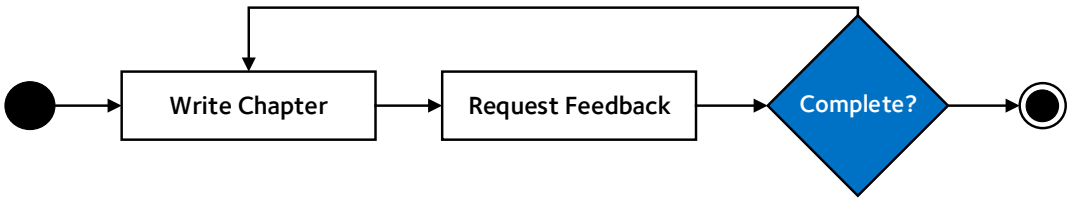
Chapter 4

The Complex Searcher Model

In this chapter, we present the *Complex Searcher Model (CSM)*. The CSM is an updated, conceptual searcher model[1] that is one of the major contributions of this thesis. It is an amalgamation and development of prior, established searcher models. These models capture the complex sequence of interactions that take place between a searcher and a retrieval system over the course of a search session. As such, this chapter provides a partial answer to our first high-level research question, **HL-RQ1**.



As discussed in Section 2.3.5, earlier examples of searcher models include the Markov-based approach presented by Baskaya et al. (2013), and the model proposed by Thomas et al. (2014). These searcher models (along with others) are in broad agreement with the general sequence of events that take place within the IIR process – from issuing a query to examining documents for relevance.

[1] The CSM can also be considered as a *browsing model*, as per Carterette et al. (2011).

Given the aforementioned searcher models outlined in Section 2.3.5, the CSM offers a number of advancements in modelling searcher and retrieval system interactions. In this chapter, we provide:

- the **flow** of the proposed CSM, outlining the different steps and decisions that those subscribing to it undertake (Section 4.1);
- a discussion of the **stopping decision points** that the CSM considers (Section 4.2);
- a summary of the key **advancements** that the CSM provides (Section 4.3); and
- an outline of the key **assumptions** that we consider as part of the CSM (Section 4.4).

We also briefly outline the specifics for evaluating the CSM as a viable searcher model (Section 4.5). Specific details of the implementation of the CSM are discussed in our general methodology (Section 6.4, page 157). We begin this chapter with a discussion of the flow of the CSM, discussing the different steps and decisions that searchers will make.

4.1 Model Flow

The CSM is illustrated as a flowchart in Figure 4.1. It is comprised of a number of different activities denoted by boxes, and decisions represented as blue diamonds. The flowchart is divided up into a number of different blocks, labelled **A** to **F**. Each of the blocks denotes a logical set of interactions – block **B**, for example, considers all of the actions and decisions a searcher is likely to consider in relation to querying. In this section, we outline the flow of the CSM, discussing the key activities and decisions that searchers would undertake when subscribing to it. This is done in relation to the six labelled blocks that are discussed below.

A **Topic Examination** A searcher subscribing to the CSM would begin the search process with some information need. This would typically be provided as a *topic*, with

