 5, I would recommend this website to my family and friends.
Perceived Usability (PUs)	1, I felt frustrated while using this website. 2, I found this website confusing to use. 3, I felt annoyed while using this website. 4, I felt discouraged while using this website. 5, Using this website was mentally taxing. 6, This experience was demanding. 7, I felt in control of while using this website. 8, I could not do some of the things I needed to do while using this website.

Reverse coding is applied to the following items: PU-1, PU-2, PU-3, PU-4, PU-5, PU-6, PU-8, and EN-3.

sub-scale is the general evaluation of the experience and future intentions, and we keep it as the total score of engagement. The *Perceived Usability (PUs)* sub-scale contains 8 items, measuring the challenges users face when interacting with the system, and whether the user could conduct the task using the system the way they wanted to. Overall, these items assessed users' perceived effort in performing the required tasks by the system and reflected the users' affective and cognitive response to the system (O'Brien and Toms, 2010a). The reason for selecting PUs is that usability has been a major measure for user experience (Kelly, 2009) in information retrieval, and this sub-scale appears to be stable in both UES structures (O'Brien and Toms, 2010a; O'Brien et al., 2018). Each sub-scale represents one dimension of user perception of engagement. In order to differentiate it from the six sub-scales, we refer to the four selected sub-scales as four dimensions of user perception of engagement. Table 3.2 displays the items used in each dimension.

To assign a single score for each dimension, the items within one dimension were averaged as recommended in O'Brien et al. (2018). Although the two datasets used different scales for

the UES (7-point scale for wikiSearch (browsing) and 5-point scale for CHiC (searching)), we preserve the user group labels by normalizing our data related to the median of the score of each engagement dimension and divided the users in to two groups: *high* and *low* engagement. We refer to the group labels as engagement labels. Assigning binary classes provides a clear and simple overview of the label distribution and choosing the median as the threshold makes the positive / negative distribution unbiased, which allows more freedom in the choice of evaluation metrics, makes our baselines more relevant, and in general provides a task that is more fairly designed for the purposes of binary classification.

3.4.2 User Behavioural Variables

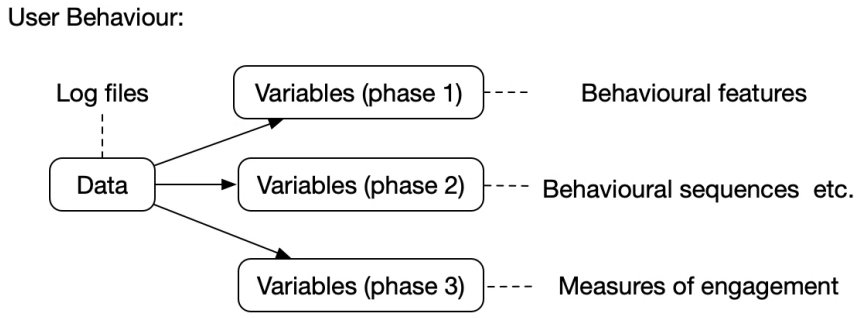


Figure 3.3: Data, and variables used to access user behaviour.

Figure 3.3 describes the data, and variables used to assess user behaviour. We used system log files as the data source of user behaviour, which is collected through the web analytics technique (discussion in section 2.5.2). System log files contain the users' actions (e.g., mouse clicks, scrolling, keystrokes, and associated parameters) issued while interacting with the system with associated timestamps. An *action* is any physical input that the user generates and passes to the system and can be recorded via the use of system logs throughout the duration of the task. Table 3.3 shows an example of the pre-processed log files from the CHiC dataset. We know from the example that the user (ID: 01) submitted a query 'castle' to the system right after the session started, and then clicked one document from the search result page (SERP) at the 23.33 second. The URL of the document was recorded in the associated parameter field. After scrolling down the document, the user saved the URL to the bookbag at the 108.4

second. The main benefit of using log files lies in reducing the interruption of a user's natural interaction with the system, and thus provides a situation closer to the real world task.

Table 3.3: An example of the log files.

User ID	Time	Action Type	Parameter
01	0	Start	
01	19.86	Query	castle
01	23.33	Click_SERP	http://...
01	69.63	Scroll_down	
01	108.40	Add_Bookbag	http://...
...	...	...	...

Each phase contains different sets of variables as different user behaviour information is being focused on in each phase. Following are descriptions for each phase and a few representative examples.

For phase 1, we extracted *behavioural features* based on existing studies from log files to suit our research needs via aggregation and selection. As illustrated in figure 3.3, behavioural features are variables and are directly calculated from log files. Back to the example in table 3.3, the user (ID: 01) performed four actions, query, Click_SERP, Scroll_down, Add_Bookbag. We say behavioural features are *bags of actions*, which are engineered without the information of the order of interaction. Examples of behavioural features and their values based on the logs files are present in table 3.4. More detailed descriptions of the behavioural features used and the steps to extract them are provided in section 4.2.1 in phase 1.

Table 3.4: An example of behavioural features extracted from the log files in table 3.3.

Behavioural feature	Value
Number of actions	4
Number of pages viewed	1
Time to click the first search result	3.47
Number of pages added into the bookbag	1

For phase 2, we design an analysis that contains several steps to extract variables from the log files automatically. Details of this procedure are given in chapter 5, while here we take a closer look at the variable itself, the so-called *behaviour sequence*. Behaviour sequences focus on the order of user interaction. An example of such sequences based on the log files is

`Query-Click_SERP-Scroll_down-Add_Bookbag` (table 3.3). The process we designed aims to extract small fractions from the behaviour sequence, namely *subsequences* that are useful to describe user perception of engagement. An example of such subsequence based on the log files is `Query-Click_SERP` (table 3.3). All the formal definitions of behaviour sequence and related concepts are in the definition in section 5.2, and the method to extract them is in section 5.3 in chapter 5.

In phase 3, we engineered more variables, namely *measures of engagement*, based on findings from phases 1 and 2. As we have not presented the findings yet, the rationale and details of those variables are described in section 6.2 in phase 3.

3.5 Research Phase Design

In order to address the five research objectives posed in section 1.2 and figure 3.1, we designed three research phases to address a total of 10 research questions. Figure 3.4 illustrates the relationships between the five general objectives, how each motivates several questions and how, together they compound into the design of our research phases. Tables 3.5, 3.6, 3.7 present the design for phases 1, 2, and 3 separately, including the analysis used, detailed research objectives for each analyses and research questions addressed in each study.

3.5.1 Phase 1: Behavioural Features - Engagement Relationship

This phase was designed to extract behavioural features that are correlated to the four engagement dimension scores and test engagement prediction using these features. Table 3.5 presents the research design of this phase. Four research questions are answered:

RQ.1 What are the behavioural features used in previous studies to describe user perception of engagement?

RQ.2 To what extent can individual features predict user perception of engagement?

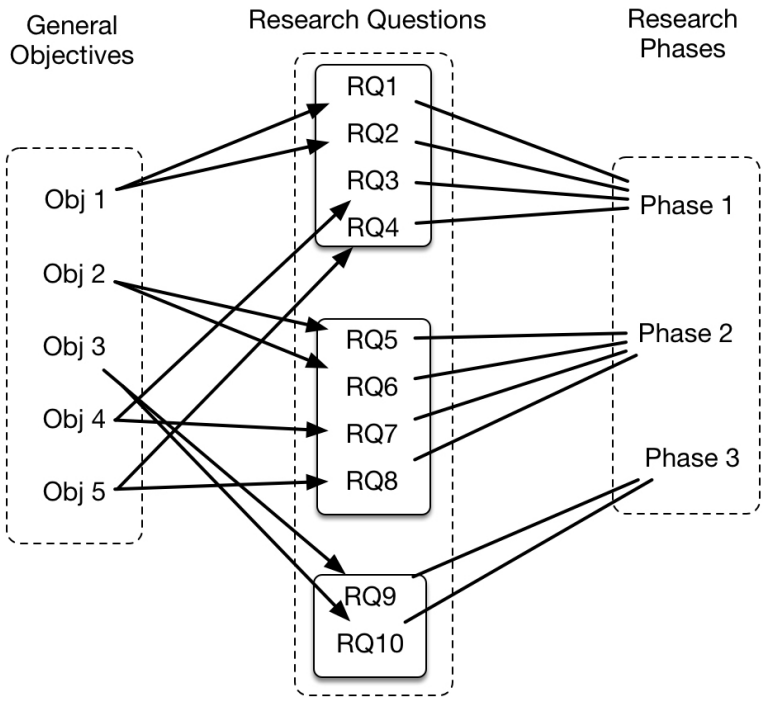


Figure 3.4: The relationships between general objectives, research questions and research phases.

RQ.3 How do the relationships between behavioural features and user perception of engagement vary between dimensions?

RQ.4 How do the relationships between behavioural features and user perception of engagement differ between browsing and searching?

This phase is divided into three studies:

Study A. Behavioural feature selection. Previous information retrieval studies (e.g., (Song et al., 2013; Dupret and Lalmas, 2013; Drutsa et al., 2015a; Lehmann et al., 2012)) have suggested behavioural features as proxies of engagement - highly correlated variables that are used to represent the target variables in a quantitative way (Dodge, 2006). To validate the assumed relationship, we used a forward and backward snowballing approach to summarize the types of behavioural features used in the existing literature. We then select a set of features to test via feature importance analysis (study B) and engagement prediction (study C) and extract them from the system log files for each user. This aims to answer *RQ1*.

Study B. Feature importance analysis. By *importance* here we mean a score which mea-

Phase 1: Behavioural Features - Engagement Relationship (Chapter Four)			
Study	**Objectives**	**Research Questions**	**Analysis**
A. Behavioural feature selection	Identify the user behaviour used in previous studies as engagement proxies.	RQ1: What are the behavioural features used in previous studies to describe user perception of engagement?	Comparison and manual classification
B. Feature importance analysis	Determine the extend of contribution of individual behavioural feature in inferring engagement prediction.	RQ2: To what extent can individual features predict user perception of engagement? RQ3: How do the relationships between behavioural features and user perception of engagement vary between dimensions?	Descriptive and analytical statistics, feature selection analysis
C. Engagement prediction using behavioural features	Validate the use of behavioural features in predicting user perception of engagement.	RQ2: To what extent can individual features predict user perception of engagement? RQ3: How do the relationships between behavioural features and user perception of engagement vary between dimensions?	Classification analysis
RQ4 (How do the relationships between behavioural features and user perception of engagement differ between browsing and searching?) is addressed by comparing results from study A, B, and C between datasets.			

Table 3.5: Diagram of research design of research phase 1.

sures the drop in predictive accuracy when out-of-bag samples of a feature are randomized. In order to examine the predictive power of individual behavioural features with respect to each UES dimension (*RQ2*), we average users' score across each engagement dimension. We then report the descriptive statistics of each dimension and the correlation analysis with Spearman Rank Coefficient, and divide users into two groups by the median of each of the resulting engagement dimension scores - *high* engagement and *low* engagement. We refer *high* and *low* as engagement labels. Regarding behavioural features extracted from study A, we report the cross-correlation within the set of behavioural features. We then perform a Pearson point-biserial correlation hypothesis test between behavioural features and the distribution of users across each engagement group. We conduct feature selection through machine learning, using Mean Decrease in Accuracy (MDA) of a Random Forest model (Breiman, 2001) in order to rank the behavioural features with respect to the engagement labels (detailed description in section 4.3.1). Comparison of the top 10 behaviour measures ordered by MDA among different

engagement dimensions is discussed (*RQ3*).

Study C. Engagement prediction using behavioural features. To test how the behavioural features predict user perception of engagement dimensions, a binary classification problem was defined: predicting the *high* and *low* engagement labels through behavioural features extracted from study A. Four classifiers, a baseline of the majority label, two Random Forest models and a Support Vector Machine, were implemented to predict each of the four UES dimensions. We evaluate the performance of the four classifiers according to precision, recall, F-Measure, accuracy, and area under curve (AUC). A k-fold cross-validation paired t-test (Dietterich, 1998) was used to assert statistically significant improvements of each classifier over the baseline. Comparison of the performance among different engagement dimensions is discussed (*RQ3*).

All three studies were conducted using both datasets, and differences between results using CHiC (browsing) and wikiSearch (searching) were discussed to answer *RQ4*. Evidence of a statistically significant correlation between user behavioural features and user perception of engagement in this domain emerged from this analysis.

3.5.2 Phase 2 : Behaviour Sequences - Engagement Relationship

This phase examines the sequence of user behaviour rather than discrete behavioural features tested in phase 1. Table 3.6 presents the research design of this phase. In this phase, we addressed four research questions:

RQ.5 What is the most general set of actions which suffices to describe user interaction with information retrieval systems?

RQ.6 What is the relationship between user behaviour sequences and user perception of engagement?

RQ.7 How do the relationships between user behaviour sequences and user perception of engagement vary between dimensions?

RQ.8 How do the relationships between user behaviour sequences and user perception of engagement differ between browsing and searching?

Phase 2: Behaviour Sequence-Engagement Relationship (Chapter Five)			
Study	**Objectives**	**Research Questions**	**Analysis**
A. Behaviour sequence extraction	Identify the user actions from ISP models and common IR system interface components.	RQ5: What is the most general set of actions which suffices to describe user interaction with information retrieval systems?	Comparison and manual classification
B. Sequence analysis	Determine the extend of behaviour sequential patterns that have a significant association with user perception of engagement.	RQ6: What is the relationship between user behaviour sequences and user perception of engagement? RQ7:How do the relationships between user behaviour sequences and user perception of engagement vary between dimensions?	Descriptive and analytical statistics, frequent subsequence extraction, χ^2 hypothesis test of independence
C. Engagement prediction using sequences	Validate the use of behavioural sequential patterns in predicting user perception of engagement.	RQ6: What is the relationship between user behaviour sequences and user perception of engagement? RQ7: How do the relationships between user behaviour sequences and user perception of engagement vary between dimensions?	Classification analysis
RQ8 (How do the relationships between user behaviour sequences and user perception of engagement differ between browsing and searching?) is addressed by comparing results from study A, B, and C between datasets.			

Table 3.6: Diagram of research design of phase 2.

This phase was divided into three studies.

Study A. Behaviour sequence extraction. *RQ5* is addressed by presenting the criteria employed in selecting actions that best describe the information seeking process. We further ground our selection based on the ISP model (Marchionini, 1995) and common IR system interfaces. Behaviour sequences were formed by first extracting actions from the system log files and then concatenating consecutive equivalent actions. A detailed justification of the selection criteria is in section 5.3.

Study B. Behaviour sequence analysis. We use the engagement labels assigned in phase

1, study B and the behaviour sequences extracted in phase 2, study A in order to examine the behaviour sequences with the four engagement dimensions (*RQ6*). Descriptive and analytical statistics of the behaviour sequences, including user action distribution and behaviour sequence length, were first presented. We then use a sliding window method (Bishop, 2006) to extract *frequent subsequences* (definition in section 5.2) in order to identify the common patterns across all users. To identify the most discriminating patterns in terms of user perception of engagement labels, a χ^2 hypothesis test of independence was conducted between the engagement labels and the presence of each frequent subsequence. The frequent subsequences that have a significant χ^2 score are then called *discriminative subsequences*. Comparisons of the identified discriminative subsequences between different dimensions are presented in order to answer *RQ7*. A detailed description of the method study B is presented in section 5.4.1.

Study C. Engagement prediction using behaviour sequences. In this section we tested whether the added sequential information inherent to the user's behaviour provides any improvement over the use of discrete behavioural features. Sequences entail a certain degree of aggregation, but otherwise we are evaluating the *bag of* versus the *ordered set of* methods of feature extraction. To test how the identified discriminative subsequences (phase 2, study B) can predict engagement, a binary classification problem in which the goal is to predict if a given participant will perceive *high* or *low* engagement was defined. We train two Support Vector Machine models with different sets of features based on discriminative subsequences extracted from phase 2, study B - one set consisting of only the sequential features, and a second comprised of the discrete behavioural features together with the sequential features. The performances of these predictive models were compared to the best classifiers tested in phase 1, study C. We conclude with a k-fold t-test (Dietterich, 1998) of statistical significance in the performance of classifiers obtained.

The analysis was conducted in both wikiSearch (searching) and CHiC (browsing) datasets, and the comparison of the results was discussed to answer *RQ8*. The results were used to identify a set of essential behavioural patterns that can describe user perception of engagement.

3.5.3 Phase 3 : Implementing measures of engagement

The third phase focuses on designing measures that best represent the Endurability (EN) dimension by incorporating the relationships between it and user behaviour identified in the previous two phases. We selected the Endurability (EN) dimension as the target because this dimension represents the final evaluation of the session and is the overall score of the user experience (table 3.2). The fitness of this set is evaluated by learning frameworks, namely linear and SVM regression, in order to assess how much of the variance in EN can be explained by the users behaviour via these designed measures. Table 3.7 present the research design of this phase. We present our outcomes as answers to the following two research questions:

RQ.9 What are the properties that empirically computable measures of engagement should possess?

RQ.10 Which of the developed measures improve user perception of engagement prediction?

Phase 3: Implementing measures of engagement (Chapter Six)			
Study	**Objectives**	**Research Questions**	**Analysis**
A. Measure development	Identify properties to represent the behaviour - perception relationships and develop measures based on the properties.	RQ9: What are the properties that empirically computable measures of engagement should possess?	Manually classification and Comparison
B. Evaluation of developed measures	Evaluate the use of implementing developed measures in predicting user perception of engagement.	RQ10: Which of the developed measures improve user perception of engagement prediction?	Regression analysis.

Table 3.7: Diagram of research design of phase 3.

This phase can be divided into two studies:

Study A. Measure development. In this study, we engineer, out of our pre-acquired knowledge of the interaction between user perception of engagement and user behaviour, a set of features that directly correlate with engagement, so that they act as *measures*. We developed a set of five measures of engagement which build upon and augment a set of two baselines, one

of which is the utility metric developed in Machmouchi et al. (2017), the other being simply the mean. This process is guided by a theoretical set of six properties that we derive from phases 1 and 2 and argue for in detail (*RQ9*). Not all of these properties are achievable under all measures, the intricacy of the task leaving much room for exploration. The measures are based on time which is motivated by the findings from phase 1, and designed via two overarching principles: to be extensions of the utility metric (Machmouchi et al., 2017), by either considering different types of actions, weighting the time contributions differently or applying different summarization techniques to the action space, or to be simple functions of the user's action sequence motivated by the findings from phase 2.

Study B. Evaluation of developed measures. This study is concerned with validating our previous intuition. To what extent was our knowledge and the properties derived thereafter accurate? Since this study concerns itself with quantitative measures of engagement, we frame a regression problem, predicting the Endurability dimension using each measure developed in phase 3, study A. We extract our measures from the behavioural features in phase 1, study A and the behaviour sequences presented in phase 2, study A. Strong positive correlation of our variables motivates the use of a linear model, which would, in effect produce a set of weights that express engagement as a weighted average of our features. We employ an elastic net here, due to its robustness and essentially free feature selection method which will allow us to draw conclusions on our data. We compare it with a less interpretable but very standard learning model, the SVM. The performances of these predictive models were compared to answer *RQ10*.

3.6 Datasets

In order to address the research questions above, two datasets collected in previous studies were chosen to be used in the research based on four criteria:

1. The dataset should contain sufficient number of participants and activities so that the analysis will be reliably valid.
2. The dataset should contain both user perception of engagement data (e.g., UES scale) and user

behaviour data (e.g., rich log files).

3. The dataset should cover either browsing or searching.
4. The dataset should be available for re-use according to research ethics regulations.

The two datasets which met the criteria are:

1. *CHiC*: collected in the CLEF 2013 Cultural Heritage Track with non-purposeful browsing task (Hall et al., 2013).
2. *wikiSearch*: collected using the wikiSearch system with goal-based searching tasks (Toms et al., 2013).

As mentioned in section 1.4, the two datasets were collected at two different times (see details in (Toms et al., 2013; Hall et al., 2013)) and so two different ethics applications were made before the collection. As part of the informed consent process, participants gave permission for their data to be reused in future research. Table 3.8 displays the key characteristics of these two datasets.

Table 3.8: Description of CHiC and wikiSearch datasets.

	CHiC	**wikiSearch**
Information object	Image	Text
Task type	Browsing	Searching
# Participants	180	447
# Valid participants	157	377
Average duration per participant	8 minutes 34 seconds	23 minutes 57 seconds
# behaviour logs	15,396	85,857
Collection method	In-lab and Online	In-lab

The two datasets were chosen to comply with the fundamental requirements for significant statistical analysis and also reflect two types of information retrieval, searching and browsing. They achieve a high and dense enough statistical sample in both of the information retrieval contexts we deal with in this study (e.g, they both contain a sufficient number of participants and each of them performed a large number of interaction). Both user perception of engagement data and user behaviour log files were collected in the two datasets. Moreover, as we plan to investigate how engagement can be described by user behaviour in the two online information retrieval contexts, searching and browsing, we select one dataset for each respectively. The

differences between the task types corresponding to these two datasets allow for replications and investigation under different settings, which will contribute to our understanding of generalisability of modelling engagement across information retrieval types. A brief description for each dataset is provided as follows. Also, a controlled study with a well-defined setting would help us to examine only the named context. The experimental conditions of collecting the two datasets were controlled, via regulated experiment time and detailed instructions. Especially, the wikiSearch study only contained in-lab participants, and the CHiC study contained both in-lab and online participants. Another frequent question concerning the data collected from human participants is regarding the sample size, as we want to make inferences of the population. An adequate number of participants would lead to precise modelling (e.g., by the law of large numbers). Thus, we select the datasets with relatively large numbers of participants compared to other available user studies in the information retrieval field.

3.6.1 Dataset 1: CHiC (browsing)

The study (Hall et al., 2013) that collected the CHiC dataset focused on collecting and analysing interactive information retrieval (IIR) behaviour in a Digital Cultural Heritage collection. The purpose of the study was to develop a dataset describing undirected exploration and browsing in such a collection and understand how users interact with the system.

Application System

The CHiC dataset was collected in the CLEF 2013 Cultural Heritage Track (Petras et al., 2013). The system, an image Explorer based on Apache Lucene 2.2 [1], contained about one million records from the Europeana Digital Library English language collection. The Explorer was accessed using a custom-developed interface (see figure 8 in (Hall et al., 2013)), adapted from the wikiSearch study (Toms et al., 2013) mentioned previously. The interface (figure 3.5) displays potential interests in a single window, and was also divided into three main parts by 'column': 1) task description and menus of the broad topic, which is the hierarchy, 2) search

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

box and SERP, 3) details of the clicked document and bookbag. Participants had a full visual display of all options available, enabling the sort of visual searching and priming that browsing requires, which fits Bates (2007)'s notion of the idea of a natural browsing system. The interface allows the user to explore the collection in three ways: 1) clicking a topic in the hierarchical category browser, 2) submitting terms in the search box, and 3) click an item in the metadata filter based on the Dublin core ontology [2], in which the labels were modified for better user understanding. Using one of the three access methods, participants searched or browsed the content, added interesting items into a bookbag, and at the same time provided information about why the object was added using a pop up box. The items added into the bookbag are meant to be interesting to the participants.

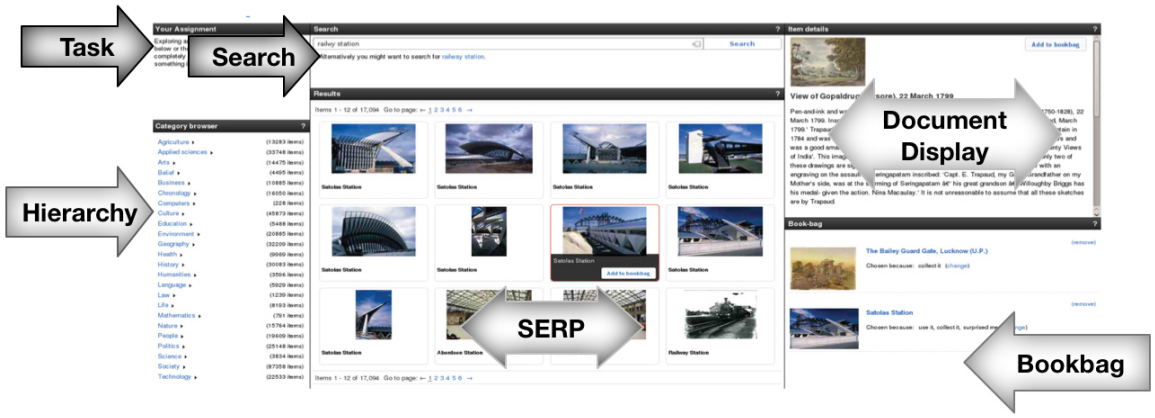


Figure 3.5: CHiC interface.

Participants

180 participants were recruited via a volunteers mailing list, with 160 on-line participants and 20 in-lab participants. Twenty-three participants did not engage with the task as expected (e.g., missing response to the questionnaire) and their data was disregarded. Of the remaining 157 participants, 68.8% were under 35 years old (N=60, 18-25 years old; N=48, 26-35 years old). Ninty-five had completed at least a Bachelor degree. The sample had an unbalanced gender distribution (N=110, 70% female). Eighty participants (51%) were students, and 70 were also employed (44.6%). 145 participants used English to search the web in daily life.

[2] http://dublincore.org/documents/dcmi-type-vocabulary/

Task

The study used one non-purposeful task. Participants first read the scenario: "Imagine you are waiting to meet a friend in a coffee shop or pub or the airport or your office. While waiting, you come across this website and explore it looking at anything that you find interesting, or engaging, or relevant..." The next display is the system interface in figure 3.5, which presents the browse task with no explicit goals in the upper left corner: "Your Assignment: explore anything you wish using the Categories below or the Search box to the right until you are completely and utterly bored. When you find something interesting, add it to the Book-bag."

Procedure

Figure 3.6 details the procedure of the study. The application system is embedded in a web-based experimental system described in (Hall and Toms, 2013), which guided the participants (both in-lab and online) through the process from consent to end, without any interaction with researchers. The only difference between the two participant groups is that in-lab participants were interviewed after task completion, but the interview transcripts are not included in this analysis. It started with an explanation of the experiment, acquired informed consent, and asked for a basic demographic profile and questions about European culture before presenting the CHiC Explorer and the task to participants. System log files were recorded from the beginning when the participant started the browsing task. Once participants had executed the task, indicated they were finished, they moved on to the UES questionnaire (O'Brien and Toms, 2010a) and also other post-study questionnaires about their perceptions of the search experience and the interface. They also provided a brief explanation of objects in the bookbag, the selected metadata and the interface.

Data

From the data collected, we extracted system log files with timestamps and UES questionnaire which contained the user perception of engagement data. The stages in which these two types

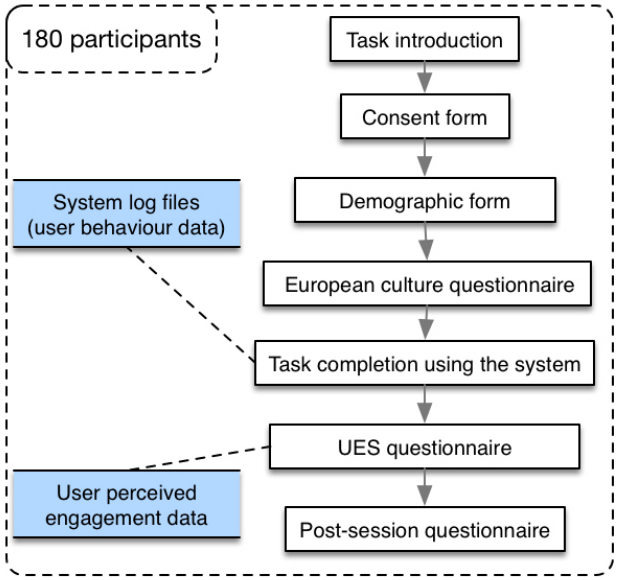


Figure 3.6: CHiC study procedure. Blue blocks indicate the stage that the data was collected.

of data are collected are labelled in figure 3.6. Overall, the data used in this study consists of the system log files, UES questionnaire data, demographic information. The average duration of the task and average number of actions issued per participant are summarized in table 3.8.

3.6.2 Dataset 2: wikiSearch (searching)

The purpose of the study (Toms et al., 2013) that collected the wikiSearch dataset is to understand how people search for information online and make decisions related to the information found.

Application System

The wikiSearch system also used Lucene, an open source search engine, contained information extracted from Wikipedia, as well as a custom-developed interface (figure 3.7) designed to support the search work flow. The interface provided a bird's eye view of the tasks by using a single display panel that brought items to the surface, leaving the interface structure constant. It was divided into three main parts by column: task, search and document display. The task column contained a task box providing participants with instructions for the current task. Below

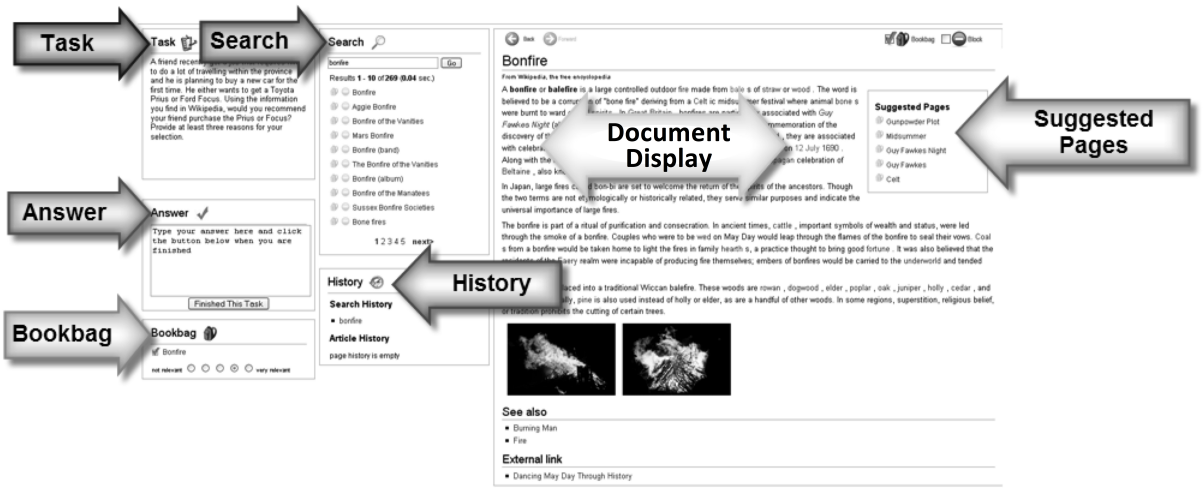


Figure 3.7: wikiSearch interface (Fig 1 in (Toms et al., 2013). Reprint with permission).

the task box, there was an answer box for participants to input their answer in respect to the task shown above, and a bookbag function to store documents that participants think they will use to respond to the tasks. The search column contained a search box for submitting queries, while a history box contained previous searches and previous results that were viewed. For the document display column, the suggested pages provided links to a list of other documents that are relevant to the current document displayed next to it. The page display box had a scrollable article selected by participants through search results, history, bookbag or the suggested pages box. Using one of the four methods to access documents, participants searched through the content, added related items into a bookbag, and rated the documents' relevance on a five-point scale. The items added into the bookbag are meant to be used to answer the task questions.

Participants

447 users participated in the study, and all of them were recruited via a volunteers' mailing list. But after assessing the data, the number was reduced to 377, eliminating pilot participants, those with partial or incomplete data, and those who experienced technical problems. Most of the participants (N=281, 74.5%) were 18-24 years old. The sample had a balanced gender distribution, with 52.5% male participants. The majority of participants (N=327) were students, but 13.2% were also employed in some capacity. 90.7% out of 377 participants used

search engines at least once daily, and most were frequent users of Wikipedia. Almost 73% of participants used Wikipedia at least once per week (N=201; 53.3%) or daily (N=73; 19.4%).

Tasks

In total, there were 12 decision-making tasks (see task details in (Toms et al., 2013)) that asked the participants to choose between two options. The tasks were designed to cover a range of topics, and each of which presented two options to choose from based either holistically or using a set of pre-ordained criteria. Each task followed a similar pattern that started from a brief background, which is followed by two options. The participants were asked to make a decision between the two options using the information found in Wikipedia. Each participant was assigned three tasks from different topics. Tasks were randomly assigned but the topics were counter-balanced across the participant group.

Procedure

Figure 3.8 shows the procedure of the study, in which the data was collected. Before starting the task, each participant was assigned 3 of the 12 designed tasks with a unique identification number in order to track their search activities. Under laboratory settings, the wikiSearch system embedded in WiIRE (Toms et al., 2004) guided participants through the procedure as a series of web pages, from introduction of the study, consent form and demographics questionnaire, tutorial of the system, task assignment and completion. System log files were recorded from the beginning when participants started the task. After they completed the three tasks, they moved on to the User Engagement Scale (UES) (O'Brien and Toms, 2010a) questionnaire and other post-session questionnaires. Details of the UES and the reasons for using it are provided in the later section 3.4.1, and examples of system log files are provided in section 3.4.2. Ethical assessment was done before collecting the data.

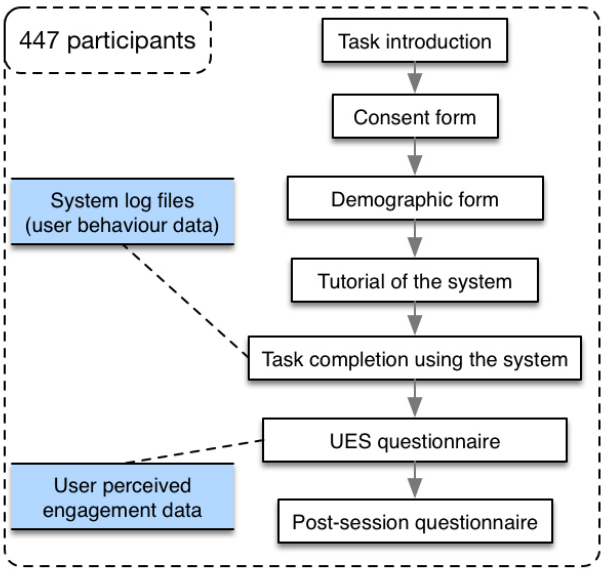


Figure 3.8: wikiSearch study procedure. Blue blocks indicate the stage that the data was collected.

Data

From the data collected, we extracted system log files with timestamps and UES questionnaire which contained the user perception of engagement data. The stages in which these two types of data are collected are labelled in figure 3.8. The data we received is robustly anonymised, thus the original participants cannot be identified. In total, we acquire the system log files, UES questionnaire responses, demographical, cultural background and experience for all participants from the wikiSearch study. The average duration of the task and average number of actions issued per participant are summarized in table 3.8.

3.7 Summary

A correlational research design based on a positivist philosophy was used to analyse and model the relationships between user behaviour and user perception of engagement in information retrieval. The research is split into three phases in order to answer ten research questions. We select the measures of user perception of engagement and briefly introduce the system log files as the source of user behaviour and how we extracted the variables for each phase.

Phase 1 contains three studies and investigates the role of discrete behavioural features in inferring user perception of engagement and the differences between the behavioural features - engagement relationships of four engagement dimensions in both browsing and searching. Phase 2 contains three studies and investigates the potential added benefits of exploring the sequential relationship between user actions with respect to user perception of engagement with statistical hypothesis testing. Phase 3 takes the information gained from the previous two phases and investigates how one might develop context specific measures of engagement through more advanced feature engineering. Two datasets were chosen for analysis based on their sample size, collected measurements, context representation and the availability according to research ethics regulation.

Chapter 4

Phase 1: Behavioural Features - Engagement Relationship

4.1 Overview

This chapter reports the first phase of this research, in which we investigated how individual behavioural features are linked with different dimensions of user perception of engagement. The phase was designed to answer four research questions:

RQ.1 What are the behavioural features used in previous studies to describe user perception of engagement?

RQ.2 To what extent can individual features predict user perception of engagement?

RQ.3 How do the relationships between behavioural features and user perception of engagement vary between dimensions?

RQ.4 How do the relationships between behavioural features and user perception of engagement differ between browsing and searching?

In chapter 2 we discussed how over the past decades, the research meant to evaluate and characterize user engagement in information retrieval based on behavioural features, has largely

been centred around domain specific studies motivated by practical necessities. In the following chapter we start the beginning of an investigation into the interaction of these elusive concepts from a statistically grounded perspective.

Very little work has been presently undertaken on concurrently testing the proposed link between behavioural features and user perception of engagement. Moreover, limited empirical studies were conducted on differentiating the dimensions of engagement by how user behaviour reflects them. As a by-product of resolving our research questions we aim to address these problems concomitantly.

This phase is comprised of three studies:

Study A. Behavioural feature selection, in which we focus on extracting a comprehensive set of features from system log files based on existing studies. We selected the behavioural features that were assumed to be a proxy of user perception of engagement in previous information retrieval studies, and gave the reasons why these features were chosen (section 4.2.2). (*RQ.1*)

Study B. Feature importance analysis, in which we compute the mean decrease in accuracy of our features with respect to a Random Forest model and rank them with respect to the four engagement dimensions (see definition in section 3.4.1). (*RQ.2*, *RQ.3*)

Study C. Engagement prediction using behavioural features, in which we formulate a classification problem for the engagement labels of user groups and test the performance of several learning models in solving this problem using the behavioural features. (*RQ.3*)

Finally we compare the three sets of results on the two datasets corresponding to our two information retrieval contexts, browsing and searching, respectively, and outline how the differences in our results reflect the inherent disparities between the contexts themselves. (*RQ.4*)

A pictorial summary of the above description can be found in figure 4.1.

The measures, approaches, and findings are presented below in separate sections, one for each study. The analyses were conducted using both datasets. The final analysis compares results

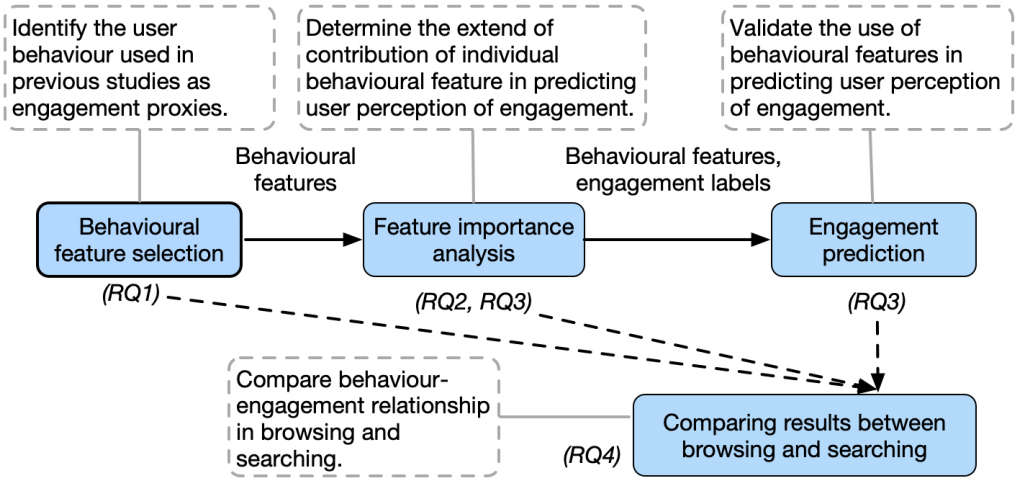


Figure 4.1: Design of research phase 1, which contains four main parts: behavioural feature selection (section 4.2), feature importance analysis (section 4.3), engagement prediction (section 4.4) and results comparison (section 4.5).

from both datasets, resulting in a thorough discussion of the differences between the two settings. Outcomes from this phase lead to the identification of the key behavioural features for each engagement dimension, as well as evidence to suggest properties of an ideal measure of engagement based on behavioural features.

4.2 Study A. Behavioural features selection

The current section is devoted to the first of our three studies in this phase. The purpose of this study is to identify the behavioural features used in previous studies as engagement proxies. We start by identifying the categories of behavioural features that are used in existing literature, then extract the selected features from log files in order to prepare data for the next analysis in this phase. Figure 4.2 presents the steps, data and variables used in this section. The methods and data used in this study are described in section 4.2.1. The results of this study, which are the behavioural features and their categories, are presented in section 4.2.2.

We restate below the research question associated to this study:

RQ.1 What are the behavioural features used in previous studies to describe user perception of engagement?

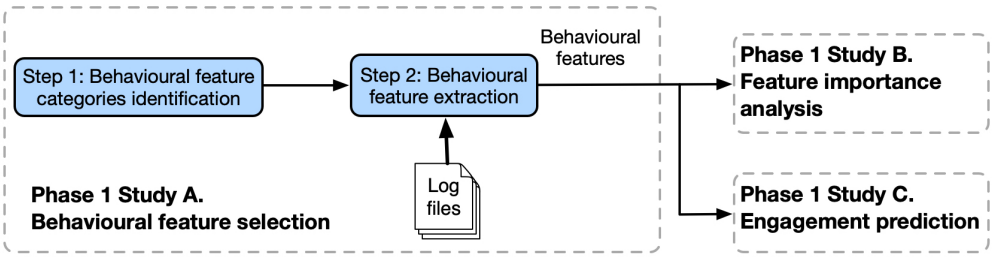


Figure 4.2: Analysis steps, data and variables used in section 4.2.

4.2.1 Method

As outlined in chapter 2, a user's interaction with the system produces a retrievable signature in the system's log files. While some information loss occurs against the real-world set of actions, *behavioural features* nevertheless encode much of the information contained in the process. All the analyses in this study were written in R[1].

In this section, and throughout the text, we will denote by *behavioural features* this numerical footprint of the user's interaction as observed within system log files. They are collected independently, usually by cumulation (averages or counts) and disregard the order in which the actions took place.

Behavioural feature categories identification (step 1):

Previous information retrieval studies (e.g., (Song et al., 2013; Dupret and Lalmas, 2013; Drutsa et al., 2015a; Lehmann et al., 2012)) have suggested behavioural features as proxies of engagement. To validate the assumed relationships, we identified the types of behavioural features used in the existing literature and selected a set of features to test in our study. This fits into the study of *RQ.1*. The review of relevant studies is conducted following the stages suggested in Bryman (2016).

1: "Define the purpose and scope of the review.": the purpose of reviewing the behavioural features used as engagement proxies in the IR literature was to identify the associated categories and behavioural features as candidates for describing user perception of engagement. The

[1]https://www.r-project.org

previous relevant studies were reviewed so that the candidates could be based on what are generally considered by the field as good user behaviour representations of user perception of engagement. For these reasons only studies with empirical evaluation with sufficient number of participants were reviewed, and the scope of the review was limited to key studies.

2: "Seek out studies relevant to the scope and purpose of the review.":

The same searching engine, database, and key publisher websites used in section 2.2 are also used here. The forward and backward snowballing approach (Wohlin, 2014) is used to search for the related works, as we started from a review in information seeking behaviour authored by Case and Given (2016). In total, 257 studies were selected at this stage.

3: "Assess the relevance of each study for the research question(s)."

Studies were considered relevant if user behavioural features as engagement proxies were described and empirical evaluation of implying such proxies was performed. In total, 42 studies were selected at this stage.

4: "Appraise the quality of studies from Step 3."

Only studies influential in the field were selected. Studies should be in published in journals or top conferences or be cited by others. In total, 21 studies were selected at this stage.

5: "Extract the results of each study and synthesise the results."

We first identify the behaviour categories from these studies. This provides a thematic grouping of our features according to 4 general types of interactions. It furthermore allows us to draw conclusions about the qualities of the categories themselves when evaluating the results of our correlational analysis. Lastly it acts as a form of summarization, reducing the dimension of our feature space, when we construct measures of engagement.

Behavioural feature extraction (step 2):

After identifying the categories of behavioural features, we further selected behavioural features from the existing studies and tailor one set of features for browsing and searching each, that are available for common information retrieval systems, and extracted them from system log files. The main benefit of using log files lies in less interruption of the user's natural interaction with the system, and thus allows the user interaction to proceed unfettered. This provides a context closer to the real world task.

4.2.2 Behavioural features and categories

Suggested features cover a wide range of interactions and mainly map into four categories, namely click-related features, query-related features, result-related features, and time-related features (table 4.1).

Table 4.1: Categories of behaviorual features used as engagement proxies.

Feature category	**Examples**	**Example of studies**
Click-related	Number of cursor clicks on interface components; ratio of certain type of clicks	Ponnuswami et al. (2011); Lehmann et al. (2012); Song et al. (2013); Kim et al. (2013); Drutsa et al. (2015a,b); Yamamoto et al. (2016); Teo et al. (2016); Ma et al. (2016); Wu et al. (2017); Bai et al. (2017)
Query-related	Length of query submitted; number of queries	Dupret and Lalmas (2013); Song et al. (2013); Dave et al. (2013); Barreda-Àngeles et al. (2015); Drutsa et al. (2015b)
Result-related	Number of pages viewed; depth of pages viewed.	Yamamoto et al. (2016); Teo et al. (2016); Ma et al. (2016); Li et al. (2016)
Time-related	Time spent on viewing pages; time spent on issuing queries.	Lehmann et al. (2012); Song et al. (2013); Kim et al. (2013); Drutsa et al. (2015b); Yamamoto et al. (2016); Teo et al. (2016); Dave et al. (2013); Wu et al. (2017); Budylin et al. (2018)

Click- and query- related behaviours are most commonly used. Considering the breadth of literature in information retrieval, it is almost standard to include these features (e.g., Click-Through-Rate (CTR)) in describing user engagement during or between the information re-

trieval sessions, such as website popularity (Lehmann et al., 2012), user growth in social media (Yamamoto et al., 2016), the periodicity of the user engagement in general search (Drutsa et al., 2015a), the effect of search relevance on user engagement (Song et al., 2013), and also disengagement, such as fatigue (Ma et al., 2016). Time-related features have also been studied widely, either to infer users' loyalty between sessions (e.g., absence time (Dupret and Lalmas, 2013; Budylin et al., 2018)) or to infer users satisfaction (e.g., time spent on page content (Dave et al., 2013)). A popular assumption used in previous studies (Song et al., 2013; Kiseleva et al., 2016a; White and Dumais, 2009) is that the longer the user spent on the session, the more engaged or satisfied the user feels. In the context of information retrieval, time spent has often been interpreted as effort on the part of the user. However, the allocation strategy of this resource on different actions has not been sufficiently examined. As a simple example, a long time spent on finding a document without success may indicate a negative experience, while a long time spent on reading a document may suggest a positive one. In this phase, we measure time-related features based on different interface components which are contained in common interfaces. In addition, we select features associated with document pages and search engine result pages (SERPs), which are the direct retrieval outcomes from the system. Features associated with those pages were used to infer users' perception of usefulness of the query results (Mao et al., 2016) and detect fake social engagement (Li et al., 2016).

