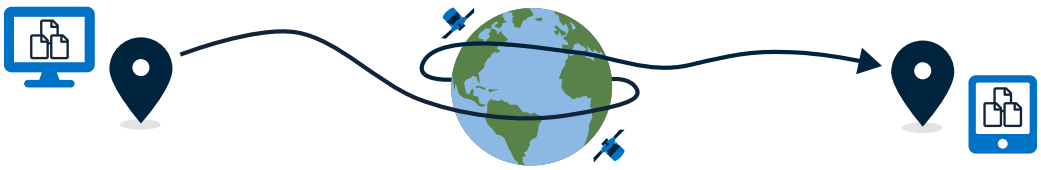
Chapter 1

Introduction

Today, we live in the *Information Age*, an era of human history characterised by the rapid development of technology. This allows for the creation, transmission and retrieval of large volumes of information. Two key developments that have permitted an increase in information generation are the electronic computer and the associated technologies that allow for near-instantaneous communication with devices all around the planet, including the *Internet* and *World Wide Web (WWW)* (Berners-Lee et al., 1994).



Since the early 1990's, the WWW has emerged as the dominant means of publishing information over the Internet, replacing obsolete technologies such as the *Gopher* protocol.[1] As the amount of information available on the WWW grew, so too did the paradigms that were employed by those wishing to seek information on it.

[1] Gopher was designed primarily with a menu-driven interface in mind (i.e. selecting options from a series of choices). The Gopher ecosystem provided the foundations for the *HyperText Transfer Protocol (HTTP)* protocol, which the WWW today utilises.

Information seekers would traditionally *surf* the WWW, starting from a particular domain. From there, they would navigate through the WWW via a series of *hyperlinks* within web pages (or *documents*). This proved practical, as portal websites typically presented categorised lists of websites, much like a telephone directory. However, as the volume of content available on the WWW grew ever larger, this approach became impractical. The development of *search engines* – referred to as *retrieval systems* in this thesis – provided information seekers with the ability to search the ever-increasing universe of documents available at their disposal (refer to Figure 1.1).[2]

This is not to say that surfing no longer occurs. Information seekers today will often use a retrieval system to find a particular domain. From there, they may then begin surfing within the said domain to find the information that they seek – if such information was not found immediately by the retrieval system. Retrieval systems are however today the most effective way to locate information. Helping searchers realise this by developing efficient retrieval systems is seen as the *raison d'être* of the study of IR.

> *"...but perhaps the key technology that took the web from a useful supplement of current information practice to become the default communication medium is search."*
>
> Wilson et al. (2010)

Contemporary commercial retrieval systems such as *Google* and *Bing* are considered to offer an effective means of finding the proverbial needle in the haystack (Wilson et al., 2010), where near perfect accuracy is regularly attained for popular *queries* (Vaughan, 2004). These retrieval systems, along with the many others in existence today (for use in a variety of contexts[3]), are the product of the collective work undertaken in the field of IR, as we discuss in more detail in Chapter 2.

[2] McBryan (1994) considered a retrieval system as a means of *taming* the considerable number of documents online.

[3] Google and Bing may be the most popular retrieval systems for *general web queries*, but other contexts, for example, can include academic search, enterprise search, multimedia search and patent search.