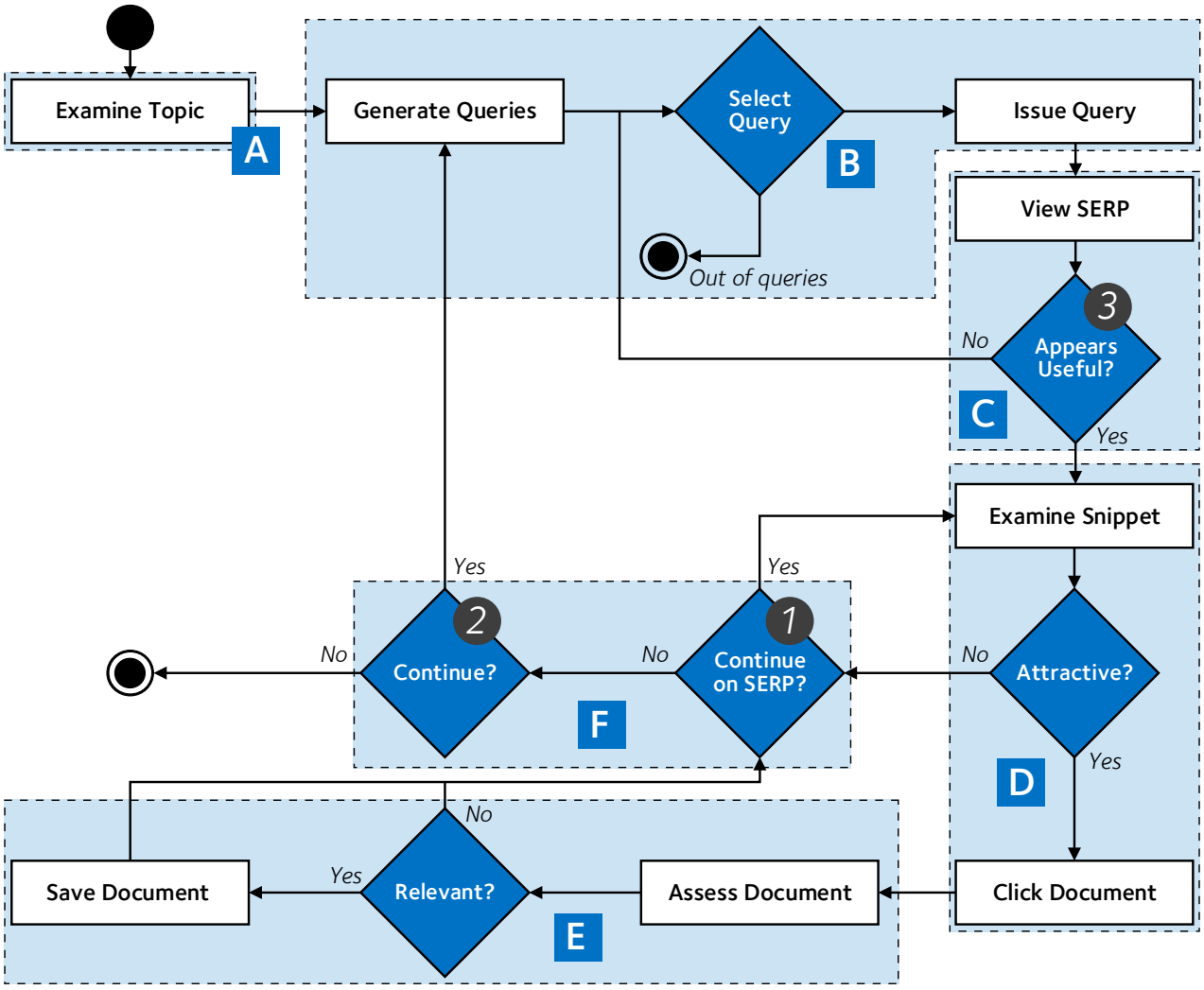


Figure 4.1 A flowchart of the *Complex Searcher Model (CSM)*. A cornerstone of this thesis, the CSM is extensively used as the grounding model for simulations of interaction that we report on in subsequent chapters. The main logical components of the CSM as discussed in Section 4.1 are labelled **A** to **F**, complete with surrounding boxes. The *three* stopping decision points are highlighted with numbers ❶, ❷ and ❸ (refer to Section 4.2).

a topic description outlining said information need. From this topic description, various entities can be extracted and used for the generation of queries, as described in block **B**.

B **Querying** Once the information need has been established, the searcher will then move onto the *querying* block. Here, a number of different activities and a decision point are considered. Within the block, the first activity that the searcher will un-

dertake is the **generation of queries**. Given the information need from block **A**, a searcher will formulate a number of *candidate queries* that they could issue to the underlying retrieval system. This is achieved through the use of some form of *querying strategy* that generates the said candidate queries. The searcher then must make a decision as to what query they should issue. A query is **selected** from the candidate queries list by some process (e.g. some form of ranking). This query is the one the searcher believes is most likely to return relevant documents. The query is then **issued** to the underlying retrieval system[2], with the searcher proceeding to **C**.

As stated previously, IIR is an iterative process where multiple queries can be issued in a single search session. The CSM provides support for this – as can be seen from the flowchart line from block **F** to querying block **B**. At each point, the list of candidate queries generated could theoretically be regenerated, thus supporting query reformulation. If a searcher then finds that the candidate queries list has been exhausted, a stopping point is provided for this scenario.

C **SERP Examination** With the query now issued, the retrieval system will then return a SERP for the searcher to examine. From here, the searcher is able to **view the SERP** – that is, to obtain an *initial impression* of the SERP by examining the various *proximal cues* (Chi et al., 2001) presented. If the SERP does not appear to look promising, or gives the answer straight away, the searcher will abandon the SERP and proceed to issue a further query from the list of candidate queries as described in block **B**. If the SERP however does look **useful**, the searcher will then *enter* the SERP and proceed to examine individual result summaries in detail.

D **Result Summary Examination** Result summaries are presented to the searcher within the SERP. Searchers can take individual result summaries in turn, examining the title and snippet text provided for **attractiveness**. If deemed to be sufficiently attractive to warrant further examination, the searcher will then click on the provided link. This

[2] As the CSM considers interactions with a retrieval system only, we assume that a searcher will have already selected a retrieval system to use beforehand as discussed in Section 4.4.2.

link will then take the searcher to the associated document for further examination. If the searcher does not deem the summary to be sufficiently attractive to warrant further examination, he or she will then move to block F. As described below, the searcher in this block must decide whether to continue examining results on the SERP – and if not, whether to continue with the search session.