Moreover, in order to tailor the analysis to fit the two different information retrieval contexts - browsing and searching, different sets of behavioural features were selected from the four categories for each dataset. In total, 43 behavioural features were extracted for CHiC, and 34 behavioural features were extracted for wikiSearch. Some features related to more than one category, for example, clicks on document pages is both a click-related and result-related feature. Such features are put into the non result-related category. As all the features related to SERPs belongs to more than one category, only features related to document pages such as the page length viewed by users are put into the result-related category.

We also labelled behavioural features that are likely to be associated with "satisfied" or "dissatisfied" feelings because satisfaction is a widely studied perception concept associated with user behaviour, and is also covered by the definition of engagement (see discussion of these two

Table 4.2: Description of 44 behavioural features used for browsing, and 34 behavioural features used for searching.

Click Feature	**Feature description**	**Browsing**	**Searching**
NumClicks	Number of clicks	✓	✓
NumClicksOnSERPs	Number of clicks on SERPs	✓	✓
AveNumClicksPerSERP	Average number of clicks per SERP	✓	✓
NumClicksOnPages	Total number of clicks on pages	✓	✓
AveNumClicksPerPage	Average number of clicks per page	✓	✓
Ratio⟨*sat*/*all*⟩Clicks	Satisfied clicks: all clicks ratio	✓	✓
Ratio⟨*dissat*/*all*⟩Clicks	Dissatisfied clicks: all clicks ratio	✓	✓
NumClicksPerQuery	Number of clicks per query	✓	✓
AveCTR	Average Click Through Rate	✓	✓
NumClicksSug.	Number of clicks on system suggested contents	✓	-
NumClicksOnSERPsSug.	Number of clicks on SERPs from system suggested contents	✓	-
AveNumClicksPerSERPSug.	Average number of clicks per SERP from system suggested contents	✓	-
NumClicksOnPagesSug.	Number of clicks on pages from system suggested contents	✓	-
AveNumClicksPerPageSug.	Average number of clicks per page from system suggested contents	✓	-
Query Feature	**Feature description**		
NumQuery	Number of queries	✓	✓
NumUniqueQuery	Number of unique queries	✓	✓
Rati⟨*unique*/*all*⟩Query	Unique:all query ratio	✓	✓
LengthQuery	Total number of words in query	✓	✓
AveLengthQuery	Average number of words in query	✓	✓
AveIntervalQuery	Average time interval between two queries	✓	✓
MaxIntervalQuery	Maximum time interval between two queries	✓	✓
MinIntervalQuery	Minimum time interval between two queries	✓	✓
NumQueryWithClicks	Number of queries leading to clicks	✓	✓
NumQueryNoClick	Number of queries without leading to any clicks	✓	✓
Result Feature	**Feature description**		
NumPages	Number of pages viewed	✓	✓
NumUniquePages	Number of unique pages viewed	✓	✓
Ratio⟨*unique*/*all*⟩Pages	Unique:all pages ratio	✓	✓
NumPagesPerQuery	Number of pages viewed per query	✓	✓
NumUniquePagesPerQuery	Number of unique pages viewed per query	✓	✓
NumPagesSug.	Number of pages viewed from system suggested contents	✓	-
Time Feature	**Feature description**		
TimeOnTask	Total time on this task	✓	✓
LogTimeOnTask	Logarithm of total time on this task	✓	✓
TimeOnSERPs	Total time on SERPs	✓	✓
AveTimePerSERP	Average time per SERP	✓	✓
TimeOnPages	Total time on pages	✓	✓
AveTimePerPage	Average time per page	✓	✓
TimeOn⟨*sat*⟩Pages	Total time on satisfied result pages	✓	✓
AveTimeOn⟨*sat*⟩Page	Average time on satisfied result page	✓	✓
TimeOn⟨*dissat*⟩Pages	Total time on dissatisfied result pages	✓	✓
AveTimeOn⟨*dissat*⟩Page	Average time on dissatisfied result page	✓	✓
TimeOnSug.	Total time on query from system suggested contents	✓	-
TimeOnSERPSug.	Total time on SERPs from system suggested contents	✓	-
TimeOnPagesSug.	Total time on pages from system suggested contents	✓	-

All features are extracted from system log files.
✓means the corresponding behavioural features is used for this domain.

concepts in section 2.4). We catagorised several features, including clicks and time spent on certain contents, into satisfied and dissatisfied. A popular way to define "likely to be satisfied" actions is users exceeding a threshold of dwell time, usually 30 seconds, while "likely to be dissatisfied" refers to actions with a dwell time less than 15 seconds (Fox et al., 2005). The threshold might not be constant due to various influential factors (e.g., topics, content length) (Kim et al., 2014). Still, it was widely adapted (Song et al., 2013; Kiseleva et al., 2016a; White and Dumais, 2009) to reduce noise of the correlations between actions and satisfaction.

After sorting the system log files into participant groups containing participant id, time stamp, action type and parameter, behavioural features (table 4.2) for browsing and searching were extracted. Following is a brief description of the behavioural features based on the four categories:

Click: These features describe click behaviour. Click is a term used to describe the action of pressing a mouse button one or more times, which is often associated with an interface component (e.g., click on a document page). Click behaviour refers to how the action of clicks is interspersed within the user's interaction with the system. Here, we divided the clicking activities into the ones on search engine result pages (SERPs) and on results pages. A click is said to be satisfied if the leading SERP/Page was attended for more than 30 seconds; and dissatisfied if the leading SERP/Page was attended for less than 15 seconds. The Click Through Rate of the task (AveCTR) is the average CTR of all queries issued during the task. CTR for a query is 1 if there are one or more clicks leading by the query (0 otherwise). Apart from the 9 features used in both browsing and searching, 5 features about clicks on suggested content (e.g., clicks on built-in hierarchy, and clicks on metadata link.) were selected for browsing only. There are 14 features for browsing, and 9 for searching.

Query: A query is the action of entering text into a field or option used to locate information within a database or another location. When not being precise, we may also refer to a query as the text itself. These 10 features describe query behaviour for browsing and searching. Some users in our dataset re-visited previous query results, and we include features such as the number of unique queries to capture this redundancy. Query interval is time lag between two

consecutive user queries. Some of the features overlapped with click features mentioned above, which was also reflected in the correlation value between behavioural features in section 4.3.

Result: These features are related to retrieved document pages viewed by user, not the SERPs. NumPagesPerQuery is the number of pages viewed per query, and is also defined as the length of query sub trails (Yuan and White, 2012). We put time spent on document pages related measure into the Time category, as time was proved to be crucial in measuring both intra- and inter- session engagement (Dave et al., 2013; Dupret and Lalmas, 2013; Lehmann et al., 2012), and we prefer to examine it separately. We also include the number of suggested pages viewed by user for browsing. There are 6 features for browsing, and 5 for searching.

Time: These features measure time spent by user during search. Previous research has shown that time spent by the user on interface components (Arapakis and Leiva, 2016; Dave et al., 2013) and between actions (Dupret and Lalmas, 2013) can provide information for user engagement. Here, we computed time spent on task, time spent on SERPs and document pages, and also cumulative time spent on satisfying and dissatisfying result pages. Similarly to the threshold used for click behaviour, a page is said to be satisfying if the user stays for more than 30 seconds; and dissatisfying if the user stays for less than 15 seconds. Three features about time spent on suggested content were used for browsing only. There are 13 features for browsing, and 10 for searching.

4.2.3 Differences between features for browsing and searching

Comparing the 43 behavioural features extracted for browsing, and the 34 extracted for searching, the main difference is that nine features about suggested content were included for browsing only. This is due to the intrinsic nature of each type of interaction with the system: in a browsing task, the users do not have a defined goal and can explore the collection freely, therefore there is a possibility that the user runs out of query objectives, while remaining overall interested in the collection. As a result, studies have been working on functions to assist the user browsing experience (e.g., personalization and result diversification for item discovery (Teo et al., 2016)). The CHiC system contains a hierarchical menu, which assists the user with checking results by

topic. Another function, metadata links, allows the user to quickly access similar results that contain the same metadata. These two functions provide user some system suggested content and thus help the users explore the collection. Therefore, these nine extra features were selected to capture the user interaction with the suggested content in the browsing context.

4.3 Study B. Feature importance analysis

The purpose of this study is to determine the extent to which individual behaviour features (extracted from phase 1, study A) contribute to predicting user perception of engagement. We gauge this contribution by the *feature importance* assigned to each dependent variable by a Random Forest model (Breiman, 2001). Prediction is done by splitting users into two groups across the median of their reported engagement score distributions, collected in the UES questionnaire. By *importance* here we mean a score which measures the drop in predictive accuracy when out-of-bag samples of a feature are randomized. This type of importance is thus dubbed *permutation importance* and the method used to compute it is an algorithm know as MDA (mean decrease in accuracy) evaluation. The latter is tailored to Random Forests, whereas the notion of permutation importance applies to any ensemble of estimators based off bagging. Figure 4.3 presents a pictorial description of the analysis steps, data and variables used in this section. We restate below the research questions associated to this study:

RQ.2 To what extent can individual features predict user perception of engagement?
RQ.3 How do the relationships between behavioural features and user perception of engagement vary between dimensions?

4.3.1 Method

In this study we simultaneously employ user perception of engagement data, collected through the UES questionnaire, and behavioural features extracted in phase 1, study A. In total, user perception of engagement data are collected from 157 participants in browsing and 377 in

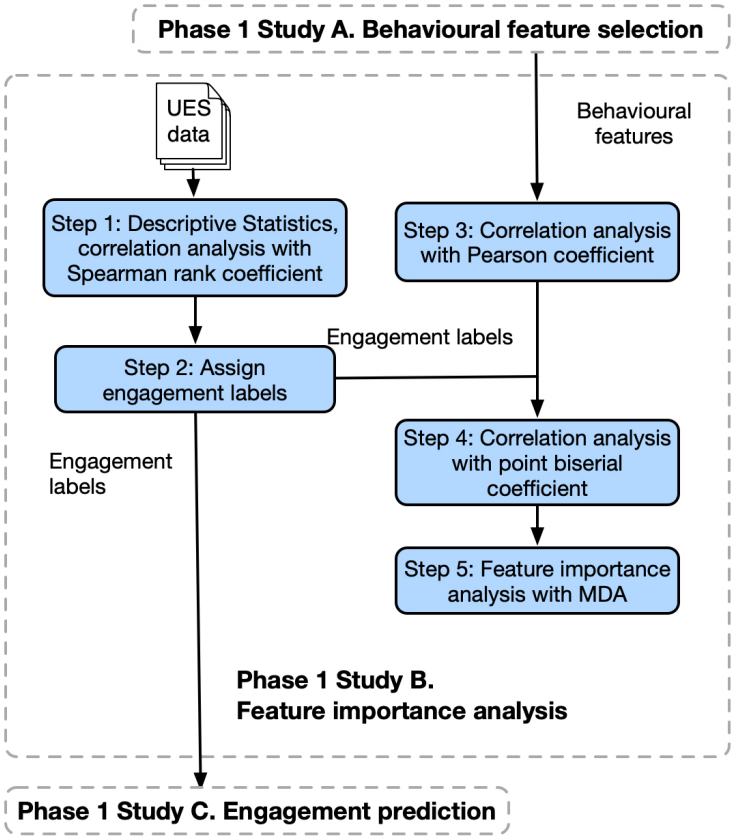


Figure 4.3: Analysis steps, data and variables used in section 4.3.

searching, and 43 behavioural features are extracted for the 157 participants in browsing and 34 behavioural features are extracted for the 377 participants in searching. Each of the four UES sub-scale represents one dimension of engagement. These four dimensions (details in section 3.4.1) are Novelty (NO), Felt Involvement (FI), Endurability (EN), and Perceived Usability (PUs). To assign scores for each dimension, all the items within one dimension were aggregated. All the analyses in this study were written in R.

Analysing user perception of engagement (step 1 and step 2):

Descriptive statistics and correlation analysis (step 1): we first examine user perception of engagement data and present the statistics of the four engagement dimensions. In order to examine the relationships between the four user perception of engagement dimensions, descriptive statistics were provided based on the original scale and correlation analysis using Spearman's ρ

(Spearman, 1904) was undertaken in order to test the distinctiveness between each dimension. We choose Spearman's ρ as it measures rank correlation, and the four engagement dimensions are collected using a 5 point scale. Correlation coefficients measure the degree of correlation between two variables. In a correlation analysis, the null hypothesis, H_0 is that there is no correlation between the two variables (X and Y).

$$H_0 = \{\text{Corr}(X, Y) = 0\} \tag{4.1}$$

When a p-value is less than 0.05, we say the observed correlation coefficient is statistically significant, or more specifically, the correlation is significant at the 0.05 level. Different types of correlation coefficients have been introduced based on the variable types (e.g., Spearman rank coefficient ρ, Pearson correlation coefficient r). In our case, since there are 4 engagement dimensions, we test 6 hypotheses, one for each pair.

Assigning engagement labels (step 2): The median of each engagement dimension score is used to divide users into a two groups: high and low engagement, in order to pose a class-balanced classification problem. We refer to the group labels as engagement labels in figure 4.3. We also present a descriptive statistic of the intersection of the user sets in the *high* engagement group for each dimension, in order to compare overlaps between these test groups.

Analysing behavioural features (step 3):

Correlation analysis (step 3): subsequently, we analyse behavioural features alone; Pearson correlation coefficient (r) was used to analyse the linear relationship between each pair of behavioural features and thus view the redundancy within the set of behavioural features as a whole. Hypothesis tests of uncorrelation are conducted between these each pair of features.

Analysing behavioural feature importance with respect to user perception of engagement (step 4 and step 5):

Correlation analysis (step 4): after that, we examined user perception of engagement and behavioural features together. We first reported the correlation between the two using point biserial correlation coefficient (Tate, 1954), as the engagement dimensions are dichotomous (i.e., *high* and *low*). Again, a null hypothesis, H_0 of uncorrelation was tested for each pair of $(feature, dimension)$. The coefficients together with the significance level are reported.

Feature importance analysis with MDA (step 5): to examine the importance of individual behavioural features with respect to each engagement dimension (*RQ.2*), we measured feature contribution by a commonly used technique, Mean Decrease in Accuracy (MDA) of a Random Forest model (Breiman, 2001). One binary Random Forest model was trained to predict each of the four engagement dimensions using all behavioural features, and output the ranking of behavioural features according to their MDA. Each tree in random forest is constructed using a different bootstrap sample from the whole dataset, and the data left out are called out-of-bag (OOB) data, which is used to get estimates of feature importance. The mean decrease in accuracy of the random forest model is calculated by random shuffling only the single selected feature in the feature set, and compare the model performance on the OOB data using the original feature set and the shuffled one. Thus, a negative MDA is possible due the random shuffling as it may make the feature more useful on the OOB data. The behavioural features are ranked by their MDA for each engagement dimension. Subsequently, we compared the top 10 behavioural features of all four engagement dimensions according to the best practise (e.g., (White and Horvitz, 2015; Arapakis and Leiva, 2016; Yu et al., 2018)). This was to answer how variable are the engagement dimensions in terms of behaviour signals (*RQ.3*).

4.3.2 Results (browsing)

This section describes the results of study B using the CHiC dataset. We first present the analysis employing only user perception of engagement data (step 1, and step 2), and then

report the results of the analysis using behavioural features alone (step 3). Subsequently, we report results for a feature contribution analysis in which we examined the MDA of behavioural features in predicting user perception of engagement (step 4 and step 5).

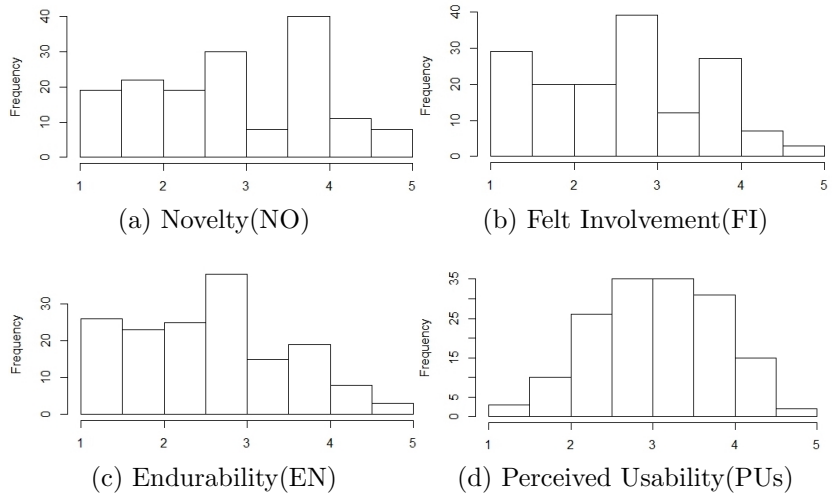
User Perception of Engagement

We examined the correlation between the four engagement dimensions first and separate the users into two groups for each dimension based on their scores. Table 4.3 presents the descriptive statistics of the four chosen engagement dimensions on a 5-point scale. We used the average score of the questionnaire items for each dimension. The mean value of three dimensions was slightly below 3, which is the middle of the 5-point scale. The standard deviation values for all dimensions was in the interval of $[0.74, 1.09]$. Correlation analysis with Spearman's ρ revealed that there were low to moderate ($0.3 < \rho < 0.5$, $p < 0.001$) correlations between PUs-EN. Significant correlation coefficient more than 0.5 were observed for pairs of FI-EN, FI-NO, and EN-NO, suggesting that the two dimensions in those pairs are similar ($p < 0.001$). Correlation coefficient between PUs-FI (0.26) is low ($\rho < 0.3$, $p < 0.001$) and the correlation coefficient between PUs-NO ($\rho = 0.13$, $p = 0.097$) is not significant. The reliability analysis resulted in Cronbach's $\alpha = 0.79$ to 0.88, indicating good internal consistency for each engagement dimensions (table 4.3); values between 0.7 and 0.9 are considered optimal (DeVellis, 2003). For NO and EN, the square roots of the dimension's average variance extracted (AVE) are not greater than all the dimension's correlations with other dimensions, which does not support the discriminant validity (Fornell and Larcker, 1981). We recall that in a subsequent recent revision of the UES scale (O'Brien et al., 2018), dimensions EN, FI and NO were merged into a single factor, a modification which is consistent with the current results (table 4.3). However, due to our focus on individual dimensions of the UES scale in an attempt to explain how they relate to user behaviour in browsing and searching, we opt to preserve the original scale. We shall see that NO, FI and EN carry independent characteristics in terms of be how they relate to user behaviour. Whether the UES is a stable measurement of engagement should also be tested in further research.

Table 4.3: Descriptive statistics of engagement dimensions (browsing).

Engagement dimension	Mean(SD)	Median	Cronbach's α	NO	FI	EN	PUs
Novelty (NO)	2.92 (1.09)	2.67	0.79	*0.75*	0.82***	0.71***	0.13
Felt Involvement (FI)	2.63 (1.02)	2.67	0.87		*0.83*	0.82***	0.26***
Endurability (EN)	2.59 (0.96)	2.6	0.88			*0.79*	0.5***
Perceived Usability (PUs)	3.11 (0.74)	3.13	0.83				*0.62*

Significance level (2-tailed): *** = p <0.001).
Italic numbers on the diagonal are the square roots of the AVE for the dimension.



(a) Novelty(NO) (b) Felt Involvement(FI)

(c) Endurability(EN) (d) Perceived Usability(PUs)

Figure 4.4: Frequency distribution of user scoring on four engagement dimensions(browsing). The x-axis of the histograms represent bins of intervals rather than discrete numbers.

Figure 4.4 shows the frequency distribution of all four engagement dimensions. For all four dimensions, the distribution was relatively balanced. The median number of each engagement dimension is used to divide the 157 users into a binary categorisation: *high* and *low*, in order to achieve a relatively balanced sample for each category. For PUs, 83 users were labeled as high, and 74 users were labeled as low, while for FI, EN, and NO this distribution was 88 and 69; 83 and 74; 97 and 60, respectively.

Behavioural Features

Pearson r correlation coefficient of each pair of behavioural features was calculated. Figure 4.5 summarises the Pearson correlation coefficients of each pair of the 43 behavioural features. Blue indicates positive correlation coefficients, while red indicates negative correlation coefficients. Only a few pairs, such as the pair NumClicksOnSERPsSug and Ratio$\langle sat/all \rangle$Clicks, have

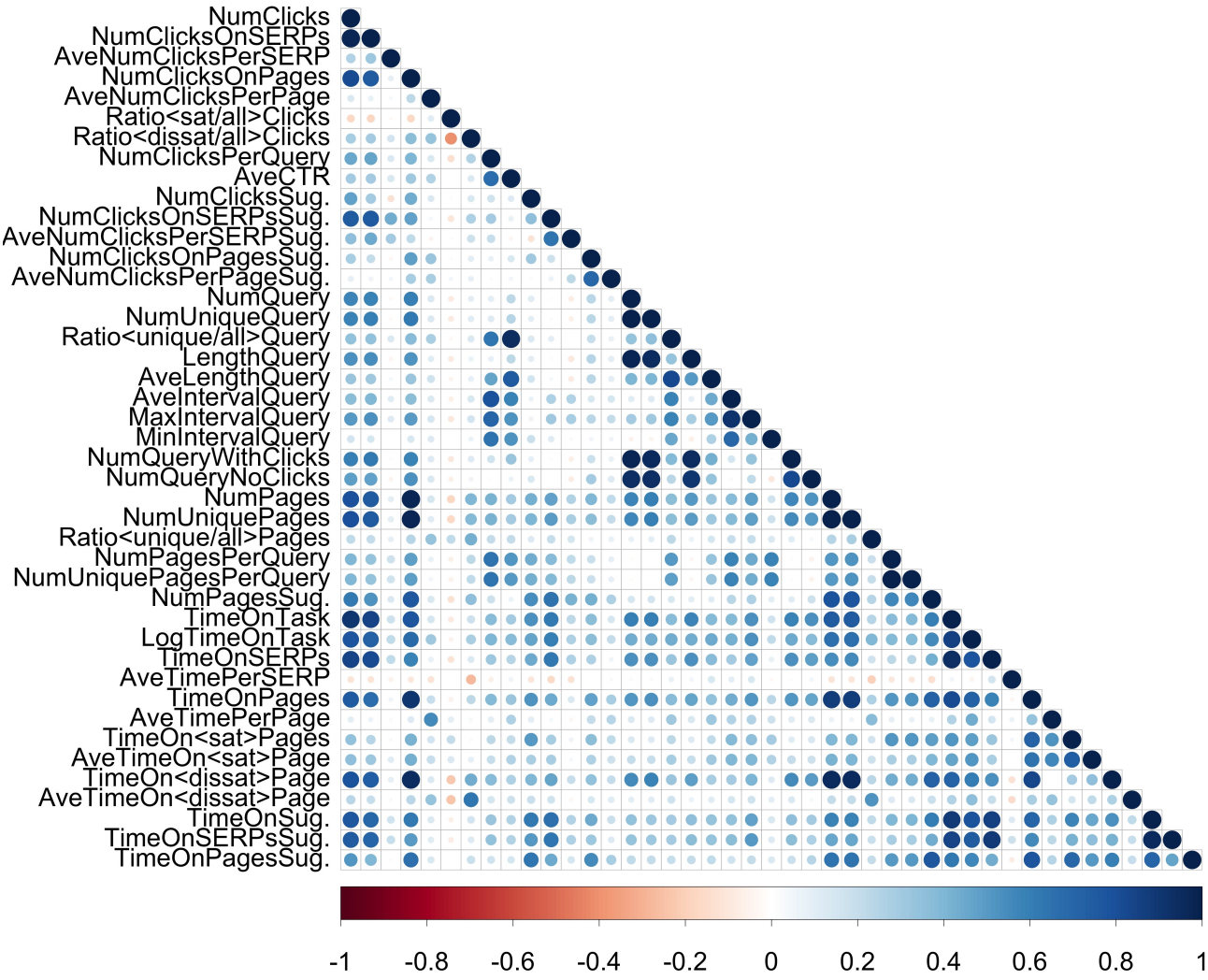


Figure 4.5: Correlation coefficient between behavioural measures (browsing). A colour gradient is used to encode the $[-1, 1]$ interval.

a negative r value and most of the relationships are not strong (0 $>r$ >-0.2). The pair Ratio$\langle sat/all \rangle$Clicks and Ratio$\langle dissat/all \rangle$Clicks has a moderate negative correlation (r = -0.41 ,p <0.001) as they are contra measures. Similarly, a few strong positive correlations (r >0.8) can be explained by the meaning of the features, such as NumClicks and NumClicksOnSERP (r >0.97,p <0.001), and NumClicksOnPages and NumUniquePages (r >0.97,p <0.001). Only 14 users submitted duplicate queries (8.92%) and they rarely clicked on anything on the returned SERP from the duplicated queries. There was also a sizeable amount of low to moderate correlation coefficients (0.3 <$|r|$ <0.8) for features, such as NumClicksOnPages and AveCTR, that quantify different aspects of user behaviour, suggesting that these facets present some moderated degree of mathematical independence and are thus expected, in their totality, to

better capture the variance in the response variable.

Feature Importance

After examining engagement dimensions and behavioural features separately, we examine the contribution of individual behavioural features with respect to each engagement dimension (*RQ.2*). Firstly, correlation analysis was conducted between the two using point-biserial correlation coefficients and then feature importance for each behavioural feature was measured by Mean Decrease in Accuracy (MDA) of a Random Forest model. Table 4.4 presents the correlation coefficients between the engagement dimensions and the behavioural features. Each category contains behavioural features that correlate to at least one engagement dimension significantly, which appear to be in line with those reported in previous studies (e.g., Lehmann et al. (2012); Drutsa et al. (2015b); Teo et al. (2016)). The FI and NO dimensions were significantly correlated with the largest set of behavioural features, with statistically significant moderate correlation coefficients ($0.2<|r_{pb}|$). Only a few behavioural features, (e.g., Ratio$\langle dissat/all \rangle$Clicks, AveTimePerPages, and AveNumClicksPerSERPsSug.) have a weak correlation ($|r_{pb}|<0.2$, p <0.05) with dimension PUs. It is likely that behavioural features reflect NO and FI dimensions more than the other two.

To examine the importance of individual behavioural features with respect to each engagement dimension, all 43 behavioural features were used to train a Random Forest classifier to predict each of the binary engagement labels and feature importance were measured by MDA. Figure 4.6, 4.7, 4.8, and 4.9 show the MDA of all 43 features for the four engagement dimensions (see definitions in table 3.1). For brevity, we present only the top 10 most important behavioural features ordered by MDA for each engagement dimension in table 4.5. The 43 behavioural features were selected based on existing studies, and MDA only tests the effect of individual features on the model conditional on the whole feature set. Some features are found to have a negative MDA for some engagement dimensions. This implies that the average accuracy of the model increases when permuting the distribution of these features across the out-of-bag (OOB) samples for each tree in our random forest model. This is potentially dependent on the number

Table 4.4: Point-biserial correlation coefficients between behavioural features and four engagement dimensions (browsing).

	Features	NO	FI	EN	PUs
Click-related	NumClicks	0.343**	0.348**	0.214**	0.042
	NumClicksOnSERPs	0.332**	0.312**	0.202*	0.055
	AveNumClicksPerSERP	0.137	0.108	0.143	-0.002
	NumClicksOnPages	0.363**	0.373**	0.222**	0.017
	AveNumClicksPerPage	0.209**	0.185*	0.171*	-0.056
	Ratio$\langle sat/all \rangle$Clicks	0.014	-0.044	0.05	-0.026
	Ratio$\langle dissat/all \rangle$Clicks	0.243**	0.316**	0.247**	0.170*
	NumClicksPerQuery	0.230**	0.151	0.083	0.043
	AveCTR	0.225**	0.158*	0.002	-0.092
	NumClicksSug.	0.135	0.244**	0.096	0.009
	NumClicksOnSERPsSug.	0.247**	0.251**	0.223**	0.094
	AveNumClicksPerSERPSug.	0.145	0.067	0.165*	0.199*
	NumClicksOnPagesSug.	0.163*	0.190*	0.179*	0.02
	AveNumClicksPerPageSug.	0.158*	0.151	0.200*	0.08
Query-related	NumQuery	0.194*	0.185*	0.04	-0.107
	NumUniqueQuery	0.198*	0.194*	0.062	-0.094
	Ratio$\langle unique/all \rangle$Query	0.252**	0.186*	0.027	-0.099
	LengthQuery	0.169*	0.159*	0.01	-0.117
	AveLengthQuery	0.161*	0.105	-0.054	-0.138
	AveIntervalQuery	0.246**	0.198*	0.151	0.08
	MaxIntervalQuery	0.295**	0.273**	0.193*	0.115
	MinIntervalQuery	0.148	0.046	0.035	-0.059
	NumQueryWithClicks	0.236**	0.223**	0.052	-0.099
	NumQueryNoClicks	0.112	0.107	0.018	-0.108
Result-related	NumPages	0.348**	0.358**	0.205*	0.022
	NumUniquePages	0.356**	0.358**	0.210**	0.022
	Ratio$\langle unique/all \rangle$Page	0.304**	0.302**	0.273**	0.045
	NumPagesPerQuery	0.234**	0.192*	0.121	-0.008
	NumUniquePagesPerQuery	0.233**	0.190*	0.126	-0.001
	NumPagesSug.	0.267**	0.307**	0.221**	0.065
Time-related	TimeOnTask	0.342**	0.351**	0.232**	-0.027
	LogTimeOnTask	0.414**	0.433**	0.280**	-0.028
	TimeOnSERP	0.278**	0.286**	0.189*	-0.03
	AveTimePerSERP	-0.096	-0.101	-0.095	-0.152
	TimeOnPages	0.355**	0.373**	0.234**	-0.014
	AveTimePerPages	0.141	0.076	0.055	-0.199*
	TimeOn$\langle sat \rangle$Page	0.202*	0.217**	0.185*	-0.017
	AveTimeOn$\langle sat \rangle$Page	0.187*	0.228**	0.102	-0.077
	TimeOn$\langle dissat \rangle$Page	0.331**	0.319**	0.166*	0.028
	AveTimeOn$\langle dissat \rangle$Page	0.349**	0.389**	0.258**	0.035
	TimeOnSug.	0.289**	0.320**	0.224**	0.002
	TimeOnSERPSug.	0.269**	0.290**	0.215**	-0.001
	TimeOnPagesSug.	0.221**	0.260**	0.155	0.009

The darker shading indicates the correlation coefficients r_{pb} is greater than 0.2.
Significance level (2-tailed): ** = p <0.01, and * = p <0.05.

of trees and the number of OOB splits, as well as the size of the bag. In general, these are expected to be the most irrelevant features with respect to the accuracy of our Random Forest model.

Table 4.5: Top-10 behavioural features with respect to each engagement dimension according to the MDA (browsing).

NO		FI		EN		PUs	
Feature	MDA	Feature	MDA	Feature	MDA	Feature	MDA
TimeOnPages	12.61	TimeOn⟨*dissat*⟩Page	16.50	AveTimeOn⟨*dissat*⟩Page	8.59	AveNumClicksPerSERPSug.	12.85
NumClicksOnPages	10.61	NumClicksOnPages	10.97	TimeOn⟨*dissat*⟩Page	7.54	AveNumClicksPerSERP	8.69
TimeOn⟨*dissat*⟩Page	7.37	LogTimeOnTask	9.43	TimeOnPages	6.74	MaxIntervalQuery	5.28
Ratio⟨*unique*/*all*⟩Page	7.17	NumUniquePages	8.99	NumUniquePages	5.55	Ratio⟨*dissat*/*all*⟩Clicks	4.30
LogTimeOnTask	6.82	AveTimeOn⟨*dissat*⟩Page	8.66	AveTimePerPages	5.06	AveTimePerPages	3.10
NumClicks	6.72	TimeOnPages	8.49	Ratio⟨*unique*/*all*⟩Page	4.86	NumClicksSug.	1.94
NumUniquePages	6.67	Ratio⟨*unique*/*all*⟩Page	8.09	NumClicksSug.	4.70	TimeOnSERP	1.81
AveNumClicksPerPage	6.61	NumClicksOnSERP	7.81	NumPages	4.40	LengthQuery	1.64
AveTimeOn⟨*dissat*⟩Page	6.58	NumClicks	7.60	TimeOnSug.	4.34	TimeOn⟨*sat*⟩Page	1.23
NumClicksOnSERP	6.07	NumPages	7.46	TimeOnPagesSug.	4.31	NumQueryWithClicks	1.22

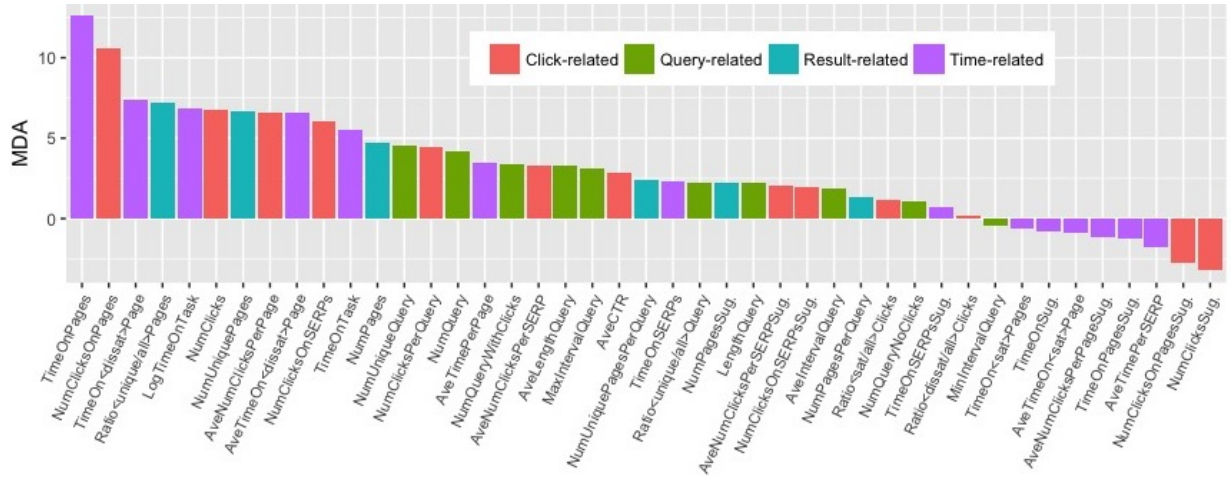


Figure 4.6: Feature importance for NO (browsing).

Novelty (NO). The importance of all behavioural features with respect to NO is illustrated in figure 4.6. Although moderate correlated with NO (r= 0.289, p <0.01, table 4.4), feature TimeOnSug. has an almost zero MDA value. This is because MDA measures the contribution a single feature could make in a feature set and thus is affected by the interaction between features. It will have a zero or even negative value if the information contained by the single feature is also contained by other features in the set. Four out of ten features are time related. Features that suggest a negative experience (AveTimeOn⟨*dissat*⟩Page, and TimeOn⟨*dissat*⟩Page) are also ranked high for the NO dimension.

Felt Involvement (FI). As illustrated in figure 4.7, the most important feature for FI was

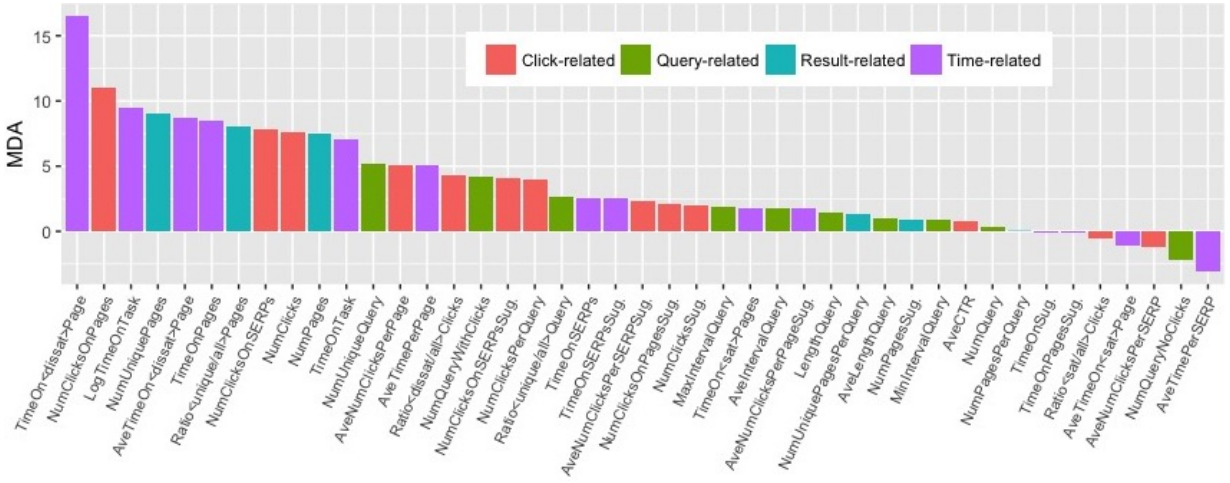


Figure 4.7: Feature importance for FI (browsing).

TimeOn⟨*dissat*⟩Pages, suggesting a negative experience according to the MDA. A similarly negative signal is also observed by feature AveTimeOn⟨*dissat*⟩Page at the result level. These suggested that users might be sensitive to negative, dissatisfied interaction. Although the top-ranked features belong to three categories, seven out of ten features were associated with the document pages, which suggests that the interaction with the document pages plays a main role in making user feel involved. Four of the top-10 features belonged to the time category, and three of them belong to the click category.

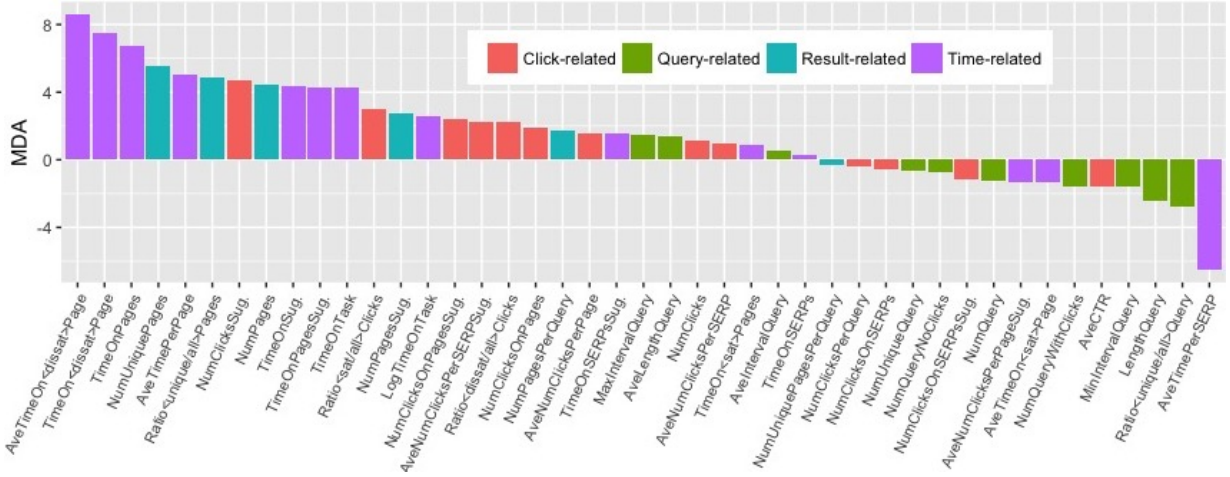


Figure 4.8: Feature importance for EN (browsing).

Endurability (EN). Figure 4.8 presents the features ranked by MDA with respect to EN. Feature AveTimeOn⟨*dissat*⟩Page ranked top for this dimension, followed by TimeOn⟨*dissat*⟩Page. These two features indicate a possible negative experience. Regarding feature categories, six

out of top-10 features are time-related, which suggests time-related information is important for measuring EN in the browsing context. Three of the other four top-ranked features belong to the result category and one belongs to the click category, confirming the intuition that the quality of results obtained has an impact on EN. Same as the NO and FI dimensions, none of the query-related features are ranked in the top 10. A possible explanation is that in the browsing task, users follow the suggested topics or links more than submitting a query on their own.

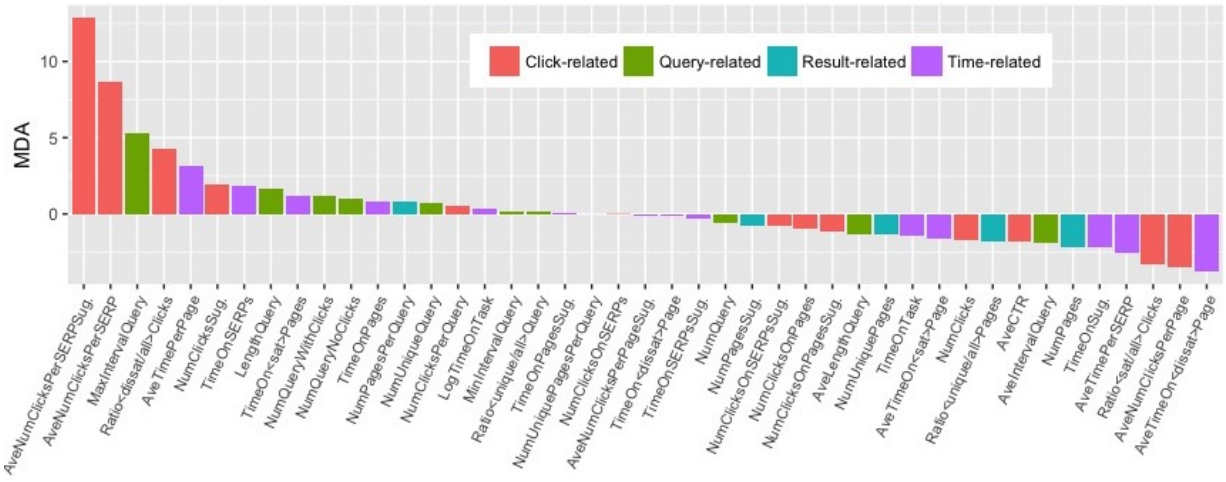


Figure 4.9: Feature importance for PUs (browsing).

Perceived Usability (PUs). Figure 4.9 presents the features ranked by MDA with respect to PUs. Compared to the other three engagement dimensions, the number of features which have a positive MDA for PUs is the smallest, while a similar situation was observed in table 4.4 that PUs has a low correlation coefficient with all behavioural features. The most important feature is AveNumClicksPerSERPSug. Given that users performed a non-purposeful browsing task, the ability of the system to provide interesting suggestions or inspiration is important as it motivates the user to continue interacting with the system. Thus, the click-related features on the suggested content (e.g., AveNumClicksPerSERPSug., and NumClicksSug.) are accordingly ranked high. Clicks on the SERP (AveNumClicksPerSERP), representing the average number of interactions users performed on the SERPs, is ranked second. This addresses the frequency that users find the items on the SERPs are potentially interesting to them.

A number of features made the top-10 for at least two dimensions (table 4.6). Notably,

Table 4.6: Comparison of Top-10 behavioural features between engagement dimensions (browsing).

Features	NO	FI	EN	PUs
AveTimePerPages			✓	✓
NumClicksSug.			✓	✓
AveTimeOn⟨*dissat*⟩Page	✓	✓	✓	
TimeOn⟨*dissat*⟩Page	✓	✓	✓	
TimeOnPages	✓	✓	✓	
NumUniquePages	✓	✓	✓	
Ratio⟨*unique*/*all*⟩Page	✓	✓	✓	
NumPages		✓	✓	
NumClicksOnPages	✓	✓		
LogTimeOnTask	✓	✓		
NumClicksOnSERP	✓	✓		
NumClicks	✓	✓		

✓means the corresponding behavioural features is ranked in the top 10 for this dimension.

five features, namely AveTimeOn⟨*dissat*⟩Page, TimeOn⟨*dissat*⟩Page, TimeOnPages, NumUniquePages, and Ratio⟨*unique*/*all*⟩Page appear in the top-10 list for three engagement dimensions NO, FI and EN. Three of these features listed in table 4.6 originated from the time-related feature category, and two of them are from the result-related feature category. These two types of features potentially are better in describing engagement in general. When assessing the overlap between the behavioural feature lists with high relative MDA, for the four engagement dimensions, NO and FI shared nine out of the top 10 features, suggesting these behaviour features reflect these two dimensions similarly in browsing. This large overlap was also observed in the pair's correlation coefficient (ρ=0.82, p <0.001, table 4.3). An explanation is that browsing is curiosity-driven, the feeling of novelty the user experiences directly affecting how much they is motivated to continue using the system and thus leads the user into an involved state. The strong connection also supports the design in the UES, in which NO contributes to FI directly (O'Brien and Toms, 2010a). Moreover, dimensions NO and EN share five out of the top 10 features (Spearman ρ=0.71, p <0.001), and dimensions FI and EN share six out of the top 10 features (Spearman ρ=0.82, p <0.001). Regarding the feature category for each dimensions, time-related features and result-related features reflect dimension NO, FI and EN the most, while PUs do not share much top-ranked behaviour features with any other dimensions. This

is also observed in the correlation analysis (table 4.3): PUs has a low to moderate correlation ($\rho = 0.5$, $p < 0.001$ for EN; $\rho = 0.26$, $p < 0.001$ for FI; $\rho = 0.13$, $p = 0.097$ for NO) with the other three dimensions.