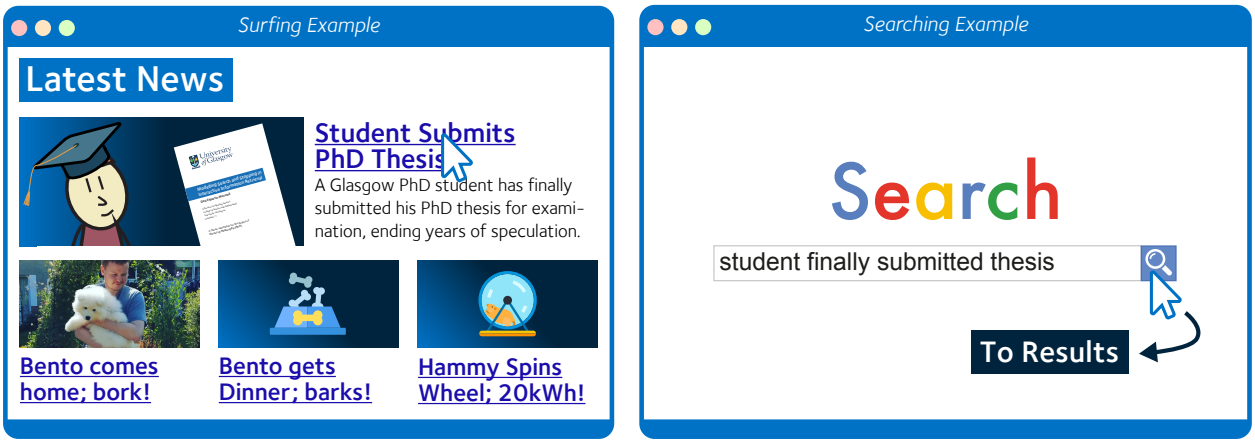


Figure 1.1 The paradigms of surfing and searching. On the left, a seeker will navigate through a series of documents via *hyperlinks* (perhaps without a specific *information need* in mind), while a searcher (right) will issue a query articulating their information need, relying on a *retrieval system* to retrieve a series of documents that are judged to be useful to the seeker.

Retrieval systems aim to make it easier for searchers to satisfy their underlying *information need*. A searcher will develop an information need from a perceived problem – either from a knowledge gap, an internal inconsistency, or a conflict of evidence. This state has been referred to as the *Anomalous State of Knowledge (ASK)* (Belkin, 1980). A searcher, once they have realised this information need, will formulate a **query** – an expression of what they are looking to seek (Borlund, 2003), typically consisting of a number of different terms. This query is then submitted to the retrieval system, before a potentially relevant set of documents – as judged by the retrieval system – are returned to the searcher. From this set of documents, the searcher can then begin the process of examining them for relevance.

A number of complex interactions take place between an individual seeking information and the retrieval system being utilised (Ingwersen and Järvelin, 2005). This interactive process, where the searcher engages in dialogue with the retrieval system, is considered the study of *Interactive Information Retrieval (IIR)* (Borlund, 2003). One of the fundamental aspects of IIR is that of **stopping** – where, for example, a searcher must decide when to stop examining the list of results returned to him or her.

Examining stopping behaviour is one of the many different aspects of interaction that have

been examined to help us better understand a searcher's behaviours. This knowledge can be used to make the search process a more seamless experience for the individuals using a retrieval system. As we discuss in the next section, much of the research in both IR and IIR has been limited in terms of examining stopping. Subsequently, these limitations provide motivation for the work that we present in this thesis.

1.1 Motivation and Context

Central to much of the work undertaken in the field of IR is the Cranfield paradigm, a term denoting a standardised approach of IR evaluation (Cleverdon et al., 1966). Primarily credited to Cyril Cleverdon at Cranfield University[4], the paradigm revolves around the notion of standardised test collections – standardised corpora of documents that can be used by different researchers, providing a uniform foundation for IR experimentation.

While the basic principles of the Cranfield paradigm have remained in place since it was established in the 1960's, aspects of the approach have evolved over the years to cater for the ever increasing complexity of the tasks trialled (Harman, 2010). The approach is widely used in evaluation forums, such as *NTCIR (NII Testbeds and Community for Information access Research)* and *CLEF (Conference and Labs of the Evaluation Forum)*. However, one of the best-known evaluation forums following the paradigm is the *National Institute of Standards and Technology (NIST)* sponsored Text REtrieval Conference (TREC) (Harman, 1993). Indeed, the work reported in this thesis extensively utilised material generated as part of TREC efforts, provided as part of different *TREC Tracks* over a number of years.