E **Document Examination** Once a searcher clicks on an attractive result summary, he or she will then **assess** the associated document for relevance, after which a further decision must be made. *Is this document relevant to the information need?* If so, the document is **saved**, meaning that it is added to a list of saved documents, as we describe below. The searcher then proceeds to block F. If not considered relevant, the searcher then proceeds directly F.

F **Deciding to Stop** Regardless of how the searcher reaches this block (either from block D or E), a searcher here can make two key stopping decisions. The first considers whether he or she should remain on the present SERP. If this is decided to be the case, the searcher will then move to the next result summary presented within it, and begin to examine that for attractiveness. If it is decided not to remain on the SERP, the searcher will then move to a final decision – *should I stop this search session, or continue?* If the searcher decides to continue the search session, he or she will then move back to the query generation activity in block B, beginning the cycle again.

Of particular interest to the work in this thesis are the *stopping decision points*, as discussed above – and shown in blocks C and F in Figure 4.1.

4.2 Stopping Decision Points

Outlined previously in Section 3.1.1, established searcher models consider stopping from two key perspectives: *result summary level stopping*, and *session level stopping*. The two estab-

lished stopping decision points are included within the CSM, and are labelled ❶ and ❷ in Figure 4.1 respectively. They are also briefly outlined below.

❶ **Result Summary Level Stopping** This stopping decision point concerns the depth at which a searcher will stop examining a list of ranked results for a given query, assuming that results are ranked in a particular order. After stopping at this point, the searcher can continue the search session by issuing a further query.

❷ **Session Level Stopping** This second stopping decision point considers the point at which a searcher will stop their search session in its entirety. As an example, a searcher will stop their search session when they believe that they have satisfied their search goal, for example.

The CSM however includes a third, *SERP level stopping decision point,* highlighted as stopping decision point ❸ within block C of Figure 4.1.

❸ **SERP Level Stopping** With this new stopping decision point, a searcher can abandon a SERP before *entering* it to examine result summaries in detail.

This new stopping decision point permits searchers subscribing to the CSM to become savvier with their interactions. By gauging the SERP, a searcher can make an informed decision as to the quality of said SERP before making the decision to invest more time examining its contents, or simply cutting their losses and abandoning it – for better or for worse. The new stopping decision point is one of the key advancements that the CSM provides, and is discussed in more detail in Section 4.3.1.

4.3 Model Advancements

The CSM provides two novel advancements in modelling interactions between a searcher and retrieval system. These are highlighted as blocks B and C in Figure 4.1, and ad-

vances our understanding of the *querying* process – as well as including the aforementioned third stopping decision point. In this section, we discuss each in turn. While the advances to querying are novel, they are not the core focus of this work, and thus discussion of the new SERP level stopping decision point will be in greater depth.

4.3.1 SERP Level Stopping

This new stopping decision point – illustrated in block C of the CSM (Figure 4.1) – is motivated by the idea of the information scent (refer to Section 3.3.1.1 on page 92) present on a given SERP. This section also introduced the concept of *proximal cues* (Chi et al., 2001), cues that provide insights into whether the presented SERP will yield information that will aid the searcher in satisfying their underlying information need. This has been demonstrated in prior studies (Wu et al., 2014; Ong et al., 2017) – and in Chapter 7 of this thesis.

By operationalising information scent as the perceived performance of a given SERP, we allow a searcher to obtain an *impression* of the SERP before deciding to *enter* it and examine presented content in detail – or *abandon* the SERP altogether, and move to the next activity. The notion of forming an impression is similar to the summary impressions formed by searchers subscribing to the model defined by Thomas et al. (2014), as detailed in Section 2.3.5. In their model, a searcher would not form an overview of the SERP, but rather an impression of each individual result summary. The impression can then be used as a means of gauging whether further examination would be worthwhile.

This new stopping decision point is analogous to the well-studied phenomenon of *SERP abandonment* in which limited interaction occurs with the searcher. This has been typically assumed to provide an indication of the searcher's *dissatisfaction* with the presented results (Das Sarma et al., 2008; Chuklin and Serdyukov, 2012; Kiseleva et al., 2015), or *satisfaction* (through the concept of *good abandonment*) (Loumakis et al., 2011; Wu et al., 2014).[3]

[3]We discuss this in more detail in Section 4.4.4.

Thus, we provide, to the best of our knowledge the first searcher model that incorporates a path for a searcher to leave a SERP that appears to be of poor quality (or a *low scent*), or even satisfies their information need outright.