4.3.3 Discussion of feature importance in browsing

The experimental results reveal several insights which we use to first answer the research questions. In general, results support the thesis that a relationship exists between user perception of engagement and behavioural features in browsing (*RQ.2*). Certain categories of features present higher negative or positive correlations with the four engagement dimensions. NO exhibits the highest correlations with result-related and time-related features. For FI, it is time-related features that stand out. EN seems to combine the two responses and benefits from average correlations from time-related features, result-related features, and click-related features. PUs do not exhibit positive correlations with any category. The differences among four dimensions answer *RQ.3*, that dimension NO, FI and EN share the largest similarity.

We then discuss the detailed findings that support our answers. First, regarding feature categories, time-related features and result-related features are relatively the best in describing user perception of engagement dimensions NO, FI and EN (table 4.5). Click-related features are also important for NO. Intuitively, these features are also central in predicting FI and EN. Similar results were observed in satisfaction prediction (Mehrotra et al., 2017) that wrapping the time information in models improves prediction performance. However, due to the known difficulty in picking the right threshold for time (Kim et al., 2014), and the overwhelming choices of models, how to use the time information is a delicate issue. A new metric (Machmouchi et al., 2017) adapting time as units has been designed for the general search engine case, and shows good sensitivity in online A/B testing and accuracy in predicting search success. Apart from that, none of the query-related features is ranked in the top 10 for NO or FI or EN in the feature importance analysis. This could potentially be because of the small amount and short length of queries issued in the session as it is not mandatory for users to submit queries. Among 157 users, 87 (55.4%) submitted at least one queries with an average length of 1.37 words (min=1,

max=3.1). Thus, the query-related features are less distinguishable and therefore, less indicative of user perception of engagement in this context. Result-related features simulate users interaction with the collection, which are of potential interest for users.

Regarding the differences between behavioural and engagement relationships between four dimensions (*RQ.3*), most of the features considered were found to have a weak or moderate correlation with the NO, FI and EN dimensions, and a very low correlation with the PUs dimension (table 4.4). Similar relationships were observed in other analyses such as the feature importance analysis (figure 4.9), where the number of features with a positive MDA for predicting PUs is the lowest among all four dimensions. Intuitively, this could be due to the system interface (figure 3.5) which is similar to a general commercial search interface. Users might be used to the interface and simply do not feel strongly about the functions. This could also be caused by the task type, as in a browsing scenario without defined search goals, the users might not expect the functions to assist them to achieve such a goal. Moreover, the NO and FI dimensions share nine top-ranked behavioural features, suggesting that user behaviour reflects these two dimensions very similarly. These two dimensions also share at least 5 top ranked behavioural features with EN, suggesting NO and FI might be the main contributor to the overall evaluation (EN) than PUs given the browsing context of this dataset. This could also be the reason why PUs has a lower correlation with all other dimensions (table 4.3), whereas all the other pairs have a strong correlation coefficient ($r > 0.6$, $p < 0.001$).

4.3.4 Results (searching)

This section describes the results of study B using the wikiSearch dataset. We first present the analysis employing only user perception of engagement data (step 1 and step 2), and then report the results of the analysis using behavioural features alone (step 3). Subsequently, we report results for a feature contribution analysis in which we examined the MDA of behavioural features in predicting user perception of engagement (step 4 and step 5).

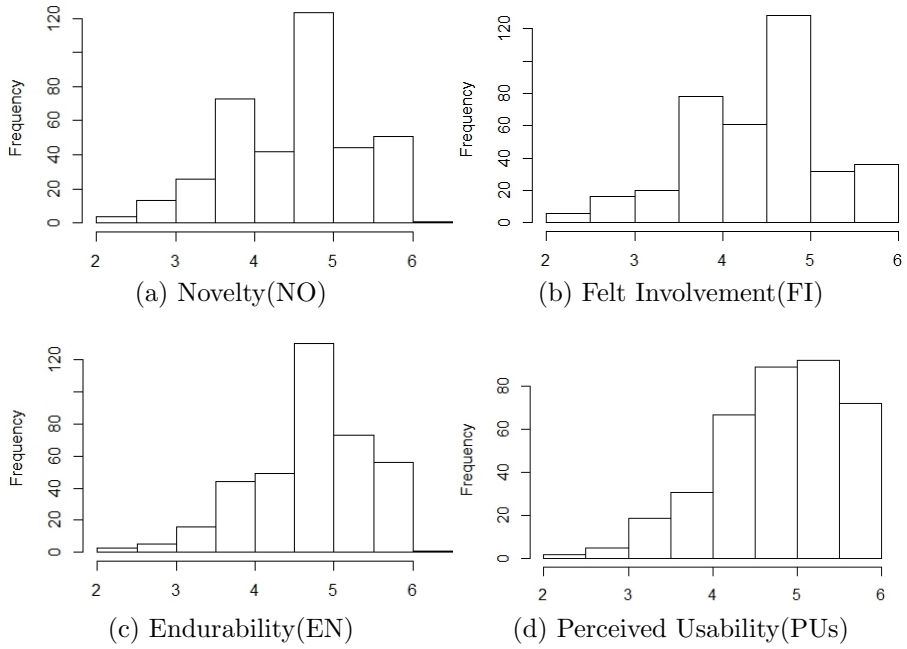
User Perception of Engagement

Table 4.7: Descriptive statistics of engagement dimensions (searching).

Engagement dimension	Mean(SD)	Median	Cronbach's α	NO	FI	EN	PUs
Novelty (NO)	4.59 (0.81)	4.67	0.73	*0.66*	0.63***	0.62***	0.4***
Felt Involvement (FI)	4.49 (0.79)	4.67	0.72		*0.68*	0.57***	0.35***
Endurability (EN)	4.78 (0.74)	5	0.8			*0.66*	0.63***
Perceived Usability (PUs)	4.85 (0.76)	4.88	0.86				*0.64*

Significance level (2-tailed): *** = p <0.001

Italic numbers on the diagonal are the square roots of the AVE for the dimension.



(a) Novelty(NO) (b) Felt Involvement(FI)

(c) Endurability(EN) (d) Perceived Usability(PUs)

Figure 4.10: Frequency distribution of user scoring on four engagement dimensions (searching). The x-axis of the histograms represent bins of intervals rather than discrete numbers.

Table 4.7 presents the descriptive statistics of the four chosen engagement dimensions in the searching context. The mean values of all four dimensions were above 4, which is the mean of this 7-point scale. The standard deviation values for all dimensions were below one, indicating consistency amongst users' responses. Correlation analysis with Spearman's ρ revealed that there were low to moderate (0.3<r <0.5) correlations between two pairs of dimensions: PUs-FI and PUs-NO. Correlation values more than 0.5 were observed for pairs of PUs-EN, FI-EN, FI-NO, and EN-NO. The reliability analysis resulted in Cronbach's $\alpha = 0.72$ to 0.86 indicating good internal consistency for each engagement dimensions (table 4.3), which are optimal (0.7 $<\alpha <$0.9) (DeVellis, 2003). The square roots of each dimension's average variance extracted

(AVE) are greater than the dimension's correlations with any other dimensions, supporting that the discriminant validity (Fornell and Larcker, 1981) is established in searching.

Figure 4.10 shows the frequency distribution of all engagement dimensions. For all the four dimensions, the distribution was skewed towards the high-end of the scale. All histograms began with "2", as no users responded to any of the dimensions with an average of "1". As the tested search system was usable (Toms et al. (2013) found this system effectively supports search tasks), it was expected that the user feedback on engagement would likely be biased towards positive scores. The median number of each engagement dimension is used to divide users into a binary categorisation: *high* and *low*. Users assigned high were ones that experienced equal or above the median score in this engagement dimension, while low group refers to the ones that experienced below median in this engagement dimension. For PUs, 213 users were labelled as high, and 164 users labelled as low, while for FI, EN, and NO this distribution was 196 and 181; 189 and 188; 219 and 158, respectively.

Behavioural Features

In order to probe for the redundancy and relevance of the selected 34 behavioural features, we conducted Pearson's correlation analysis between all the behavioural feature pairs. Figure 4.11 summarises the Pearson correlation coefficients of each pair of behavioural features. Clearly, a few correlation coefficients were negative, and strong negative correlation coefficients could be attributed to the mathematical relationship between the measures. For instance, a strong negative correlation is found between AveCTR and NumQueryNoClicks (r = -0.82, p <0.001), as well as Ratio$\langle sat/all \rangle$Clicks and Ratio$\langle dissat/all \rangle$Clicks ($r$ = -0.84, p <0.001). Similarly, a few positive correlations were strong (more than 0.8) by definition (e.g., NumClicks, NumClicksOnPages, NumClicksOnSERP). One notable pair with high correlation coefficient (r=0.92, p <0.005) was TimeOn$\langle sat \rangle$Pages and TimeOnTask, and this suggests that users' experience was towards positive as their likely-to-be satisfied time accounted for a large amount of the total time. On the other hand, there were a sizeable amount of low to moderate correlation coefficients ($0.3 < |r| < 0.8$) for some pairs of features, suggesting that these facets present some moderated degree of mathematical independence.

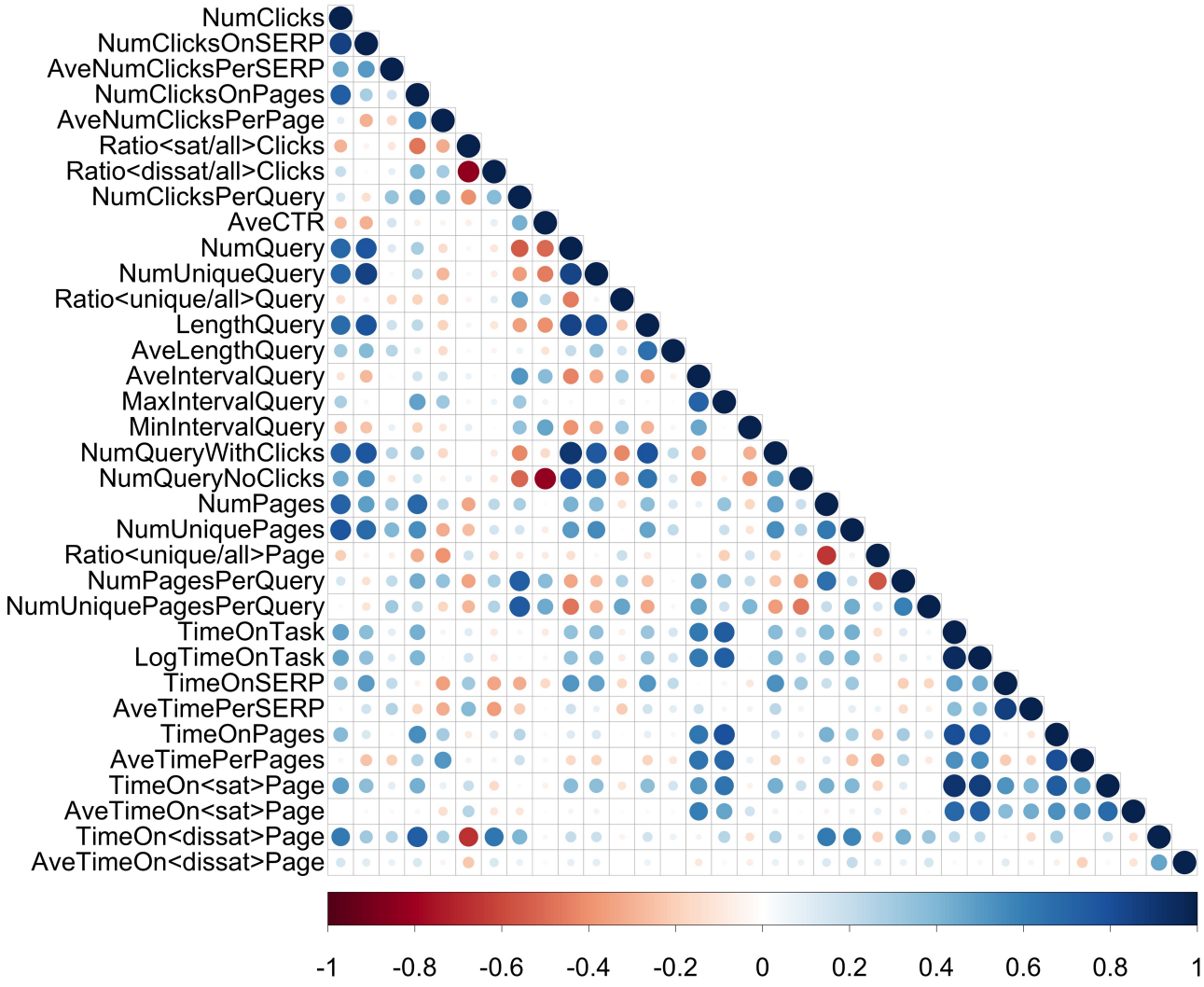


Figure 4.11: Correlation between beahvioural features (searching). A colour gradient is used to encode the $[-1, 1]$ interval.

Feature Importance

In order to examine the importance of individual behavioural features with respect to each engagement dimension (*RQ.2*), we first reported the correlation between the two using point-biserial correlation coefficients. Then, we measured feature importance by Mean Decrease in Accuracy (MDA) of a Random Forest model.

Table 4.8 shows the point-biserial correlation value between behavioural features and the four engagement dimensions. The darker shading indicates the correlation value r_{pb} is greater than 0.2 and is statistically significant. The PUs dimension is significantly correlated with the most behavioural features with low correlation values ($|r_{pb}|$ <0.2). The correlated features primarily

Table 4.8: Point-biserial correlation between behavioural features and four engagement dimensions (searching).

	Features	NO	FI	EN	PUs
Click-related	NumClicks	-0.043	-0.114*	-0.161**	-0.192**
	NumClicksOnSERP	-0.079	-0.133**	-0.185**	-0.196**
	AveNumClicksPerSERP	0.017	-0.068	-0.048	-0.043
	NumClicksOnPages	0.031	-0.029	-0.051	-0.096
	AveNumClicksPerPage	0.049	0.004	0.108*	0.006
	Ratio$\langle sat/all \rangle$Clicks	0	0.077	0.007	-0.003
	Ratio$\langle dissat/all \rangle$Clicks	0.022	-0.073	-0.03	0.014
	NumClicksPerQuery	0.092	0.023	0.084	0.107*
	AveCTR	0.125*	0.121*	0.148**	0.144**
Query-related	NumQuery	-0.089	-0.135**	-0.167**	-0.243**
	NumUniqueQuery	-0.109*	-0.123*	-0.189**	-0.202**
	Ratio$\langle unique/all \rangle$Query	0.028	0.053	0.018	0.136**
	LengthQuery	-0.09	-0.143**	-0.152**	-0.234**
	AveLengthQuery	-0.064	-0.071	-0.075	-0.095
	AveIntervalQuery	0.126*	0.116*	0.133**	0.059
	MaxIntervalQuery	0.063	0.074	0.076	-0.007
	MinIntervalQuery	0.143**	0.077	0.121*	0.102*
	NumQueryWithClicks	-0.06	-0.089	-0.132*	-0.219**
	NumQueryNoClicks	-0.1	-0.154**	-0.160**	-0.196**
Result-related	NumPages	0.041	-0.059	-0.049	-0.093
	NumUniquePages	0.084	0.015	0.047	-0.022
	Ratio$\langle unique/all \rangle$Page	-0.045	-0.064	-0.172**	-0.141**
	NumPagesPerQuery	-0.101*	0.004	-0.091	-0.015
	NumUniquePagesPerQuery	0.109*	0.029	0.097	0.106*
Time-related	TimeOnTask	0.07	0.038	0.021	0.096
	LogTimeOnTask	0.022	0.014	-0.047	-0.155**
	TimeOnSERP	-0.082	-0.116*	-0.180**	-0.252**
	AveTimePerSERP	-0.047	-0.061	-0.109*	-0.167**
	TimeOnPages	0.035	0.022	0.014	-0.056
	AveTimePerPages	0.064	0.032	0.122*	0.033
	TimeOn$\langle sat \rangle$Page	0.011	-0.009	-0.083	-0.187**
	AveTimeOn$\langle sat \rangle$Page	0.088	0.056	0.005	-0.041
	TimeOn$\langle dissat \rangle$Page	-0.019	-0.086	-0.129*	-0.098
	AveTimeOn$\langle dissat \rangle$Page	-0.104*	-0.078	-0.138**	-0.05

The darker shading indicates the correlation coefficients r_{pb} is greater than 0.2.

Significance level (2-tailed): ** = p <0.01, and * = p <0.05.

belong to the Time and Query categories, suggesting that the interactions that require users to perform an action might be more indicative for engagement in the search task. The EN dimension is significantly correlated with at least one feature from each category. Only a couple of features, such as AveCTR and NumUniqueQuery, have a significant weak correlation with dimension NO and FI.

Table 4.9: Top-10 behavioural features with respect to each engagement dimension according to the MDA (searching).

NO		FI		EN		PUs	
Feature	MDA	Feature	MDA	Feature	MDA	Feature	MDA
LengthQuery	12.70	Ratio⟨*dissat*/*all*⟩Clicks	4.48	AveCTR	5.66	TimeOnSERP	12.06
AveIntervalQuery	6.57	MaxIntervalQuery	1.81	TimeOnSERP	4.75	AveLengthQuery	8.73
NumPagesPerQuery	5.50	NumClicksOnPages	1.75	NumQueryNoClicks	4.70	AveTimePerSERP	7.38
MinIntervalQuery	4.71	AveTimeOn⟨*dissat*⟩Page	1.45	NumUniqueQuery	4.50	NumClicks	7.27
AveTimeOn⟨*dissat*⟩Page	3.68	NumClicksOnSERP	1.22	LengthQuery	3.67	LengthQuery	6.37
NumQuery	3.68	LogTimeOnTask	1.18	NumClicksPerQuery	3.57	Ratio⟨*unique*/*all*⟩Page	6.02
NumUniquePagesPerQuery	2.34	AveCTR	1.11	Ratio⟨*unique*/*all*⟩Page	3.41	TimeOn⟨*sat*⟩Page	5.36
TimeOnTask	2.06	TimeOn⟨*dissat*⟩Page	1.04	AveTimeOn⟨*sat*⟩Page	2.56	NumPagesPerQuery	5.28
NumQueryNoClicks	2.06	AveIntervalQuery	0.92	NumClicksOnSERP	2.56	NumPages	3.58
LogTimeOnTask	1.87	AveTimeOn⟨*sat*⟩Page	0.86	MinIntervalQuery	2.26	Ratio⟨*unique*/*all*⟩Query	3.48

All 34 behavioural features were used to train a Random Forest model to predict each of the binary engagement dimensions. Figure 4.12, 4.13, 4.14, and 4.15 show the feature importance of all 34 behaviour features for each engagement dimension (see definition in table 3.1). As in the previous section, since the 34 features were selected based on previous studies, some of them are not useful in predicting all engagement dimensions. For brevity, the top 10 most important behavioural measures ordered by MDA for all four engagement dimensions are presented in Table 4.9.

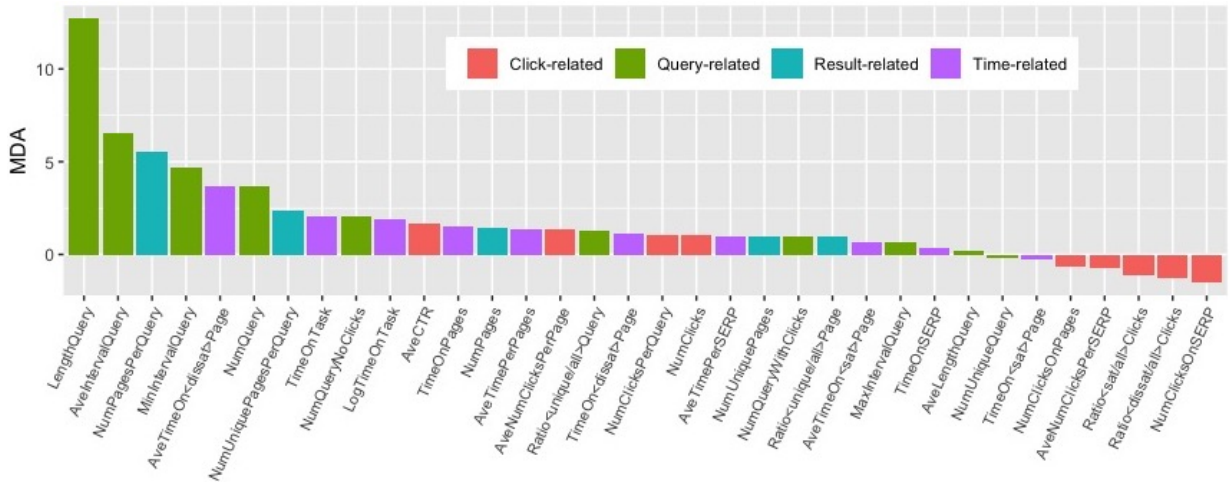


Figure 4.12: Feature importance for NO (searching).

Novelty (NO). Figure 4.12 presents the features ranked by MDA with respect to NO. In general, behavioural features do not reflect this dimension strongly. Feature LengthQuery, which is related to query complexity, ranked top in figure 4.12. Five out of the top-10 features were query-related features, and three were result-related. This aligns with our expectation, since the quality of content reflects more on relevance and only marginally on novelty in the search context. The quality of results is a prior to the user's interaction with the system, whereas users evaluate the search experience primarily on the ease of retrieval. Although NO is about the quality of content, it mainly measures the curiosity evoked during the session, thus features associated with the results were not ranked highly.

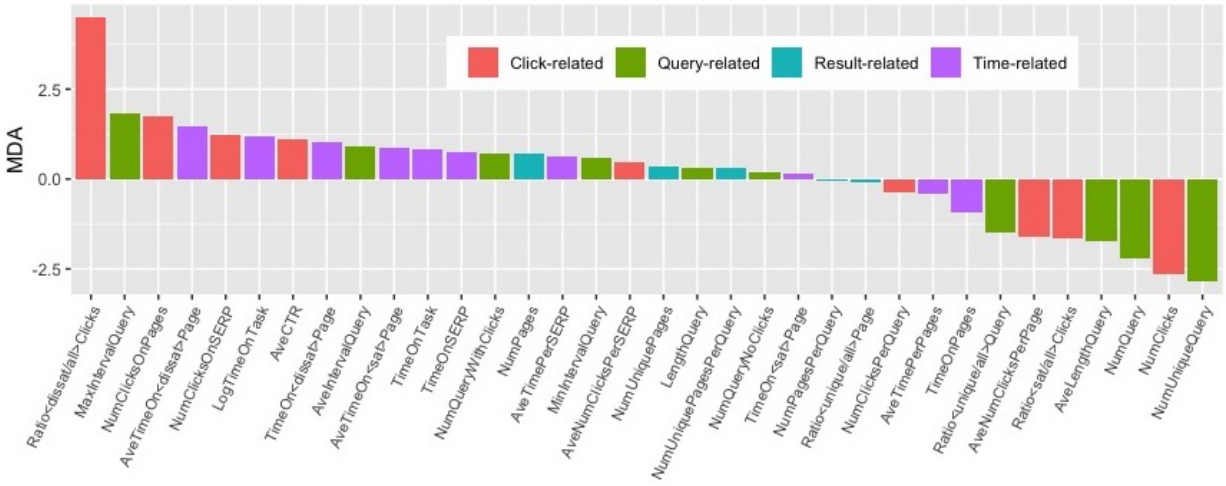


Figure 4.13: Feature importance for FI (searching).

Felt Involvement (FI). Figure 4.13 presents the features ranked by MDA with respect to FI. The top ranked feature is Ratio $\langle dissat/all \rangle$Clicks, suggesting that the negative effect of unsuccessful clicks plays a relatively important role. A similarly negative effect is also observed on the document page level, by features AveTimeOn $\langle dissat \rangle$Page and TimeOn $\langle dissat \rangle$Page. These suggested that users' felt involvement is likely sensitive to negative, dissatisfied interaction. Two out of ten features were query related. Four of the top-10 features belonged to the time category, and three other features, Ratio $\langle dissat/all \rangle$Clicks, MaxIntervalQuery, AveIntervalQuery, were also calculated based on the time spent on actions, demonstrating that time in the searching context also serves a crucial role in reflecting FI.

Endurability (EN). As illustrated in figure 4.14, the top-10 features were balanced in terms

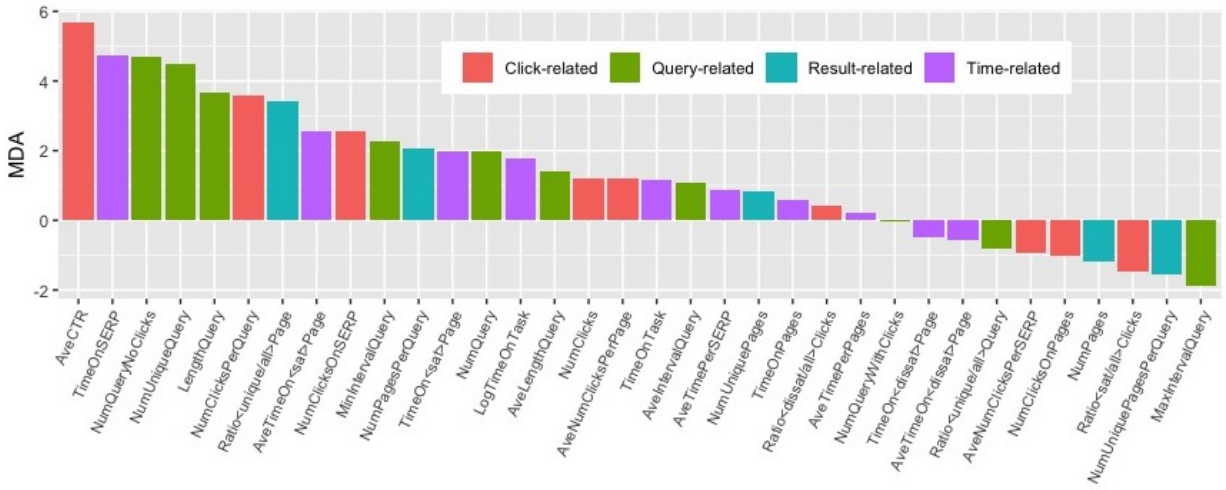


Figure 4.14: Feature importance for EN (searching).

of types: three were from the time category, two from click category, three from result category, and three from query category, suggesting that EN reflects on a wide range on search behaviour compared to other dimensions in searching. Feature AveCTR ranked top for this dimension. CTR is an indicator of user perceived relevance (Joachims, 2002) of a result and is influenced by the position of the result. AveCTR describes CTR at session level, in other words, the user perceived relevance of the submitted queries and associated SERPs.

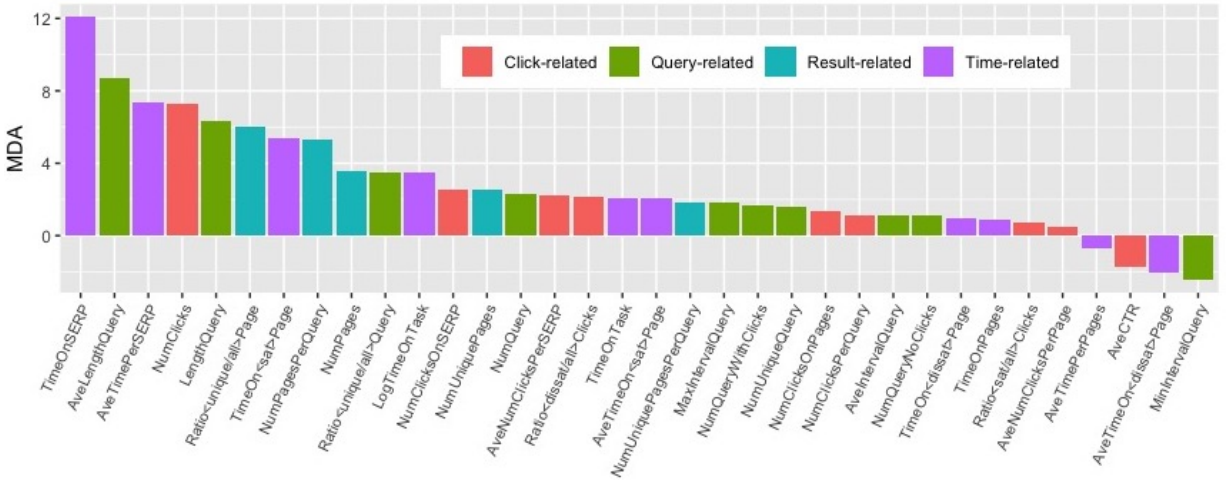


Figure 4.15: Feature importance for PUs (searching).

Perceived Usability (PUs). As illustrated in figure 4.15, the most important feature was TimeOnSERPs. Given that the task requires users to search for information to support their decision-making, the effort they spent on locating information, here measured by time on the SERP, is the most indicative of the user perception on the usability of the system. Besides,

three of the top-10 features belonged to the time-related category, covering time spent on SERP, on satisfied result pages, and on the whole task. LengthQuery and AveLengthQuery stand for query complexity, and are known to be positively related with user satisfaction. That is to say, the longer the query users submitted, the more precise the system returned results will be, thus leading to a positive user perceived experience (Belkin et al., 2003).

Table 4.10: Comparison of top-10 behavioural features between engagement dimensions (searching).

Features	NO	FI	EN	PUs
LengthQuery	✓		✓	✓
AveLengthQuery	✓	✓		
NumPagesPerQuery	✓			✓
MinIntervalQuery	✓		✓	
AveTimeOn $\langle dissat \rangle$Page	✓	✓		
NoQueryNoClicks	✓		✓	
LogTimeOnTask	✓	✓		
NoClicksOnSERP		✓	✓	
AveCTR		✓	✓	
AveTimeOn $\langle sat \rangle$Page		✓	✓	
TimeOnSERPs			✓	✓
Ratio$\langle unique/all \rangle$Page			✓	✓

✓means the corresponding behavioural features is ranked in the top 10 for this dimension.

To compare the similarities between how behaviour reflects engagement dimensions differently (*RQ.3*), table 4.10 presents the behaviour features ranked top for at least two engagement dimensions. Only a few features failed to contribute to any dimensions (outside the top-10), and conversely, 12 features were ranked among the top 10 for more than one dimension (table 4.10). Notably, LengthQuery, representing query complexity contributed to three engagement dimensions. Other features forming more diverse groups, such as AveCTR, NumPagesPerQuery, and TimeOnSERPs, all contributed to at least two dimensions. These measures originated from the Time, Query, and Click categories, and are therefore potentially important in covering different dimensions of engagement in search.

Regarding the similarity between dimensions, four pairs of engagement dimensions NO-FI, NO-EN, FI-EN, EN-PUs, shares three top ranked features, one pair, NO-PUs, only share two top ranked features. This suggests that behaviour signals reflect engagement dimensions very

differently.

4.3.5 Discussion of feature importance in searching

Our analysis provides several insights into the engagement dimensions in the searching context. We start by answering the research questions. Overall, our results support the connection between user perception of engagement and behavioural features in searching *RQ.2*. All feature categories reflect PUs best, but the correlations are moderate at most. In fact, all high correlations for this dimension are negative, suggesting that an affluence of interactions on the user's part are interpreted as a sign of an unusable system. In particular query related features exhibit the highest negative correlation with PUs, which implies in turn that the more queries users are forced to emit to the system to achieve their goal, the lower their satisfaction with the system becomes. All dimensions are relatively different from each other in terms of correlated behavioural features (*RQ.3*).

We then discuses the detailed findings that support our answers. Foremost, regarding the similarity between dimensions (*RQ.3*), we found that the sets of behavioural features that are relatively more informative with respect to each of the four engagement dimensions vary, suggesting that behaviour reflects different dimensions of engagement differently in searching. For FI, four out of its top 10 important behavioural measures came from the Time category, including the time on task, SERPs, and document pages. While, for other dimensions, the distribution of behavioural measures in terms of the four defined categories did not follow a discernible pattern. This is a result of the differences between engagement dimensions in searching. As revealed by the correlation analysis, these four dimensions have only a moderate relationship with each other, which suggests they cover different aspects of engagement, and therefore provide a more complete picture of the user experience.

Regarding feature contribution, in the point-biserial correlation analysis, some of the behavioural features selected were found to have a statistically significant weak correlation with the PUs and EN dimensions, but very few features have such correlation with the FI or NO dimensions. Intuitively, this could be due to the task type. As users were supposed to complete

search tasks, they might feel potentially under pressure and therefore they do not spare any attention on the potential additional interestingness of the content and do not allow themselves to be drawn into exploring it past the finite goal of the search task. This could also be the reason that search behaviour reflects the FI dimension the poorest, as in the feature importance analysis (figure 4.13) the number of features with a positive MDA for predicting FI is the lowest among all four dimensions.

Regarding feature categories, time-related features, click-related features and query-related features are relatively the best in characterizing user perception of engagement. Their importance was also observed with respect to satisfaction (Al-Maskari and Sanderson, 2010; Fox et al., 2005; Belkin et al., 2003), an important component of engagement. Previous studies (Hassan et al., 2010) have shared the same mindset while using Markov models with Click-related and Query-related actions as crucial states to predict successful search. However, the interpretation and direction of effect varies (e.g., considered as user effort (Lancaster, 1981) with negative influence(Al-Maskari and Sanderson, 2010), or as interests with positive influence(Fox et al., 2005)), which leaves room for exploring a more robust interpretation.

4.4 Study C. Engagement prediction using selected behavioural features

The purpose of this study is to determine how well behavioural features (from phase 1, study A) perform in predicting user perception of engagement. By prediction, in this setting we mean a classification problem of predicting either high or low engagement labels (extracted in phase 1, study B) of users using their behavioural features. In total we train 4 classifiers, comprising one baseline, a Support Vector Machine model and 2 Random Forest models once for each dimension, and once for each dataset totalling 16 fits. We report standard metrics of performance: precision, recall, F-1 score, accuracy and AUC. For each metric and for each dataset the three classifiers are evaluated against the baseline by determining statistically significant improvements in performance via a paired t-test. Figure 4.16 presents the analysis steps, data

and variables used in this section. We restate below the research question associated to this study:

RQ.3 How do the relationships between behavioural features and user perception of engagement vary between dimensions?

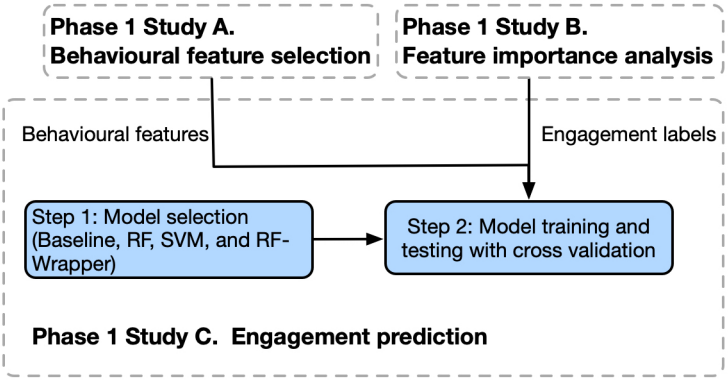


Figure 4.16: Analysis steps, data and variables used in section 4.4.

4.4.1 Method

To test how well the behavioural features extracted in phase 1, study A, predict user perception of engagement (*RQ.2*), a binary classification problem is defined in which the goal is to predict if a user will have *high* or *low* engagement levels given the selected behavioural features. *High* or *low* engagement levels are determined by splitting user groups against the median of each engagement score, as in phase 1, study B. The numbers of behavioural features used are 43 in browsing and 34 for searching. All the analyses in this study were conducted in Weka [2], which contains implementation of all selected models.

Model selection (step 1):

Four binary classifiers were then trained using behavioural features to predict each engagement dimension in this phase: (i) Baseline, that always predicts the majority class (`Baseline`); (ii) Random Forest classifier with all behavioural features (`RF-All`); (iii) Support Vector Machines

[2]https://www.cs.waikato.ac.nz/ml/weka/

(SVM) with all the selected behavioural features; (iv) Random forest classifier with a wrapper method to select a subset of features that maximises the model's performance (RF-Wrapper) (Kohavi and John, 1997). We employ two learning frameworks that have been proven to be accurate and robust in multiple scenarios. Support Vector Machines (SVM) (Cortes and Vapnik, 1995) solve a linear maximal margin optimization problem. Coupled with transformations of the data by various kernel functions, SVMs have the power to fit complex decision boundaries. Random Forests (RF) (Breiman, 2001) are widely employed as one of the most successful ensemble bagging methods, resistant to overfitting but still able to capture intricate relationships between features. The wrapper methods consider how the selected algorithm and the feature set interact, and apply feature elimination to achieve best performances. We applied the wrapper methods to Random Forest models in order to test the improved prediction performance that can be achieved using subsets from selected behavioural features.

Model training and testing using cross-validation (step 2):

The prediction performances were obtained by 5-fold cross-validation with the CHiC dataset and 10-fold cross-validation with the wikiSearch dataset.

To measure the performance of the trained classifiers, we used the standard metrics of precision, recall and accuracy. However, metrics like accuracy can be deceiving in certain situations and are highly sensitive to the distribution of data (Powers, 2011). Therefore, we also computed the F-Measure ($\beta = 1$), which combines precision and recall as a measure of the effectiveness of classification, and AUC (area under ROC). Four of the five performance measures, namely precision, recall, F-Measure, accuracy, are reported as weighted averages as we applied cross-validation. A k-folder cross-validation paired t-test (Dietterich, 1998) was used to compare the three classifiers, (ii) RF-All, (iii) SVM, and (iv), RF-Wrapper against the Baseline classifier. The general null hypothesis (H0) of this test is that the mean difference between the performance of these two classifiers is equal to zero. The p-value of the test is the probability that the true mean difference between the two classifiers would be as large or more extreme than the actual observed difference, assuming the null hypothesis is true. In this study, we say the null

hypothesis is rejected when p-value is less than 0.05 and this two classifiers have significantly different performances. Subsequently, we compared the performance of the classifiers among four engagement dimensions in order to answer *RQ.3*.

4.4.2 Results and Discussion (Browsing)

Table 4.11: Performance metrics using four different classifiers(browsing).

	Performance	Baseline	RF-All	SVM	RF-Wrapper
PUs	(Weighted Avg.) Precision	0.279	0.552***	0.537***	**0.696*** **
	(Weighted Avg.) Recall	0.529	0.554	0.541	**0.675***
	(Weighted Avg.) F-1	0.366	0.552**	0.534***	**0.66*** **
	(Weighted Avg.) Accuracy	0.529	0.554	0.541	**0.675***
	AUC	0.471	0.558	0.534	**0.683***
FI	(Weighted Avg.) Precision	0.314	0.692***	**0.756*** **	0.738***
	(Weighted Avg.) Recall	0.561	0.694**	0.732***	**0.739*** **
	(Weighted Avg.) F-1	0.403	0.692**	0.717***	**0.738*** **
	(Weighted Avg.) Accuracy	0.561	0.694*	0.732***	**0.739****
	AUC	0.484	0.715	0.707***	**0.747*** **
EN	(Weighted Avg.) Precision	0.279	0.558**	0.633**	**0.695*** **
	(Weighted Avg.) Recall	0.529	0.561	0.624	**0.694***
	(Weighted Avg.) F-1	0.366	0.554*	0.609**	**0.693*** **
	(Weighted Avg.) Accuracy	0.529	0.561	0.624	**0.694***
	AUC	0.471	0.599	0.613*	**0.685***
NO	(Weighted Avg.) Precision	0.382	0.599***	0.71***	**0.738*** **
	(Weighted Avg.) Recall	0.618	0.611	0.707	**0.739****
	(Weighted Avg.) F-1	0.472	0.602**	0.681***	**0.738*** **
	(Weighted Avg.) Accuracy	0.618	0.611	0.707**	**0.739****
	AUC	0.489	0.691*	0.645***	**0.758*** **

A bold typeface denotes the best result in a row.

Significance level (2-tailed): $* = p<0.05$; $** = p<0.01$; $*** = p<0.001$).

Table 4.11 presents the performance metrics for each engagement dimension in comparison to four classifiers, and their significance level compared to the `Baseline` classifier. PUs was relatively the most difficult to predict, while FI, EN and NO were relatively easier. For PUs, the `RF-Wrapper` classifier was the best among all other four classifiers (compared to `Baseline`, F-measure improved by 29.4%, and AUC improved by 14.6%). For predicting FI, the `RF-Wrapper` classifier and the `SVM` were the best among all three classifiers (compared to `Baseline`, F-measure improved by 33.5% using `RF-Wrapper`, and accuracy improved by 17.8%

using `RF-Wrapper`), but the differences between these two classifiers was small. With respect to EN, the `RF-Wrapper` classifier again was the best among all three feature sets (compared to `Baseline`, F-measure improved by 32.7%, accuracy improved by 16.5%, and AUC improved by 21.4%). For NO, the `RF-Wrapper` classifier outperformed all three feature sets (compared to `Baseline`, F-measure improved by 26.6% and AUC by improved 29.6%; accuracy improved by 12.1%).

In general, (`RF-Wrapper`) classifier outperformed the other three classifiers in all dimensions, suggesting that predicting user engagement through behavioural features is possible in the browsing task.

4.4.3 Results and Discussion (searching)

Table 4.12: Performance metrics using four different classifiers (searching).

	Performance	Baseline	RF-All	SVM	RF-Wrapper
NO	(Weighted Avg.) Precision	0.337	0.507**	0.549	**0.633*** **
	(Weighted Avg.) Recall	0.581	0.531	0.581	**0.639**
	(Weighted Avg.) F-1	0.427	0.509*	0.461	**0.631*** **
	(Weighted Avg.) Accuracy	0.581	0.53	0.581	**0.639**
	AUC	0.493	0.531	0.507	**0.636****
FI	(Weighted Avg.) Precision	0.27	0.507***	0.527***	**0.54*** **
	(Weighted Avg.) Recall	0.52	0.509	0.531	**0.541**
	(Weighted Avg.) F-1	0.356	0.507**	0.52***	**0.539*** **
	(Weighted Avg.) Accuracy	0.52	0.509	0.53	**0.541**
	AUC	0.491	0.483	0.525	**0.551**
EN	(Weighted Avg.) Precision	0.492	0.555***	0.584***	**0.634*** **
	(Weighted Avg.) Recall	0.496	0.554	0.584	**0.634****
	(Weighted Avg.) F-1	0.445	0.554***	0.584***	**0.634*** **
	(Weighted Avg.) Accuracy	0.496	0.554	0.584	**0.634****
	AUC	0.493	0.578	0.584	**0.66****
PUs	(Weighted Avg.) Precision	0.319	0.581***	**0.646*** **	0.629***
	(Weighted Avg.) Recall	0.565	0.589	**0.637***	0.631
	(Weighted Avg.) F-1	0.408	0.579***	0.603***	**0.63*** **
	(Weighted Avg.) Accuracy	0.565	0.589	**0.637***	0.631
	AUC	0.488	0.601**	0.601**	**0.641****

A bold typeface denotes the best result in a row.

Significance level (2-tailed): $* = p<0.05$; $** = p<0.01$; $*** = p<0.001$).

Table 4.12 presents the performance metrics for each engagement dimension in comparison to four classifiers, and their significance level compared against `Baseline`. Overall, FI was relatively the most difficult to predict, while for PUs, EN and NO, our models significantly outperformed the `Baseline` in accuracy. As in the previous study we remark on a possible influence of the task type, as in searching, users are *required* to complete a clearly-defined task and thus they might lack the mood to further prolong the experience. For PUs, the `RF-Wrapper` classifier was the second best among all the other three classifiers, but F-Measure and AUC also showed improvement over the `Baseline` (compared to `Baseline`, F-measure improved by 22.2%, and AUC improved by 15.3%). The `SVM` classifier outperformed others for accuracy (improved by 7.2%). The performance of `RF-Wrapper` and `SVM` did not change much. For predicting FI, the `RF-Wrapper` classifier was the best among all three feature sets, but the differences were small (compared to `Baseline`, F-measure improved by 18.3%; accuracy improved by 2.1%, and AUC improved by 6%, but the differences were not significant). Similar results were reported for accuracy and AUC. With respect to EN, the `RF-Wrapper` classifier again was the best among all three feature sets (compared to `Baseline`, F-measure improved by 18.9%, accuracy improved by 13.8%, and AUC improved by 16.7%). Similar to FI and EN, the `RF-Wrapper` classifier outperformed among all others with moderate improvements in predicting NO (compared to `Baseline`, F-measure improved by 20.04%, and AUC by improved 14.3%; accuracy improved by 5.8%, but the difference was not significant).

In general, predicting engagement in searching using behavioural features is difficult. The explanation may be that as the user has a clearly-defined task to complete, all the users, regardless of their perception of engagement, might need to perform some similar interactions (e.g., search for the keywords in the task).

4.5 Comparison of results between browsing and searching

We compared the results obtained from study A, B and C between browsing and searching in order to answer the following research question:

RQ.4 What are the differences that user behaviour features exhibit in their relationship to user perception of engagement in browsing and searching?

Analysis on the two datasets provides information for comparing the differences caused by the two information contexts. We observed very different relationships between behavioural features and user perception of engagement in browsing and searching. This suggests that the interaction between behavioural features and user perception of engagement is, a priori, context dependant. This motivates adequate, context-aware feature selection and engagement modelling in subsequent studies. We now discuss the observations that lead to this answer in details:

Regarding the user perception of engagement data, the collected data in the browsing task is not as skewed as in the searching task (figure 4.4 and figure 4.10). These could be caused by the differences between the system, as the wikiSearch system is confirmed to be usable in completing complicated information retrieval tasks (Toms et al., 2013) and there was no similar test for CHiC system. It could also be caused by the type of the tasks. In the searching task, the users have one clear goal in mind, which is to complete a preassigned task. Being able to finish the task might bring them a general positive experience and thus allow them to perceive the system as helpful. This is in contrast with the setting in a browsing task, where the users do not have a goal to achieve. In browsing, three engagement dimensions (NO, FI and EN) are more similar to each other compared to PUs, while in searching, all dimensions seem to share some similarity but also exhibit some particularities. In general, based on the experimental results using the two datasets, the behavioural features which can describe user perception of engagement dimensions and how the behavioural features reflect the dimensions are different from each other.