With the Cranfield paradigm, significant advances have been made possible regarding the evaluation of IR systems. However, the approach can be argued to be somewhat limited from the context of IIR as it highly abstracts the interactions that take place between a

[4]Cranfield University is located at Cranfield, Bedfordshire, England. It is a unique university in that it has a semi-operational airport, given its heritage with aeronautics research.

searcher and a retrieval system (Borlund, 2000; Ingwersen and Järvelin, 2005). In other words, the paradigm broadly fails to consider the complexities of the IIR process. As an example of such a complexity, searchers could issue multiple queries during the course of a search session. Subsequently, they would adapt their interactions based upon the perceived quality of presented ranked result lists (Moffat et al., 2013).

A key example of such behaviour adaption is the searcher's stopping behaviour. For example, a poor set of results may mean that searchers would stop examining results comparatively early than a set of results perceived to be of good quality. Searchers also often stop once they feel that they have found sufficient information to satisfy their information need (Zach, 2005). Indeed, selecting good terms to use within a query is difficult yet important (Efthimiadis, 2000). The initial query posed in a search session often acts as an entry to the search system, followed by phases of browsing and query reformulations (Marchionini et al., 1993). Searchers also will typically abide by the *principle of least effort*, whereby they strive to minimise the expected rate of work expenditure over time (Zipf, 1949).

The experimentation paradigms that have evolved from Cranfield make a series of different assumptions that are largely at odds with how searchers interact with retrieval systems. Namely, these assumptions state that a searcher will:

- issue a *single query* over the course of a search session;
- *examine documents to a fixed depth* (typically 1,000 in TREC experimentation); and
- *assess all documents to the fixed depth.*

While providing a simple platform for performing retrieval system evaluation, such assumptions are unrealistic. Herein lies a fundamental disconnect between the studies of IR and IIR – the naïve assumptions made of searchers within IR experimentation listed above do not hold when considering the complex interactions that actually take place during the IIR process (Ingwersen and Järvelin, 2005). In order to address the fundamental disconnect between the two fields, we need to create more realistic searcher models that better

articulate what real-world searchers actually do. A better searcher model would ultimately mean a better understanding of the complex interactions that take place, which would lead to an improved understanding of how to assist searchers. Work to improve our understanding has been undertaken in the field of IIR to address this, examining searcher behaviours under a number of different phenomena – including (but not limited to) the following:

- *query formulation and suggestions* (Azzopardi, 2009; Azzopardi et al., 2007; Baskaya et al., 2013; Carterette et al., 2015; Jordan et al., 2006; Keskustalo et al., 2009; Verberne et al., 2015);
- *browsing behaviours* (Carterette et al., 2015; Chuklin et al., 2015; Guo et al., 2009; Pääkkönen et al., 2015; Smucker, 2011);
- the influence of *costs and time* (Azzopardi, 2011; Baskaya et al., 2013); and
- *performance over search sessions* (Luo et al., 2014, 2015).

When considering how we model searcher interactions, a further (and particularly important) phenomenon largely overlooked in the above is a **searcher's stopping behaviour**. Indeed, given its title, this is what we consider in this thesis – *how can we make improvements to searcher models when considering stopping behaviours?* This phenomenon is now seeing an increasing amount of time devoted to its examination. In the following subsection, we provide an argument as to why examining this phenomenon is important.

1.1.1 Considering Stopping Behaviours

Knowing when to stop is a fundamental aspect of animal – and by definition, human – thinking and behaviour. There must come a time when an animal must stop what it is doing. In the natural world, for example, a honeybee, when *foraging* for pollen, will eventually make a decision to stop collecting pollen on the flowerhead it finds itself on and flies away

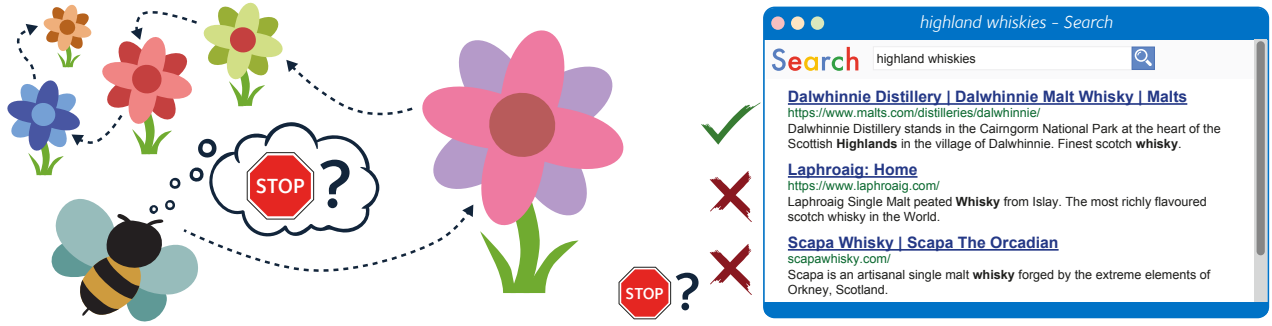


Figure 1.2 Examples of stopping. On the left, when will the bee move from one flowerhead to the next? On the right, under the context of information seeking, how far down a list of ranked results will a searcher go before he or she decides to stop examining content? In the example above, Search has failed to return a comprehensive list of highland single malt whiskies. Will the searcher become frustrated with this, and stop examining results early?

to another flowerhead. The honeybee is in essence attempting to maximise the amount of *gain* (pollen) she accumulates over time on each *patch* (flowerhead) that is visited.

If we consider stopping from an information seeking context, there are many different examples we can use to demonstrate why this behaviour is of great importance. For example, a searcher may decide to stop searching for information when the documents presented show a large volume of non-relevant material, frustrating the searcher (Cooper, 1973b) – perhaps because the retrieval system failed to gauge the searcher's *query intent* (Ashkan et al., 2009), as demonstrated in Figure 1.2. Searchers could also stop examining content after they have become satisfied with the information found previously in a search session (Cooper, 1973a; Gibb, 1958; Simon, 1955), or if they feel that the information being presented is too similar to what has been found earlier (Nickles, 1995).

A number of different *external factors* can influence the decision of when one should stop. Examples of these include the bee finding a flowerhead with no pollen, or time pressures when searching for information. However, Nickles (1995) argues that knowing when to stop is largely determined by a series of *internally defined stopping criteria* that the decision maker employs, just like the examples defined above. Therefore, this internal construct makes stopping a phenomenon that is difficult to model in an effective way. Given that

internal factors are a major drive in determining when to stop, studies have largely been unable to quantify *why* searchers stop, other than what they find during the search process gives them the feeling that the located information is *"good enough"* (Zach, 2005).

In contrast to this vague definition of stopping behaviour, several researchers have attempted to create a series of reasoning- and judgement-based stopping heuristics that attempt to formally define when a searcher should stop. It is these stopping heuristics that we will primarily consider in this thesis. These heuristics can then be integrated within a wider searcher model, allowing us to determine whether they improve or worsen approximations of actual searcher stopping behaviours. From here, we can then begin to ascertain potential answers to what the feeling of *"good enough"* (or even *not* good enough!) may entail. The searcher model can incorporate stopping behaviours at a variety of different stopping decision points – such as at an individual result summary level *(how far down this list of ranked results should I go?)*.

Examining stopping behaviours during search is important because it considers the judgements of a searcher as part of their interactions. For example, it would be prudent of a searcher examining a ranked list of results that are mostly non-relevant to stop early, thus saving time and effort (thus making the searcher more *efficient*). Stopping behaviour is also implicitly or explicitly encoded within a variety of different IR and IIR measures. Obtaining a better understanding of when searchers stop means that we can encode this information within measures of search (improving their credibility), and provides an evidence-based approach to mapping these measures with what actually takes place in reality.

1.2 High-Level Research Questions

Having set out the problem space above, we can now begin to formulate the four high-level research questions that the work in this thesis addresses, denoted as HL-RQx. Our first research question considers the concept of modelling searchers, and how, with an emphasis

on examining stopping decision points, we can improve current models to better reflect actual searcher behaviours – in particular, their stopping behaviour.

- **HL-RQ1** How can we improve searcher models to incorporate different stopping decision points?