Starting from the feature category, time-related features and result-related features are the best for predicting user perception of engagement in browsing as they occupied the top ranked behavioural features for NO, FI and EN (figure 4.6, 4.7, and 4.8), whereas a much diverse feature groups showed up in searching: time-related features and query-related features are relatively placed very high, but features from other categories also contribute. Time-related and query-related features have been discussed as effort in information retrieval for a long time (e.g.,(Kim et al., 2014; Azzopardi, 2011)). In a searching task, users have a defined goal. Regardless of whether the goal is correct, they need to issue queries or click on pages to achieve this goal, while this is not the case in browsing, where no clear defined goal exists at the beginning. Without that in mind, users are not in a hurry to acquire a specific piece of information, but just wandering in the collection.

Regarding the similarities between dimensions, all the four dimensions have a moderate relationship with each other in searching, whereas NO, FI and EN are more similar to each other compare to PUs in browsing. This is reflected first in the correlation analysis (table 4.3 and table 4.7), where PUs has moderate correlations with any of the three in searching, but only low correlations in browsing. Also, the differences between correlation coefficients of each engagement pair is smaller in searching ($0.35 > r > 0.63$) than browsing ($0.13 > r > 0.82$). This is further aligned with the feature contribution analysis (table 4.6 and table 4.10), in which the NO, FI and EN dimensions share at least 5 top ranked behavioural features (NO-FI: 9; NO-EN: 5; FI-EN: 6) in browsing, but just a few overlaps were observed between any pairs of engagement dimensions in searching. The big overlap between NO and FI in browsing supports the original design of the UES (O'Brien and Toms, 2010a), in which NO predicts FI as NO is users' assessment of the content, which may result in users' attention and more interaction with the content. NO is even likely to be the prior criterion of FI in browsing as the users become curious of the collection and then may interact more, which may lead to a feeling of being drawn in by the interaction, measured by FI in this case. As browsing is motivated by users' interests, their perception of novelty of the content and being involved directly indicates their overall feelings towards the session measured as EN. On the other side, as we discussed, searching is affected more by external factors, and thus an overall evaluation might include

more concepts rather than a single one and thus no big overlap was observed.

Dimensions NO, FI and EN have a stronger relationship with behavioural features than PUs in browsing and PUs and EN have a stronger link with behavioural features than FI and NO in searching. This might be because the engagement dimensions with stronger links are the main criteria by which the users characterise their interaction with the system.

Regarding the prediction task, the ability of predicting engagement using selected behavioural features in browsing is better than the ones in searching. This is not unexpected, as in the browsing context, the correlation coefficients between the engagement dimension and the search behavioural features are larger than the ones in the searching context. This might be caused by the nature of task. In the browsing context, the users are curiosity driven (Bates, 2007), and have more flexibility to explore the interface and content. They stop once they get bored (disengaged), and therefore the main influential factor of their behaviour is whether they enjoy the experience or not. While for the search task, the users do need to follow a certain pattern (e.g., input query, examining highly ranked results). Thus, the differences between individuals might be smaller and harder to observe. Furthermore, the search task is goal-driven. The users' behaviour is motivated by more external factors, such as finishing the given problem, rather than how they enjoy the experience only. These two reasons might make the engagement prediction in searching more difficult.

4.6 Summary and next steps

In phase 1, we reviewed two large sets of behavioural features used in literature as engagement proxies and grouped them into four categories (*RQ.1*). We further tested those behavioural features with user perception of engagement in browsing and searching, revealed the differences among four engagement dimensions, and demonstrated the possibility to predict user perception of engagement by behavioural features. We answered four research questions.

Our results, in both non-purposeful browsing context and goal-based searching context, support the connection between user perception of engagement and user behaviour (*RQ.2*). This is fur-

ther demonstrated in the prediction task of user perception of engagement via user behavioural features.

More specifically, we show that time and result-related features are best suited for predicting user perception of engagement in browsing, and time, click and query-related features are the best for predicting user perception of engagement in searching. This also answers (*RQ.4*).

We also found that different behavioural features better reflect specific dimensions (*RQ.3*), which support the multi-dimensional definition of engagement. Moreover, the relationships of the four selected dimensions might be different in different contexts, and each dimension may serve as the principal descriptor of engagement accordingly. This suggests that it will be beneficial to be selective and specific about which dimension is measured in further studies.

Chapter 5

Phase 2: Behaviour Sequence - Engagement Relationship

5.1 Overview

This chapter reports the second phase of this research, in which we investigated how the sequential information of user behaviour are linked with different dimensions of user perception of engagement. The phase was designed to answer four research questions:

RQ.5 What is the most general set of actions which suffices to describe user interaction with information retrieval systems?

RQ.6 What is the relationship between user behaviour sequences and user perception of engagement?

RQ.7 How do the relationships between user behaviour sequences and user perception of engagement vary between dimensions?

RQ.8 How do the relationships between user behaviour sequences and user perception of engagement differ between browsing and searching?

In the phase 1 (chapter 4), we identified and extracted key behavioural features correlated with user perception of engagement in both browsing and searching contexts. However, user

interactions with the system are dynamic and thus the sequence of interactive actions contains additional information from the static set of behavioural features relevant to predicting user perception of engagement. The value of considering behaviour sequences has been outlined by previous studies such as satisfaction prediction (Mehrotra et al., 2017). This information is challenging to fully capture by discrete behavioural features alone. By an *action* undertaken by a user in their information retrieval process we mean a representation of a *physical operation undertaken on or in relation to the system and its outputs*. This implies that each action is a part of a more ample physical process which we call the *interaction* of the user with the system. Each action comes a type (such as clicking a result) and comes equipped with a time-stamp for its commencement. We first define the terminology in section 5.2 surrounding *behaviour sequences*.

This phase is comprised of three studies:

Study A. Behaviour sequence extraction, in which we focus on extracting a set of behaviour sequence from system log files that describe users' information retrieval process. We first ground our selection based on the ISP model (Marchionini, 1995) and common IR system interface. The selected actions mainly captures user interacting with the search box, SERPs and document pages by clicking and querying, which belongs to the behavioural features identified to have a strong relationships with user perception of engagement. This follows naturally from phase 1 of our study (chapter 4). Behaviour sequences were formed by first extracting action from the system log files and then concatenating consecutive equivalent actions. ($RQ.5$)

Study B. Behaviour sequence analysis, in which we use the engagement labels assigned in phase 1, study B and the behaviour sequences extracted in phase 2, study A in order to analyse and discuss around various informative sequential patterns that can differentiate high and low engagement using χ^2 tests of independence. ($RQ.6$, $RQ.7$)

Study C. Engagement prediction using behaviour sequences, in which we further test our intuition of the usefulness of sequential patterns through formulating a classification problem on the engagement labels of user groups and test the improved performance of adding features based on the sequential patterns identified in phase 2, study B to behavioural features

extracted in phase 1, study A. ($RQ.6$, $RQ.7$)

Finally we compare the three sets of results on the two datasets corresponding to our two information retrieval contexts, browsing and searching, respectively, and outline how the differences in our results reflect the inherent disparities between the contexts themselves. ($RQ.8$)

Figure 5.1 presents the overarching design of this research phase.

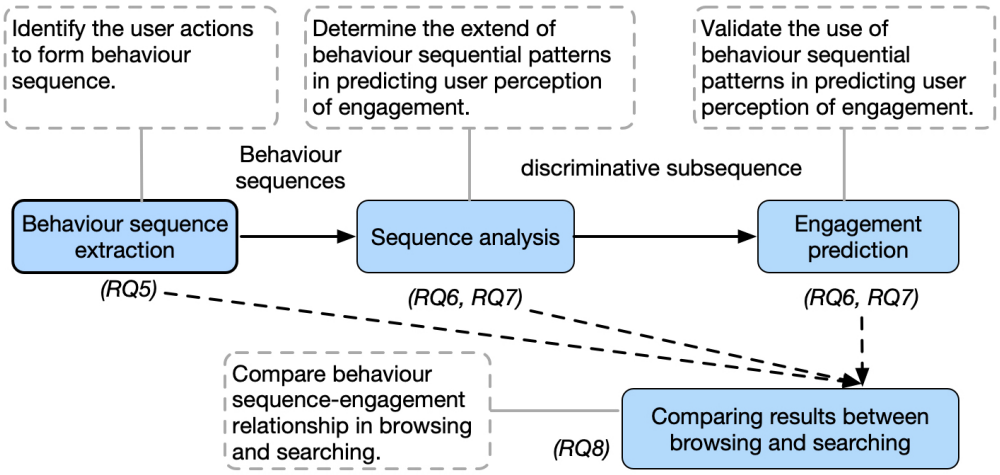


Figure 5.1: Design of research phase 2, which contains four main parts: behaviour sequence extraction (section 5.3), sequence analysis (section 5.4), engagement prediction (section 5.5) and results comparison (section 5.6).

In this section, and throughout the text, we will denote by actions and behaviour sequence this numerical footprint of the users interaction as observed within system log files. The definitions, measures, approaches, and findings are presented below in separate sections, one for each study. The analyses were conducted using both datasets. The final analysis compares results from both datasets, resulting in a thorough discussion of the differences between the two settings. Outcomes from this phase lead to the identification of the key behaviour sequential patterns for each engagement dimension, as well as evidence to suggest properties of an ideal measure of engagement based on behaviour sequence.

5.2 Definitions

The purpose of the section is to provide clear definitions for the concepts and terms mentioned in this chapter, and formal mathematical formulas which help to explain the data analysis in the following studies. We relate the definitions of critical terms to an example in the introduction (figure 1.1, section 1.1, chapter 1) to further elucidate them in the context of IR process. In study A of this phase (section 5.3.2), along with data extraction and preparation, we also provide a more detailed example based on our study.

We first defined **action** as following:

Action: In this study, an action is any physical input that the user generates and passes to the system that and can be recorded via the use of system logs throughout the duration of the task. This includes pressing keys, and clicking the mouse buttons. In this chapter, where not otherwise specifically indicated, A will refer to a set of actions, i.e. a collection of elements of the above type. Each action from A may occur at multiple time instances, and hence, will arise in practice, labeled by a unique timestamp which together with its label as an element in A acts as a unique identifier. However, in this research phase, timestamps are only used to order the actions.

Next, we present four general terms regarding *sequence analysis*:

Sequence: A finite sequence of elements from a set X is a function $f : [n] \to X$, where $[n] = \{1, 2, 3, ..., n\}$. We say that n is the *length* of the sequence. Let $S(X)$ denote a set of sequences of elements from X, and write $S(X) = \{s_i\}_{i=1}^{|S|}$.

Subsequence of a sequence: A subsequence of a sequence f as above is the sequence defined as $f \circ g$ where $g : [m] \to [n]$ is a strictly increasing function. We write $s_j \leq s_i$ to mean that a sequence s_j is a subsequence of a sequence s_i. Note that each sequence is a subsequence of itself.

Frequent subsequence: Given a predefined set of sequences $S(A)$ of elements from A, lets denote s_j as a subsequence of at least one sequence s_i in $S(A)$. The support of s_j in a set $S(A)$

is the frequency of the sequences that contains s_j in the set, $\text{supp}_S(s_j) := \frac{|\{s \in S(A) | s_j \leq s\}|}{|S(A)|}$. We say s_j is frequent if $\text{supp}_S(s_j)$ is larger than a defined threshold.

Discriminative Subsequence: Given a frequent subsequence s_j of set $S(A)$, and class labels of each s_i in $S(A)$, we say the s_j is a discriminative subsequence of $S(A)$ if significant relationship is found between the presence of s_j and the class label by performing the χ^2 test of independence. The details of performing the χ^2 test is described in the methods section later (section 5.4.1).

We now look at an example that revisits the concepts above and hopefully sheds light on their usefulness. The set $A = \{a, b, c, d\}$ consists of 4 different actions. Infinitely many sequences can be formed out of these action labels. Suppose we are given a set of sequences $S(A)$ of size two, formed by the elements in A, say $S(A) = \{s_1, s_2\}$. Suppose now that $s_1 = ab\mathbf{bcca}d$, and $s_2 = adb\mathbf{bcca}dd$. As an example of subsequence s_3 of s_1 we take $bcca$; notice that it is also a subsequence of s_2. The support value of s_3 on set $S(A)$ is 1, $\text{supp}_{S(A)}(s_3) = 1$, as it is a subsequence of all the sequences contained in $S(A)$ ($s_3 \leq s_1$, $s_3 \leq s_2$).

As our analysis focuses on user behaviour, we define a **behaviour sequence** as:

Behaviour sequence: A finite sequence of elements from the set of actions A, where consecutive equivalent actions are concatenated. The index of the sequence will always be in the increasing order of the associated timestamps. Thus a sequence fully represents a user's interaction for the session, making the correspondence between users and behaviour sequences for this study one-to-one. Note that the even though the *(timestamp, action)* pairs are unique, each action may be repeated multiple times within one user interaction.

We choose to merge consecutive equivalent actions because it reduces the number of redundant subsequences, and extracts patterns at a higher level. Even though this procedure has the potential to lose some information about the full action sequence, we employ this reduction in order to focus on broader, more interpretable patterns, rather than micro-scale variations in the action sequence. We have also tried other coding of the sequence (e.g., only merge three or more consecutive equivalent actions), but this method reserves most comfortable format for interpreting user's interaction, and we reserve consideration of other sequence coding for further

studies.

We now revisit the example of A and $S(A)$ above. By concatenating consecutive equivalent actions, the two sequences in $S(A)$, s_1 and s_2, become behaviour sequences $\bar{s_1}$ and $\bar{s_2}$, where $\bar{s_1} = abcad$, and $\bar{s_2} = adbcad$. It should be apparent to the reader, without the need for a formal proof that, for any two sequences, s and v, if $s \leq v$ then also $\bar{s} \leq \bar{v}$, the only confusion appearing at the beginning and end of the sequence where we have to accept $a \leq aa$ implies $\bar{a} \leq \bar{a}\bar{a} = \bar{a}$ without equality.

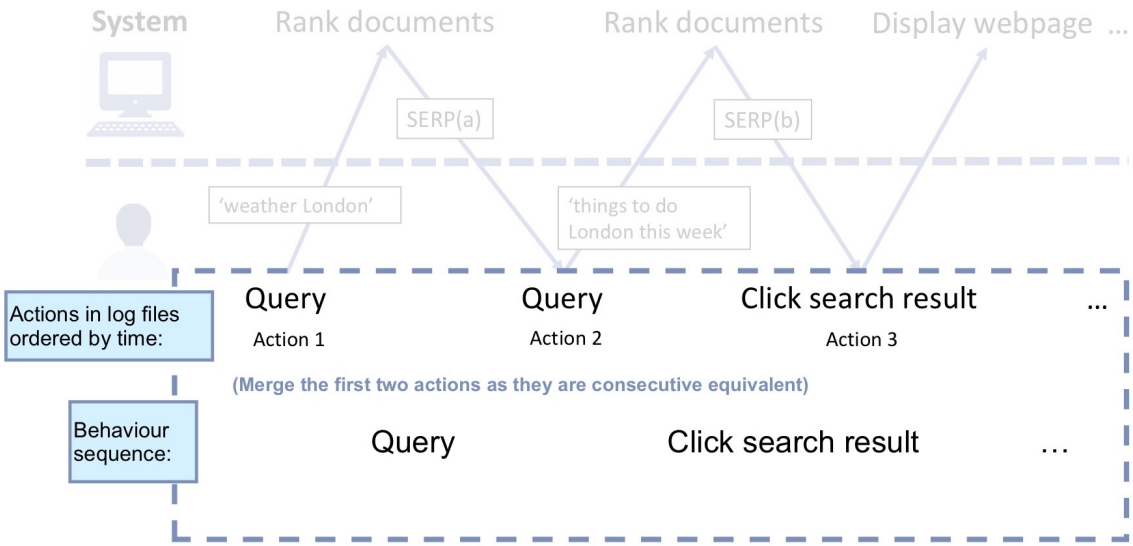


Figure 5.2: Example of actions and behaviour sequence.

To illustrate this point further, we return to a previously discussed example (figure 1.1) of a search process provided in the introduction (chapter 1) and extract a behaviour sequence from the process (figure 5.2). The user queries the weather in London online, then searches for activities of interest in London during the current week and clicks the first link on the returned SERP. The three actions were logged in the system log files and ordered by their associated timestamps. The action sequence representing these three actions is `Query - Query - Click search result`. Then, in order to construct a behaviour sequence, we merge the two query actions as they are consecutive and equivalent, and obtain a behaviour sequence `Query - Click search result`. The concatenation of a different action sequence `Query - Query - Query - Click search result` is indistinguishable from that of our previous sequence under our definition. Note that the equivalence of two behaviour sequences cannot identify the user

behaviour uniquely.

5.3 Study A. Behaviour sequence extraction

The current section is devoted to the first of our three studies in this phase. The purpose of this study is to identify the user actions from information retrieval process models and common IR system interface components to form behaviour sequences. We start by discussing the selection criteria, presenting the selected actions, and extracting the behaviour sequence from log files in order to prepare data for the next analysis in this phase. Figure 5.3 presents the steps, data and variables used in this section. All the analyses in this study were written in R.

We restate below the research question associated to this study:

RQ.5 What is the most general set of actions which suffices to describe user interaction with information retrieval systems?

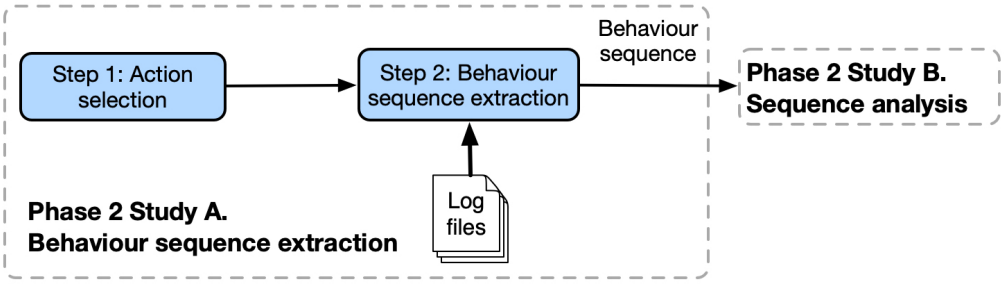


Figure 5.3: Steps, data and variables used in section 5.3.

5.3.1 Actions selection (step 1)

In this section, we first discuss the criteria we use to select the actions which best describe user interaction with information retrieval systems. As found in chapter 4, user behaviour captured by time-, click-, query- and result- related behavioural features are informative in predicting user perception of engagement. We select the actions that can capture users' click and query behaviour that interact with the SERPs or document pages. We also select the ones that fit

into the Information Seeking Process (ISP) model (Marchionini, 1995), and are available in the common IR interface components to increase the generality of our study. The resulting sets of actions are then contrasted in the two information retrieval contexts, browsing and searching, with the purpose of uncovering characteristic differences between the two.

Selection criteria

In order to address the research question above, actions were chosen to be used in our research based on three criteria:

1. The actions must fit into the sub-processes described in the ISP model (Marchionini, 1995) (see detailed discussion of this model in literature review (section 2.3.2)).
2. The actions capture users click and query behaviour that interact with the SERPs or document pages.
3. The actions are available in the common interface components for both browsing and searching.

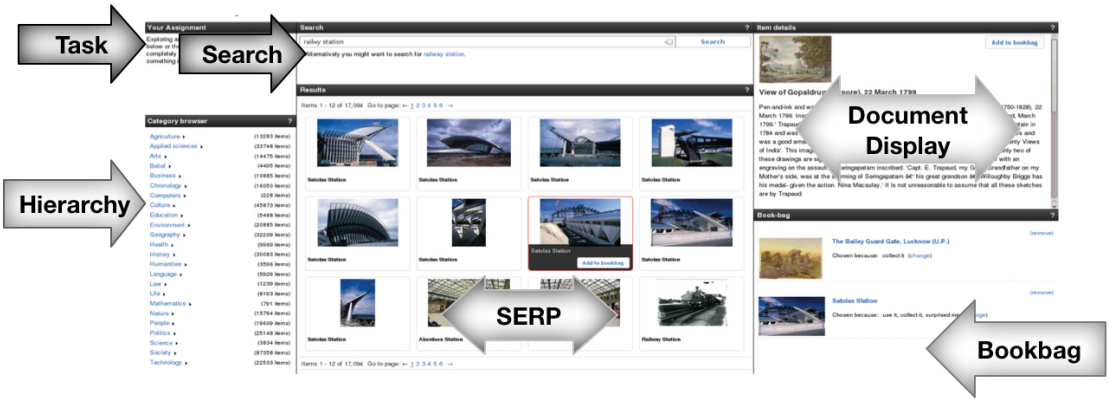
The rationale behind choosing the ISP model (Marchionini, 1995) resides in the fact that the model captures both searching and browsing contexts, and breaks the process into a set of eight key subprocesses that a user progresses through: "Recognize/Accept","Define problem", "Select source","Formulate query", "Execute query","Examine results","Extract information", and "Reflect/Stop". The iteration of these subprocesses is different in each of the two contexts which Marchionini (1996) describes as discovery and search: "Search implies an effort to locate a known object; the information seeker has in mind specific characteristics or properties of the object and these characteristics are used to specify and guide search activity. Discovery implies an effort to explore some promising space for underspecified or unknown objects; the information seeker has in mind general characteristics or properties that outline an information space in which perceptual and cognitive powers are leveraged to examine candidate objects." (p.1). Differences between search and discovery in terms of user behaviour are explained as contrasting strategies: on the one part searching, in which strategies are formal, and planned,

while browsing strategies are informal, casual, and opportunistic. We select the actions based on the ISP model (Marchionini, 1995) and capture the transitions by the order of actions.

A further motivation in selecting sets of actions that are common to both IR tasks that are available to the user in both interfaces is to increase the generality of our study. A fine-grained summarization of behaviour will certainly provide more information, but also might need extra functions to capture (e.g., enabling Javascripts in the browser to record mouse motifs), which may need user consent, and are not always available in ordinary use - cases. Although in these two datasets, CHiC and wikiSearch, we collected very fine-grained behaviour from users, we decided to model their interactions at a general level. Moreover, interactions associated with common user interface components (e.g., SERP, document display) are widely considered and proved to be useful in predicting user perception (e.g., (Ponnuswami et al., 2011; Thomas et al., 2014)). Figure 5.4 displays the interface components considered for the two datasets.

Also, we conceptually divided all actions into one of two types: *exploration* and *immersion*. This is inspired by Bates (2007)' definition of browsing, and the IR researchers treat CTR as a signal of success, that the trigger of looking into certain objects closely as a sign of potential interest. Thus the action before and after that point, should be analysed differently. We also base our discussion around this point. Exploration actions are the ones which examine an information space such as creating queries, examining results lists, etc. On the other hand, immersion actions are the ones which relate to comprehension and conceptualization: reading and comprehending the contents of an information object, such as reading the document page fall into this category. This distinction will become apparent in the discussion of the results of sequence analysis (section 5.4.3, and section 5.4.5).

In addition to the selected actions, we use 'Start' to mark the start of a task, which differentiates the behaviour sequences happen right after the beginning from the one happens in the middle. As 'Start' only appears once for each user participating in the browsing session and exact three times for the users taking the searching session as each of the user was assigned three tasks, we did not report the statistics. In order to tailor to fit the two different information retrieval contexts, different sets of actions were extracted for each.



(a) CHiC (browsing)

(b) wikiSearch (searching)

Figure 5.4: Interface components considered for two datasets.

Selected actions

Table 5.1 presents 11 actions (A), which were considered and used to form behaviour sequences. The IR interface components associated with each action are also listed in table 5.1.

Seven actions are selected for browsing and nine actions are selected for searching. Selected actions cover a wide range of interactions and mainly mapped into five out of eight subprocesses in the ISP model (Marchionini, 1995). Table 5.2 presents the connections between the selected actions and all the eight ISP sub-processes and the two action types, *exploration* and *immersion*. Following is a description of the actions selected based on their groups:

- First three subprocesses "Recognize/Accept","Define problem", and "Select source" of the

Table 5.1: Description of actions selected, with related interface components.

Action	Description	Start	Return	System	
				CHiC	wikiSearch
`Query`	Type text in the search box.	Search	SERP	✓	✓
`Click_Hie`	Click a link in the hierarchy.	Hierarchy	SERP	✓	
`Click_HisQue`	Click on a query in the history section.	History	SERP		✓
`Click_HisRes`	Click on a result in the history section.	History	Doc. Display		✓
`Click_NextS`	Click next page on the SERP.	SERP	SERP	✓	✓
`Click_Nav`	Click a navigation button to view viewed document.	Doc. Display	Doc. Display		✓
`Click_SERP`	Click a result from the SERP.	SERP	Doc. Display	✓	✓
`Click_Link`	Click an URL on the document display.	Doc. Display	Doc. Display	✓	✓
`Click_Metadata`	Click a metadata on the document display.	Doc. Display	SERP	✓	
`Click_Bookbag`	Click a saved URL from the bookbag.	Bookbag	Doc. Display	✓	✓
`Answer`	Answer the assigned question.	Answer	Answer		✓

✓means the corresponding behavioural features is used for this domain.
Column *Start* refers to the interface component the action happens on.
Column *Return* refers to interface components the system updates in response to the user action.
Column *System* indicates the context to which each action pertains.

ISP model mainly contain cognitive activities. "Recognize/Accept" and "Define problem" can not be captured by the system log files easily. "Select source" can be monitored by the switch between systems in the real world. But, as this research focuses on intra-session engagement, and selected the datasets with the information retrieval systems provided to the participants for completing instructed tasks, no action was selected for these three subprocesses.

- The subprocesses "Formulate query" and "Execute query" can be observed by the users typing in queries in the search box, or picking the queries suggested by the system. In browsing, two actions are associated with these two sub-processes: inputting query (`Query`) and clicking an item from the built-in hierarchy (`Click_Hie`). Clicking on a hierarchy item will be treated as a query request, and the system will return a SERP in response to the content of the clicked item. Only `Query` is selected for searching as the system does not have the hierarchy function. All the actions in this group are classified as *exploration* as they serve a common purpose, requesting a SERP from the system.

- Subprocesses "Examine results" and "Extract information" happen during users' interaction

Table 5.2: ISP model (Marchionini, 1995) and corresponding actions selected.

ISP subprocess	**Actions** CHiC (browsing)	wikiSearch (searching)	**Type**	**Notes**
Recognize/Accept Define problem Select source	N/A	N/A	N/A	N/A
Formulate query	`Query`	`Query`	Exploration	Issuing a request to get a related SERP.
Execute query	`Click_Hie`		Exploration	
Examine results	`Click_NextS`	`Click_NextS`	Exploration	Issuing a request to get more ranked documents or view details of one document.
Extract information	`Click_SERP`	`Click_SERP`	Immersion	
	`Click_Link`	`Click_Link`	Immersion	
	`Click_Metadata`		Exploration	
Reflect/Stop	`Click_Bookbag`	`Click_Bookbag`	Immersion	Revisiting a document/SERP, or summarizing the information absorbed.
		`Click_HisQue`	Exploration	
		`Click_HisRes`	Immersion	
		`Click_Nav`	Immersion	
		`Answer`	Immersion	

with two main interface components: the SERP and the document display. An example of typical interaction is the user examining the ranked documents displayed in the SERP, which includes the title or sometimes a short summary of each document, and deciding which document to click on. Once the user clicked a document on the SERP, the system will display the linked page. In browsing, four actions are associated with these two sub-processes: 1, clicking the next page option on SERP (`Click_NextS`), followed by the system displaying the next n documents, ordered by the system in response to the same query; 2, clicking one document on the SERP (`Click_SERP`); 3, clicking a hyper link on the document (`Click_Link`), followed by the system displaying the linked document; 4, clicking the metadata on the document page (`Click_Metadata`), followed by the system returning a SERP in response to the metadata. For wikiSearch, `Click_Metadata` is not selected as the system does not have the metadata function. `Click_NextS` and `Click_Metadata` are *exploration* type actions, as a new SERP will be returned and the user is still as the stage of examining the relatively shallow information to locate the more detailed one. All the other actions in this group are classified as *immersion* as they lead to a closer look at the content of the document.

- For the final subprocess "Reflect/Stop", clicking the document page saved in the book bag (`Click_Bookbag`) is selected for browsing, whereas for searching, three more actions are se-

lected. This is made possible by the wikiSearch interface's particular functionality which implements a history display (figure 5.4) highlighting the queries the users issued and the documents the users clicked on. Two actions were tailored to capture the associated interaction: 1, clicking on a query in the history section (`Click_HisQue`), and followed by the system returning a SERP in response to the query. 2, clicking on a document in the history section (`Click_HisRes`). The users can also flip the documents they viewed by clicking the navigation buttons in the document display area (`Click_Nav`). Action (`Click_HisQue`) belongs to the *exploration* category as it leads to a SERP being viewed by the participants, and the other actions in this group belong to the *immersion* category as they either present the participant with the content of a document (e.g., `Click_Bookbag`) or indicate the participant is comparing (e.g., `Click_Nav`) or summarizing (e.g., `Answer`) the information.

Differences between selected actions for browsing and searching.

The sets of actions selected for each dataset differ to a moderate extent (table 5.2). Five actions, `Query`, `Click_SERP`, `Click_NextS`, `Click_Link`, and `Click_Bookbag` appear in both sets. `Query`, `Click_SERP`, and `Click_NextS` are the very basic information retrieval actions, while `Click_Link` is associated with the hyperlink function in the document, and `Click_Bookbag` is associated with the bookbag function. Both of these functions are expected to assist the user in viewing relevant documents or revisiting saved documents. In browsing, the hierarchy function is designed to assist the user in case she runs out of ideas or wants to visualize documents by category. The metadata function also serves a similar purpose, allowing the user quick access to similar documents with the same metadata. Thus, `Click_Hie` and `Click_Metadata` are selected. The reason for designing a history function in the searching is to allow fast access to the SERPs or documents the user viewed. As users were given a binary decision-making task, they might want to make comparison between document pages, and thus return to the previous pages. Three actions, `Click_HisQue`, `Click_HisRes`, and `Click_Nav`, were selected to capture this type of behaviour. `Answer` is the action users perform when submitting their response to the assigned question, suggesting that the user has found enough information to complete it. Although the two sets of actions are not identical, they are tailored to the information retrieval

context.

5.3.2 Behaviour sequence extraction (step 2)

Table 5.3: Example of user actions ordered by timestamps and associated behaviour sequence.

User's actions ordered by time: `Click_Hie` → `Click_NextS` → `Click_NextS` → `Click_NextS` → `Query` → `Click_SERP`
Behaviour sequence s_1**:** `Click_Hie` → `Click_NextS` → `Query` → `Click_SERP`

We recall the definition of a set $S(A)$ of behaviour sequences based on the selected actions in section 5.3.1. We further present an example of behaviour sequence that we hope sheds some light on the a priori assumptions of our analysis. Table 5.3 shows details of one user's actions ordered by time, and the associated behaviour sequence based on the actions defined in the table 5.1. As is apparent, the user clicked a link from the hierarchy, then checked three SERPs returned by the system in response to the query request. Subsequently, the user inputted a query in the search box and clicked one document from the returned SERP. Note that action `Click_NextS` appears three times in a row, which suggests the user is broadly examining the SERPs without investigating a specific document. In essence, these actions are homologous from the stand point of our analysis, due to our aim of capturing not fine grained document - level information, but broad strokes in behaviour patterns. In this instance it should be apparent that the order in which these three actions are performed is not essential to the user (nor is it clear whether the ordering was the product of the user's volition or merely a consequence of the search system's relevance rating). It may be the case that the user was always intending to check out the three SERPs or indeed his decision to switch to a (`Query`) was motivated by information extracted along the third SERP making the investigation of the first two superfluous. We factor out this uncertainty and normalize our data by concatenating the three `Click_NextS` actions into one. In fact, we are saying the user clicked *some* SERPs, then proceeded with his investigations. Another motivating factor, though secondary in this case, is that the number of actions, and dwelltime spent on each type of action are partially examined in research phase 1 (chapter 4). Here we aim to extract new information from high-level sequential patterns without duplication of effort and overlap of means. Finally, we

extracted 157 behaviour sequences for browsing, and 377 behaviour sequences for searching.

5.4 Study B. Sequence analysis

This section reports study B of this research phase. The purpose of this study is to determine the extent to which behaviour sequence patterns (extracted from phase 2, study A) contribute to predicting user perception of engagement, represented through the engagement labels assigned in phase 1, study B. We gauge these statistically significant relationships through a χ^2 test of independence that examines whether there is a significant association between the presence of *frequent subsequences* (definition in section 5.2) and engagement labels. The frequent subsequences are extracted from the behaviour sequences (extracted from phase 2, study A) using a sliding window method. Figure 5.5 presents a pictorial description of the steps, data and variables used in this section. We restate below the questions associated to this study:

RQ.6 What is the relationship between user behaviour sequences and user perception of engagement?

RQ.7 How do the relationships between user behaviour sequences and user perception of engagement vary between dimensions?

5.4.1 Method

In this study we focus on analysing behaviour sequences extracted from phase 2, study A through three steps, namely descriptive statistics, frequent subsequence extraction and discriminative subsequence extraction. In total, 157 behaviour sequences are formed to represent user interaction in browsing and 377 for searching. Each behaviour sequence has associated engagement labels obtained from phase 1, study B. All the analyses in this study were written in R[1].

[1]https://www.r-project.org

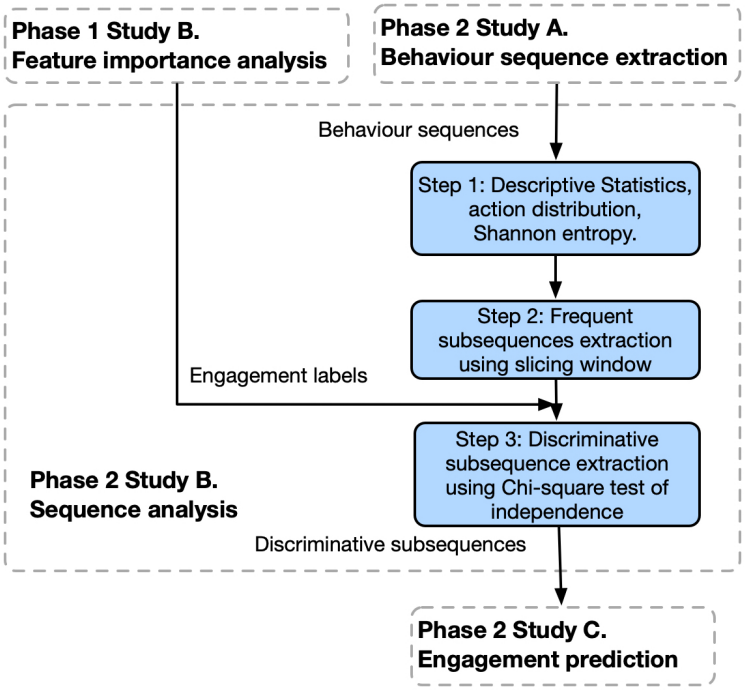


Figure 5.5: Steps, data and variables used in section 5.4.

Descriptive statistics, action distribution, and Shannon entropy (step 1)

We proceed by reporting the statistical description of the set of behavioural sequences. For each user and for each of the two datasets we assign a single sequence describing the interaction for the entire session. For example $S(A_{Browsing})$ has 157 elements corresponding to each user ID and indexed by the order of their interactions, and $S(A_{Searching})$ has 377 elements.

First, we report the average occurrence of each action. For a set $S(A)$ of sequences and for $a \in A$, we define the mean of action a, $\mu(a)$, to be the count of occurrences of a among all sequences in $S(A)$, counted with multiplicity, and divided by the size of $S(A)$. For example, if $A = \{a, b\}$ and $S(A) = \{aba, a\}$ then $\mu(a) = 1.5$.

Furthermore we present the positional distribution of each actions across the process. We define the positional distribution of actions at step t to be the vector-valued function which, at index t outputs the probability that a user is undertaking a certain action, for all actions in A. This statistic elucidates two points. Firstly, it allows us to observe the rate with which users complete their sequences. Effectively, this is the function which, at step t, outputs the probability that the user is still performing actions (any action) and has not stopped. A normalized scaling of this

function is in fact a certain probability distribution itself: the distribution of the stopping time of a behaviour sequence. Heuristically this is expected to correspond to a distribution located on some exponential family with parameter λ, due to the analogy one can make between stopping time for such sequences and the stopping time for a sequence of Bernoulli trials with constant probability of success: p (each action in the sequence can be modeled as an independent trial, stopping occurring when an answer to the task objective is attained and the user reports completion, respectively when the trial is successful). An unbiased maximum likelihood estimation (MLE) for $1/\lambda$ is the empirical mean of the distribution (Ross, 2014). A simple test of this estimator produces a 95% confidence interval of $(\hat{\lambda}(1 - \frac{1.96}{\sqrt{N}}), \hat{\lambda}(1 + \frac{1.96}{\sqrt{N}}))$ in this case, where N is the number of data points.

We further report the entropy (Shannon, 1948) of the positional distribution of actions at each step t, as a measure of uncertainty, or *lack of accord* between the user's behaviours. We expect this statistic to start out rather low and increase to a maximum rather quickly before decaying to 0 as the users exhaust their actions and the tails of the sequences become more predictable.

Frequent subsequence extraction (step 2)

We proceed to extract frequent subsequences in order to identify the common patterns that occur in the behaviour of our users in relation to the system. We use sliding windows with size greater than 2 across all sequences in $S(A)$ to generate potential candidates for frequent subsequences. All subsequences extracted this way have non zero *support* (definition and formula in section 5.2), a length of at least 2, and consist of *consecutive* actions from their original sequence. For example, `Click_Hie` $\rightarrow$ `Click_NextS` is a subsequence of s_1: `Click_Hie` $\rightarrow$ `Click_NextS` $\rightarrow$ `Query` $\rightarrow$ `Click_SERP` in the previously mentioned behaviour sequence example (table 5.3). We avoid enforcing a maximum window length in order to capture a larger variety of candidates. Notice that in this study this is feasible due to the relatively small number of datapoints. In the general case one obtains $O(Nn^2)$ subsequences where N is the number of datapoints and n is the maximum length of a sequence. An alternative, more computational expensive, way to generate the subsequences is to allow actions to be non-consecutive with a predefined

maximum gap between 2 actions, or a predefined maximum gap in the whole sugsequence (to reduce sparsity), while still preserving the order of actions in the original sequence. In this case, `Click_Hie`→ `Query` would be a subsequence of s_1. As concatenating actions can obscure the gaps created between non-consecutive actions and thus unnaturally identify patterns that in fact represent very different interactions with the system, we depart from exploring this method further at this point. A study in which more granular feature extraction is desired across a dataset ample enough to accommodate adequate sparsity issues could make use of our latter technique.

To highlight patterns that occur at the start of a sequence (and hence the user's interaction), the start of a session is included as an action with length 1. This will help differentiate patterns that appear at the start of a session from identical ones in the middle of a session and are especially relevant for subsequences of small lengths. As an example consider `Start`→ `Query`→ `Click_SERP` and `Query`→ `Click_SERP`. Although these two sequences represent the user clicking a document from the SERP returned in response to her query, we suspect they indicate user perception of engagement very differently as the user having a query in mind right at the start of a session may indicate the user is motivated by a self-formulated goal, while the other might suggest a motivation acquired in a serendipitous fashion during the session.

The support of a subsequence s_k in $S(A)$ is $\text{supp}_S(s_k) := \frac{|\{s \in S(A) | s_k \leq s\}|}{|S|}$ (detailed definition in section 5.2). We say s_k is frequent if $\text{supp}_S(s_k)$ is larger than a defined threshold, which is 0.05 in this study. Varying this threshold would produce a different filtration on the subset of frequent subsequences, but here, a low threshold is used in order to accommodate as many candidates as possible. The subsequences with low support values will be penalized in the computation of their discriminatory power at a later stage. We hope the introduction of the notion of frequent subsequence is self explanatory in the attempt to capture essential discriminatory high level behaviour patterns between users. In total, we extracted 418 frequent subsequences for browsing and 662 frequent subsequences for searching.

Discriminative subsequence of engagement extraction using χ^2 test of independence (step 3)

The purpose of this step is to determine the extent to which each frequent subsequence contributes to predicting user perception of engagement. Given the 418 frequent subsequences extracted for browsing and the 662 frequent subsequences extracted for searching, and the engagement labels (*high* and *low*) of each behaviour sequence in $S(A)$, we use the Chi-square (χ^2) test of independence to analyse the most discriminating patterns with respect to the labels, in order to answer *RQ.6*. High and low engagement labels are assigned in phase 1 study B.

Our measure of discriminatory power of each sequence will be its χ^2 score. Generally, the χ^2 test of independence is used to test the relationship between two groups. The general null hypothesis (H_0) for this test is there is a no detectable differences in the frequency of the variable between these two groups, which, in our case translated into there not being a significant difference in the presence of a given subsequence between the user groups with high and low engagement respectively. We are interested in the subsequences that reject the null hypothesis (H_0) in the test, which means that the presence of a certain subsequence is significantly correlated to the engagement labels. Thus, those subsequences are the patterns that are indicative of high and low engagement. Given the candidate set, which contains the frequent subsequences generated from the previous step, we computed the χ^2 score for each frequent subsequence s_k based on:

$$\chi^2_{s_k} = \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}\,, \tag{5.1}$$

where O denotes the observed counts of s_k in $S(A)$; E denotes the expected counts of s_k; n denotes the number of class labels, which is 2 for high and low engagement labels, m denotes the number of possible statuses, which is 2 for with/without the subsequence s_k. The degree of freedom is 1. A zero $\chi^2_{s_k}$ score means there is no difference between the two classes in terms of s_k, and thus the H_0 is not rejected. After collecting the test statistics we display the frequent subsequences ranked by their discriminatory power. The test also provides detailed information on exactly which categories account for any differences found by Pearson residual,

calculated by $\frac{(O-E)}{\sqrt{E}}$, for each group. To conclude, top ranked subsequences, which we refer to as discriminative subsequences for the corresponding engagement dimension, were compared with respect to the four different engagement dimensions in order to answer *RQ7*.

5.4.2 Results (browsing)

This section described the results of study B using the CHiC dataset. We first present the descriptive statistics of behaviour sequences, and then report the frequent subsequences extracted from the behaviour sequences. We report the discriminative subsequences selected through the frequent subsequences using χ^2-test of independence.

Descriptive statistics of behaviour sequences

After extracting the actions listed in table 5.1 and forming 157 behaviour sequences, we present the descriptive statistics of this set of sequences.

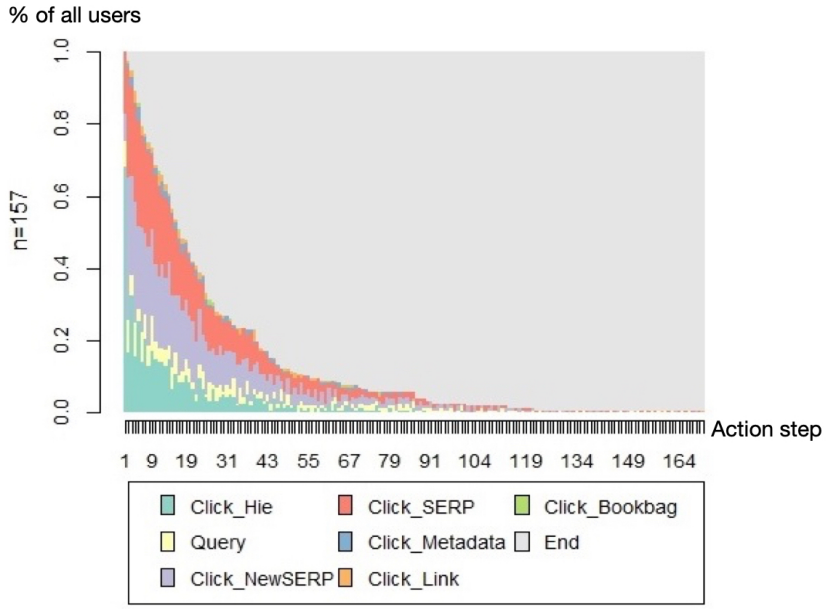


Figure 5.6: Distribution of behaviour sequences (browsing).

All users issued at least 2 actions, while the sequence of maximum length among these users was comprised of 172 actions. The average number of actions issued per user is 25.99 (SD=26.16), with a median number of 18 (the 1st quartile (Q_1): 9, the 3rd quartile (Q_3): 33). Figure 5.6

presents the actions distribution by the order they were issued in the behaviour sequences. The gray area represents the proportion of participants who left the session, marked as 'End' in figure 5.6. The parameter (λ) of the exponential distribution lies within the following 95% confidence interval: (0.0325, 0.0445).

Table 5.4: Descriptive statistics of actions in the behaviour sequences (browsing).

Action	**Mean**	**Median**
`Query`	2.53	1
`Click_Hie`	4.88	3
`Click_NextS`	8.29	5
`Click_SERP`	8	5

The average number of actions of exploration type, namely `Click_Hie`, `Click_NextS` and `Query`, is 15.7, and for the immersion type, namely `Click_SERP`, `Click_Link`, `Click_Metadata`, and `Click_Bookbag`, it is 9.23, indicating the users performed more actions to locate a document page they felt interested in than reading the content of documents (the ratio of exploration and immersion types of actions: 1.7). Certain actions are more likely to happen at the start and their proportion drops significantly as the session carried on (e.g., `Click_Hie`). Table 5.4 shows the descriptive statistics of four actions with a mean value more than 1. For `Click_NextS`, the average count is 8.29, while for `Click_Hie`, `Click_SERP` and `Query`, the number is 4.88, 8 and 2.53. All the other type of actions occurred less than once per user.

Figure 5.7 shows the entropy plot of actions spread over positions in the sequences. A higher score suggests, at that position, the user issued a more diverse set of actions. At the very start of the behaviour sequences, there is a lower entropy, which implies that users tend to do the similar action (e.g., `Click_Hie`) at the start of the task. The entropy score increases rapidly, suggesting users issued very different type of actions as their search processes branch out. The distributions in later positions of the graph have relatively low entropy value compared to those in initial positions, as entropy is reduced with the coalescence of the users' actions towards the point of reaching their goals. The main takeaway from the analysis of entropy is that users progress through their tasks in a manner which tends to minimize entropy over time (perhaps after the initially disorderly explosion). This is due to the uniformizing goal of *task completion*

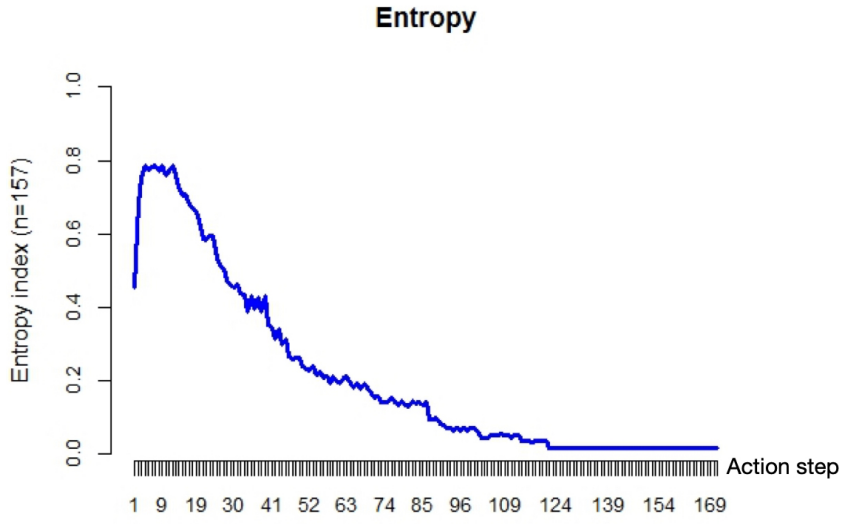


Figure 5.7: Entropy of actions spread over positions in behaviour sequences (browsing).

which, inevitably brings the system to a state of null entropy. Also, note that the maximal entropy of a system on N states is $\log(N)$ which in our case is 0.903 (8 in total: 7 user emitted actions plus the state of task completion). The distribution of the 8 states (figure 5.7) almost approached a uniform one, which accounts for the maximum value of our entropy curve of 0.76.

Frequent subsequence

After reporting the basic statistics of behaviour sequences, we extract the frequent subsequences to prepare for the next analysis. Frequent subsequences from the behaviour sequences were extracted by shifting a sliding window across all the session. In total, 418 subsequences passed the support threshold of 0.05. Table 5.5 displays the top 20 frequent subsequences for browsing ranked by their support value (definition in section 5.2).

The top frequent subsequences (table 5.5) are the common patterns shared by users and most of them are short ones with 2 or 3 actions (length <4). The frequent subsequence with the highest support is the `Click_SERP`→ `Click_NextS`, with a support value of 0.707, which encapsulates the pattern of viewing a document page followed by returning to the SERP and clicking on the next SERP. It suggests that after a successful triggering of the user's interest in investigating more details of one document page, 70.7% of users continue exploring similar documents, which

Table 5.5: Top-20 frequent subsequences extracted from behaviour sequences according to the support value (browsing).

Rank	Sup.	Length	Subsequence	Type
1	0.707	2	Click_SERP → Click_NextS	E, I
2	0.694	2	Click_Hie → Click_NextS	E
3	0.694	2	Click_NextS→ Click_SERP	E, I
4	0.682	2	Start → Click_Hie	E
5	0.669	2	Click_NextS→ Click_Hie	E
6	0.592	3	Click_NextS→ Click_SERP→ Click_NextS	E, I
7	0.573	2	Click_Hie→ Click_SERP	E, I
8	0.573	3	Click_SERP→ Click_NextS→ Click_SERP	E, I
9	0.522	3	Click_Hie→ Click_NextS→ Click_Hie	E
10	0.484	2	Click_SERP→ Click_Hie	E, I
11	0.465	4	Click_SERP→ Click_NextS→ Click_SERP→ Click_NextS	E, I
12	0.452	4	Click_NextS→ Click_SERP→ Click_NextS→ Click_SERP	E, I
13	0.439	3	Click_NextS→ Click_Hie→ Click_NextS	E
14	0.433	3	Click_SERP → Click_NextS→ Click_Hie	E, I
15	0.414	2	Query→ Click_SERP	E, I
16	0.382	3	Click_NextS→ Click_Hie→ Click_SERP	E, I
17	0.376	3	Click_Hie→ Click_NextS→ Click_SERP	E, I
18	0.376	5	Click_NextS→ Click_SERP→ Click_NextS→ Click_SERP → Click_NextS	E, I
19	0.369	5	Click_SERP→ Click_NextS→ Click_SERP→ Click_NextS→ Click_SERP	E, I
20	0.363	3	Click_Hie→ Click_SERP→ Click_NextS	E, I

Rank refers to the rank in terms of the support value of the subsequence.
Sup. denotes the support value.
Length denotes the number of actions in the subsequence.
Type refers to the types of actions contained in the subsequence.

belong to the same search request. Click_Hie→ Click_NextS and Click_NextS→ Click_SERP are both ranked second, and are contained by 69.4% of the behaviour sequences. Click_Hie→ Click_NextS suggests that the user looked through the first SERP returned by clicking on an item from the system built-in hierarchy without clicking on any documents, but continues to the next SERP, which means 69.4% of the users failed to find interesting documents on the first SERP returned by the hierarchy function. Click_NextS→ Click_SERP shows that 69.4% of the users found potential interests on the second or latter SERPs returned by the system. In total, 68.2% of the users started their session with clicking on the built-in hierarchy. The frequent subsequences provide an overview of the frequency of behaviour patterns performed by the users.

Discriminative subsequence of engagement

After preparing the behaviour data into frequent subsequences, we computed the discriminatory power for all the frequent subsequences using the χ^2 test for each engagement dimension in order to understand how these common patterns are associated with user perception of engagement (*RQ.6*).

Table 5.6: Discriminative subsequences for Novelty (browsing).

χ^2	R.	Sup.	Len.	Re.L	Re.H	Subsequence	Type
29.19***	3	0.694	2	-2.43	1.91	Click_NextS → Click_SERP	E, I
22.72***	22	0.357	2	-3.11	2.45	Query → Click_NextS	E
18.34***	8	0.573	3	-2.28	1.80	Click_SERP → Click_NextS → Click_SERP	E, I
16.67***	11	0.465	4	-2.44	1.92	Click_SERP → Click_NextS → Click_SERP → Click_NextS	E, I
15.76***	19	0.369	5	-2.58	2.03	Click_SERP → Click_NextS → Click_SERP → Click_NextS → Click_SERP	E, I
14.04***	18	0.376	5	-2.43	1.91	Click_NextS → Click_SERP → Click_NextS → Click_SERP → Click_NextS	E, I
13.62***	6	0.592	3	-1.94	1.52	Click_NextS → Click_SERP → Click_NextS	E, I
13.61***	40	0.255	4	-2.63	2.07	Query → Click_SERP → Click_NextS → Click_SERP	E, I
12.82***	1	0.707	2	-1.60	1.26	Click_SERP → Click_NextS	E, I
12.67***	56	0.204	3	-2.64	2.08	Click_NextS → Query → Click_NextS	E

R. denotes the rank of the subsequences in the frequent subsequence set.
Sup. denotes the support value. *Len.* denotes the number of actions in the subsequence.
Re.L and *Re.H* denote the Pearson residuals of the subsequence for low NO, and high NO respectively.
Type refers to the types of actions in the subsequence.
Degree of freedom: df = 1. Significance level: ***= p <0.001.

To compare with the previous frequent subsequence ranking (table 5.5), the rank of the subsequences in the frequent subsequence set and their support values are also listed. Table 5.6 shows the top 10 discriminative subsequences for high and low Novelty (NO). From the 418 subsequences extracted, the numbers of subsequences that reject the null hypothesis for each p-value are (p <0.001: $N = 15$; p <0.01: $N = 42$; p <0.05: $N = 101$). Subsequences are ranked by their χ^2 values, which are also reported along with the Pearson residual for each group. Click_NextS → Click_SERP is ranked at the very top, representing the user clicks one document from the non-first SERP, and switches her action from exploration to immersion. Similarly, we observed the combination of exploration and immersion types of actions in the subsequences ranked 3rd to 9th in table 5.6. This also suggests that once the users feel the content is interesting and evokes their curiosity, they tend to switch between action types and perform an in-depth check such as clicking on more than one documents from SERPs about the

same request (e.g., Click_SERP → Click_NextS → Click_SERP, ranked 3rd). This property also motivates the extraction and further examination of an exploration/immersion based score for each user, which leverages the alternating patterns between the two (section 6.2.3).

Three subsequences containing the action Query are ranked in the top 10. Although querying is considered a form of exploration of the website because the action results in the retrieval of a SERP rather than detailed content of a selected document, it requires more effort from the users (e.g., Query → Click_NextS, ranked 2nd). Other actions, except Click_Bookbag, are also contained in the subsequences that can discriminate NO significantly (e.g., Click_Hie appears in the list from rank 26, p <0.01; Click_link appears in the list from rank 35, p <0.01; Click_Metadata appears in the list from rank 49, p <0.05), indicating these actions are also informative in describing the users' feeling of perceived Novelty. None of the significant subsequences contains action Click_Bookbag, which is potentially caused by the low overall frequency of this action (mean = 0.16). Although adding items into the bookbag was suggested in the task description, and the interface displays the bookbag content, the participants rarely performed this action in this browsing session (highest support value of subsequences that contains action Click_Bookbag : 0.064).

Table 5.7: Discriminative subsequences for Felt Involvement (browsing).

χ^2	R.	Sup.	Len.	Re.L	Re.H	Subsequence	Type
23.95***	8	0.573	3	-2.47	2.19	Click_SERP → Click_NextS → Click_SERP	E, I
22.57***	12	0.452	4	-2.72	2.41	Click_NextS → Click_SERP → Click_NextS → Click_SERP	E, I
21.89***	3	0.694	2	-2.01	1.78	Click_NextS → Click_SERP	E, I
19.88***	18	0.376	5	-2.74	2.42	Click_NextS → Click_SERP → Click_NextS → Click_SERP → Click_NextS	E, I
18.83***	1	0.707	2	-1.83	1.62	Click_SERP → Click_NextS	E, I
16.53***	22	0.357	2	-2.54	2.25	Query → Click_NextS	E
16.46***	11	0.465	4	-2.31	2.05	Click_SERP → Click_NextS → Click_SERP → Click_NextS	E, I
16.39***	6	0.592	3	-2.01	1.78	Click_NextS → Click_SERP → Click_NextS	E, I
15.96***	19	0.369	5	-2.47	2.19	Click_SERP→ Click_NextS → Click_SERP → Click_NextS → Click_SERP	E, I
11.67***	69	0.185	4	-2.45	2.17	Click_SERP → Click_NextS → Click_Hie → Click_SERP	E, I

R. denotes the rank of the subsequences in the frequent subsequence set.
Sup. denotes the support value. *Len.* denotes the number of actions in the subsequence.
Re.L and *Re.H* denote the Pearson residuals of the subsequence for low FI, and high FI respectively.
Type refers to the types of actions in the subsequence.
Degree of freedom: df = 1. Significance level: ***= p <0.001.

Table 5.7 shows the top 10 discriminative subsequences for the high and low Felt Involvement

(FI) dimension. The numbers of subsequences reject the null hypothesis for each p-value are (p <0.001: $N = 13$; p <0.01: $N = 43$; p <0.05: $N = 94$). Although 8 out of 10 subsequences are the same in the top 10 list for Novelty (table 5.6), their orders do not match, which testifies to the different aspects of user perception of engagement that these dimensions reflect. `Click_SERP` → `Click_NextS` → `Click_SERP` ranked at the top, which encapsulates the pattern of clicking at least one document from a SERP, and then checking the next SERP(s) from the same request and clicking at least one document from it / them. Similar patterns, which are an in-depth check with clicking more than one document page from SERPs about the same request, are indicative of high FI and are represented by other highly ranked subsequences (subsequences with support value ranked 8, 12, 18, 11, 6, and 19). The scenario in which the user finds more than one document page worth investigating in relation to the same topic captures a significantly higher commitment of time and effort (the user's main resources) on the browsing task and possibly contributes to higher FI. Subsequences with a high support value (e.g., `Click_NextS` →`Click_SERP`, and `Click_SERP` →`Click_NextS`) are also ranked high. The combination of exploration and immersion types of actions in the subsequences occupied nine out of the top ten in table 5.7. Again the alternation of exploration and immersion seems to be relevant to this dimension.

Subsequences containing action `Query` or action `Click_Hie` are ranked in the top 10. Action Click_link is also contained in the subsequences that can discriminate FI significantly (e.g., appears at rank 94, p <0.05). None of the subsequences that contain actions `Click_Bookbag`, or `Click_Metadata` exhibit a significant difference between the high and low FI groups.

Table 5.8 shows the top 10 discriminative subsequences for the Endurability (EN) dimension. For EN, all the subsequences have a relatively lower discriminative power (χ^2 <15) comparing to the two dimensions NO and FI. This strikes a parallel to research phase 1 (chapter 4), where individual behavioural features have a relatively low MDA in predicting high and low EN groups (table 4.5). Overall, the numbers of subsequences that reject the null hypothesis are (p <0.001: $N = 2$; p <0.01: $N = 11$; p <0.05: $N = 44$). The top of the list for EN (table 5.8) contains patterns that are also highly-ranked for NO and FI (table 5.6, and table 5.7), which represent either an in-depth check (e.g., subsequences with support value ranked 8, 12, 11, and 6), or the

Table 5.8: Discriminative subsequences for Endurability (browsing).

χ^2	R.	Sup.	Len.	Re.L	Re.H	Subsequence	Type
14.85***	8	0.573	3	-1.91	1.80	Click_SERP→ Click_NextS→ Click_SERP	E, I
14.24***	3	0.694	2	-1.59	1.50	Click_NextS→ Click_SERP	E, I
10.25**	12	0.452	4	-1.81	1.71	Click_NextS→ Click_SERP → Click_NextS→ Click_SERP	E, I
10.09**	11	0.465	4	-1.77	1.68	Click_SERP→ Click_NextS → Click_SERP → Click_NextS	E, I
9.60**	1	0.707	2	-1.29	1.22	Click_SERP→ Click_NextS	E, I
8.22**	87	0.146	5	-2.08	1.96	Click_NextS → Query → Click_SERP → Click_NextS → Click_SERP	E, I
7.54**	80	0.159	5	-1.98	1.87	Click_SERP→ Click_NextS→ Click_SERP→ Click_NextS→ Query	E, I
7.52**	18	0.376	5	-1.67	1.58	Click_NextS→ Click_SERP→ Click_NextS→ Click_SERP→ Click_NextS	E, I
7.35**	6	0.592	3	-1.33	1.26	Click_NextS→ Click_SERP→ Click_NextS	E, I
6.96**	74	0.172	4	-1.89	1.78	Click_NextS→ Query→ Click_SERP→ Click_NextS	E, I

R. denotes the rank of the subsequences in the frequent subsequence set.
Sup. denotes the support value. *Len.* denotes the number of actions in the subsequence.
Re.L and *Re.H* denote the Pearson residuals of the subsequence for low EN, and high EN respectively.
Type refers to the types of actions in the subsequence.
Degree of freedom: df = 1, Significance level: ***= p <0.001; **= p <0.01

popular ones shared by majority of the users (e.g., subsequences with support value ranked 1 and 3). Apart from these popular ones, the other three top ranked subsequences all contain the action `Query` (e.g., `Click_NextS` → `Query` → `Click_SERP` → `Click_NextS`, ranked 10th). These subsequences do not have a high discriminatory power because they are penalized by their low support values. Two other actions are also contained in the subsequences that discriminate EN significantly (e.g., `Click_Hie` appears in the list from rank 15, p <0.05; `Click_Link` appears in the list from rank 16, p <0.05) indicating these actions are also eloquent in describing users' EN. Again, none of the subsequences which contain `Click_Bookbag` and `Click_Metadata` have significant discriminative power according to the χ^2 test.

Table 5.9 shows the top 10 discriminative subsequences for Perceived Usability (PUs), and the list is different from all the other three with relatively low overlaps. In general, the top ranked subsequences all have a relatively low discriminative power and support values (support <0.15). Only 16 subsequences reject the null hypothesis (p <0.01: $N = 2$; p <0.05: $N = 16$). The average length of the top ranked subsequences is also comparatively long (e.g, length of the subsequence ranked 1st: 12). Some subsequences (e.g., support value ranked 199, 204, 244, 185, 161, 271) encapsulate similar behaviours to the in-depth check pattern observed previously, but contain more actions. It is currently unknown to us why the subsequences ranked highly in

Table 5.9: Discriminative subsequences for Perceived Usability (browsing).

χ^2	R.	Sup.	Len.	Re.L	Re.H	Subsequence	Type
7.21**	199	0.083	12	-2.07	1.96	Click_NextS→ Click_SERP→ Click_NextS→ Click_SERP→ Click_NextS→ Click_SERP→ Click_NextS→ Click_SERP→ Click_NextS→ Click_SERP	E, I
7.21**	204	0.083	13	-2.07	1.96	Click_SERP→ Click_NextS→ Click_SERP→ Click_NextS→ Click_SERP→ Click_NextS→ Click_SERP→ Click_NextS→ Click_SERP→ Click_NextS→ Click_SERP	E, I
5.35*	228	0.076	2	1.83	-1.72	Start→ Query	E
5.33*	148	0.108	5	1.76	-1.66	Click_SERP→ Click_Hie→ Click_SERP→ Click_Hie→ Click_SERP	E, I
5.33*	244	0.070	13	-1.84	1.74	Click_NextS→ Click_SERP→ Click_NextS→ Click_SERP→ Click_NextS→ Click_SERP→ Click_NextS→ Click_SERP→ Click_NextS→ Click_SERP→ Click_NextS	E, I
5.29*	185	0.089	12	-1.79	1.69	Click_SERP→ Click_NextS→ Click_SERP→ Click_NextS→ Click_SERP→ Click_NextS→ Click_SERP→ Click_NextS→ Click_SERP→ Click_NextS	E, I
4.56*	161	0.102	12	-1.65	1.56	Click_SERP→ Click_NextS→ Click_SERP→ Click_NextS→ Click_SERP→ Click_NextS→ Click_SERP→ Click_NextS→ Click_SERP	E, I
4.44*	89	0.146	4	1.57	-1.48	Click_SERP→ Click_Hie→ Click_SERP→ Click_Hie	E, I
4.43*	271	0.064	9	-1.71	1.62	Query→ Click_SERP→ Click_NextS→ Click_SERP→ Click_NextS→ Click_SERP→ Click_NextS→ Click_SERP→ Click_NextS	E, I
4.43*	280	0.064	7	-1.71	1.62	Click_NextS→ Click_SERP→ Click_NextS→ Click_SERP→ Click_NextS→ Query→ Click_NextS	E, I

R. denotes the rank of the subsequences in the frequent subsequence set.
Sup. denotes the support value. *Len.* denotes the number of actions in the subsequence.
Re.L and *Re.H* denote the Pearson residuals of the subsequence for low PUs, and high PUs respectively.
Type refers to the types of actions in the subsequence.
Degree of freedom: df = 1, Significance level: *= p <0.05; **= p <0.01

terms of support do not feature highly on this list. A possible explanation lies in the nature of the PUs, which is intrinsically more system (or interface) - dependent and is more sensitive to user preference. As such this should be treated as an indication on low impact in the behaviour sequence approach to inferring PUs in browsing. Another possible explanation is that the user, who feels she can not perform what she wants, does not stay long in the session, and thus does not issue many actions. Interestingly, querying right after the start of the session ranked 3rd, with a positive residual value on the low PUs group, which potentially represents the scenario in which the user feels nothing is interesting enough to click on from the system suggestions and then issues a query.

5.4.3 Discussion of sequential patterns in browsing

Seven actions, namely `Query`, `Click_NextS`, `Click_Hie`, `Click_SERP`, `Click_Link`, `Click_Metadata`, and `Click_Bookbag`, were selected to describe the user interaction in a browsing context, and each action is dichotomized into either *exploration* or *immersion*. The sequence analysis results reveal a couple of insights. We first answer the research questions. In general, our results reveal the statistically significant relationships between user perception of engagement and behaviour sequence (*RQ.6*). Two patterns indicate the user perceives high NO, FI and EN: alternating between actions of immersion and exploration types, on the one hand, and, on the other, checking SERPs about the same request in-depth. Moreover, three dimensions, NO, FI and EN exhibit similar behaviours all with highly ranked sequences in terms of support, discriminating high engagement dimension levels positively (*RQ.7*).

We then discuses the detailed findings that support our answers. Regarding the action distribution reported in descriptive statistics of behaviour sequences, the users issued some actions (e.g., clicking on system build-in hierarchy) more than others (e.g., clicking on results saved in bookbag), which is natural, but also potentially identifies an action bias in this particular type of system interaction.

Glancing at the top frequent subsequences (table 5.5) and the top discriminative subsequences of any engagement dimensions (table 5.6, 5.7, 5.8, 5.9), one can conclude that the lists present significant differences and certain subsequences with a smaller support value are better indicators of high/low engagement for certain dimensions. Our results show that 6 out of 7 selected actions, except `Click_Bookbag`, are contained in the subsequences that can discriminate at least one high and low engagement dimension significantly. In spite of the fact that adding the results that one feels interesting was suggested in the task description, re-checking those results was not mandatory. The users did not click results from the bookbag many times, which is evident from the fact that the mean occurrence of this action is less than one. This further negatively affects the discriminative power of the subsequences that contain this action.

Comparing the top discriminative subsequences between engagement dimensions, dimension

Novelty (NO), Felt Involvement (FI) and Endurability (EN) shared a number of popular subsequences (NO-FI: 7 out of 10; NO-EN: 6 out of 10; FI- EN: 7 out of 10) whereas PUs correlates most highly with only a few subsequences with very low support. Six subsequences contributed to all three dimensions (with support values ranked 1, 3, 6, 8, 11, and 18). Two behaviour patterns are represented by these six subsequences: firstly, switching between exploration and immersion action types, suggesting that the user checks the collection at a shallow level and proceeds by more closely investigating the results and secondly, checking SERPs about the same request in-depth, describing a behaviour where the user is interested in the current request and subsequently wants to examine more similar items. The presence of these two patterns was found to be indicative of the user experiencing high NO, and also high FI and high EN. We recall the following interpretation from Bates (2007): "Browsing is the activity of engaging in a series of glimpses, each of which exposes the browser to objects of potential interest; depending on interest, the browser may or may not examine more closely one or more of the (physical or represented) objects; this examination, depending on interest, may or may not lead the browser to (physically or conceptually) acquire the object."

Dividing the process of browsing into three levels according to the state of engagement with the object of information we remark that the progression leads the user to an increasingly narrower space along the investigation - acquisition axis, the full conceptual engagement being realized only in the latter stage. Therefore, the switching between exploration actions, which exist at the first level, and immersion actions, which exist at the latter two levels, suggests the user is able to progress, and is intent on what she is doing.

Interestingly, similarly to this alternating behaviour, which represents a deep interaction at the session level, checking SERPs about the same request in-depth represents a deep interaction, but at the request level. In this study, there are three ways to issue a request: submitting a query (`Query`), clicking a link in the hierarchy (`Click_Hie`), or clicking a metadata link (`Click_Metadata`) on the document page. Even though a single browsing session, which only ends when the user feels bored, may contain more than one request, this pattern suggests that a single engaged request indicates a high engagement at the session level, which provides empirical evidence of the impact of such patterns in browsing. Additionally, evaluation at the

request level is discussed predominantly in the context of searching rather than browsing, and primarily refers to query level evaluation. Like browsing, a user's search session may contain more than one query. A query session usually ends when the user finds what they want or feels bored or tired (Järvelin and Kekäläinen, 2000), but a single successful query does not lead to the task completion as a task is potentially comprised of multiple queries (Järvelin et al., 2008). Neither does an unsuccessful query mean a failure as users can acquire the answer by other queries or simply reformulate the current one. Thus a representation at the query-level is not capable of describing the session-level (e.g., satisfaction (Al-Maskari et al., 2007; Mao et al., 2016; Kiseleva et al., 2016b), engagement (Song et al., 2013)) and how it contributes to the session-level varies across scenarios (Kiseleva et al., 2016a), and is used as an additional metric in information retrieval evaluation (Song et al., 2013; Mao et al., 2016). As opposed to the weak connection between query-level and session-level assessment of user experience in searching, a pattern which is indicative of high engagement at the query (request) level also indicates high engagement at the session level in browsing.

In addition to the common patterns, there are a couple of differences among all four engagement dimensions. Although EN is originally designed as the overall evaluation of the interaction experience and future intention to return to this service (O'Brien and Toms, 2010a), the behaviour subsequences have a lower discriminative power for this dimension compared to NO and FI. Looking back at the definition of browsing, out of the three levels of interactions "glimpses" , "examine more closely" and "acquire the object", distinguishing the fist level from the second and third levels by user behaviour is not difficult as it can be estimated by the clicking depth and dwelltime. However, the third level, "acquire the object", may require mental activities such as extracting or adapting the information, which place it on the solitary side of a cognitive rift from the previous levels. EN maps mainly onto the third level, as it is the overall evaluation, while NO and FI are potentially responsible for the transitions the user makes into more immersive states such as from level 1 to level 2, as this is when the users' curiosity is evoked and she is expected to continue to use the system or examine information further: a key indicator of feeling involved is losing control of time. It is also embedded in the UES design (O'Brien and Toms, 2010a) (e.g., "I was so involved in my task that I lost track of time", and

"The time I spent searching just slipped away." for FI; "I would continue to use wikiSearch out of curiosity." for NO). For PUs, the dimicriminative power is very low and the top ranked subsequences suggest a low PUs. As PUs was designed to evaluate the function of the system, and the browsing task has no specific goal, the user might just not feel strongly towards it either way. These differences also suggest that the four dimensions capture engagement from a multitude of perspectives.

5.4.4 Results (searching)

This section describes the results of study B using the wikiSearch dataset. We first present the descriptive statistics of behaviour sequences, and then report the frequent subsequences extracted from the behaviour sequences. We report the discriminative subsequences selected through the frequent subsequences using the χ^2-test of independence.

Descriptive statistics of behaviour sequences

After extracting the actions listed in table 5.1 and forming 377 behaviour sequences, we present some descriptive statistics of our dataset as for the previous study. The layout of this section follows that of the previous one, in order to better allow the reader to contrast results between the searching and browsing contexts.

Table 5.10: Descriptive statistics of actions in the behaviour sequences (searching).

Action	Mean	Median
`Query`	7.70	7
`Click_SERP`	8.73	8
`Click_Link`	6.95	6
`Click_Bookbag`	1.91	1

All participates issued at least 13 actions, and at most 99. This differs slightly from the previous setting (browsing) where some behaviour sequences had a very short length of only 2 actions. The average number of actions issued per user is 34.88 (SD= 10.67), with a median number

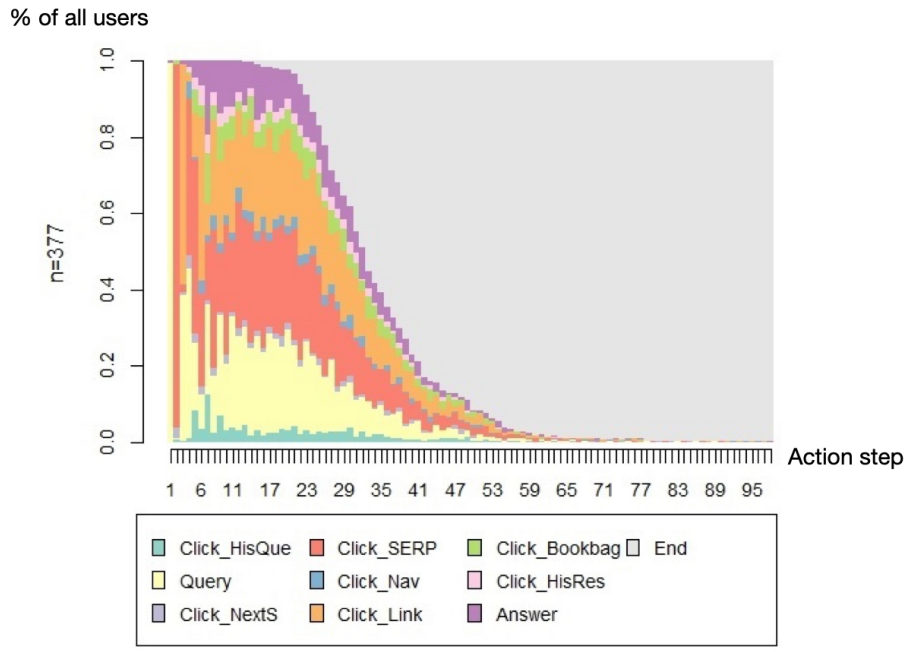


Figure 5.8: Distribution of behaviour sequences (searching).

of 33 (the first quartile Q_1: 27; the third quartile Q_3: 40). Figure 5.8 presents the actions distribution by the order they were issued in the behaviour sequences. The gray area again represents the proportion of participants who left the session, marked as 'End' in figure 5.8. The parameter (λ) of the exponential distribution lies within the following 95% confidence interval: (0.0258, 0.0316). We can clearly observe from these statistics that certain actions, `Query`, `Click_SERP`, and `Click_Link` occurred more often than the others. Table 5.10 shows the descriptive statistics of four actions with a mean value more than 1. The three actions appeared the most are `Query` (mean = 7.7, median = 7), `Click_SERP` (mean = 8.73, median = 8), and `Click_Link` (mean = 6.95, median = 6). Action `Click_Bookbag` has a mean number of 1.91 and a medium number of 1. As all the users were assigned three questions for the searching task, they all performed action `Answer` at least three times (min=3, median = 3), but some of them reformulated their answers a couple of times before submitting (max= 11, mean=3.15). More than half of the users did not perform all the other type of actions (Median=0). Action `Click_Nav` has a mean value of 0.83, and median of 0 (max=19), indicating although more than half of the users (N=254, 67.37%) did not perform this action but some of the user did a lot. Users performed on average 22.51 immersion actions (`Click_Link`, `Click_Nav`, `Click_SERP`, `Click_Bookbag`, and `Click_HisRes`) and 9.37 exploration actions (`Query`, `Click_HisQue`, and

Click_NextS), and the exploration and immersion ratio is 0.4163.

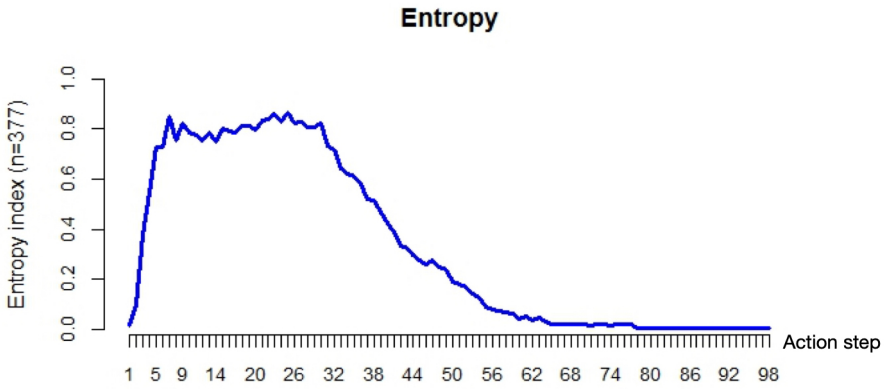


Figure 5.9: Entropy of actions spread over positions in behaviour sequences (searching).

Figure 5.9 shows the entropy of the above action distributions at each step in the sequence. At the very start of the behaviour sequences, there is a lower entropy, which implies that users tend to do the similar actions at the start of the task. These actions are querying and clicking the SERP. The entropy score increases rapidly and then decays as actions coalesce as users tend to emit the same actions at the ends of their sequences. The possible maximal entropy in this case is 1, and the distribution of the 10 states (9 user emitted actions plus the state marks task completion) almost approaches a uniform one towards the head of the sequence (accounting for a maximally disordered state), with the maximal value of our entropy curve being 0.87.

Frequent subsequence

After the basic statistics of behaviour sequences, we extract the frequent subsequences in order to prepare for the next analysis using a sliding window method. In total, 662 subsequences passed the 0.05 support threshold.

Table 5.11 displays the top 20 frequent subsequences extracted. The top frequent subsequences (table 5.11) are common patterns shared by users. The most frequent subsequence is the Query → Click_SERP as all the users performed it, representing that after issuing a query, the user clicks a document page from the returned SERP. A low mean value of Click_NextS (mean = 0.56) is observed, which suggests that users tend to click at least one document from the first

Table 5.11: Top-20 frequent subsequences extracted from behaviour sequences according to the support value (searching).

Rank	Sup.	Length	Subsequence	Type
1	1.000	2	Query → Click_SERP	E, I
2	0.979	2	Click_SERP → Click_Link	I
3	0.958	3	Query → Click_SERP → Click_Link	E, I
4	0.931	2	Click_Link → Query	E, I
5	0.918	3	Click_Link → Query → Click_SERP	E, I
6	0.918	3	Click_SERP → Click_Link → Query	E, I
7	0.897	4	Click_SERP → Click_Link → Query → Click_SERP	E, I
8	0.883	2	Click_Link → Answer	I
9	0.873	4	Click_Link → Query → Click_SERP → Click_Link	E, I
10	0.870	2	Click_SERP → Query	E, I
11	0.854	3	Click_SERP → Query → Click_SERP	E, I
12	0.844	5	Click_SERP → Click_Link → Query → Click_SERP → Click_Link	E, I
13	0.594	3	Click_SERP → Click_Link → Answer	I
14	0.589	3	Query → Click_SERP → Query	E, I
15	0.570	4	Query → Click_SERP → Click_Link → Query	E, I
16	0.544	4	Query → Click_SERP → Query → Click_SERP	E, I
17	0.538	5	Query → Click_SERP → Click_Link → Query → Click_SERP	E, I
18	0.477	6	Query → Click_SERP → Click_Link → Query → Click_SERP → Click_Link	E, I
19	0.477	2	Click_SERP → Answer	I
20	0.469	4	Query → Click_SERP → Click_Link → Answer	E, I

Rank refers to the rank in terms of the support value of the subsequence.
Sup. denotes the support value.
Length denotes the number of actions in the subsequence.
Type refers to the types of actions contained in the subsequence.

SERP in response to their queries, without checking the next SERP. Click_SERP → Click_Link is ranked second with a support value of 0.979, suggesting most users clicked on a link from the content of the document. The top ranked frequent subsequences are mostly a combination of three popular actions Query, Click_SERP, and Click_Link, as is also easy to see from figure 5.8.

Discriminative subsequence of engagement

After preparing the behaviour data into frequent subsequences, we computed the discriminatory power for all the frequent subsequences using the χ^2 test for each engagement dimension in order to understand how these common patterns are associated with user engagement ($RQ.6$) in searching. To compare with the previous frequent subsequence ranking (table 5.11), the rank of the subsequences in the frequent subsequence set and their support values are also listed.

Table 5.12: Discriminative subsequences for Novelty (searching).

χ^2	R.	Sup.	Len.	Re.L	Re.H	Subsequence	Type
6.87 **	8	0.883	2	-0.72	0.62	Click_Link → Answer	E, I
4.38 *	44	0.255	4	-1.46	1.24	Click_SERP → Query → Click_SERP → Click_Bookbag	E, I
4.24 *	27	0.334	2	-1.35	1.15	Click_SERP → Click_Bookbag	I
3.86 *	3	0.958	3	-0.35	0.30	Query → Click_SERP → Click_Link	E, I
3.72	32	0.297	3	-1.30	1.11	Query → Click_SERP → Click_Bookbag	E, I
3.63	269	0.058	5	1.57	-1.34	Click_SERP → Query → Click_SERP → Click_Link → Click_Nav	E, I
3.54	30	0.302	2	-1.27	1.08	Click_Bookbag → Answer	I
3.47	93	0.151	5	-1.41	1.20	Click_SERP → Query → Click_SERP → Click_Bookbag → Answer	E, I
3.28	225	0.069	2	-1.48	1.26	Click_HisQue → Click_Bookbag	E, I
2.84	9	0.873	4	-0.50	0.43	Click_Link → Query → Click_SERP → Click_Link	E, I

R. denotes the rank of the subsequences in the frequent subsequence set.
Sup. denotes the support value. *Len.* denotes the number of actions in the subsequence.
Re.L and *Re.H* denote the Pearson residuals of the subsequence for low NO, and high NO respectively.
Type refers to the types of actions in the subsequence.
Degree of freedom: df = 1, Significance level: **= p <0.01, *= p <0.05.

Table 5.12 shows the top 10 discriminative subsequences for high and low Novelty (NO). The discriminative power for the subsequences ranked top for this dimension are the lowest among all four engagement dimensions, suggesting that NO is the least sensitive engagement dimension to user behaviour patterns in searching. From the 662 subsequences extracted, the number of subsequences that reject the null hypothesis is only 4 (p <0.01: $N = 1$; p <0.05: $N = 4$). This is potentially a consequence of the task issued, which is goal-driven rather than curiosity-driven. As NO is not a concept that the users would actively think about while completing the search tasks, the dimension reflects on searching behaviour poorly.

Table 5.13 shows the top 10 discriminative subsequences for the high and low Felt Involvement (FI). In total, 18 subsequences reject the null hypothesis (p <0.01: $N = 3$; p <0.05: $N = 18$). The top ranked subsequence is Click_SERP → Click_Bookbag, which encapsulates the behaviour of checking a document page and revisiting another document page saved in the bookbag. Other subsequences indicative of high FI also contain the action Click_Bookbag (e.g., the ones ranked 4th, 5th and 7th). As suggested in the task description, users are encouraged to save the documents they think are helpful into the bookbag, which means, the action Click_Bookbag is a sign that user revisited a useful result. When this action appears before action Answer, it is an optional self-validation step. The users who have the option not

to do this, but choose to invest the effort regardless, might be more involved in this experience. Apart from it, there are a couple of subsequences which indicate low FI significantly. Two subsequences, `Query` → `Click_SERP` → `Answer` → `Click_SERP`, and `Click_SERP` → `Answer` → `Click_SERP`, suggest that the user checks another document from the previous SERP after inputting the answer, indicating the user is not fully confident about what she writes. Another pattern, `Click_Nav` → `Click_Link` → `Click_Nav`, represents the user flipping documents. Subsequences containing action `Click_HisQue` appears at rank 15 (p <0.05). The other two actions `Click_HisRes` and `Click_NestS` with a similar number of occurrence (mean ≤ 1) are not contained by any of the significant subsequences.

Table 5.13: Discriminative subsequences for Felt Involvement (searching).

χ^2	R.	Sup.	Len.	Re.L	Re.H	Subsequence	Type
10.74**	27	0.334	2	-1.99	1.91	`Click_SERP` → `Click_Bookbag`	I
8.68**	243	0.064	4	2.20	-2.12	`Query` → `Click_SERP` → `Answer` → `Click_SERP`	E, I
8.31**	154	0.095	3	2.10	-2.01	`Click_SERP` → `Answer` → `Click_SERP`	I
6.47*	32	0.297	3	-1.61	1.54	`Query` → `Click_SERP` → `Click_Bookbag`	E, I
6.28*	44	0.255	4	-1.63	1.57	`Click_SERP` → `Query` → `Click_SERP` → `Click_Bookbag`	E, I
5.78*	124	0.111	3	1.74	-1.68	`Click_Nav` → `Click_Link` → `Click_Nav`	I
5.04*	58	0.212	3	-1.52	1.46	`Click_SERP` → `Click_Bookbag` → `Answer`	I
4.86*	65	0.194	3	1.51	-1.45	`Click_Link` → `Click_SERP` → `Click_Link`	I
4.65*	196	0.077	4	1.63	-1.57	`Click_Nav` → `Click_Link` → `Click_Nav` → `Click_Link`	I
4.59*	142	0.101	4	1.58	-1.52	`Click_Link` → `Click_Nav` → `Click_Link` → `Click_Nav`	I

R. denotes the rank of the subsequences in the frequent subsequence set.
Sup. denotes the support value. *Len.* denotes the number of actions in the subsequence.
Re.L and *Re.H* denote the Pearson residuals of the subsequence for low FI, and high FI respectively.
Type refers to the types of actions in the subsequence.
Degree of freedom: df = 1, Significance level: **= p <0.01, *= p <0.05.

Table 5.14 shows the top 10 discriminative subsequences for Endurability (EN). For this engagement dimension, the top ranked subsequences have a relatively higher discriminative power compared to the two previous dimensions, NO and FI. In total, 34 subsequences reject the null hypothesis (p <0.001: $N = 3$; p <0.01: $N = 12$; p <0.05: $N = 34$), and the occurrence of all the top ranked subsequences indicates low EN. `Query` → `Click_SERP` → `Click_Link` → `Click_Nav` is ranked at the very top, representing the user checking at least two documents after querying and switching between them. The action `Click_Nav`, representing the switching behaviour, also appears in the other seven subsequences listed in table 5.14.

Looking at an example from the seven subsequences containing `Click_Nav`: `Click_Link` →

Table 5.14: Discriminative subsequences for Endurability (searching).

χ^2	R.	Sup.	Len.	Re.L	Re.H	Subsequence	Type
12.55***	87	0.159	4	2.39	-2.38	Query → Click_SERP → Click_Link → Click_Nav	E, I
11.72***	33	0.294	2	2.10	-2.10	Click_Link → Click_Nav	I
11.25***	211	0.074	2	2.42	-2.41	Click_Nav → Answer	I
9.79**	124	0.111	3	2.20	-2.19	Click_Nav → Click_Link → Click_Nav	I
9.66**	196	0.077	4	2.25	-2.24	Click_Nav → Click_Link → Click_Nav → Click_Link	I
8.56**	142	0.101	4	2.08	-2.07	Click_Link → Click_Nav → Click_Link → Click_Nav	I
8.39**	35	0.281	4	1.81	-1.80	Click_SERP → Query → Click_SERP → Query	E, I
8.13**	52	0.228	5	1.85	-1.84	Click_SERP → Query → Click_SERP → Query → Click_SERP	E, I
7.54**	120	0.117	5	1.93	-1.93	Query → Click_SERP → Click_Link → Click_Nav → Click_Link	E, I
7.17**	46	0.241	2	1.73	-1.72	Click_Nav → Click_Link	I

R. denotes the rank of the subsequences in the frequent subsequence set.
Sup. denotes the support value. *Len.* denotes the number of actions in the subsequence.
Re.L and *Re.H* denote the Pearson residuals of the subsequence for low EN, and high EN respectively.
Type refers to the types of actions in the subsequence.
Degree of freedom: df = 1, Significance level: ***= p <0.001, **= p <0.01.

Click_Nav → Click_Link → Click_Nav, which is ranked 6th, a possible scenario is that the user clicks a link on document 1, and views document 2. Then she flips back to document 1 and clicks a link from it which leads to document 3. But after viewing the third document, she again flips back to document 1. The user issues at least 4 actions and views at least three documents. This suggests that the user is redoubling many search efforts and switching from one document to another multiple times, all of which could contribute to a feeling of confusion or lack of progress. As action Click_Nav is performed by 32.63% of the users, this pattern is wide-spread enough to carry a significant impact on EN.

Table 5.15 shows the top 10 discriminative subsequences for the Perceived Usability (PUs). The discriminative power of the top ranked subsequences is the highest among all four dimensions. All the top ranked 10 subsequences indicate low PUs, and interestingly the action Click_Bookbag is missing in the whole table. This is similar to the observation for the EN dimension. In total, 55 subsequences reject the null hypothesis (p <0.001: $N = 10$; p <0.01: $N = 26$; p <0.05: $N = 55$). Query → Click_SERP → Click_Link → Click_Nav again is ranked at the very top, and is followed by Click_Link → Click_Nav. Again, the iterative pattern was observed. 7 out of the top 10 ranked subsequences for PUs also ranked in the top 10 for EN, suggesting the PUs and EN reflect on user behaviour sequences more similarly in searching

comparing to FI and NO.

Table 5.15: Discriminative subsequences for Perceived Usability (searching).

χ^2	R.	Sup.	Len.	Re.L	Re.H	Subsequence	Type
19.12***	87	0.159	4	3.11	-2.73	Query → Click_SERP → Click_Link → Click_Nav	E, I
13.66***	33	0.294	2	2.41	-2.11	Click_Link → Click_Nav	I
13.51***	120	0.117	5	2.71	-2.38	Query → Click_SERP → Click_Link → Click_Nav → Click_Link	E, I
13.12***	72	0.183	4	2.55	-2.24	Click_SERP → Click_Link → Click_Nav → Click_Link	I
12.23***	49	0.239	3	2.37	-2.08	Click_SERP → Click_Link → Click_Nav	I
12.00***	196	0.077	4	2.64	-2.32	Click_Nav → Click_Link → Click_Nav → Click_Link	I
11.41***	46	0.241	2	2.29	-2.01	Click_Nav → Click_Link	I
11.41***	124	0.111	3	2.51	-2.20	Click_Nav → Click_Link → Click_Nav	I
11.06***	35	0.281	4	2.19	-1.92	Click_SERP → Query → Click_SERP → Query	E, I
10.86***	214	0.074	5	2.53	-2.22	Click_SERP → Click_Link → Click_Nav → Click_Link → Click_Nav	I

R. denotes the rank of the subsequences in the frequent subsequence set.
Sup. denotes the support value. *Len.* denotes the number of actions in the subsequence.
Re.L and *Re.H* denote the Pearson residuals of the subsequence for low PUs, and high PUs respectively.
Type refers to the types of actions in the subsequence.
Degree of freedom: df = 1, Significance level: ***= p <0.001.