As previously stated, being able to improve upon the current searcher models from the perspective of stopping should allow those subscribing to such a model to become more efficient as to how they search. Closely related to this advancement in modelling this process is the consideration of the various stopping heuristics.

- **HL-RQ2** Given the stopping heuristics defined in the literature, how can we encode these heuristics into a series of *operationalised*, programmable **stopping strategies** that can be subsequently incorporated into the searcher model and evaluated?

Stopping heuristics that we detail later in Section 3.2 are high-level in nature and do not provide an explanation as to how they can be operationalised within a wider system. The challenge that must be addressed in order to answer this second high-level research question will be how we can operationalise such stopping heuristics.

With a more realistic searcher model from **HL-RQ1** and a series of stopping strategies defined by addressing **HL-RQ2**, how well does this combination perform?

- **HL-RQ3a** Given the aforementioned operationalised stopping strategies, how well does each one perform?
- **HL-RQ3b** How closely do the operationalised stopping strategies compare to the actual stopping behaviours of real-world searchers?

These questions are of course of a very broad nature, and it is simply not possible to evaluate them in every conceivable search context. As such, we will examine different contexts that are likely to impact upon searcher stopping behaviours. Specifically, we will examine topical *interactive search* in the domain of news, where we will consider various conditions: search goals and task types; retrieval systems; and result summary length. In the following section, we expand upon these conditions to provide a concrete set of thesis contributions.

1.3 Thesis Contributions

This thesis presents a number of key contributions. Listed below, we consider primary contributions from conceptual, theoretical, methodological and empirical standpoints.

Conceptual **Complex Searcher Model** Our first contribution is a new searcher model. Taking current searcher models, we propose an updated, high-level model of the search process called the *Complex Searcher Model (CSM)*. This provides us with a solution for addressing **HL-RQ1**. Outlined in Chapter 4 (page 107), the conceptual CSM outlines a series of different activities and decision points that searchers undertake throughout the search process, and establishes a flow of interaction based upon established models. Within the CSM are a number of different innovations, key of which is the new stopping decision point. For example, this improvement allows us to ascertain a better understanding of the search process, and the complex interactions that occur between a searcher and retrieval system. Being a conceptual model, we can take the CSM and instantiate it in a number of different ways. The stopping strategies that we consider in this thesis, for example, provide a means for instantiating stopping decision points within the CSM.

Theoretical **Stopping Strategies** As previously discussed, there is a range of different stopping heuristics that have been defined in the literature that provide an explanation for when searchers should stop examining content. The second major contribution of this thesis is the development of twelve operationalised stopping strategies. These may then be

subsequently deployed as the logic underpinning a stopping decision point of the CSM (as defined above). These twelve strategies encode a total of seven different stopping heuristics and IR measures. The operationalised stopping strategies provide a solution to HL-RQ2.

Methodological The proposed CSM and the twelve stopping strategies that we operationalise need to be evaluated, such that we can then subsequently address the two remaining high-level research questions, HL-RQ3a and HL-RQ3b. To do this, a general methodology outlines an approach undertaken for user studies. **Simulation** is then used to determine how the different stopping strategies perform over each of the different search contexts trialled, and how the stopping strategies compare to actual searcher behaviour.

Empirical **Varying Result Summary Length** We report on a study where the length of individual result summaries presented to searchers are varied to determine what impact that this has on searcher stopping behaviours. As we modify the length of result summaries, we also argue that we influence the overall quality of result summaries. We then perform a simulated analysis examining each of the stopping strategies, determining what strategies perform best and offer the closest approximations to real-world stopping behaviours.

Empirical **Varying Goals, Tasks and Systems** We report on an additional user study, examining the impact of stopping behaviours when the search task and goals are changed. For this, we consider topical *ad-hoc retrieval*[5], along with a diversified search task, changing the overall goal of what searchers are looking to find. This is then complemented by a further simulated analysis, examining the individual stopping strategies like above.

Empirical **New Stopping Decision Point** The final empirical contribution complements the conceptual contribution of this thesis, addressing HL-RQ1. We perform a further simulated analysis, examining how well the new stopping decision point performs when incorporated within the CSM – and whether it offers better approximations to actual searcher stopping behaviours.