5.4.5 Discussion of sequential patterns in searching

Nine actions, namely Query, Click_HisQue, Click_NextS, Click_SERP, Click_Nav, Click_HisRes, Click_Link, Click_Bookbag, and Answer were selected to describe the user interaction in a searching context, and each action is dichotomized into either *exploration* or *immersion*. The sequence analysis results reveal a couple of insights. We first answer the research questions. In general, our results support that a relationship exists between user perception of engagement and behaviour sequences in searching (*RQ.6*). One pattern of high uncertainty behaviour that suggests the user is flipping through viewed documents forward and backward, indicates low EN and PUs. Moreover, two dimensions, PUs and EN exhibit similar behaviours all with highly ranked sequences in terms of support discriminating low engagement dimension levels positively (*RQ.7*). We then discuses the detailed findings that support our answers.

The users issued some actions (e.g., Query, Click_SERP, and Click_Link) more than others (e.g., Click_NextS). The low number of action Click_NextS issued can be explained by the users either quickly retrieving the information they require of the first looked-up SERP, or

alternatively, having failed to do so on the first try, resubmitting a query rather than continuing to investigate the next SERP. This corroborates with the position bias effect in search, for example, the second SERP receives very little traffic (Joachims et al., 2017).

The top frequent subsequences (table 5.11) contain mainly the expected popular searching actions, and are very different from the top discriminative subsequences for any engagement dimensions (tables 5.12, 5.13, 5.14, 5.15). Our results show that eight out of nine selected actions, except `Click_HisRes`, are contained in the subsequences that can discriminate at least one high and low engagement dimension significantly.

Regarding the actions contained in the most discriminative subsequences, we observe that the occurrence of action `Click_Nav` suggests the user is experiencing low PUs and EN. Action `Click_Nav` represents flipping backwards and forwards througt documents viewed and is performed by 32.63% of the users. With a clear search goal in mind, users recognize achieving this goal as a success and thus perceive the experience positively. However, evaluating search satisfaction is a more ephemereal task which combines multiple aspects of the user experience. Sucessfully locating a page has been investigated in Hassan et al. (2010), but as claimed in other studies (Kim et al., 2014; Odijk et al., 2015), one can only make a vague decision on whether the user finds the webpage useful or satisfying. Our insight, based on the analysis of user behaviour sequences, and supported by the discriminative power of sequences containing `Click_Nav` to indicate low engagement levels is that, the positive perception of the user is not conditioned only on the success of their retrieval task, but also on what they perceive to be superfluously allocated effort (not total allocated effort, as the previous studies corroborate, effectively invested effort counts positively towards the experience evaluation). This is an elusive component of behaviour to measure, but in this study it can be observed in the amount of *uncertainty* experienced by the user as they are forced to revisit documents often, loop back in their search process and retrace steps or redo analyses. This pattern forces the user to disconnect and context switch multiple times per session, which confining the user to a less engaged state. This pattern, and others like it, due to the negative impact uncertainty have on the user, we record and refer to henceforth as HUPs (high uncertainty patterns). We hope to corroborate more evidence towards this interpretation in future studies.

Comparing the top dictiminatiove subsequences between dimensions, EN shared 7 out of the 10 top ranked subsequences (with a support value ranked 33, 35, 46, 87, 120, 124, and 196) with PUs, and all the other pairs of engagement dimensions has a very low overlap. The overlap between PUs and EN suggests that the PUs and EN characterize on user behaviour sequences similarly in searching. With a clear destination to reach, the users value and appreciate the support from the system. Thus, how users perceived the usability of the search system has an major impact on their overall evaluation of the search experience.

5.5 Study C. Engagement prediction using behaviour sequences

The purpose of this study is to determine how well discriminative behaviour subsequences (extracted in phase 2, study B) perform in predicting user perception of engagement. By prediction, in this setting we mean a classification problem of assigning either high or low engagement labels (phase 1, study B) to users based on their discriminative behaviour subsequences. Figure 5.10 presents the steps, data and variables used in this section. We restate below the research questions associated with this study:

RQ.6 What is the relationship between user behaviour sequences and user perception of engagement?

RQ.7 How do the relationships between user behaviour sequences and user perception of engagement vary between dimensions?

5.5.1 Method

To test how complementary the discriminative subsequences extracted in phase 2, study B are in predicting engagement, we frame a classification problem for each engagement dimension using features engineered from the discriminative subsequences and behavioural features extracted in

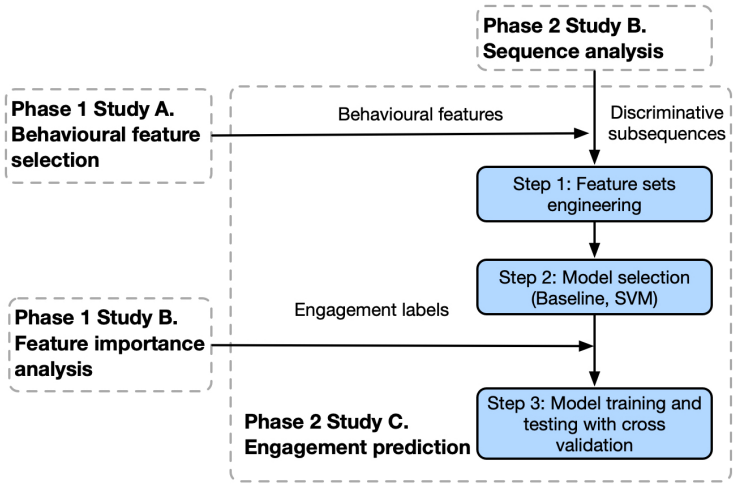


Figure 5.10: Steps, data and variables used in section 5.5.

phase 1, study A (*RQ.6*). The binary classification problem is defined in which the goal is to predict if a user will have *high* or *low* engagement levels given the engineered features. *High* or *low* engagement levels are determined by splitting user groups against the median of each engagement score, as in phase 1, study B. The feature engineering (step 1) is conducted in R[2], and all the other analyses in this study were conducted in Weka [3].

Feature engineering (step 1)

In this study, we tested three feature sets: i) behavioural features (table 4.2); ii) sequential features; and iii) behavioural features and sequential features combined. However, as the overlap between subsequences is nontrivial (e.g., `Click_NextS`→ `Click_SERP`→ `Click_NextS` and `Click_NextS`→ `Click_SERP`), the information on discriminating high and low engagement levels may become redundant. Thus, focusing only on the top-ranked ones (e.g., top 10) might not be generally optimal. To solve this, all the significant discriminative subsequences were used in the prediction tasks. Sequential features were constructed based on the inclusion of the discriminative subsequences that significantly rejected the null hypothesis with a binary label marking the presence of such a subsequence in the behaviour sequence. A feature selection with the best first strategy was performed on the features to select the best set by using the

[2]https://www.r-project.org
[3]https://www.cs.waikato.ac.nz/ml/weka/

discriminative subsequences with a p value less than 0.001 first, then adding the ones with a p value less than 0.01, and finally the ones with a p value less than 0.05.

Model selection (step 2)

We trained SVM classifiers using these three different sets of sequential features. We choose SVM for consistent parallelism to the previous section and due to its strong performance all round in the previous task 4.4) as well as its usage as part of the machine learning status quo. Apart from reporting the performance of SVM classifiers using sequential features (`SVM-Seq`) , we also combined the best performed sequential feature set with the behavioural features (table 4.2) and trained an SVM classifier (`SVM-All`) for each engagement dimension. This allows us to test the improved prediction performance that can be achieved by adding the sequential features over the behavioural features tested in the previous phase (chapter 4), if any. In order to make a comparison, two classifiers `Baseline`, that always predicts the majority class, and `SVM`, trained with all the behaviour features tested in the previous research phase (chapter 4) were included as well.

Model training and testing (step 3)

The performance scores for all classifiers were obtained by cross-validation and were measured by the same metric used in phase 1 (section 5.5): precision, recall, accuracy, F measure and AUC. Finally, a k-fold cross validation paired t-test (Dietterich (1998)) was applied to compare the performance of classifiers using sequential features, which are `SVM-Seq` and `SVM-All`, against `SVM`.

We point out that the main comparison of interest is not with `Baseline` anymore, but with `SVM`. We have previously compared `SVM` against `Baseline` in the first research phase (section 5.5) and `SVM` outperformed `Baseline` significantly for most of the performance metrics, including all precision, F-1 and AUC, and chiefly, we want to test how much of an improvement the sequential information can provide on top of the behavioural features in terms of predict-

ing engagement. Subsequently, we compared the performance of the classifiers among four engagement dimensions in order to answer *RQ.7*.

5.5.2 Results and discussion (browsing)

Table 5.16 presents the performance of four classifiers obtained by performing 5-fold cross-validation, and the significance level of `SVM-Seq` and `SVM-ALL` compared against `SVM`.

Table 5.16: Performance metrics using three different feature sets (browsing).

	Performance	Baseline	SVM	SVM-Seq	SVM-All
NO	(Weighted Avg.) Precision	38.2%	71.0%	**73.0%**	72.9%
	(Weighted Avg.) Recall	61.8%	70.7%	**73.2%**	**73.2%**
	(Weighted Avg.) F-1	47.2%	68.1%	**73.1%**	72.9%
	Accuracy	61.8%	70.7%	**73.2%**	**73.2%**
	(Weighted Avg.) AUC	48.9%	64.5%	**71.0%**	70.7%
FI	(Weighted Avg.) Precision	31.4%	**75.6%**	71.2%	74.3%
	(Weighted Avg.) Recall	56.1%	**73.2%**	66.9%	**73.2%**
	(Weighted Avg.) F-1	40.3%	71.7%	66.4%	**72.2%**
	Accuracy	56.1%	**73.2%**	66.9%	**73.2%**
	(Weighted Avg.) AUC	48.4%	70.7%	68.7%	**71.1%**
EN	(Weighted Avg.) Precision	27.9%	63.3%	**69.2%**	68.3%
	(Weighted Avg.) Recall	52.9%	62.4%	61.8%	**68.2%**
	(Weighted Avg.) F-1	36.6%	60.9%	59.0%	**67.8%**
	Accuracy	52.9%	62.4%	61.8%	**68.2%**
	(Weighted Avg.) AUC	47.1%	61.3%	63.3%	**67.6%**
PUs	(Weighted Avg.) Precision	27.9%	53.7%	**67.7%***	59.3%
	(Weighted Avg.) Recall	52.9%	54.1%	**63.7%**	59.2%
	(Weighted Avg.) F-1	36.6%	53.4%	**60.3%**	58.1%
	Accuracy	52.9%	54.1%	**63.7%**	59.2%
	(Weighted Avg.) AUC	47.1%	53.4%	**62.1%**	58.3%

A bold typeface denotes the best result in a row.
Corrected paired t-test to compare the `SVM-Seq` and `SVM-All` against the `SVM` classifier (underlined).
Significance level (2-tailed): * = p <0.05

First, we observed that including sequential information can improve the prediction performance of engagement dimensions as `SVM-Seq` is the best for predicting PUs and NO, while `SVM-All` is the best for predicting FI and EN. Paired t-tests show that for most of the improvements there are no significant differences between `SVM` and `SVM-Seq` or `SVM-All` apart from the precision between `SVM` and `SVM-Seq`. Overall, PUs is still the most difficult to predict, while FI, EN

and NO are relatively easier. For PUs, the best performing sequential feature set used in `SVM-Seq` marks the presence of a discriminative subsequence at a significance level of 0.05, and has the best performance among all three feature sets. Compared to `Baseline`, F-measure is improved by 23.7%, and accuracy is improved by 10.8%, and also improvements are shown over the `SVM` (F-measure improved by 6.9%, and accuracy improved by 9.6%). The performance of `SVM-All` is slightly worse than `SVM-Seq`. For predicting FI, the `SVM` classifier and `SVM-All` (compared to `Baseline`, F-measure improved by 31.9%; accuracy improved by 11.7%, and AUC improved by 22.7%) are the best among all feature sets, and the differences between these two classifiers are small. The best performing feature set for `SVM-Seq` contains features of discriminative subsequences at a significance level of 0.01. However, the `SVM-Seq` does not outperform the classifiers using behavioural features. With respect to EN, the `SVM-All` classifier is the best among all (compared to `Baseline`, F-measure improved by 31.2%, accuracy improved by 15.3%, and AUC improved by 20.5%; compared to `SVM`, F-measure improved by 6.9%, accuracy improved by 5.8%, and AUC improved by 6.3%), but the improvements over `SVM` are not significant. The best performing sequential feature set captures the presence of subsequences at a significance level of 0.05. For NO, the `SVM-Seq` classifier outperforms all others with moderate improvements over `Baseline` (F-measure improved by 25.9%, and AUC by improved 22.1%; accuracy improved by 11.4%), and small improvements over `SVM` (F-measure improved by 5%, and AUC by improved 6.5%; accuracy improved by 2.5%), which are not significant. The best performing sequential feature set contains the presence of subsequences at a significance level of 0.001. The differences between `SVM-Seq` and `SVM-All` are small.

As the `SVM` classifier for PUs did not improve much over the `Baseline`, discriminative subsequences are more useful in indicating PUs than behavioural features. Comparing the information contained in the sequence features to the 43 behavioural features, sequence features capture the order of user actions while time spent on action is only contained ion the behavioural features. For FI, the performance of `SVM` is as good as `SVM-All`, and both of them outperformed `SVM-Seq`. These again suggests time-related information is very important in describing FI. This aligns with the results in the research phase 1, that time-related features (table 4.5, 4.4) are the best type of features in informing user perception of engagement, especially for NO and

FI dimensions in browsing.

5.5.3 Results and discussion (searching)

Table 5.17 presents the performance of four classifiers obtained by performing 10-fold cross-validation, and the significance level of `SVM-Seq` and `SVM-ALL` compared against `SVM`.

Table 5.17: Performance metrics using three different feature sets (searching).

	Performance	Baseline	SVM	SVM-Seq	SVM-All
NO	(Weighted Avg.) Precision	33.7%	54.9%	**61.0%**	**61.0%**
	(Weighted Avg.) Recall	58.1%	58.1%	**60.7%**	**60.7%**
	(Weighted Avg.) F-1	42.7%	46.1%	**53.7%**	**53.7%**
	Accuracy	58.1%	58.1%	**60.7%**	**60.7%**
	(Weighted Avg.) AUC	49.3%	50.7%	**54.7%**	**54.7%**
FI	(Weighted Avg.) Precision	27.0%	52.7%	57.5%	**61.0%***
	(Weighted Avg.) Recall	52.0%	53.1%	57.6%	**61.0%***
	(Weighted Avg.) F-1	35.6%	52.0%	57.0%	**60.8%***
	Accuracy	52.0%	53.0%	57.6%	**61.0%***
	(Weighted Avg.) AUC	49.1%	52.5%	57.1%	**60.7%***
EN	(Weighted Avg.) Precision	49.2%	**58.4%**	57.1%	**58.4%**
	(Weighted Avg.) Recall	49.6%	**58.4%**	56.8%	58.1%
	(Weighted Avg.) F-1	44.5%	**58.4%**	56.1%	57.6%
	Accuracy	49.6%	**58.4%**	56.8%	58.1%
	(Weighted Avg.) AUC	49.3%	**58.4%**	56.7%	58.1%
PUs	(Weighted Avg.) Precision	31.9%	**64.6%**	62.8%	64.2%
	(Weighted Avg.) Recall	56.5%	63.7%	63.1%	**64.5%**
	(Weighted Avg.) F-1	40.8%	60.3%	61.4%	**63.0%**
	Accuracy	56.5%	63.7%	63.1%	**64.5%**
	(Weighted Avg.) AUC	48.8%	60.1%	60.5%	**62.0%**

A bold typeface denotes the best result in a row.
Corrected paired t-test to compare the `SVM-Seq` and `SVM-All` against the `SVM` classifier (underlined).
Significance level (2-tailed): * = p <0.05

Including sequential information can improve the performance of predicting PUs, and FI, as `SVM-All` outperformed `SVM` and `SVM-Seq`. With respect to EN, both classifiers using sequential information (`SVM-Seq` and `SVM-All`) did not improve over `SVM`. For PUs, the improvements of `SVM-All` over `SVM` are small (F-measure improved by 2.7%, and Accuracy improved by 0.8%) and not significant. The best performed sequential feature set marks the presence of subsequences

at a significance level of 0.01. For predicting FI, the `SVM-All` (compared to `Baseline`, F-measure improved by 25.2%; accuracy improved by 9%) were the best among all classifiers, and significantly outperforms the `SVM` classifier (F-measure improved by 8.8%; accuracy improved by 8%, and AUC improved by 8.2%)). `SVM-Seq` outperformed `SVM`, and the best performed feature set for `SVM-Seq` contains the presence of subsequences at a significance level of 0.05. For NO, the `SVM-Seq` classifier and `SVM-All` have the same performance with moderate improvements over `Baseline` (F-measure improved by 11%, and AUC improved by 5.4%; accuracy improved by 2.6%), and small improvements over `SVM` (F-measure improved by 7.6%, and AUC improved by 4%; accuracy improved by 2.6%). The improvements over `SVM` are not significant, and the best performed sequential feature set contains the presence of subsequences at a significance level of 0.01.

In general, leveraging the discriminative subsequences only improves significantly in predicting FI (table 5.17). One possible explanation is that when users are given a well-defined task, they are expected to perform certain actions and spend a least amount of effort, which may lead to small differences in their behaviour sequences.

5.6 Comparison of results in browsing and searching

We compared the results obtained from study A, B and C between browsing and searching in order to answer the following research question:

RQ.8 How do the relationships between user behaviour sequences and user perception of engagement differ between browsing and searching?

Analysis on the two datasets provide information for comparing the differences between the two caused by the two information contexts. We observed very different relationships between behaviour sequence and user perception of engagement in browsing and searching. This motivates adequate, context-aware engagement modelling in subsequent studies. We now discuss the observations that lead to this answer in details:

Although in both contexts the distribution of actions and the distribution of the stopping times for each user should theoretically align, the distribution of actions in the behaviour sequence in browsing has a longer tail (figure 5.6), while the distribution in searching starts decaying after a longer period of high entropy (figure 5.8). This might be a result of the goal of the two given tasks: in order to complete a detailed decision-making task, the users have to complete a minimum number of expected actions, whereas in the browsing task is absolutely based on their own will, and thus without a minimum number of actions. This also reflects how irregular browsing behaviour compares to searching behaviour in a more general setting. Users performed relatively more exploration actions in browsing and more immersion actions in searching (the ratios of exploration and immersion actions: browsing : 1.7 and searching: 0.4163).

Regarding the discriminative subsequences, overall they have a relatively low χ^2 value in searching, suggesting that engagement dimensions are not captured in the user behaviour sequences as much as in browsing. This is also observed in the previous research phase, that selected behavioural features have a lower correlation value with all four engagement dimensions in searching than in browsing (table 4.8 and table 4.4), and the top ranked behavioural features have a low contribution in predicting user perception of engagement in searching (table 4.9 and table 4.5).

Also, in browsing, we can clearly identify three of the dimensions, NO, FI and EN, as similar in as much as they correlate to user behaviour as expressed through behaviour sequences. In searching, dimensions PUs and EN are relatively more similar from this perspective. This is indicative of a heuristics which argues that usability is a core component of a searching system and the users reflect negatively on a system that needlessly increases the amount of effort required for the task, while novelty and felt involvement are moot considerations to this type of interaction. On the opposite side of the spectrum, involvement and novelty play a pivotal role in browsing, which we have seen to carry a more exploratory component and benefit highly from factors such as serendipity (McCay-Peet and Toms, 2015) and spur of the moment insights the user acquires during the session.

The observed two patterns in browsing: switching between immersion and exploration actions

and in-depth investigation of one request point to the following interpretation: the user is willing to devote more effort on the task at hand, regardless of whether it is a detailed check of a document or more actions on the same topic. Above all else, they suggest a high engagement in dimensions NO, FI and EN. In contrast, the HUPs pattern we observed in searching are indicative of a low engagement, especially low EN and PUs.

Overall, it is harder to predict engagement in the searching context than in the browsing context, as was expected from the first research phase using selected behavioral features (table 4.11 and table 4.12).

5.7 Summary

The key objective of this phase was to assess the relationship between behaviour sequences and user perception of engagement as we answer four research questions. We selected two sets of actions for browsing and searching (*RQ.5*), based on the ISP model (Marchionini, 1995) and common IR system interface components, and formed behaviour sequences based on these actions. We then tested the behaviour sequences in terms of discriminating high and low engagement labels; this revealed the differences among four engagement dimensions.

Our results, in both browsing and searching contexts, support that a relationship exists between user perception of engagement and behaviour sequences (*RQ.6*). More specifically, we identified two behaviour patterns, switching between exploration and immersion actions, checking SERPs about the same request in-depth, suggest that the user may experience high NO, FI and EN in browsing. On the other hand, in searching, the HUPs described in section 5.4.5 is indicative of a low EN and PUs. This also answers *RQ.8*.

Moreover, NO, FI and EN are comparatively more similar to each other in reflecting on behaviour sequences than PUs in browisng, where as in searching, PUs and EN are relatively more similar to each other in reflecting on behaviour sequences than NO and FI (*RQ.7*).

We further leveraged the subsequences to predict engagement, which outperformed behavioural

features in predicting NO, EN, and PUs in browsing, and in predicting NO in searching (*RQ.7*, *RQ.8*). Although the improvements are not statistically significant (at the 0.05 level) for most of the performance metrics, it is nevertheless empirically evident that incorporating the information captured by behaviour sequences is relevant in improving performance and can potentially be a subsequent source of insight for future studies and trials.

Chapter 6

Phase 3: Implementing measures of engagement

6.1 Overview

In the previous research phases, we investigated the important behavioural features in terms of predicting user perception of engagement in phase 1 (chapter 4) and the discriminative behaviour subsequences that are associated with high or low engagement in phase 2 (chapter 5). At this stage, there comes a natural question: can we leverage the relationships identified to design new context aware measures of user perception of engagement based on user behaviour? These measures should be sensitive to patterns that we identified as relevant in previous phases and be at least as accurate as the state of the art.

This phase was designed to answer this question through, firstly, suggesting the ideal properties based on the findings of the previous two phases and proposing measures of engagement that implement those properties, and secondly, evaluating these measures of engagement. We choose to focus on Endurabrility as it measures not only the feeling of reward that users receive in the current session but also their intention of returning to this system or recommending this system to others. This dimension is important as it extends the intra-session engagement concept (measures of the current session) to inter-session engagement (intentions in the future

sessions), and thus represents the overall evaluation of user engagement (O'Brien and Toms, 2010a). We provide detailed reasoning for proposing these measures, together with our intuition for the insight they bring into the research questions. We strive to make the measures simple and interpretable, and our computations easily replicable in an online setting. Ultimately, this phase produced evidence for how incorporating the user behaviour and user perception of engagement relationships into our design can improve engagement prediction by answering these two research questions:

RQ.9 What are the properties that empirically computable measures of engagement should possess?

RQ.10 Which of the developed measures improve user perception of engagement prediction?

This phase can be divided into two studies:

Study A. Measure development, in which we identified a set of six properties that are guided by the findings from phases 1 and 2. We then developed a set of five measures of engagement, and gave the reasons why these measures capture the properties identified (section 6.2.3). (*RQ.9*)

Study B. Evaluation of developed measures, in which we formulated a regression problem for the Endurability score computed from the user perception of engagement data, and tested the performance of two learning models in solving this problem using the proposed measures of engagement. (*RQ.10*)

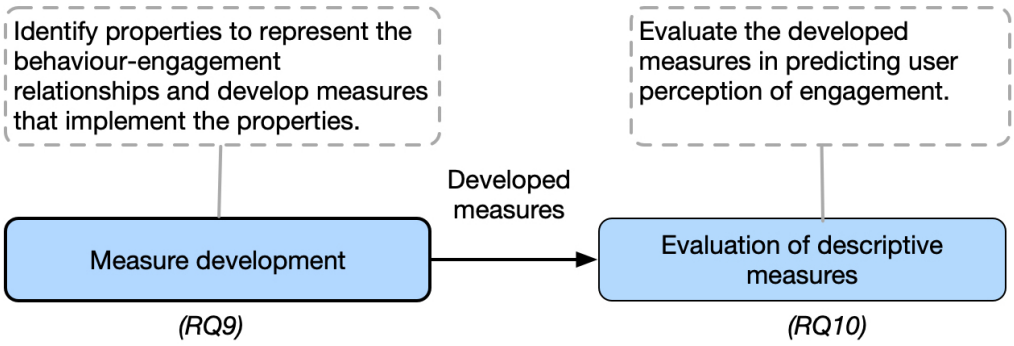


Figure 6.1: Design of research phase 3, which contains two main parts: measure development (section 6.2), evaluation of developed measures (section 6.3).

Figure 6.1 presents the design of this research phase. We selected the Endurability (EN) dimension as the engagement score because this dimension represents the final evaluation of the session and is the overall score of the user engagement (table 3.2). Details on how the developed measures capture the properties identified were provided (section 6.2.3). We further formed the evaluation tasks as a regression problem, in which we trained two models for each measure to predict EN (section 6.3). We replicated the state-of-art intra-session framework (Machmouchi et al., 2017) that quantifies the user perception of information retrieval as one of the baselines, and also compare the difference of the performance between the two settings, browsing and searching. Outcomes from this phase lead to the more comprehensive understanding of the performance of measuring user perception of engagement through user behaviour in browsing and searching.

6.2 Measure development

The current section is devoted to the first of our two studies in this phase. The purpose of this study is to identify the properties of an ideal measure of engagement as revealed by the findings in phase 1 and phase 2, and develop measures that implement these properties. We start by recalling our definition of measure and introducing the definition of *property* in section 6.2.1. Then we discuss the properties representing the behaviour - perception of engagement relationships suggested in the previous two phases in isolation in order to give a general intuition which is independent from our particular design, and then propose measures which implement those properties in a later section. Figure 6.2 presents the steps conducted and variables resulted in this section. We restate below the research question associated to this study:

RQ.9 What are the properties that empirically computable measures of engagement should possess?

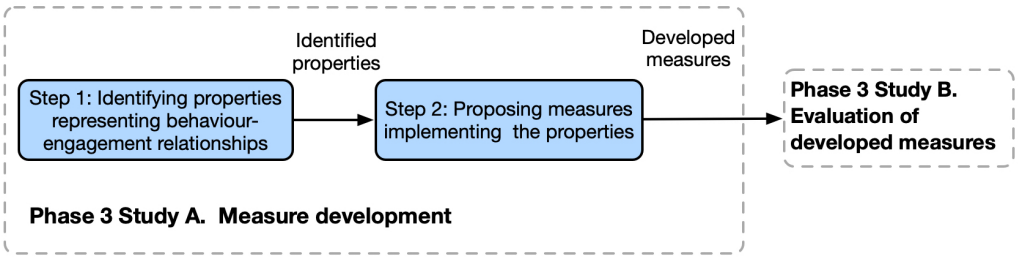


Figure 6.2: Steps and variables used in section 6.2.

6.2.1 Method

This study is mainly a theoretical consequence of the findings from phases 1 and 2, which we recall in brief. Our aim is to identify the properties for an ideal measure of user perception of engagement based on user behaviour and develop measures that implement these properties. We do not compute measures in this study as some of them are designed to be extracted in the prediction task. Thus, neither user behaviour nor perception of engagement data were processed in this study.

Recalling our discussion in chapter 3 section 3.4, a measure of engagement is a random variable which is a function of a certain set of features of user behaviour. This function should be easily computable, and the output should correlate positively to user perceived engagement. By a *property* of such a measure we mean a qualitative description of its behaviour with respect to certain types of variations of features in its domain. The quantitative description is how its outputs correlate to partitions in the domain generated by this variation. Here is a primitive example. Let ι_A be the indicator variable of the action A in the feature space. If m is a measure, $m(\iota_A, \cdot, ..., \cdot) \in \mathbb{R}$, we say the measure is insensitive to ι_A if:

$P_0(A)$ The marginal distribution of m given all of its other parameters is independent from ι_A.

More complex properties can be listed given more complex combinations of features, asking for the partial correlation coefficients to satisfy certain criteria, etc. It is infeasible to check all of these combinations, and in practice we only have access to samples of these underlying distributions and checking for independence would entail statistical tests at the level of each

feature. Our properties are constructed from features determined in phases 1 and 2, where these correlation tests have been performed.

Identifying properties representing behaviour - perception of engagement relationships (step 1)

In the first phase we tested the intuition that behavioural features are indicative of user perception of engagement, while in the second phase, it was the variance of certain discriminative patterns in the behaviour sequence that were found to be relevant. These two studies motivate the principle that time spent on each action / state and the relative ordering of the actions / sequence both carry important information which should be encoded in our measures. In section 6.2.2, we have selected properties that ask for measures to respond appropriately to the variance in certain subsets of our features or behaviour signal motivated by the above principle, and the examples of patterns and features extracted from the previous two phases.

Proposing measures implementing the identified properties (step 2)

After identifying the properties of an ideal measure of engagement, we developed a set of five measures of engagement, which implements these properties. We first introduce the utility metric (Machmouchi et al., 2017), that our measures were build upon and augment. The developed measures are based on time which is motivated by the findings from phase 1, and designed via two overarching principles: to be extensions of the utility metric (Machmouchi et al., 2017), by either considering different types of actions, weighting the time contributions differently or applying different summarization techniques to the action space, or to be simple functions of the user's action sequence motivated by the findings from phase 2. Not all of these properties are achievable under all measures, the intricacy of the task leaving much room for exploration. The outcomes of this study are six identified properties (*RQ.9*), and five measures which are computed and evaluated in study B.

6.2.2 Properties representing behaviour - perception of engagement relationships

In total, we identified six properties from the findings of phase 1 and 2. Table 6.1 present them with associated sections of the findings. We denote the properties as P_m^n, in which m is the phase in which the property is identified and n is an index. We now discuss the findings that lead to these properties in details:

Table 6.1: Six properties representing the behaviour-engagement relationships found in phase 1 and 2.

Phase	No.	Description
1	P_1^a	The measure should capture the time spent information.
	P_1^b	The measure should capture the quality of the user's interaction with a query or a document page.
2	P_2^a	The measure should be sensitive to changing between actions of type immersion and type exploration.
	P_2^b	The measure should capture the order of each user action.
	P_2^c	The measure should capture the depth of one user request.
	P_2^d	The measure should capture the patterns of attempts with/without progressing.

Properties suggested in findings of Phase 1

In phase 1, we have investigated how discrete behavioural features can contribute to engagement prediction. In the browsing task, time-related features contribute to EN the most in feature importance analysis, which followed by result-related features (figure 4.8). In the searching task, the most important features are more diverse (figure 4.14) as the top-10 features belongs to all four categories. In particular, features representing the quality of the query, such as average click-through-rate and number of queries without clicks, are ranked the top.

To conclude, the properties suggested from phase 1 are :

(P_1^a) The measure should capture the time spent information.

This property is motivated by the observations in phase 1 study B that in both browsing and searching time-related features are important in predicting EN, such as TimeOnPages

for browsing (figure 4.8) and TimeOnSERPs for searching (figure 4.14). Moreover, they are in general important in describing more than one dimensions of engagement (table 4.6 and table 4.10). The usefulness of time-related features is not new to the research community as observed in user satisfaction evaluation (Machmouchi et al., 2017; Liu et al., 2018; Garcia-Gathright et al., 2018), classifying search success or abandonment (Diriye et al., 2012; Ageev et al., 2011). The detailed observations that support this property have been discussed in section 4.3.3 and section 4.3.5.

(P_1^b) The measure should capture the quality of the user's interaction with a query or a document page.

It is first inspired by the observation, in searching, that features which reflect query quality (e.g., AveCTR, and NumQueryNoClicks) are important in describing EN (figure 4.14). Second, features representing document page quality, are ranked top in browsing as well, such as average time spent on dissatisfied pages (figure 4.8). We merge these two type of behaviours into one property because they represent users' attempts in progressing in both contexts. In searching, each query is an attempt from the user to locate the useful document. In browsing, the time spent on document pages is the allocation of users' resources to find potentially interesting documents. Similar ideas were mentioned in (Ageev et al., 2011; Kim et al., 2013), as they proposed models that can estimate good and bad queries. While, measuring the quality of a query or a document page remains challenging, as discussed in the literature review (section 2.5.2), mapping behaviour signals to a subjective response is difficult.

Properties suggested in findings of Phase 2

In phase 2, we have examined how complementary behaviour sequences are in predicting user perception of engagement. In browsing, two patterns, switching between exploration and immersion actions, and checking SERPs about the same request in depth, suggest that the user may experience high EN. In searching, the behaviour pattern HUPs (section 5.4.5), which the user shows high uncertainty in locating a document, indicates low EN. In general, taking the

sequential information of user behaviour into account improved user perception of engagement prediction. The properties suggested from this phase are:

(P_2^a) The measure should be sensitive to changing between actions of type immersion and type exploration.

In phase 2 study B. sequence analysis (section 5.4), we observed that the alternation of exploration and immersion types of actions in the subsequences can discriminate high and low EN (table 5.8). This observation motivates the extraction and further examination of the alternating pattern between exploration and immersion types of actions. As action types partition immersion and exploration into subtypes, the requirement that a measure is at least sensitive to the more general category is strictly weaker than it being sensitive to the reordering of action types. The stronger statement is property (P_2^b). As the immersion and exploration types are introduced in chapter 5, to our best knowledge, there are no models based on these two high-level types.

(P_2^b) The measure should capture the order of each user action. In particular, it should be sensitive to shuffling the order of actions in a behaviour sequence.

This property is motivated by the improved prediction performance (table 5.16) of combining the features based on discriminative subsequences with the behavioural features in browsing. This is the ideal pattern of response of a measure, and it contains more detailed requirements compares to P_2^a. This design is naturally embedded in all sequential behaviour models, such as predicting satisfaction using Hawke processes (Mehrotra et al., 2017), using neural networks (Williams and Zitouni, 2017).

(P_2^c) The measure should capture the depth of one user request. In particular, it should evaluate differently on users that emit many queries but pursue them to a shallow extent and users who emit a few queries but investigate a deep document path after each.

The depth of one request refers to the number of document items or document pages that the user clicked on from the SERPs returned by a single request. This also means the immersion that a user invested after one request. This property is motivated by the

observation that the pattern, checking SERPs about the same request, indicates a high EN in browsing (table 5.8). This property is well captured by the length of query trails (Yuan and White, 2012).

(P_2^d) The measure should capture the patterns of attempts with/without progressing.

This property is inspired by the observation that certain patterns (such as HUPs, in searching (section 5.4.5)) indicate the user is not progressing in the information retrieval task and correlate with low engagement. A challenge is to define what the terms 'progress' or 'successful' mean, and related discussion varies (e.g., (Ageev et al., 2011; Liu et al., 2018)). Multiple patterns in this study potentially satisfy this criterion, but this property aims to capture this behaviour at a higher level. In particular, a measure that is sensitive to the amount of time spent redoubling efforts (checking the same page or document multiple times, re-emitting queries etc.) should capture these patterns a priori. We remind the reader that these patterns may naturally be different in browsing versus searching.

6.2.3 Measures of engagement that implement the identified properties

In this section we construct five measures based on the properties above. Not all measures will capture all of these properties, and defining better measures as well as more refined feature sets is a constantly evolving task. All measures constructed are measures of the EN dimension of engagement, as it is the most descriptive of the overall user experience during a session (section 3.4.1). We first introduce the utility metric (Machmouchi et al., 2017) which our measures were build upon and augment. Then we propose the measures of engagement that implement the properties identified in section 6.2.2.

As a crucial advance in the development of measures that quantify the users perception of search within the session, Machmouchi et al. (2017)'s Utility Metric stands out. This simple but powerful construction leverages the intuition that the time users spend on a certain type

of action is particularly telling of their perception of search, which they used satisfaction as the target of interest. If one could only decide on the relative importance of each action with respect to one-another, one could compute a measure of the target of interest as a weighted average of these time spans. Therefore, this metric is transferable in measuring engagement. Below we describe how to compute this measure in detail since is serves as inspiration for our own development.

Utility Metric (Machmouchi et al., 2017), is a normalized and weighted sum of the total time spent on each user action, namely UtilityRate. Lets recall the notation used in the research phase 2 (section 5.2). A refers to a set of actions a user issued in the session, which lasted time T, where $A = \{a_i\}_{i=1}^{|A|}$. $t(a_i)$ refers to the total time spent on one action a_i in the session, and $\sum_{a_i \in A} t(a_i) = T$. The total utility of the user session is Utility $= \sum_{a_i \in A} w(a_i) * t(a_i)$, in which $w(a_i)$ is obtained by training a linear regression to a selected measurement of user experience, which in our case is EN. The normalized utility of the user session is the utility rate : UtilityRate $= \frac{\sum_{a_i \in A} w(a_i) * t(a_i)}{T}$. In order to compare all the metrics with the same training methods, we do not only train a linear regression to learn the weights as suggested (Machmouchi et al., 2017), but rather treat $t(a_i)$ as a set of features and train the weights $w(a_i)$ for each type of action a_i using two different models described in section 6.3.

Proposed measures of engagement

In this section, we propose five measures of engagement each contains a set of features. We first describe the five measures and then argue for the properties captured by each one.

1. Behavioural features (`BF`): it is the set of behavioural features (table 4.2) extracted in phase 1 study A. These features are selected because they are used in previous studies to describe user perception of engagement.
2. Time spent on immersion and exploration types of actions (`IETime`): this measure is based on the utility metric (Machmouchi et al., 2017), in which we group the actions into only two types, immersion and exploration. Two features are extracted for this measure,

which are the total time the user spent on the two types of actions: $t(Immersion)$ and $t(Exploration)$.

3. Time spent on discriminative subsequences (`SeqTime`): in this measure, we extract features that are the time spent on each discriminative subsequences (definition in section 5.2) for EN at a significance level 0.001 ($p-$value). As a result of our statistical testing and our discussion in chapter 5, these are the patterns that capture the most relevant patterns of the user behaviour action stream to predict EN. These patterns are in general simple and occur often so that the question of how much time users spend engaged in these patterns becomes relevant. While it is never possible to capture the entire variety of user behaviour from a user sample, we have taken as much care as possible to ensure these patterns are as general and replicable as possible.

4. The area under the curve of immersion and exploration action path (`IEPath`): To construct the path, we first group the actions into only two types, immersion and exploration. The area between a curve and the x-axis is a definite integral, and the concept, such as the area under the Receiver Operating Characteristic (ROC) curve, has been used as an popular evaluation measure of IR systems (Ingwersen and Järvelin, 2006). A user spent a finite time T in their session, which counts as resource. We assume the user makes the decision to spent the resource on immersion or exploration type of actions because she think it is beneficial for the current task, and thus the interaction of these two types of actions and the dwelltime associated to each type carry information. To capture this interaction, we create the IE path for each user, by plotting their behaviour sequence in a two-dimensional space. If the action belongs to immersion type, we made a step forward on the x-axis. If the action is exploration type, the step will be made on the y-axis. The length of the step is equal to the dwelltime the user spent on the action. A behaviour sequence $s(A) = \{action_j\}_{j=1}^{l}$ with length l is formed by a set of actions, $a \in A$, to which corresponds a list of dwelltimes $time_j$ associated with $action_j$, where $j \in 1, ..., l$. The curve is generated as follows :

   After constructing the IE path from the sequence, the area between this curve and the x-axis is computed. Two features are included in this set: first, the area under curve of

Algorithm 1 Creating IE path.

Input: action-dwelltime pairs ($action_j$, $time_j$), int l, int j
Output: (l+1)*2 matrix representing the IE Path: $path$

$path_1 \leftarrow [0, 0]$
for j from 1 to l **do**
  if $action_j$ is immersion type **then**
    $path_{j+1} \leftarrow path_j + [time_j, 0]$
  else
    $path_{j+1} \leftarrow path_j + [0, time_j]$
  end if
end for
return $path$

the path IE_{AUC}, and second, the total time of the session T, where $\sum_{i=1}^{l} time_i = T$.

5. Exponential weighted time spent on actions (`ExpActionTime`): This set of features is engineered in much the same way as that for the utility metric (Machmouchi et al., 2017), but here we wish to test the following piece of intuition. We wish to verify whether patterns which occur at the end of a user's behaviour sequence impact their assessment of engagement more than the ones closer to the start. This is perhaps due to the memory latency effect, or potentially due to users being more forgiving in making negative evaluations through dissatisfied interactions at the start of the sequence. We avoid extracting the heuristic argument for which this measure is designed into one of our ideal properties, simply because it is not supported by significant statistical evidence from previous studies. We believe the merits of this approach to be perhaps beyond the simplistic attempt we make here, but wish to collect our view on this approach here. To be precise, these features incorporate an exponential tail for time spent on action. If action a_i, where $a_i \in A$, occurs from times t_{a_i} to times $t_{a_i} + \Delta t$ then the contribution of this feature is $w_{a_i} e^{t_{a_i}} (e^{\Delta t} - 1)$, where w_{a_i} are weights that depend only on the action type which we train via a regression model.

Then we discuss the identified properties in the previous step captured in each the proposed measure. Table 6.2 presents how the proposed five measures fulfil the properties. In the following we go through the list and attempt to explain these allocations in better detail. All

of our measures trivially capture the total time spent on the task, so we forego mentioning this further.

1. Behavioural features (`BF`): The quality of queries or pages is captured through the ration of $TimeOn\langle sat\rangle$ and $TimeOn\langle dissat\rangle$ types of features. The depth of one request is participially captured by features such as NumPagesPerQuery and NumUniquePagesPerQuery.
2. Time spent on immersion and exploration types of actions (`IETime`): This is a strictly weaker measure than the utility metric (Machmouchi et al., 2017) as it groups the actions into two types, and hence has no hope of capturing more.
3. Time spent on discriminative subsequences (`SeqTime`): All our properties were motivated by the behaviour sequence. Indeed an ideal measure should be fully faithful on the user behaviour sequence. As an immediate proxy, discriminative frequent subsequences capture all of these properties, albeit not in a necessarily explicit manner. (P_2^b) is captured by definition.
4. The area under the curve of immersion and exploration action path (`IEPath`): Captures the alternation between immersion and exploration types by definition (P_2^a). These broad types, as explained where introduced (section 5.3.1) are intrinsically related to the user's progression through a task. In browsing, the types map onto the three levels of interaction with the system (section 5.4.3), whereas in searching, a successful information retrieval task always starts with an exploratory phase and ends with an immersion phase in which the user is spiralling down a proverbial rabbit hole of references and connections in their attempt to acquire the object of search. Hence, without providing a concrete measure for progress, we conclude this measure is not insensitive to (P_2^d).
5. Exponential weighted time spent on actions (`ExpActionTime`): While this does not capture more than the utility metric (Machmouchi et al., 2017), it was engineered to capture a behaviour we do not reflect in the table of properties.

Table 6.2: Proposed measures against the idea properties.

Measure	Description	P_1^a	P_1^b	P_2^a	P_2^b	P_2^c	P_2^d
`BF`	Behavioural features	•	•			○	
`IETime`	Time spent on immersion and exploration types of actions	○					
`SeqTime`	Time spent on discriminative subsequences	•	○	○	•	○	○
`IEPath`	The area under the curve of immersion and exploration action path	•		•			○
`ExpActionTime`	Exponential weighted time spent on actions	•					

The description of all properties (P_m^n) are presented in table 6.1.
• represent properties that the measure is specifically engineered to capture.
○ represent properties that the measure is not insensitive to, but cannot be guaranteed to be faithful to in full generality.

6.3 Evaluation of developed measures

The purpose of this study is to evaluate the measures (section 6.2.3) developed in phase 3, study A in predicting user perception of engagement. By prediction, in this setting we mean a regression problem of predicting the Endurability (EN) score computed from the user perception of engagement data. We selected the EN dimension as the target because this dimension represents the final evaluation of the session and is the overall score of the user experience (table 3.2). Figure 6.3 presents the steps, data and variables used in this section. We restate below the research questions associated to this study:

RQ.10 Which of the developed measures improve user perception of engagement prediction?

6.3.1 Method

In order to answer *RQ.10*, we test how well the measures developed in the previous study predict user perception of engagement. The regression problem is defined in which the goal is to predict the EN score of a user given the measures. Measures are extracted based on the behaviour features obtained in phase 1, study A (table 4.2), discriminative subsequences for EN obtained in phase 2, study B (section 5.4), log files and the user perception of engagement data, collected through the UES questionnaire. As some measures are designed to be extracted

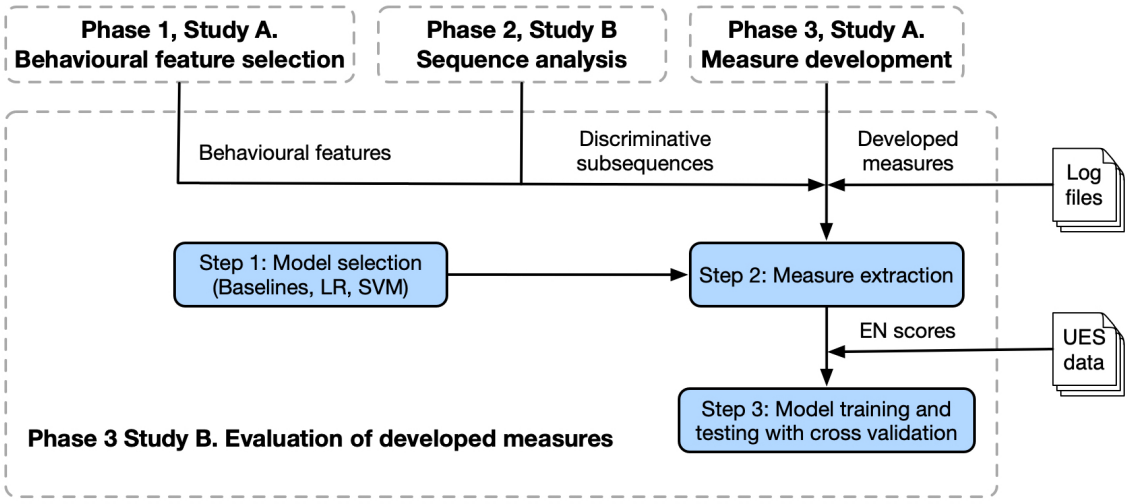


Figure 6.3: Steps, data and variables used in section 6.3.

inside the model, we describe the model selection (step 1) before measure extraction (step 2), and then model training and testing (step 3). All the analyses in this study were written in python [1].

Model selection (step 1):

We frame this evaluation task as a regression problem: to predict EN using two baselines and five developed measures from phase 1, study A. The two baselines and the proposed five measures are:

1. `Baseline-M`, which always predicts the mean value of EN in the training set. This baseline is independent from user behaviour as it only needs the EN scores in the training set without additional informations.
2. `Baseline-U`, which is the Utility metric described in section 6.2.3.
3. `BF`, which are the selected behavioural features (table 4.2) extracted in phase 1, study A.
4. `IETime`, which is the dwell time spent on immersion and exploration type of actions.
5. `SeqTime`, which is the time spent on discriminating subsequences.
6. `IEPath`, which contains the AUC of IE path and the total time spent on the session.

[1] https://www.python.org

7. `ExpActionTime`, which is the exponentially weighted version of `Baseline-U`.

We select two models: (i), Linear Regression (LR), and (ii) Support Vector Machine (SVM), in accordance with the models used to train the `Baseline-U` (Machmouchi et al., 2017) and the technique used in our previous studies.

The LR is implemented with Elastic net regularization (Zou and Hastie, 2005). This method is well known for its robustness and accuracy due to its internal feature selection method by cross-validation. It becomes a powerful tool in selecting features which combine to explain the variance of our dependent variable and provides a clear, descriptive, interpretable solution to the regression problem. One can easily read off the weights assigned to each feature, and perform the appropriate statistical analysis of simple liner regression. Furthermore, this method is meant to match that which the `Baseline-U` was established with. We use an implementation for python [2].

SVM is implemented with sequential minimal optimization (Platt, 1998). It is selected mainly because our datasets have a high participants and features ratio. For example, there are 157 participants in the CHiC dataset, and the number of behavioural features is 43 (ratio 27.38%). Models using boosting methods such as gradient boosted trees will potentially suffer from over-fitting problems, while this is less of a concern for SVM. We use an implementation for python [3].

Extracting measures (step 2):

In order to extract the two baselines and five proposed measures, user behavioural features from phase 1, study A, discriminative subsequences of EN from phase 2 study B, user perception of engagement data, collected through the UES questionnaire, and log files were used. In total, there are 43 behavioural features are extracted for the 157 participants in browsing and 34 behavioural features are extracted for the 377 participants in searching; two discriminative subsequences for browsing and three discriminative subsequences for searching. We only use

[2]https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.ElasticNet.html
[3]http://www.csie.ntu.edu.tw/ cjlin/libsvm

Endurability (EN) dimension in this study. To assign scores for the EN dimension, all the items within EN were aggregated, which we refer to as *EN scores*. Two baselines and five proposed measures are extracted as follows:

1. `Baseline-M`: it always predicts the mean value of the variable of interest, which is computed from the EN score in the training set. The mean value of EN scores in browsing is 2.59 (SD=0.96), and in the searching is 4.78 (SD=0.74).
2. `Baseline-U`: it captures actions selected in phase 2, study A (table 5.1), which is seven actions for browsing and nine for searching. The associated time $t(a_i)$ of each action is extracted from the log files.
3. `BF`: it contains the behavioural features extracted in phase 1, study A (table 4.2).
4. `IETime`: this measure captures actions by grouping into two types, immersion and exploration, as described in phase 2, study A (table 5.1). The associated time of each type of actions, $t(Immersion)$ and $t(Exploration)$, are extracted from the log files.
5. `SeqTime`: this measure is extracted by computing total time spent on the identified discriminative subsequences (table 5.8 and table 5.14) from the log files.
6. `IEPath`: this measure contains two features, namely the AUC of the IE curve and the total time spent on the session. The IE curve is obtained through applying algorithm 1 on the log files.
7. `ExpActionTime`: this measure also captures actions selected in phase 2, study A (table 5.1), which is seven actions for browsing and nine for searching. The associated exponential weighted time spent on each action is extracted from the log files.

Model training and testing using cross-validation (step 3):

The prediction performances were obtained by performing 5-fold cross-validation with the CHiC dataset and 10-fold cross-validation with wikiSearch dataset. To measure the performance of the trained models, we used standard metrics such as Pearson's correlation coefficient (CC), the root mean square error (RMSE), and the mean absolute error (MAE). RMSE and MAE measure

the differences between values predicted by the model and the true EN score. The correlation coefficient (CC) between predicted score and true EN score measures the extent to which the predictions and the true score are linearly related. A low CC suggests that the model explains the data loosely, and a CC with value 0 represent complete absence of correlation. A perfect fit will have a CC closer to 1, and, RMSE and MAE of 0. Two of the three performance measures, namely RMSE, and MAE, are reported as weighted averages as we applied cross-validation. A paired t-test (Dietterich, 1998) is applied to test the significance of the difference between measures that have improvements in MAE and RMSE over `Baseline-U`, against `Baseline-U` using the LR model. The Williams test (Williams, 1959) evaluates significance in a difference in dependent correlations (Steiger, 1980), and it is applied to test the difference of the significant CC against `Baseline-U` using the LR model.

6.3.2 Results and discussion (browsing)

Table 6.3 presents the performance of measures using the CHiC dataset. `Baseline-M` is the very naive model that always predicts the mean of the training data, therefore, a low CC is expected. In general, apart from `BF` and `ExpActionTime`, all the other measures showed improvements over `Baseline-M`. Moreover, measures `IETime` and `IEPath` outperformed `Baseline-U`. The RMSE and MAE of `Baseline-M` are 0.9624 and 0.7864 respectively. `Baseline-U` performs better using LR than SVM, which is the suggested learning method in Machmouchi et al. (2017), with an improved performance (RMSE= 0.9329, and MAE =0.7549) compared to `Baseline-M`. Not surprisingly, behavioural features (`BF`) inspired from existing study (table 4.2) did not outperform `Baseline-U` as they are just suggested to be proxies rather than verified measures. The noise among features in `BF` is partially handled by SVM as the results using SVM is better than using LR. Grouping time spent on two types of actions, exploration and immersion, `IETime` improves the prediction against `Baseline-U` about 0.017 of RMSE using LR model, suggesting simply modelling actions on a higher level leads to a improvements in browsing. `SeqTime` is the metric of time spent on top-ranked discriminative patterns, and the performances of using SVM and LR are about the same. But neither of them outperforms `Baseline-U`. `IEPath` has

the best performance among all seven metrics, with an improvement on RMSE of 0.0343 over `Baseline-U`. Considering the order of the actions across the session (`ExpActionTime`) does not boost the prediction even comparing the `Baseline-M` on both RMSE and MAE. However, none of the CC or the improvements in MAE and RMSE are statistically significant at least on the 0.05 level against `Baseline-U`.

Table 6.3: Performance of proposed measures using two different models (browsing).

Measure	#Features	Model	CC	RMSE	MAE
`Baseline-M`	0	-	-0.0681	0.9624	0.7864
`Baseline-U`	7	SVM	0.3003***	0.9388	0.7744
		LR	0.29***	0.9329	0.7648
`BF`	43	SVM	0.2912***	1.0724	0.8117
		LR	0.1851*	1.6764	0.9356
`IETime`	2	SVM	0.3067***	0.9157	0.7566
		LR	0.3034***	0.916	0.7552
`SeqTime`	2	SVM	0.2259**	0.9423	0.7673
		LR	0.2104**	0.9437	0.7671
`IEPath`	2	SVM	**0.3475*** **	0.9094	**0.7549**
		LR	**0.3551*** **	**0.8986**	0.7586
`ExpActionTime`	7	SVM	0.0933	1.1324	0.8407
		LR	0.0531	0.9635	0.7897

A bold typeface denoted the best result in a column.
Significance level of CC (2-tailed): * = p <0.05; ** = p <0.01; *** = p <0.001.

The performance of all the five measures leads to a couple of questions about the properties of engagement measures in general. First of all, `BF` and `ExpActionTime` perform worse than `Baseline-M` in terms of RMSE and MAE while exhibiting slightly better correlation. Similar results are observed in the first research phase, where not all features in the `BF` set are indicative of EN (figure 4.8). By comparing the performance, even though features in `BF` capture a lot more information in detail such as time and the quality of each query, the result of `BF` is not optimal. Different context aware feature selection methods may need to be employed in this setting. The poor performance of `ExpActionTime` does not validate our prior belief that the latter actions in a user's sequence carry more significance to their engagement score. In browsing, as there is no defined goal, the satisfaction can be achieved in any time of the session, which makes the latter parts not necessarily important. The property P_2^b is only included in `SeqTime`, which performs better than `Baseline-M` but worse than `Baseline-U`, although not significantly. In this setting we observe that the simple measure of `IEPath` outperforms the

rest, which validates our belief that a high-level summarization of the data which is sensitive to the user's behaviour sequence is a judicious choice for a measure of engagement.

6.3.3 Prediction and discussion (searching)

Table 6.4: Performance of proposed measures using two different models (searching).

Measure	#Features	Model	CC	RMSE	MAE
Baseline-M	0	-	-0.0711	0.7361	0.5795
Baseline-U	9	SVM	0.1452**	0.7425	0.5716
		LR	0.1584**	0.7294	0.571
BF	34	SVM	0.1504**	0.7494	0.5863
		LR	0.1249*	0.7615	0.5989
IETime	2	SVM	0.1863***	0.731	0.5601
		LR	0.1907***	0.7216	0.5651
SeqTime	3	SVM	0.1784***	0.7423	**0.5551**
		LR	0.1773***	0.7248	0.563
IEPath	2	SVM	0.1947***	0.7285	0.5653
		LR	**0.2144*** **	**0.7183**	0.564
ExpActionTime	9	SVM	0.1318*	0.7435	0.5705
		LR	0.144**	0.733	0.5749

A bold typeface denoted the best result in a column.
Significance level of CC (2-tailed): * = p <0.05; ** = p <0.01; *** = p <0.001.

Table 6.4 present the performance of measures using the wikiSearch dataset. In general, apart from BF and ExpActionTime, all the other measures showed improvements over Baseline-M with at least one model. Surprisingly, Baseline-M has a comparable performance against Baseline-U, SeqTime and ExpActionTime, as the results of these three using SVM are slightly worse than Baseline-M, where the results using LR are slightly better. We again observed Baseline-M with a low CC and RMSE of 0.7361 and MAE of 0.5795. The RMSE and MAE of Baseline-U using LR are 0.7294 and 0.571 respectively, which are slightly better than Baseline-M but not statistically significant. This means the state-of-art framework did not outperform the naive estimator significantly. Three measures, IETime, IEPath and SeqTime improve the performance over the Baseline-U using at least one model. However, the improvements are still very limited. The best RMSE is achieved by IEPath using LR (RMSE=0.7183, MAE=0.564), and the best MAE is acquired by SeqTime using SVM

(RMSE=0.7423, MAE=0.5551). None of the CC or the improvements are statistically significant at least on the 0.05 level against `Baseline-U`.

The performance of all the measures first point to an interesting observation that `Baseline-M` achieves a reasonable performance compared to other measures. This is potentially caused by the fact that more than 32% of the users have a EN score between 4.5 and 5 (mean= 4.78, distribution in figure 4.10 (c)), and the variance is also small (SD= 0.74). Again `IEPath` coupled with the linear model achieve best performance, although the sequential model `SeqTime` achieves comparable results as well. In this setting `IETime` achieves a comparative better result than `Baseline-U`, which again lends credibility to our hypothesis that a high level summarization of the user's action is a useful dimensionality reduction technique in this setting.

6.4 Discussion

We attempted to create a framework for measuring user perception of engagement from simple, interpretable feature sets that are, crucially, computable on-line and that do not necessitate a complex model training system. The aim was to extend the work of Machmouchi et al. (2017) in developing `Baseline-U` by incorporating properties of the user behaviour signal that we have seen are useful in better adjusting the decision boundary between low and high engagement levels in phase 1 (chapter 4) and phase 2 (chapter 5).

The main point of intuition was moving away from the *bag of actions* model by engineering features that are sensitive to the order that the actions are performed in, while keeping true to measuring time-on-action. We constructed our measures to progressively depart from the baseline, structuring them from weakest `IETime` to most comprehensive `SeqTime`, in an attempt to highlight incremental improvements that incorporating more and more sequential information can produce.

Our studies reveal two very important insights (*RQ.10*). One, is that our intuition that high-level summarization of our data, into either 7 or 9 actions or 2 action types - immersion and exploration - provides useful in encapsulation user behaviour. Indeed on both datasets, some

of the more high-level measures (`IEPath`) end up outperforming the rest. Furthermore the very simple and succinct `IETime` (which contains no sequential information) ends up outperforming both baselines `Baseline-M` and `Baseline-U`.

This brings us to our second point, which is that the sequential information encoded in the action path is difficult to leverage. Indeed we are not confident in asserting which of ether `IEPath` and `SeqTime` performs the best. The differences are too minor to be statistically significant, despite the `SeqTime` method incorporating much more finely grained information about the user behaviour. We wish to mention, that, in light of this, it is hard to recommend `SeqTime` over `IEPath`, owing to the latter's simplicity and easy on-line computability. We remind the reader that a sequence extraction method was required to compute `SeqTime`.

We summarize the measures we created along these two guiding principles:

1. Sequential is better than static. The time evolution in the signal is informative.
2. Action *summarization* is key in denoising the signal and extracting informative patterns from the action sequence.

`IETime` was built from `Baseline-U` by the application of the second principle. We see that in both contexts it matches or outperforms `Baseline-U` despite containing strictly less information. As an inverse example, `BF` is strictly more granular than either of these, but achieves lower performance. One may hypothesize that at hyper-large scales, models with sufficient capacity may learn better from measures such as `BF`, but this eschews the purpose of our study. Achieving such scales through user sampling is in general infeasible, and this study is not aimed at finding a model that is able to capture information from an unstructured set of features, but rather the creation of such a set in compliance with the needs of lightweight modelling frameworks such as LR and SVM.

Measures `SeqTime` and `IEPath`, were consistently the most successful. They benefit from the application of both of the above principles. `SeqTime` achieves summarizarion through the selection of the most discriminative patterns from our sequence analysis in chapter 5. Its

applicability is limited by the overall subset of discriminative sequences one can identify. The time spent on each sequence serves as a relative importance weight. The model used to engineer the measure should have enough capacity to learn the relative scale of these weights. If, in a new setting, new discriminative sequences appear, the model itself will not be able to capture them without repeating the feature extraction step. We note, however that the feature extraction can be achieved as a competitive online algorithm, and wrapped in the overall model, but this would only be meaningful at a much larger scale.

`IEPath` is one of the least conservative measures when it comes to summarization. Apart from the binary *Immersion* and *Exploration* labels, the ordering of the actions and the time spent on each action is needed in order to engineer this measure. It is surprising how in both contexts, this simple construct is sufficient to outperform even the more complex `SeqTime`, and we view this as an argument in favour of keeping the underlying signal simple.

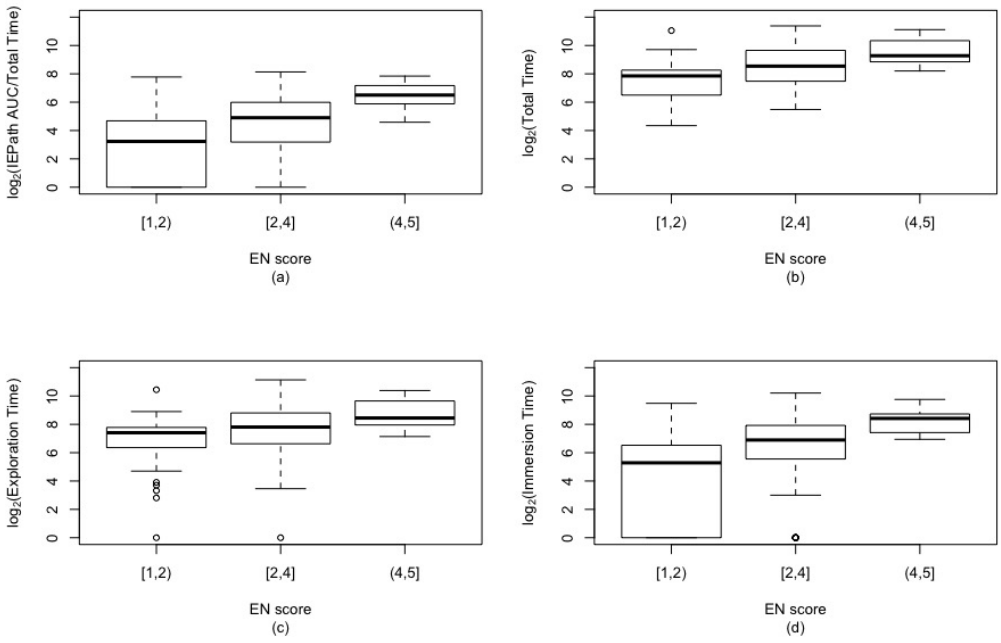


Figure 6.4: Boxplots of (a) the IEPath measure; (b) total time spent on task; (c) exploration time; (d) immersion time (browsing).

It is important to remember that the AUC of this measure holds a direct correlation with total time spent in the session. If we discretize the values of `IEPath`, `ExplorationTime`, `ImmersionTime` and `Time` into EN buckets, in both searching and browsing (figure 6.4 for

browsing, figure 6.5 for searching), we can see some of the effects of aggregation with AUC. Firstly, in browsing, there is more variance in the scale of `IEPath` then in that of `Time`. Despite all the above measures exhibiting positive trends with EN, the one for `IEPath` is clear, proving that the sequential dependence of `IEPath` on mixture of times spent on immersion and exploration actions captures is relevant in describing the user's engagement. In searching (figure 6.5), the interesting take-away is the negative correlation between `ImmersionTime` and EN. `ImmersionTime` in searching can be interpreted as user effort, that is to say, the more user have to invest in reading the document the find answer, the less engaged they are. This correlation in itself suffices to better predict EN, if prior knowledge of this dependence is assumed. One can see its detrimental effects in the aggregation performed by `IEPath`. Despite the trend, the overall relation is weaker.

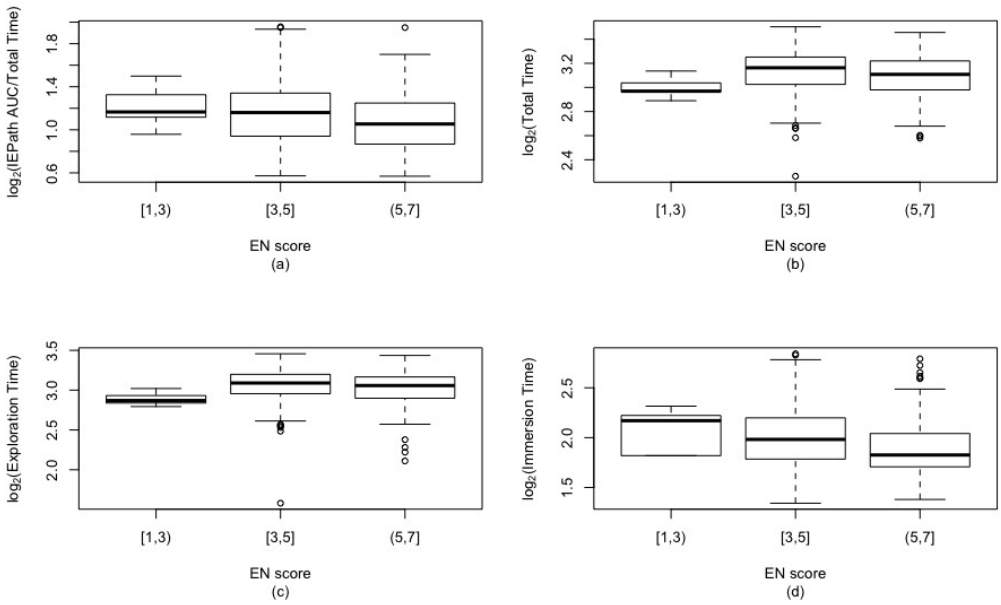


Figure 6.5: Boxplots of (a) the IEPath measure; (b) total time spent on task; (c) exploration time; (d) immersion time (searching).

The area under a curve in the plane is a traditional measure of the trade-off between two quantities in some conservative system. Its classical use is to summarize precision-recall curves, but many other applications exist. The existence of trade-off between Immersion and Exploration is evident from the point of view of the user's cost management (time and effort), and it makes sense to assume that this type of measure performs well in this setting. Interestingly, this mea-

sure is tied to the Rough Path Theoretic signature sequence, which too encodes multiple such areas taken over projections onto various 2-planes. However, the full action path is necessary to define such a measure. We discuss this in the section below.

6.4.1 A note about the full action path

It may seem intriguing that the full action path (as a higher dimensional analogue to the two-dimensional `IEPath`) was not considered in our study. In itself, we believe the full action path constitutes the most complete description of user behaviour we have available at a higher level and we would be remiss not to mention it here. The challenge encountered with this feature is the problem of simple summarization. The problem of extracting useful single invariants out of time-series or categorical paths is a classical one, and we do not attempt to circumvent it here. What does need mentioning is that after experimenting with the invariants we can extract from these paths such as Topological Data Analytic Betti sequences or Rough Path Theoretic (RPT) signature sequences (e.g., (Umeda, 2017; Xie et al., 2018)), we have found them still too complex to be trained properly on our datasets (leading to large numbers of features and heavy overfitting). On the other hand, restricting these methods to their first layers (first and second Betti numbers, for example, or second level of the RPT signature) does not capture enough information about the behaviour sequences, and the performance becomes poor. A middle ground is hard to find and this is rightly so: we have seen first hand during the sequence analysis of phase 2 (section 5.4) that user behaviours are complex enough to give rise to a multitude of discriminative sub-patterns that do not fit together in any simple formula for establishing engagement. We reserve our discussion about the full action path, but note here, that in future studies we expect the information contained within to be levied to a far greater extent and provide great improvements in our understanding of the relationship between user behaviour and user perception of engagement.

6.5 Summary

In phase 3, we identified six ideal properties for measures of engagement based on findings in phase 1 and phase 2 (*RQ.9*). From phase 1 we extracted properties that capture the desired measure's sensitivity to different types of actions and most importantly towards the time spent in each state. In phase 2 we augmented these through the knowledge that the specific patterns either at a low level or at an aggregate level (Immersion / Exploration) are also indicative of user perceived engagement. Thus it would behove our measures to be sensitive to changes in these patterns and how much users find themselves engaged in them.

We then proposed measures that implementing these properties, and evaluated these measures in browsing and searching. Our results, confirm that implementing properties through designed measures improves the performance in predicting Endurability (*RQ.10*).

More specifically, we have found that designing weighted measures of engagement by time, strictly improves performance from our baselines. In line with our intuition, we achieve best performance on the integral of the time-weighted *Immersion / Exploration* curve (`IEPath`). This is consistent amongst datasets, and also achieves the highest correlation with the response. Additionally, this measure is interpretable, easily computable, and lives in a very low dimensional feature space, which is ideal for dynamic online evaluation.

Chapter 7

Summary of findings

7.1 Overview

In this chapter, we recall the main conclusions of our studies, in order to present the reader with a clear comparative view of the results. We omit details, except where necessary for our arguments, and point the reader back to the discussion section for each study. We organize the answers to our questions by the objectives (figure 3.4):

1. *RQ.1* and *RQ.2* from phase 1 (chapter 4) belong to *Obj.1*.
2. *RQ.5* and *RQ.6* from phase 2 (chapter 5) belong to *Obj.2*.
3. *RQ.9* and *RQ.10* from phase 3 (chapter 6) belong to *Obj.3*.
4. *RQ.3* from phase 1 (chapter 4) and *RQ.7* from phase 2 (chapter 5) belong to *Obj.4*.
5. *RQ.4* from phase 1 (chapter 4) and *RQ.8* from phase 2 (chapter 5) belong to *Obj.5*.

We first recall what we did in the three research phases. The first phase investigated the role of discrete behavioural features in inferring user perception of engagement and the differences between the behavioural features - engagement relationships of four engagement dimensions in both browsing and searching.

Phase 2 investigates the potential added benefits of exploring the sequential relationship between user actions with respect to user perception of engagement. This phase is more exploratory than the previous one, which was guided by very standard statistical testing and analysis. We want to replicate as much of the previous study methods here, for comparative power. However, the main feature engineering procedure, namely discriminative subsequence extraction has no counterpart in the previous study.

In research phase 3, we took the information gained from our previous studies and investigated how one might develop context specific measures of engagement through more advanced feature engineering. We explore the general theme of introducing the variance of *time* spent in a certain state to more accurately capture user engagement. It is important in this section that the relationships between our features and the dependent variable are as simple as possible. In particular, we incorporate different amounts of sequential information in the proposed five measures.

7.2 Findings

The findings of this thesis are briefly described here and discussed from the perspective of foundational prior research.

***Obj.1*: To identify and validate the role of behavioural features in inferring user perception of engagement.**

- *RQ.1* What are the behavioural features used in previous studies to describe user perception of engagement?

  This thesis found four categories of behavioural features which are used as engagement proxies in previous studies (e.g., (Lehmann et al., 2012; Song et al., 2013; Kim et al., 2013; Drutsa et al., 2015a; Bai et al., 2017)), namely, click-related, query-related, time-related, and result-related features (table 4.2). The features were associated with common IR

system interfaces or functions, such as SERP and search box, and dwell time. These features were selected based on empirical studies of measuring an aspect of engagement (e.g., click-related features (Lehmann et al., 2012), query-related features (Drutsa et al., 2015a), and time-related features (Kim et al., 2013)) and thus we cannot assume all of them can reflect all engagement dimensions equally well. Collectively, however, these features make up a comprehensive set wherefrom positive correlations between engagement and behaviour can be extracted. This addresses the need to find an ubiquitous set of behavioural features to describe user perception of engagement, and synthesizes the results and efforts of previous studies (e.g., (Thomas et al., 2016; Lehmann et al., 2012)) aimed at feature selection in this domain.

- *RQ.2* To what extent can individual features predict user perception of engagement?

  A key assumption in the field of IR is that behavioural features are proxies of user engagement. Our results, in both browsing and searching, support that a relationship exists between user perception of engagement and user behaviour, and demonstrate the ability of behaviour signals to predict user perception of engagement through a customized task (cf. section 4.4). Certain categories of features are more indicative than others. Time and result related features (e.g., time spent on document pages, number of unique document pages viewed) are found to describe engagement the best in browsing (cf. section 4.3.3), whereas, in searching, time-related features, query-related features, and click-related features (e.g., time spent on SERPs, average CTR, and the average length of queries) contribute more (cf. section 4.3.5). Their importance was also observed with respect to satisfaction (Al-Maskari and Sanderson, 2010; Fox et al., 2005; Belkin et al., 2003; Kim et al., 2013), a crucial component of engagement. However, the correlation of the effect with satisfaction or user perception varies (e.g., the number of clicks and queries are considered as user effort (Lancaster, 1981) with negative influence (Al-Maskari and Sanderson, 2010), or as interests with positive influence (Fox et al., 2005)), which motivates a more context-based interpretation (cf. section 4.3.3 and section 4.3.5).

***Obj.2*: To identify and model user behaviour sequences that have a significant**

association with user perception of engagement.

- *RQ.5* What is the most general set of actions which suffices to describe user interaction with information retrieval systems?

  This thesis found two sets of actions (table 5.1) based on three criteria (cf. section 5.3.1), such as capturing users' click and query behaviour which is found to be informative in phase 1 (answers to *RQ.2*), and the requirement that they fit into the ISP model (Marchionini, 1995) and are available in the common search interface components. In total, seven actions are selected for browsing and nine action are selected for searching (table 5.1). Compared to previous studies using finely grained user actions (e.g., (Mehrotra et al., 2017)), this approach is based on a novel high-level user actions that also integrates with the general model of information seeking process (Marchionini, 1995), but is motivated by the natural intuition to extract signals in a sequential manner when they are presented as such (e.g., (Drutsa et al., 2015a)).

- *RQ.6* What is the relationship between user behaviour sequences and user perception of engagement?

  Our results in both browsing and searching indicate there is a significant relationship between user behaviour sequences (through their frequent subsequences) and user perception of engagement. Extracting these patterns adds to the understanding of intra-session engagement (Lalmas et al., 2014) or the activity dimension (Lehmann et al., 2012), as they extend the interpretation of 'engaged user' from discrete behavioural features to the alternations between actions. The importance in extracting recurrent patterns that are reflective of a generalized type of behaviour is well understood in the IS community (cf. Wilson (1999)), and the current results aim to reinforce this perspective. In our case the patterns describe how behaviour reflects on the user's perception of engagement.

  Three common patterns were identified. The first of these is the alternation between exploration and immersion type actions in the browsing context. This, in accordance with Bates (2007), may represent the user progressing from *glimpses* to *acquisition*, or

closer examination of the object. Its presence correlates positively with NO, FI and EN. Also in the browsing context, but, this time, at the request level, we highlight the second pattern which consists of checking many SERPs about the same request in depth, which indicates the user transitioning into a more involved state as they immerse themselves in the information retrieval task. The duality between session-level and request-level patterns manifests itself in a more complex form in searching (Järvelin et al., 2008) where, it is usually only possible to establish a weak connection (e.g., (Al-Maskari et al., 2007; Song et al., 2013)) between the session-level and request-level user experience (cf. section 5.4.3).

The last pattern is identified in searching (cf. section 5.4.5), in which the users are forced to revisit documents often using the navigation button, possibly retracing or re-analysing documents leading to the users struggling to organise their requests in such a way that a clear progression towards the goal is established. This causes frustration, uncertainty and a potential negative impact on the user's evaluation of the system. The HUPs (high uncertainty patterns) pattern is thus indicative of low engagement levels, especially low EN and PUs. In (Al-Maskari and Sanderson, 2010), the authors attempted to measure effort spent by the user in the act of searching, quantifying it as the number of queries submitted to obtain relevant documents and the rank position in the results list accessed to obtain the relevant documents. The claim made is that increased user effort has a negative impact on user experience in a searching task which our findings here support, despite the differences in how effort is measured in this context.

The added benefits of employing user behaviour sequences is demonstrated through a prediction task (cf. section 5.5). Although in both scenarios sequences struggle to provide statistically significant improvements, the empirical improvements in accuracy are consistent (table 5.16 and table 5.17).

***Obj.3*: To implement and evaluate measures of engagement.**

- *RQ.9* What are the properties that empirically computable measures of engagement should possess?

Our previous studies outline a set of six properties (table 6.1) which characterise such measures in an ideal setting. More specially, from phase 1 we extracted properties that capture the desired measure's sensitivity to different types of actions and most importantly towards the time spent in each state. In phase 2, we augmented these through the knowledge that the specific patterns, whether at a low level or at an aggregate level (Immersion / Exploration,) are also indicative of user perception of engagement. They are meant to reflect the types of behaviour that measures of engagement should be sensitive to, the same behaviour patterns revealed by our behavioural feature analysis and sequence analysis. Similar methods have been adapted in measure design for differentiating "satisfaction with failure" from "unsatisfied success" (Liu et al., 2018), in which they proposed several criteria to represent the two situations and measures were designed based on them. Our study approaches this problem in the same vein and contributes to a bottom up approach for the design of measures of engagement, starting from identifying links, extracting properties and engineering this information into a single quantitative score.

- *RQ.10* Which of the developed measures improve user perception of engagement prediction?

  Our results confirm that implementing the identified properties (section 6.2.2) through designed measures based on time (section 6.2.3) improves the performance in predicting Endurability (cf. section 6.3) with some measures, namely `SeqTime` and `IEPath`. This is consistent in both browsing and searching, and also achieves the highest correlation with the response. Although we fail to achieve significance at the 0.05 level on the improvements, our measures statistically perform at least as well as the state-of-art framework (Machmouchi et al., 2017). More importantly, the surprising merit of these measures is that they are high level summarization of selected actions, easily computable from the log files and live in a very low dimensional feature space. Thus, compared to powerful predictive models of user perception which either require capturing fine-grained behaviour which is not available in all instances, or a large enough sample size for reliable training

(e.g., (Mehrotra et al., 2017; Williams and Zitouni, 2017)), these measures are not as prone to overfitting after training an serve as a succinct but powerful insight into the relationship between the user action sequence and their reported engagement level (cf. section 6.4). We therefore manage to capture a key insight which is novel in the research literature on this topic, namely that high-level summarization of the action space may be sufficient for adequate measures when the sequential nature of behaviour is factored in.

***Obj.4*: To compare behaviour - perception relationship among four different engagement dimensions.**

- *RQ.3* How do the relationships between behavioural features and user perception of engagement vary between dimensions?

  Our findings suggests that behavioural features reflect engagement dimensions differently by comparing the overlap of top ranked behavioural features for each engagement dimension (table 4.6, 4.10).

  In browsing, dimension NO, FI and EN share the largest similarity. NO exhibits the highest correlations with result-related and time-related features. For FI, it is time-related features that stand out. EN seems to combine the two responses and benefits from average correlations from time-related features, result-related features, and click-related features. PUs do not exhibit positive correlations with either. Query related features struggle to reflect either dimension, but correlated best with NO (cf. section 4.3.3). These findings support the intuition that browsing may be curiosity-driven, the feeling of novelty the user experiences directly affecting how much they are motivated to continue using the system and thus leading the user into an involved state. It also supports the design of the UES, in which NO contributes to FI directly (O'Brien and Toms, 2010a), and EN is the overall evaluation of the experience, thus exhibiting a combination of NO and FI in this scenario.

  In searching, all dimensions are relatively different from each other, and the behavioural features are correlated with PUs the most. Interestingly, all high correlations for this

dimensions are negative, suggesting that an affluence of interactions on the user's part are interpreted as a sign of an unusable system. In particular query related features exhibit the highest negative correlation with PUs (table 4.8), which implies in turn that the more queries users are forced to emit to the system to achieve their goal, the lower their satisfaction with the system becomes. This is in stark contrast with the browsing task, where an affluence of queries did not correlate significantly with PUs (cf. section 4.3.5).

The selected dimensions, NO, FI, EN, and PUs were designed to capture different aspects of engagement and are thus interrelated (O'Brien and Toms, 2010a). Overall, EN is the most telling of a user's perceived engagement since it is meant to capture the user's global evaluation at session level and is a testament to the system's endurability and staying power. This is supported by both of our studies, with EN exhibiting a high similarity in its correlations with behaviour with other dimensions (NO and FI in browsing and PUs in searching). The result of this comparison is also a validation of the theory that user engagement is a complex multi-dimensional variable (O'Brien and Toms, 2008) whose effects are observed differently with respect to user behaviour.

- *RQ.7* How do the relationships between user behaviour sequences and user perception of engagement vary between dimensions?

  As mentioned in the answers to *RQ.5*, there are two patterns that indicate high NO, FI and EN in browsing, whereas in searching we distinguish a pattern, namely HUP, that correlates with low PUs and EN.

  The results reported in browsing align with the observation made in phase 1 (answers to *RQ.3*), that three of the dimensions, NO, FI and EN, are similar to each other from the point of view of their relationships with behavioural features. Also, this aligns with the design of the UES (O'Brien and Toms, 2010a) as discussed above. PUs seems qualitatively different with regards to behaviour sequence. Not only is it discriminated by highly infrequent sequences, but these sequences tend to be rather complex (table 5.9), and are also sometimes associated to *low* levels of PUs, indicating that user behaviour can, in this

scenario as well, serve as a critique of the system.

In searching, dimensions PUs and EN are relatively more similar from this perspective (table 5.14 and table 5.15), all their discriminative subsequences being indicative of *low* levels, and both having relative low average frequencies. NO is discriminated, (and in general positively), by the higher ranking subsequences in terms of frequency, despite their overall discriminative power being rather low (table 5.12). It is also the only dimension to exhibit this behaviour. FI has mixed responses, but we note that most of the actions contained in discriminative subsequences are of *immersion* type only, which is unique to this setting (table 5.13).

While predicting user perception of engagement based on behaviour sequences in browsing, we had already singled out PUs as having a different behaviour, and unsurprisingly, it is the one dimension for which the sequence method exhibits statistically significant gains. In searching, the unique behaviour of FI allows it to achieve significant gains through the sequence method as well. This aligns with other studies which conclude that leveraging sequential information benefits user perception prediction (e.g., (Williams and Zitouni, 2017; Mehrotra et al., 2017)). The study of behavioural sequences furthermore confirms the differences between engagement dimensions (O'Brien and Toms, 2010b), and reaffirms the need for their existence and use in order to characterize engagement through self reported methods (Lalmas et al., 2014).

***Obj.5*: To compare behaviour - perception relationship in browsing and searching.**

Together, the following questions serve to differentiate the behaviour - perception relationship in the two major information seeking contexts, browsing and searching. Toms (1998) summarized the primary differences between these two - browsing and searching - in terms of the parameters of the user interaction. On the one hand, searching consists of a focused and organized, command-driven information seeking process. Retrieval occurs through matching a set of specified terms. Browsing, on the other hand, is a less deterministic process. Following from the definition of Bates (Bates, 2007) and the berrypicking model they develop (Bates, 1989), it consists mainly of tasks with an unspecified or ambivalent goal, where the information is ac-

quired through exploration. The main retrieval method is that of recognition, either by items exciting familiarity or piquing the user's interest. We note that, certain individual differences have a larger impact in one or another context. Especially, the level of a person's interest in a given topic would affect her behaviour in browsing. However, the differential impact of these effects in browsing and searching is not investigated in this study.

Unsurprisingly, the four engagement dimensions differ quite heavily across the two contexts as well. In retrospect, we can claim to have shed light on a link between the multidimensional nature of engagement and its varying relationship with behaviour across the two contexts.

- *RQ.4* How do the relationships between behavioural features and user perception of engagement differ between browsing and searching?

  As outlined in our answers to *RQ.2* and *RQ.3* and discussed in more details the comparison section in phase 1 (section 4.5), the interaction between behavioural features and user perception of engagement is, a priori, context dependant. In browsing, dimensions NO, FI and EN exhibit the highest correlations with roughly the same set of behavioural features. These dimensions behave quite similarly. In searching, however, PUs is characterized as the dimension which exhibits the most extreme correlations with the features, but these are mostly negative, for reasons outlined in section 4.3.5. Some of this impact translates over to EN, and hence these dimensions become similar. These differences motivate adequate, context-aware feature selection and engagement modelling in subsequent studies.

- *RQ.8* How do the relationships between user behaviour sequences and user perception of engagement differ between browsing and searching?

  As outlined in our answers to *RQ.6* and *RQ.7*, and discussed in more details the comparison section in phase 2 (section 5.6), the relationships between behaviour sequences and user perception of engagement is context dependant. The most discriminative subsequences have a relatively lower χ^2 value in searching than in browsing. As in phase 1, in browsing there are 3 dimensions, NO, FI and EN, which are similarly discriminated

by frequent subsequences, whereas in searching PUs is discriminated by frequent subsequences, the presence of which indicates low levels. EN follows suite to this behaviour in searching. In general, we observe in phase 2 the similar differences between searching and browsing from phase 1, but interesting patterns of behaviour emerge at the sequence level. These are discussed in section 5.6.

7.3 Summary

This chapter has collected and discussed the findings of this thesis with respect to five objectives. What has emerged is a detailed picture of the relationships between user behaviour and user perception in measuring user engagement in the Information Retrieval domain. Results provide strong evidence that a relationship between user behaviour and user perception of engagement exists, as well as identifying a number of behavioural features and behaviour patterns that are highly correlated with user perception of engagement. We also developed five measures based on six properties that characterise measures of engagement in an ideal setting by incorporating the behaviour- perception relationships we identified. We therefore establish a link between the research questions that pertain to the study of user behaviour and those that serve to characterize user perception of engagement in a novel attempt to structure the analysis of these concepts along the vertical parallelism between them. We will round up this perspective in the following chapter.

Chapter 8

Conclusion

In the present chapter we recall our results with a view towards contributions, exploring our achievements, outlining the limitations of our studies, and preparing to close our arguments on the topic of engagement and its links to user behaviour in information retrieval. Synthesizing our results succinctly, we also look ahead at possible future applications, corollaries and consequences, whose scope is to ultimately prepare more experiments, design new studies and reveal new connections to shed light upon this intricate topic.

8.1 Key contributions

This work contributes to a better understanding of the relationships between user behaviour and user perception of engagement and the issues of measuring engagement, and offers some practical steps towards designing measures to better capture user perception of engagement. It provides contributions from three perspectives, namely theoretical, empirical and methodological. We describe the contributions in the following parts of this section.

8.1.1 Theoretical contributions

From a theoretical perspective, this work is motivated by the gap between the models of HCI in which user perception plays a major role, and current models of IR in which user perception is under emphasized, and behavioural signals dominate. Although such links are illustrated in the two levels of the general model of Information Seeking and Retrieval process (figure 1.3) (Ingwersen and Järvelin, 2006), previously not enough work had been conducted in order to confirm or describe these links. Moreover, the two major information retrieval tasks discussed in our study, browsing and searching and their connections with information behaviour have been the object of much study in the past (Toms, 1998; Marchionini, 1995; Wilson, 1999; Case and Given, 2016). It has thus become of paramount importance within the research community to understand and reveal how and to what extent user behaviour can reflect user perception of engagement within each context.

Our primary contribution is establishing the postulated link between behaviour and perception of engagement in a comprehensive analysis spanning multiple contexts and multiple dimensions of engagement. Identifying this connection can increase the understanding of engagement in a multi-dimensional manner, and supports future research in a multitude of directions such as investigating the relationships between user behaviour and user perception of engagement in a specific domain (e.g., distance learning), and examining the links between the two levels of the general model of Information Seeking and Retrieval process (figure 1.3) as outlined in Ingwersen and Järvelin (2006). In this work, we made the following theoretical contributions:

1. A significant relationship between user behaviour and user perception of engagement was demonstrated at both behavioural feature and behaviour sequence levels. The fundamental contribution of our study is a validated incentive for research to focus not on either side of this relationship but on the transition from the physical to the cognitive level and vice-versa, at least in as much as it pertains to the conscious interaction of users with the system. This contributes to a greater understanding of such relationships between the cognitive-emotional level and the social and physical level of the general model of IS&R (figure 1.3), as mapped in figure 8.1, and promotes a philosophy in dealing with problems

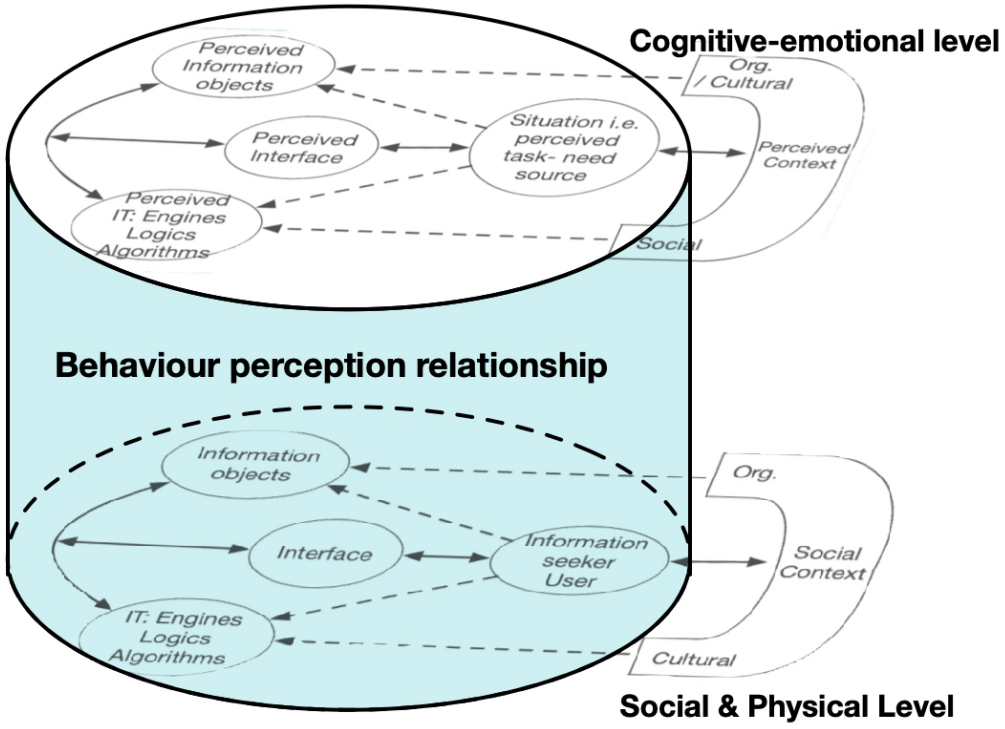


Figure 8.1: The behaviour - perception relationship mapped onto the general Model of IS&R (Ingwersen and Järvelin, 2006).

of this kind, namely that the information encoded in the user behaviour during the interaction is significant in terms of its ability to describe user perception of engagement. Our work thus draws a parallel between the two levels, with the demonstrated relationship between behaviour and perception serving as a connector or bridge between them.

2. A confirmation of the definition of browsing proposed by Bates (2007) was obtained. In particular, we associated exploration and immersion types of actions to the three levels of browsing process according to the state of engagement with the information objects, and identified that the users who are able to progress to the latter two levels as indicated by their behaviour patterns also reported high user perception of engagement (cf. section 5.4.3). These findings also extend the understanding of the connection between a positive user experience in browsing and the user behaviour sequence.