[5]The ad-hoc search task is explained in detail in Section 2.3.1.1 – it is one of many different types of search task that can be performed by searchers.

1.4 Thesis Statement

Given the above, the major claim of this thesis is that by considering stopping behaviours at different points throughout the search process, we can develop more credible and realistic models of the said search process. These more advanced models can be used as a tool for improving our understanding of stopping behaviours and other complex interactions that occur when searching. Findings from this work can then subsequently aid researchers in the development of more intuitive (and realistic) measures used to facilitate the evaluation of retrieval systems and their users.

1.5 Origins of the Material

Material presented in this thesis has appeared in several conference papers and journals throughout the duration of the author's PhD programme, from October 2013 to March 2019. All are listed in the front matter of this thesis in chronological order. In this section, we provide a narrative, explaining how the developments in the listed publications led to the contributions of this thesis. Work can be considered over three main strands:

- the development of the conceptual and theoretical contributions to this work;
- the development of the **SimIIR** framework; and
- a series of empirical studies.

Conceptual and Theoretical Work on the *Complex Searcher Model (CSM)* has been undertaken over a number of years, and were presented in various publications. Several iterations of the CSM have been developed, with each iteration offering refinements to improve its realism.[6] The first iteration of the CSM – essentially analogous to prior models of search

[6] To simplify reporting (and use) of the CSM in this thesis, we consider only the latest revision of the model.

outlined in Sections 2.3.1.2 and 2.3.5 – was used in simulated analyses, as reported in the two publications listed below.

- Maxwell, D., Azzopardi, L., Järvelin, K., and Keskustalo, H. (2015a). An initial investigation into fixed and adaptive stopping strategies. In *Proceedings of the 38th ACM SIGIR*, pages 903–906
- Maxwell, D., Azzopardi, L., Järvelin, K., and Keskustalo, H. (2015b). Searching and stopping: An analysis of stopping rules and strategies. In *Proceedings of the 24th ACM CIKM*, pages 313–322

These publications are notable for also including a number of operationalised stopping strategies, providing the foundations for the second major contribution of this thesis. The stopping strategies defined in these publications were used in subsequent publications. Further developments to the CSM were found in a subsequent publication which experimented with the notion of developing *intelligent search agents*.

- Maxwell, D. and Azzopardi, L. (2016a). Agents, simulated users and humans: An analysis of performance and behaviour. In *Proceedings of the 25th ACM CIKM*, pages 731–740

The final development of the CSM led to the inclusion of an additional stopping decision point. This new stopping decision point was tested with a thorough empirical analysis, as reported in the publication enumerated below.

- Maxwell, D. and Azzopardi, L. (2018). Information scent, searching and stopping: Modelling SERP level stopping behaviour. In *Proceedings of the 40th ECIR*, pages 210–222

SimIIR Framework One of the major pieces of scientific apparatus utilised throughout all of the aforementioned studies is the **SimIIR** framework, which we discuss in Section 6.4.1 on page 159. Conducting the extensive simulations of interaction we report in this thesis would not have been possible without it. A demonstration paper presenting the framework and the various components that could be instantiated within it has been published.

- Maxwell, D. and Azzopardi, L. (2016b). Simulating interactive information retrieval: Simiir: A framework for the simulation of interaction. In *Proceedings of the 39th ACM SIGIR*, pages 1141–1144

Empirical Studies The general methodology that we employ for the third major contribution of this thesis has been introduced and refined in the publications listed previously. In addition to this, a basic description of the methodology is provided in a Doctoral Consortium paper that the author presented at the first *ACM Conference on Human Information Interaction and Retrieval (CHIIR)* in Chapel Hill, NC, USA.

- Maxwell, D. (2016). Building realistic simulations for interactive information retrieval. In *Proceedings of the 1st ACM CHIIR*, pages 357–359

The results of two user studies have also been published, and are of direct relevance to the work detailed later in this thesis.