8.1.2 Empirical contributions

From an empirical perspective, this work is motivated by the need for more pragmatic engagement measures that do not interrupt user information retrieval process. In measuring user engagement with systems, typically measures are based on two approaches: one based on user perception of the system, and one on how the user behaves while engaging with the system. Both have been treated as discrete types of measures and rarely integrated. Both also have their limitations such that using only one type is insufficient to measure user engagement holistically. The ones based on user perception are obtrusive to user interaction, and can only be collected at acceptable intervals (Kelly, 2009; Lalmas et al., 2014), while the latter are implicit with mainly descriptive rather than explanatory power (Fazio and Olson, 2003). In many prior studies, user behaviour has been investigated as a proxy of engagement. However, the studies that have successfully verified the links between these two variables do not abound at this point in time. Amongst these few, with the risk of omission, we recall the studies in which the links between satisfaction and behaviour were discussed (Al-Maskari and Sanderson, 2010), and the conclusion of the feasibility of using behaviour features to predict or differentiate satisfaction levels were drawn (Mehrotra et al., 2017; Machmouchi et al., 2017), as well as Arapakis and Leiva (2016) where the connections of gaze/mousing movements and engagement in online news reading and the attention on interface component were established. Verifying this connection supports future research in a multitude of directions such as devising guiding principles for system design or developing tailored measures to specific information retrieval tasks. In this work, we made the following empirical contributions:

1. The key behavioural features for predicting four engagement dimension, namely Novelty, Felt Involvement, Perceived Usability, and Endurability, were identified (cf. answers to *RQ.1* and *RQ.2* in section 7.2). The features themselves belong to four categories according to the types of interaction the user executes and the relationship and prevalence of these features among these categories and how they relate to the above dimensions is discussed. Together, they present a comprehensive numerical description of user behaviour, which adds our results to the forefront of knowledge in predicting user perception

of engagement.

2. Two sequential behaviour patterns, the alternation between exploration and immersion type actions and checking many SERPs about the same request in depth, that are indicative of high engagement in browsing and the HUPs (high uncertainty patterns) pattern that are indicative of low engagement in searching were identified (cf. answers to *RQ.6* in section 7.2). These novel patterns contribute in a direct and interpretable way to understanding the relationship between user perception of engagement and user behaviour and they can serve as validation, baseline or ground truth in designing and executing other studies that look to verify this relationship.

3. Properties of an effective measure of engagement were obtained (cf. table 6.1). These properties generated by our insights showcase our design philosophy and form the basis for a generalized contribution towards our understanding of what the ideal qualities of a measure of engagement are.

4. Empirical evidences were obtained that incorporating information from the user behaviour improved user perception of engagement prediction (cf. section 4.4, section 5.5 and section 6.3). Our studies confirm the ability to predict engagement through behaviour in multiple scales which confirms the multidimensional relationship between engagement and behaviour and motivates further investigation of the predictive power of user behaviour.

5. The measure which captures the time-weighted immersion and exploration path (`IEPath`) developed in this research offers a simple way for representing user perception of engagement. Moreover, the evaluation of this measure showed that high-level measures based on user behaviour perform as well as fine grained low-level measures. We provide here a context specific answer (answers to *RQ.10* in section 7.2) which we hope will motivate future research in the area.

8.1.3 Methodological contributions

In many prior studies examining user behaviour and user perception of their experience, low-level variables of user behaviour were identified by correlation test (Al-Maskari and Sanderson, 2010; Thomas et al., 2016), and they were directly fed into a model to predict the targeted user perception (Mao et al., 2016; Mehrotra et al., 2017). However, these approaches sometimes do not benefit from the pure interpretative insight that a comprehensive analysis of the patterns obtained can provide. Two important steps in our bottom-up approach, manually summarizing the patterns identified in statistical tests, and identifying properties for an ideal measure, increases the understanding of how and why users feel engaged from the user-centred perspective, and supports future research such as measure design for systems with small application scale. In this work, we made the following methodological contributions:

1. The bottom-up approach for measure design, starting from identifying links, then extracting properties, and finally designing measures used in this research serves as a practical set of steps and proof of concept for the construction of implicit measures for user experience. The patterns summarized manually from the findings of statistical tests and the properties identified as outcomes can be used to enhance the understanding and guide measure design for systems with small application scale.

8.2 Limitations

The design of the study naturally impacts the scope of our research. Throughout the text, we have attempted to present our findings in their most appropriate context; following are the limitations of this thesis:

1. The most important caveat is the fact that the results are based on two previously conducted studies, reusing data, which is obviously bound by the respective conditions in which it was collected.

2. The two datasets were collected in two studies that differ in term of information objects, number of participants, and task types. Their potential effect is not investigated in the current study, and should be complemented by further research that considers various types of studies.

3. In this study, we select the UES questionnaire (O'Brien and Toms, 2010a) as the measure of user perception of engagement. The assumption that using UES is appropriate is based on the facts that it has been verified in various contexts including search. Whether the UES is a stable measurement of engagement in information retrieval should be tested in further research.

4. The sizes of selected datasets naturally limit the spectrum of applicable techniques we can employ in terms of modelling. In particular, models with high capacity such as neural architectures and gradient boosting tend to suffer from the overfitting problem, and their power should be complemented by further research that using larger datasets.

8.3 Future work

As previously stated, the aim of this work was two-fold: the acquisition of knowledge on the relationship between user perception of engagement and user behaviour as illustrated by two information retrieval scenarios, and the development of tools or principles aiding this research, such as measures of user perception of engagement based on user behaviour. This work could now be extended by replicating the experiments with large numbers of participants, considering other types of browsing or searching activities, further in-depth analysis of the relationships, and developing system enhancement schemes based on the findings. In more details, we mention:

1. *Replication*: Gather more data to see if the findings and our conclusions generalize.

2. *Influential factors of the behaivour - perception relationships in measuring user engagement*: more detailed analysis to get greater understanding how other factors influence the relationships between user behaviour and user perception of engagement:

(a) *Task types*: investigate further the effect of task types such as time constraints, and complexity.

(b) *Document genres*: investigate further the effect of different document genres such as multimedia.

3. *Behaviour representation*: Develop a faithful low-dimensional representation of user behaviour, which serves as a proxy for engagement, such as using clustering methods to identify discriminative behaviour motifs. This would serve as a step forward to unifying the multiple perspectives to measuring engagement and provide a one-stop resource for future research.

8.4 Summary

The importance of acquiring and retaining engaged users has been the ultimate goal of most applications and websites. Thus, having a useful and efficient approach to measuring engagement is a critical aspect of understanding how to assess and improve systems. Typically measures of engagement are based on two approaches: one based on user perception of the system, and one on how the user behaves while interacting with the system. Both have been treated as discrete types of measures and rarely integrated. Both also have their limitations to the extent that using one type is insufficient to measure user engagement holistically. The ones based on user perception are obtrusive to user interaction, and can only be collected at acceptable intervals, while the latter are implicit with mainly descriptive rather than explanatory power.

The need of integrating user behaviour and user perception in the assessment of user engagement and verifying the relationships between these two in Information Retrieval were the starting-points for this research. We revealed how user behaviour is correlated with user perception of engagement from the discrete behavioural features and the sequence of actions. The major contribution of this work is the establishing of the connection between the cognitive-emotional level and social and physical level of the Information Seeking & Searching Model and the exploration of that relationship. To achieve the latter, we identified sets of behavioural features

and sequential patterns that are indicative of users' engagement level in browsing and searching respectively. Moreover, we employ the insight we garner through the analysis of the correlating statistics in order to develop measures of engagement.

At the current standard of understanding, user engagement is not easily and uniformly defined and thus its measures based on user behaviour and/or user perception require careful context-based treatment. A goal which lies beyond the scope of our current study, but which was gently touched upon in our approach of multiple information retrieval contexts is to provide an overarching framework where from results such as ours can be extracted directly rather than extrapolated based on our insight and understanding of the practical particularities of the problem. More work will necessarily need to be undertaken in order to elucidate these relationships at a larger scale and in a setting general enough to represent a robust characterization of the concept of engagement.

Overall, our current research provides valuable insights into the theory of information seeking and retrieval. Our main contribution is to incentivise the research that incorporates the interplay between behaviour and perception in assessing user engagement. The study of this relation also motivates the construction of measures of engagement from the bottom-up perspective, where interesting properties of the behaviour signal are extracted first and engineering such measures becomes subject to the designer's ability to incorporate these insights. We hope that we have taken one step closer to understanding user engagement and measuring it in the context of information retrieval.

Bibliography

Agarwal, R. and Karahanna, E. (2000). Time flies when you're having fun: Cognitive absorption and beliefs about information technology usage. *MIS quarterly*, pages 665–694.

Ageev, M., Guo, Q., Lagun, D., and Agichtein, E. (2011). Find it if you can: A game for modeling different types of web search success using interaction data. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 345–354, New York, NY, USA. ACM.

Agichtein, E., White, R. W., Dumais, S. T., and Bennet, P. N. (2012). Search, interrupted: understanding and predicting search task continuation. In *Proceedings of the 35th International ACM SIGIR conference on Research and development in information retrieval*, pages 315–324. ACM.

Agosto, D. E. (2002). Bounded rationality and satisficing in young people's webbased decision making. *Journal of the American Society for Information Science and Technology*, 53(1):16–27.

Al-Maskari, A. and Sanderson, M. (2010). A review of factors influencing user satisfaction in information retrieval. *Journal of the American Society for Information Science and Technology*, 61(5):859–868.

Al-Maskari, A., Sanderson, M., and Clough, P. (2007). The relationship between IR effectiveness measures and user satisfaction. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 773–774. ACM.

Allport, F. H. (1955). *Theories of perception and the concept of structure: A review and critical analysis with an introduction to a dynamic-structural theory of behavior.* John Wiley & Sons Inc.

Arapakis, I., Bai, X., and Cambazoglu, B. B. (2014a). Impact of response latency on user behavior in web search. In *Proceedings of the 37th International ACM SIGIR conference on Research and development in information retrieval*, pages 103–112. ACM.

Arapakis, I., Lalmas, M., Cambazoglu, B. B., Marcos, M., and Jose, J. M. (2014b). User engagement in online news: Under the scope of sentiment, interest, affect, and gaze. *Journal of the Association*

for Information Science and Technology, 65(10):1988–2005.

Arapakis, I., Lalmas, M., and Valkanas, G. (2014c). Understanding within-content engagement through pattern analysis of mouse gestures. In *Proc. CIKM'14*, pages 1439–1448. ACM.

Arapakis, I. and Leiva, L. A. (2016). Predicting user engagement with direct displays using mouse cursor information. In *Proc. SIGIR'16*, pages 103–112. ACM.

Arapakis, I., Moshfeghi, Y., Joho, H., Ren, R., Hannah, D., and Jose, J. M. (2009). Integrating facial expressions into user profiling for the improvement of a multimodal recommender system. In *Proceedings of the 2009 IEEE International Conference on Multimedia and Expo*, ICME'09, pages 1440–1443. IEEE Press.

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

Azzopardi, L. and Zuccon, G. (2015). An analysis of theories of search and search behavior. In *Proceedings of the 2015 International Conference on The Theory of Information Retrieval*, pages 81–90. ACM.

Bai, X., Arapakis, I., Cambazoglu, B. B., and Freire, A. (2017). Understanding and leveraging the impact of response latency on user behaviour in web search. *ACM Trans. Inf. Syst.*, 36(2):1–42.

Banhawi, F. and Ali, N. M. (2011). Measuring user engagement attributes in social networking application. In *Proc. STAIR'11*, pages 297–301. IEEE.

Barreda-Àngeles, M., Arapakis, I., Bai, X., Cambazoglu, B. B., and Pereda-Baos, A. (2015). Unconscious physiological effects of search latency on users and their click behaviour. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 203–212. ACM.

Bateman, S., Teevan, J., and White, R. W. (2012). The search dashboard: How reflection and comparison impact search behavior. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, pages 1785–1794, New York, NY, USA. ACM.

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

Bates, M. J. (2007). What is browsing really? a model drawing from behavioural science research. 12.

Belkin, N. J., Kelly, D., Kim, G., Kim, J.-Y., Lee, H.-J., Muresan, G., Tang, M.-C., Yuan, X.-J., and Cool, C. (2003). Query length in interactive information retrieval. In *Proc. SIGIR'03*, pages

205–212, New York, NY, USA. ACM.

Belkin, N. J., Oddy, R. N., and Brooks, H. M. (1997). Ask for information retrieval: part I: background and theory. In Karen Sparck, J. and Peter, W., editors, *Readings in information retrieval*, pages 299–304. Morgan Kaufmann Publishers Inc.

Bennett, P. N., White, R. W., Chu, W., Dumais, S. T., Bailey, P., Borisyuk, F., and Cui, X. (2012). Modeling the impact of short-and long-term behavior on search personalization. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 185–194. ACM.

Bernard, H. R. and Bernard, H. R. (2012). *Social research methods: Qualitative and quantitative approaches*. Sage.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Machine Learning. Springer.

Bødker, S. (2006). When second wave HCI meets third wave challenges. In *Proceedings of the 4th Nordic conference on Human-computer interaction: changing roles*, pages 1–8. ACM.

Bødker, S. (2015). Third-wave HCI, 10 years later—participation and sharing. *interactions*, 22(5):24–31.

Borgman, C. L. (1989). All users of information retrieval systems are not created equal: an exploration into individual differences. *Inf. Process. Manage.*, 25(3):237–251.

Borlund, P. (2013). Interactive information retrieval: An introduction. *Journal of Information Science Theory and Practice*, 1(3):12–32.

Bota, H., Zhou, K., and Jose, J. M. (2016). Playing your cards right: The effect of entity cards on search behaviour and workload. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval*, pages 131–140. ACM.

Breckler, S. J. (1984). Empirical validation of affect, behavior, and cognition as distinct components of attitude. *Journal of personality and social psychology*, 47(6):1191.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Brooke, J. (1996). SUS-A quick and dirty usability scale. *Usability evaluation in industry*, 189(194):4–7.

Bruner, J. S. and Minturn, A. L. (1955). Perceptual identification and perceptual organization. *The Journal of General Psychology*, 53(1):21–28.

Bryman, A. (2016). *Social research methods*. Oxford university press.

Budylin, R., Drutsa, A., Katsev, I., and Tsoy, V. (2018). Consistent transformation of ratio met-

rics for efficient online controlled experiments. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 55–63. ACM.

Burton-Jones, A. (2009). Minimizing method bias through programmatic research. *MIS Quarterly*, pages 445–471.

Buscher, G., Dumais, S. T., and Cutrell, E. (2010). The good, the bad, and the random: an eye-tracking study of ad quality in web search. In *Proc. SIGIR'10*, pages 42–49. ACM.

Buscher, G., White, R. W., Dumais, S., and Huang, J. (2012). Large-scale analysis of individual and task differences in search result page examination strategies. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 373–382. ACM.

Byström, K. and Järvelin, K. (1995). Task complexity affects information seeking and use. *Information processing and management*, 31(2):191–213.

Cacioppo, J. T., Tassinary, L. G., and Berntson, G. (2007). *Handbook of psychophysiology*. Cambridge University Press.

Calisir, F. and Calisir, F. (2004). The relation of interface usability characteristics, perceived usefulness, and perceived ease of use to end-user satisfaction with enterprise resource planning (ERP) systems. *Computers in human behavior*, 20(4):505–515.

Capra, R., Arguello, J., O'Brien, H., Li, Y., and Choi, B. (2018). The effects of manipulating task determinability on search behaviors and outcomes. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 445–454. ACM.

Case, D. O., Andrews, J. E., Johnson, J. D., and Allard, S. L. (2005). Avoiding versus seeking: the relationship of information seeking to avoidance, blunting, coping, dissonance, and related concepts. *Journal of the Medical Library Association*, 93(3):353.

Case, D. O. and Given, L. M. (2016). *Looking for Information: A Survey of Research on Information Seeking, Needs, and Behavior*. Studies in Information. Emerald Group Publishing Limited.

Chapelle, O. and Zhang, Y. (2009). A dynamic bayesian network click model for web search ranking. In *Proceedings of the 18th international conference on World wide web*, pages 1–10. ACM.

Chapman, P., Selvarajah, S., and Webster, J. (1999). Engagement in multimedia training systems. In *Proceedings of the 32nd Annual Hawaii International Conference on System Science*, pages 1–9. IEEE.

Charlton, J. P. and Danforth, I. D. (2010). Validating the distinction between computer addiction and engagement: online game playing and personality. *Behaviour & Information Technology*, 29(6):601–

613.

Chuklin, A. and de Rijke, M. (2016). Incorporating clicks, attention and satisfaction into a search engine result page evaluation model. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 175–184. ACM.

Clarke, E. M. and Wing, J. M. (1996). Formal methods: state of the art and future directions. *ACM Comput. Surv.*, 28(4):626–643.

Cole, M., Liu, J., Belkin, N., Bierig, R., Gwizdka, J., Liu, C., Zhang, J., and Zhang, X. (2009). Usefulness as the criterion for evaluation of interactive information retrieval. *Proceedings of the Third Human Computer Information Retrieval Workshop*, pages 1–4.

Cole, M. J., Gwizdka, J., Liu, C., Belkin, N. J., and Zhang, X. (2013). Inferring user knowledge level from eye movement patterns. *Information Processing & Management*, 49(5):1075–1091.

Cooper, W. S. (1976). The paradoxical role of unexamined documents in the evaluation of retrieval effectiveness. *Information Processing and Management*, 12(6):367–375.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.

Dave, K. S., Vaingankar, V., Kolar, S., and Varma, V. (2013). Timespent based models for predicting user retention. In *Proc. WWW'13*, pages 331–342, New York, NY, USA. ACM.

DeVellis, R. (2003). *Scale Development*. Sage, Newbury Park, California, 2nd edition.

Diaz, F., White, R., Buscher, G., and Liebling, D. (2013). Robust models of mouse movement on dynamic web search results pages. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pages 1451–1460. ACM.

Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation*, 10(7):1895–1923.

Dillon, A. and Watson, C. (1996). User analysis in hci:the historical lessons from individual differences research. *International Journal of Human-Computer Studies*, 45(6):619–637.

Diriye, A., White, R., Buscher, G., and Dumais, S. (2012). Leaving so soon?: understanding and predicting web search abandonment rationales. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 1025–1034, 2398399. ACM.

Dodge, Y. (2006). *The Oxford dictionary of statistical terms*. Oxford University Press on Demand.

Donato, D., Bonchi, F., Chi, T., and Maarek, Y. (2010). Do you want to take notes?: identifying research missions in yahoo! search pad. In *Proceedings of the 19th international conference on World wide web*, pages 321–330. ACM.

Donker, A. and Markopoulos, P. (2002). *A comparison of think-aloud, questionnaires and interviews for testing usability with children*, pages 305–316. Springer.

Downey, D., Dumais, S., Liebling, D., and Horvitz, E. (2008). Understanding the relationship between searchers' queries and information goals. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 449–458. ACM.

Drutsa, A., Gusev, G., and Serdyukov, P. (2015a). Engagement periodicity in search engine usage: Analysis and its application to search quality evaluation. In *Proc. WSDM '15*, pages 27–36, New York, NY, USA. ACM.

Drutsa, A., Gusev, G., and Serdyukov, P. (2015b). Future user engagement prediction and its application to improve the sensitivity of online experiments. In *Proceedings of the 24th International Conference on World Wide Web*, pages 256–266. ACM.

Dupret, G. and Lalmas, M. (2013). Absence time and user engagement: evaluating ranking functions. In *Proc. WSDM'13*, pages 173–182. ACM.

Eagly, A. H. and Chaiken, S. (1998). *Attitude structure and function.* McGraw-Hill.

Ekman, P. and Friesen, W. V. (1976). Measuring facial movement. *Environmental psychology and nonverbal behavior*, 1(1):56–75.

Ekman, P. and Rosenberg, E. (2005). *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS).* Series in Affective Science. Oxford University Press.

Fazio, R. H. and Olson, M. A. (2003). Implicit measures in social cognition research: Their meaning and use. *Annual review of psychology*, 54(1):297–327.

Fenichel, C. H. (1981). Online searching: Measures that discriminate among users with different types of experiences. *Journal of the American Society for Information Science*, 32(1):23–32.

Fischer, M. H. (1999). An investigation of attention allocation during sequential eye movement tasks. *The Quarterly Journal of Experimental Psychology: Section A*, 52(3):649–677.

Forlizzi, J. and Battarbee, K. (2004). Understanding experience in interactive systems. In *Proceedings of the 5th conference on Designing interactive systems: processes, practices, methods, and techniques*, pages 261–268. ACM.

Fornell, C. and Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of marketing research*, 18(1):39–50.

Fox, S., Karnawat, K., Mydland, M., Dumais, S., and White, T. (2005). Evaluating implicit measures

to improve web search. *ACM Transactions on Information Systems*, 23(2):147–168.

Freund, L. and Toms, E. G. (July 2007). Revisiting informativeness as a process measure for information interaction. In *Web Information Seeking and Interaction Workshop, held in conjunction with the 30th Annual International ACM SIGIR Conference*, pages 33–36.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.

Garcia-Gathright, J., St Thomas, B., Hosey, C., Nazari, Z., and Diaz, F. (2018). Understanding and evaluating user satisfaction with music discovery. In *Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 55–64. ACM.

Gibson, J. J. (1966). *The senses considered as perceptual systems.* Houghton Mifflin.

Gregory, R. L. (1970). *The intelligent eye.* McGraw-Hill Companies.

Guo, Q., Lagun, D., and Agichtein, E. (2012). Predicting web search success with fine-grained interaction data. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 2050–2054. ACM.

Hall, M. M. and Toms, E. G. (2013). Building a common framework for iir evaluation. In *Information Access Evaluation. Multilinguality, Multimodality, and Visualization*, pages 17–28. Springer.

Hall, M. M., Villa, R., Rutter, S. A., Bell, D., Clough, P., and Toms, E. G. (2013). Sheffield submission to the chic interactive task: Exploring digital cultural heritage. In *Proceedings CLEF 2013, Working Note*.

Hartmann, J., Sutcliffe, A., and Angeli, A. D. (2008). Towards a theory of user judgment of aesthetics and user interface quality. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 15(4):15.

Hassan, A., Jones, R., and Klinkner, K. L. (2010). Beyond DCG: user behavior as a predictor of a successful search. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 221–230. ACM.

Hassan, A., Song, Y., and He, L.-w. (2011). A task level metric for measuring web search satisfaction and its application on improving relevance estimation. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, pages 125–134, New York, NY, USA. ACM.

Hassenzahl, M. and Tractinsky, N. (2006). User experience-a research agenda. *Behaviour & information technology*, 25(2):91–97.

Hersh, W., Turpin, A., Price, S., Chan, B., Kramer, D., Sacherek, L., and Olson, D. (2000). Do batch

and user evaluations give the same results? In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 17–24. ACM.

Huang, J., White, R., and Buscher, G. (2012). User see, user point: gaze and cursor alignment in web search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1341–1350. ACM.

Hutchins, E. L., Hollan, J. D., and Norman, D. A. (1985). Direct manipulation interfaces. *Human–Computer Interaction*, 1(4):311–338.

Huurdeman, H. C., Wilson, M. L., and Kamps, J. (2016). Active and passive utility of search interface features in different information seeking task stages. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval*, CHIIR '16, pages 3–12, New York, NY, USA. ACM.

Ingwersen, P. and Järvelin, K. (2006). *The turn: Integration of information seeking and retrieval in context*, volume 18. Springer Science & Business Media.

Ingwersen, P. E. R. (1992). *Information Retrieval Interaction.* Taylor Graham Publishing.

ISO (2010). *Ergonomics of Human-system Interaction - Part 210: Human-centred Design for Interactive Systems (ISO 9241-210:2010).* International Organization for Standardization.

Jacques, R. D. (1996). *The nature of engagement and its role in hypermedia evaluation and design.* PhD thesis, South Bank University.

Jansen, B. J. and Rieh, S. Y. (2010). The seventeen theoretical constructs of information searching and information retrieval. *Journal of the American Society for Information Science and Technology*, 61(8):1517–1534.

Järvelin, K. and Kekäläinen, J. (2000). Ir evaluation methods for retrieving highly relevant documents. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 41–48. ACM.

Järvelin, K., Price, S. L., Delcambre, L. M., and Nielsen, M. L. (2008). Discounted cumulated gain based evaluation of multiple-query ir sessions. In *European Conference on Information Retrieval*, pages 4–15. Springer.

Joachims, T. (2002). Optimizing search engines using clickthrough data. In *Proc. KDD '02*, pages 133–142, New York, NY, USA. ACM.

Joachims, T., Granka, L., Pan, B., Hembrooke, H., and Gay, G. (2017). Accurately interpreting clickthrough data as implicit feedback. *SIGIR Forum*, 51(1):4–11.

Juan, Y.-F. and Chang, C.-C. (2005). An analysis of search engine switching behavior using click streams. In *International Workshop on Internet and Network Economics*, pages 806–815. Springer.

Junco, R. (2013). Comparing actual and self-reported measures of facebook use. *Computers in Human Behavior*, 29(3):626–631.

Kelly, D. (2009). *Methods for evaluating interactive information retrieval systems with users*, volume 3 of *Foundations and Trends in Information Retrieval*. Now Publisher.

Kelly, D. and Sugimoto, C. R. (2013). A systematic review of interactive information retrieval evaluation studies, 1967–2006. *Journal of the American Society for Information Science and Technology*, 64(4):745–770.

Kim, J. Y., Cramer, M., Teevan, J., and Lagun, D. (2013). Understanding how people interact with web search results that change in real-time using implicit feedback. In *Proceedings of the 22nd ACM international conference on Conference on information knowledge management*, pages 2321–2326. ACM.

Kim, Y., Hassan, A., White, R. W., and Zitouni, I. (2014). Modeling dwell time to predict click-level satisfaction. In *Proc. WSDM'14*, pages 193–202. ACM.

Kiseleva, J., Williams, K., Awadallah, A. H., Crook, A. C., Zitouni, I., and Anastasakos, T. (2016a). Predicting user satisfaction with intelligent assistants. In *Proc. SIGIR '16*, pages 45–54, New York, NY, USA. ACM.

Kiseleva, J., Williams, K., Jiang, J., Awadallah, A. H., Crook, A. C., Zitouni, I., and Anastasakos, T. (2016b). Understanding user satisfaction with intelligent assistants. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval*, pages 121–130. ACM.

Kobayashi, T. and Boase, J. (2012). No such effect? the implications of measurement error in self-report measures of mobile communication use. *Communication Methods and Measures*, 6(2):126–143.

Kohavi, R. and John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence*, 97(1):273 – 324.

Kostkova, P. (2016). User engagement with digital health. In *Why engagement matters*, pages 127–156. Springer.

Kotov, A., Bennett, P. N., White, R. W., Dumais, S. T., and Teevan, J. (2011). Modeling and analysis of cross-session search tasks. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 5–14. ACM.

Kraft, D. H. and Lee, T. (1979). Stopping rules and their effect on expected search length. *Information Processing and Management*, 15(1):47–58.

Kruschwitz, U., Lungley, D., Albakour, M.-D., and Song, D. (2013). Deriving query suggestions for site search. *Journal of the American Society for Information Science and Technology*, 64(10):1975–1994.

Kuhlthau, C. C. (1991). Inside the search process: Information seeking from the user's perspective. *Journal of the American society for information science*, 42(5):361.

Kuhlthau, C. C. (1993). A principle of uncertainty for information seeking. *Journal of documentation*, 49(4):339–355.

Lagun, D., Ageev, M., Guo, Q., and Agichtein, E. (2014). Discovering common motifs in cursor movement data for improving web search. In *Proc. WSDM'14*, pages 183–192. ACM.

Lalmas, M., O'Brien, H., and Yom-Tov, E. (2014). Measuring user engagement. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 6(4):1–132.

Lancaster, F. W. (1981). Evaluation within the environment of an operating information service. *Information retrieval experiment*, pages 103–127.

Lang, P. J. (1995). The emotion probe: Studies of motivation and attention. *American psychologist*, 50(5):372.

Laurel, B. (1993). *Computers As Theatre*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

Lavie, T. and Tractinsky, N. (2004). Assessing dimensions of perceived visual aesthetics of web sites. *International journal of human-computer studies*, 60(3):269–298.

Law, E. L.-C., Klobučar, T., and Pipan, M. (2006). User effect in evaluating personalized information retrieval systems. In *European Conference on Technology Enhanced Learning*, pages 257–271. Springer.

Law, E. L.-C., Roto, V., Hassenzahl, M., Vermeeren, A. P., and Kort, J. (2009). Understanding, scoping and defining user experience: a survey approach. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 719–728. ACM.

Leckie, G. J., Pettigrew, K. E., and Sylvain, C. (1996). Modeling the information seeking of professionals: A general model derived from research on engineers, health care professionals, and lawyers. *The Library Quarterly*, pages 161–193.

Lefebvre, R. C., Tada, Y., Hilfiker, S. W., and Baur, C. (2010). The assessment of user engagement with ehealth content: The ehealth engagement scale. *Journal of Computer-Mediated Communication*,

15(4):666–681.

Lehmann, J., Lalmas, M., Yom-Tov, E., and Dupret, G. (2012). Models of user engagement. In *Proc. UMAP'12*, pages 164–175, Berlin, Heidelberg. Springer-Verlag.

Lewis, J. R. (1995). IBM computer usability satisfaction questionnaires: psychometric evaluation and instructions for use. *International Journal of HumanComputer Interaction*, 7(1):57–78.

Li, Y., Martinez, O., Chen, X., Li, Y., and Hopcroft, J. E. (2016). In a world that counts: Clustering and detecting fake social engagement at scale. In *Proceedings of the 25th International Conference on World Wide Web*, pages 111–120. ACM.

Liu, C., Liu, J., Belkin, N., Cole, M., and Gwizdka, J. (2011). Using dwell time as an implicit measure of usefulness in different task types. *Proceedings of the American Society for Information Science and Technology*, 48(1):1–4.

Liu, M., Liu, Y., Mao, J., Luo, C., Zhang, M., and Ma, S. (2018). "satisfaction with failure" or "unsatisfied success": Investigating the relationship between search success and user satisfaction. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, pages 1533–1542, Republic and Canton of Geneva, Switzerland. ACM.

Lopatovska, I. and Arapakis, I. (2011). Theories, methods and current research on emotions in library and information science, information retrieval and human–computer interaction. *Information Processing and Management*, 47(4):575–592.

Ma, H., Liu, X., and Shen, Z. (2016). User fatigue in online news recommendation. In *Proceedings of the 25th International Conference on World Wide Web*, pages 1363–1372. ACM.

Machmouchi, W., Awadallah, A. H., Zitouni, I., and Buscher, G. (2017). Beyond success rate: Utility as a search quality metric for online experiments. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17, pages 757–765, New York, NY, USA. ACM.

Mao, J., Liu, Y., Zhou, K., Nie, J.-Y., Song, J., Zhang, M., Ma, S., Sun, J., and Luo, H. (2016). When does relevance mean usefulness and user satisfaction in web search? In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, pages 463–472, New York, NY, USA. ACM.

Marchionini, G. (1995). *Information seeking in electronic environments*. Cambridge University Press.

Marchionini, G. (1996). Resource search and discovery. In *in Research Agenda for Networked Cultural Heritage, Getty Art History Information Program ed., Santa Monica, CA: Getty AHIP*. Citeseer.

Mauri, M., Cipresso, P., Balgera, A., Villamira, M., and Riva, G. (2011). Why is facebook so success-

ful? psychophysiological measures describe a core flow state while using facebook. *Cyberpsychology, Behavior, and Social Networking*, 14(12):723–731.

McCay-Peet, L. and Toms, E. G. (2015). Investigating serendipity: How it unfolds and what may influence it. *Journal of the Association for Information Science and Technology*, 66(7):1463–1476.

Mehrotra, R., Zitouni, I., Hassan Awadallah, A., Kholy, A. E., and Khabsa, M. (2017). User interaction sequences for search satisfaction prediction. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, pages 165–174, New York, NY, USA. ACM.

Moshfeghi, Y., Pinto, L. R., Pollick, F. E., and Jose, J. M. (2013). Understanding relevance: An fmri study. In *Proceedings of the 35th European Conference on IR Research*, pages 14–25. Springer.

Moshfeghi, Y. and Pollick, F. E. (2018). Search process as transitions between neural states. In *Proceedings of the 2018 World Wide Web Conference*, pages 1683–1692. ACM.

Moshfeghi, Y., Triantafillou, P., and Pollick, F. E. (2016). Understanding information need: An fmri study. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 335–344. ACM.

Nielsen, J. (1994). *Usability engineering*. Elsevier.

Norman, D. A. and Draper, S. W. (1986). *User Centered System Design; New Perspectives on Human-Computer Interaction*. L. Erlbaum Associates Inc.

Norman, G. (2010). Likert scales, levels of measurement and the "laws" of statistics. *Advances in health sciences education*, 15(5):625–632.

O'Brien, H. L. (2011). Exploring user engagement in online news interactions. volume 48, pages 1–10.

O'Brien, H. L. (2016). Theoretical perspectives on user engagement. In O'Brien, H. L. and Cairns, P., editors, *Why engagement matters*, pages 105–126. Springer.

O'Brien, H. L., Cairns, P., and Hall, M. (2018). A practical approach to measuring user engagement with the refined user engagement scale (UES) and new UES short form. *International Journal of Human-Computer Studies*, 112:28 – 39.

O'Brien, H. L. and Toms, E. G. (2008). What is user engagement? a conceptual framework for defining user engagement with technology. *Journal of the American Society for Information Science and Technology*, 59(6):938–955.

O'Brien, H. L. and Toms, E. G. (2010a). The development and evaluation of a survey to measure user engagement. *Journal of the American Society for Information Science and Technology*, 61(1):50–69.

O'Brien, H. L. and Toms, E. G. (2010b). Is there a universal instrument for measuring interactive information retrieval?: the case of the user engagement scale. In *Proc. IIiX'10*, pages 335–340. ACM.

O'Brien, H. L. and Toms, E. G. (2013). Examining the generalizability of the user engagement scale (UES) in exploratory search. *Information Processing and Management*, 49(5):1092–1107.

Odijk, D., White, R. W., Hassan Awadallah, A., and Dumais, S. T. (2015). Struggling and success in web search. In *Proc. CIKM'15*, pages 1551–1560. ACM.

Peterson, E. T. and Carrabis, J. (2008). Measuring the immeasurable: Visitor engagement. *Web Analytics Demystified*, 14:16.

Petras, V., Bogers, T., Toms, E., Hall, M., Savoy, J., Malak, P., Pawłowski, A., Ferro, N., and Masiero, I. (2013). Cultural Heritage in CLEF (CHiC) 2013. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 192–211. Springer.

Pirolli, P. and Card, S. (1995). Information foraging in information access environments. In *Proceedings of the conference on Human factors in computing systems*, pages 51–58. ACM.

Pirolli, P. and Card, S. (1999). Information foraging. *Psychological review*, 106(4):643.

Platt, J. (1998). Fast training of support vector machines using sequential minimal optimization.

Podsakoff, P. M., MacKenzie, S. B., and Podsakoff, N. P. (2012). Sources of method bias in social science research and recommendations on how to control it. *Annual review of psychology*, 63:539–569.

Ponnuswami, A. K., Pattabiraman, K., Brand, D., and Kanungo, T. (2011). Model characterization curves for federated search using click-logs: Predicting user engagement metrics for the span of feasible operating points. In *Proc. WWW '11*, pages 67–76, New York, NY, USA. ACM.

Powers, D. M. W. (2011). Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. *International Journal of Machine Learning Technology*, 2(1):37–63.

Prabha, C., Silipigni Connaway, L., Olszewski, L., and Jenkins, L. R. (2007). What is enough? satisficing information needs. *Journal of documentation*, 63(1):74–89.

Quesenbery, W. (2003). *Dimensions of usability*, volume 20. Lawrence Erlbaum Associates Mahwah, NJ.

Rayner, K. (2012). *Eye movements and visual cognition: Scene perception and reading*. Springer Science & Business Media.

Rice, R. E., McCreadie, M., and Chang, S.-J. L. (2001). *Accessing and browsing information and*

communication. MIT Press.

Robertson, S. E. (1977). The probability ranking principle in IR. *Journal of documentation*, 33(4):294–304.

Rodden, K., Fu, X., Aula, A., and Spiro, I. (2008). Eye-mouse coordination patterns on web search results pages. In *CHI'08 extended abstracts on Human factors in computing systems*, pages 2997–3002. ACM.

Rogers, E. M. (1986). *Communication technology*, volume 1. Simon and Schuster.

Ross, S. M. (2014). *Introduction to probability models*. Elsevier Science.

Sauer, J. and Sonderegger, A. (2009). The influence of prototype fidelity and aesthetics of design in usability tests: Effects on user behaviour, subjective evaluation and emotion. *Applied ergonomics*, 40(4):670–677.

Savenkov, D., Lagun, D., and Liu, Q. (2013). Search engine switching detection based on user personal preferences and behavior patterns. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 33–42. ACM.

Shah, C., Hendahewa, C., and González-Ibáñez, R. (2015). Rain or shine? forecasting search process performance in exploratory search tasks. *Journal of the Association for Information Science and Technology*.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell system technical journal*, 27(3):379–423.

Simon, H. A. (1955). A behavioral model of rational choice. *The quarterly journal of economics*, pages 99–118.

Song, Y., Shi, X., and Fu, X. (2013). Evaluating and predicting user engagement change with degraded search relevance. In *Proc. WWW'13*, pages 1213–1224. ACM.

Spearman, C. (1904). The proof and measurement of association between two things. *The American journal of psychology*, 15(1):72–101.

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological bulletin*, 87(2):245.

Su, L. T. (1992). Evaluation measures for interactive information retrieval. *Inf. Process. Manage.*, 28(4):503–516.

Su, L. T. (2003). A comprehensive and systematic model of user evaluation of web search engines: I. theory and background. *Journal of the American society for information science and technology*,

54(13):1175–1192.

Sutcliffe, A. (2009). Designing for user engagement: Aesthetic and attractive user interfaces. *Synthesis lectures on human-centered informatics*, 2(1):1–55.

Sutcliffe, A. (2016). Designing for user experience and engagement. In *Why engagement matters*, pages 105–126. Springer.

Tague, J. (1987). Informativeness as an ordinal utility function for information retrieval. *SIGIR Forum*, 21(3-4):10–17.

Tague-Sutcliffe, J. (1995). *Measuring information : an information services perspective*. Academic Press, San Diego ; London.

Tate, R. F. (1954). Correlation between a discrete and a continuous variable. point-biserial correlation. *The Annals of mathematical statistics*, 25(3):603–607.

Teevan, J., Liebling, D. J., and Geetha, G. R. (2011). Understanding and predicting personal navigation. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 85–94. ACM.

Teo, C. H., Nassif, H., Hill, D., Srinivasan, S., Goodman, M., Mohan, V., and Vishwanathan, S. (2016). Adaptive, personalized diversity for visual discovery. In *Proceedings of the 10th ACM Conference on Recommender Systems*, pages 35–38. ACM.

Thomas, P., Moffat, A., Bailey, P., and Scholer, F. (2014). Modeling decision points in user search behavior. In *Proceedings of the 5th Information Interaction in Context Symposium*, IIiX '14, pages 239–242, New York, NY, USA. ACM.

Thomas, P., O'Brien, H., and Rowlands, T. (2016). Measuring engagement with online forms. In *Proc. CHIIR '16*, pages 325–328, New York, NY, USA. ACM.

Toms, E. G. (1998). *Browsing digital information: Examining the 'affordances' in the interaction of user and text.* PhD thesis, University of West Ontario.

Toms, E. G., Freund, L., and Li, C. (2004). Wire: the web interactive information retrieval experimentation system prototype. *Information Processing and Management*, 40(4):655–675.

Toms, E. G., Villa, R., and McCay-Peet, L. (2013). How is a search system used in work task completion? *Journal of information science*, 39(1):15–25.

Tractinsky, N., Katz, A. S., and Ikar, D. (2000). What is beautiful is usable. *Interacting with computers*, 13(2):127–145.

Turpin, A. and Scholer, F. (2006). User performance versus precision measures for simple search tasks.

In *Proc. SIGIR '06*, pages 11–18, New York, NY, USA. ACM.

Umeda, Y. (2017). Time series classification via topological data analysis. *Information and Media Technologies*, 12:228–239.

Varian, H. R., Bergstrom, T. C., and West, J. E. (1996). *Intermediate microeconomics*, volume 4. Norton New York.

Watson, D., Clark, L. A., and Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: the panas scales. *Journal of personality and social psychology*, 54(6):1063.

Webster (2016). *The Merriam-Webster Dictionary, International Edition*. Merriam-Webster, Incorporated.

Webster, J. and Ahuja, J. S. (2006). Enhancing the design of web navigation systems: the influence of user disorientation on engagement and performance. *MIS Quarterly*, pages 661–678.

Webster, J. and Ho, H. (1997). Audience engagement in multimedia presentations. *ACM SIGMIS Database*, 28(2):63–77.

White, R., Chu, W., Hassan, A., He, X., Song, Y., and Wang, H. (2013). Enhancing personalized search by mining and modeling task behavior. In *Proceedings of the 22nd international conference on World Wide Web*, pages 1411–1420, 2488511. ACM.

White, R. and Dumais, S. T. (2009). Characterizing and predicting search engine switching behavior. In *Proc. CIKM'09*, pages 87–96, 1645967. ACM.

White, R. and Horvitz, E. (2015). Belief dynamics and biases in web search. *ACM Transactions on Information Systems (TOIS)*, 33(4):18.

White, R. W. and Kelly, D. (2006). A study on the effects of personalization and task information on implicit feedback performance. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, pages 297–306. ACM.

Wiebe, E. N., Lamb, A., Hardy, M., and Sharek, D. (2014). Measuring engagement in video game-based environments: Investigation of the user engagement scale. *Computers in Human Behavior*, 32:123–132.

Williams, E. (1959). *Regression Analysis*. Wiley publication in applied statistics. Wiley.

Williams, K., Kiseleva, J., Crook, A. C., Zitouni, I., Awadallah, A. H., and Khabsa, M. (2016). Is this your final answer? evaluating the effect of answers on good abandonment in mobile search. In *Proceedings of the International Conference on Research and Development in Information Retrieval*.

ACM.

Williams, K. and Zitouni, I. (2017). Does that mean you're happy?: Rnn-based modeling of user interaction sequences to detect good abandonment. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17, pages 727–736, New York, NY, USA. ACM.

Wilson, T. D. (1981). On user studies and information needs. *Journal of documentation*, 37(1):3–15.

Wilson, T. D. (1999). Models in information behaviour research. *Journal of documentation*, 55(3):249–270.

Wing, J. M. (1990). A specifier's introduction to formal methods. *Computer*, 23(9):8–22.

Wohlin, C. (2014). Guidelines for snowballing in systematic literature studies and a replication in software engineering. In *Proceedings of the 18th international conference on evaluation and assessment in software engineering*, page 38. ACM.

Wu, Q., Wang, H., Hong, L., and Shi, Y. (2017). Returning is believing: Optimizing long-term user engagement in recommender systems. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 1927–1936. ACM.

Wurman, R. S. (1989). *Information anxiety*. Doubleday.

Xie, Z., Sun, Z., Jin, L., Ni, H., and Lyons, T. (2018). Learning spatial-semantic context with fully convolutional recurrent network for online handwritten chinese text recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(8):1903–1917.

Yamamoto, S., Wakabayashi, K., Kando, N., and Satoh, T. (2016). Who are growth users?: analyzing and predicting intended twitter user growth. In *Proceedings of the 18th International Conference on Information Integration and Web-based Applications and Services*, pages 64–71. ACM.

Yu, R., Gadiraju, U., Holtz, P., Rokicki, M., Kemkes, P., and Dietze, S. (2018). Predicting user knowledge gain in informational search sessions. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 75–84. ACM.

Yuan, X. and White, R. (2012). Building the trail best traveled: Effects of domain knowledge on web search trailblazing. In *Proc. CHI '12*, pages 1795–1804, New York, NY, USA. ACM.

Zach, L. (2005). When is "enough" enough? modeling the informationseeking and stopping behavior of senior arts administrators. *Journal of the American Society for Information Science and Technology*, 56(1):23–35.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of*

the Royal Statistical Society: Series B (Statistical Methodology), 67(2):301–320.

Appendix A

Ethic approval letter

Downloaded: 19/12/2018
Approved: 19/12/2018

Mengdie Zhuang
Registration number: 150112547
Management School
Programme: Management (PhD/Management FT)

Dear Mengdie

PROJECT TITLE: Modeling user behaviour based on process
APPLICATION: Reference Number 021202

This letter confirms that you have signed a University Research Ethics Committee-approved self-declaration to confirm that your research will involve only existing research, clinical or other data that has been robustly anonymised. You have judged it to be unlikely that this project would cause offence to those who originally provided the data, should they become aware of it.

As such, on behalf of the University Research Ethics Committee, I can confirm that your project can go ahead on the basis of this self-declaration.

If during the course of the project you need to deviate significantly from the above-approved documentation please inform me since full ethical review may be required.

Yours sincerely

Sophie May
Departmental Ethics Administrator

Appendix B

Publications resulted from the PhD

Zhuang, M., Toms, E. G., and Demartini, G. (2018). Can User Behaviour Sequences Reflect Perceived Novelty? In CIKM 2018: 27th ACM Conference on Information and Knowledge Management, 1507-1510. ACM, Turin, Italy. doi: 10.1145/3269206.3269243 *(partially overlaps with chapter 5)*

Zhuang, M., Demartini, G., and Toms, E. G. (2017). Understanding Engagement through Search Behaviour. In CIKM 2017: 26th ACM Conference on Information and Knowledge Management, 1957-1966. ACM, Singapore. doi: 10.1145/3132847.3132978 *(partially overlaps with chapter 4)*

Zhuang, M., Toms, E. G., and Demartini, G. (2016). The Relationship Between User Perception and User Behaviour. In ECIR 2016: 38th European Conference on Information Retrieval, 293-305. Springer, Padua, Italy.