- Maxwell, D., Azzopardi, L., and Moshfeghi, Y. (2017). A study of snippet length and informativeness: Behaviour, performance and user experience. In *Proceedings of the 40th ACM SIGIR*, pages 135–144
- Maxwell, D., Azzopardi, L., and Moshfeghi, Y. (2019). The impact of result diversification on search behaviour and performance. *Information Retrieval Journal*. In press.

These studies provide the grounding for simulated analyses that we also consider later in this thesis. The data extracted from these user studies provides credibility to our simulations through the extraction of aspects such as interaction costs and probabilities.

1.6 Thesis Outline

This section provides a brief summary of the remaining parts and chapters of the thesis.

Part I The remainder of Part I concerns prior work that has been undertaken in the fields of IR and IIR. Two chapters outline the basics of IR and IIR (with particular emphasis to how models and measures that we commonly employ consider stopping), before examining the literature that has explicitly considered searcher stopping behaviours.

> **Chapter 2** Beginning on page 21, this chapter provides an overview of the key concepts of the fields of IR and IIR. We focus on core IR concepts, such as the indexing and retrieval processes (including retrieval models). We then move towards a more user-centric examination of established methods in the field of IIR, such as the consideration of various evaluation measures that are commonly used. We also outline different searcher models that have been previously defined in the literature. These capture the activities and decisions that individuals perform while searching.
>
> **Chapter 3** We then consider work that has considered stopping in relation to search. In this chapter, we begin by describing various stopping heuristics defined in the literature. We summarise previous user studies that have examined searcher stopping behaviours, and then consider key theoretical models of search that provide explanations for when individuals stop.

Part II Beginning on page 106, Part II presents the conceptual and theoretical contributions of this thesis, including a discussion of the CSM. In this part of the thesis, we also provide an outline of the general methodology that is used in Part III.

Chapter 4 This chapter introduces the CSM, discussing the advances that the conceptual model provides over contemporary searcher models. We discuss the key stopping decision points provided by the CSM that are central to this thesis, before discussing the assumptions of the model. This partly addresses **HL-RQ1** – evaluation of the model is also required, and is discussed in Chapter 9.

Chapter 5 In this chapter, we introduce and discuss the various stopping strategies that we operationalise as part of the contributions of this thesis, thus addressing **HL-RQ2**. Each of the different stopping strategies, complete with examples, are discussed in depth. The chosen stopping strategies are linked back to their originating stopping heuristics, which are detailed in Chapter 3.

Chapter 6 This chapter outlines our general methodology, detailing the high-level structure of the scientific method used in our empirical work. We also provide a discussion of common approaches that we used across all subsequent chapters.

Part III The third part of this thesis considers our empirical contributions. In this part, we present the user studies that were undertaken, as well as a number of simulated analyses that allow us to address research questions **HL-RQ3a** and **HL-RQ3b**.

Chapter 7 The first empirical chapter considers how stopping behaviours vary when the length (and thus quality) of result summary snippets are varied. We provide a discussion of a user study that examined this phenomenon, before summarising the findings of simulated analyses that were conducted in order to determine what stopping strategies offered the best performance and approximations of real-world searchers under this context.

Chapter 8 In this chapter, we report on a user study examining how a searcher's stopping behaviour varies when subjected to conditions that vary the task, goal, and system used. We then again perform simulated analyses to examine these stopping behaviours in more detail.

Chapter 9 The final chapter wherein novel findings are presented considers the new stopping decision point that is provided by the CSM. We empirically test the CSM, allowing us to determine whether the inclusion of the new stopping decision point discussed in Chapter 4 provides improvements in overall performance and approximations of actual searcher stopping behaviours. As such, this chapter provides sufficient evidence, in conjunction with Chapter 4, to address **HL-RQ1**. We utilise data from user studies discussed in Chapters 7 and 8 to ground our simulations.

Part IV The final part of this thesis consists of a solitary chapter, **Chapter 10**. The concluding chapter of this thesis provides a summary of the work that was undertaken, and the results obtained. We then discuss potential avenues for future work.