4.3.2 The Querying Process

Outlined previously, search sessions are inherently interactive (Ingwersen and Järvelin, 2005). During a session, a searcher's underlying mental model of a given information need can adapt and is likely to change as he or she examines new content for relevance (Borlund, 2003). Searchers may find more descriptive terms associated with the said information need, and incorporate these terms in a subsequent query reformulation.

From block B in Figure 4.1, the querying process has been broken up into two distinct activities and decisions: query generation (thinking of queries that could be issued) and query selection (selecting a query to issue). A searcher subscribing to the model will have the capability of revising their generated query list at each query reformulation, thus supporting the concept of the dynamic information need. Updated terms can in theory be selected from newly examined documents and incorporated within the query generation process for future reformulations.

Query selection then determines which one of the generated queries are to be issued to the retrieval system. Of course, the potential exists whereby all generated queries have been exhausted. This scenario would thus provide a natural stopping point for the searcher, as included in Figure 4.1.

4.4 Model Assumptions

When modelling a real-world phenomenon, a number of different assumptions are made about said phenomenon's exhibited behaviours (Fishwick, 1995). The CSM is no excep-

tion from this rule, and we make a number of different assumptions about a searcher's behaviours and the presentation of the retrieval system's results. This section details five key assumptions that we consider as part of the CSM.

4.4.1 Search Task

In this thesis, we are interested in the wider IIR process, considering all of the activities and decisions involved. We are particularly interested in improving our understanding of complex retrieval tasks.

The CSM provides scope for the modelling of a variety of different *interactive search tasks*. Examples of these include the aforementioned *ad-hoc,* exploratory, and *diversity tasks.* These tasks can be undertaken in different search *domains*, such as informational (refer to Section 2.3.2) or patent searching. As discussed in Section 1.2, we exclusively consider informational search in the domain of news. Tasks we consider include both ad-hoc and diversity, such that we can then examine how behaviours vary under each task. This is because while the CSM is able to model other search tasks, the selected search tasks provide for more interesting task types to examine, and consider a greater depth of activities and decisions that would not otherwise be examined by the more simplistic approach.

These tasks are interesting to examine for two key reasons:

- the search goals between each task vary; and
- from an examination of the literature (refer to Section 3.4), it is not clear when *enough information is enough.*

These reasons will undoubtedly produce interesting results between the two tasks. As the tasks are not simple lookups, a searcher will not stop once a single relevant page has been found. Instead, he or she will stop once *enough* (Zach, 2005) information has been found to satisfy their goal, or other constraints are imposed (e.g. time constraints).

4.4.2 Retrieval System Tool Choice

The searcher model proposed by Thomas et al. (2014) provides those subscribing to it with a choice as to which retrieval system they should use. As discussed earlier, we assume with the CSM that a searcher uses a single retrieval system. Our focus is therefore with the interactions that take place between the searcher and the said retrieval system.

Of course, the inclusion of such a decision point would be interesting to examine within the wider IIR process. Different retrieval systems will have benefits and drawbacks for particular domain types (e.g. a patent retrieval system would perform better for patent searching tasks). It would be interesting to examine this kind of *tool switching behaviour* – and could be considered as a further stopping decision point, or *retrieval system stopping*.[4]

4.4.3 Simple SERPs

When considering the SERP presented to the searcher as a whole, we make three simplifying assumptions within the CSM. These are enumerated and detailed below.

- **Ten Blue Links** Under the CSM, a SERP will consist purely of a set of result summaries, coined in IR literature as the *ten blue links*. Of course, we acknowledge that additional components are present in contemporary SERPs, such as multimedia content in federated search (Chen et al., 2012). These are however not considered to simplify the CSM.
- **Linear Examination Order** Once a searcher has decided to examine a SERP in detail, the result summaries presented to the searcher will be examined in a linear order. There is evidence to suggest that real-world searchers examine results from top to bottom, as demonstrated by Joachims (2002) and Joachims et al. (2005), for example.

[4] Refer to Section 10.3.1 on page 349 for a more in-depth discussion on *tool switching*.

Click models, such as the *cascade model* (Craswell et al., 2008), have been developed that utilise this assumption. Such approaches are subject to *positional bias*, where the searcher implicitly trusts the results of the retrieval model and assumes that the first result presented is the most relevant to their information need.

- **No Explicit Pagination** The CSM also assumes that the SERP presented to a searcher is of a single page, with no pagination of results. This does simplify the modelling process, with pagination activities and costs also not considered in earlier searcher models that consider the session as a whole.

4.4.4 Good and Bad SERP Abandonment

As previously mentioned, the CSM provides a third SERP level stopping decision point. Associated literature considers the notion of bad SERP abandonment, where a searcher is dissatisfied with the presented results. More contemporary research has introduced the notion of good SERP abandonment (Khabsa et al., 2016), where a searcher satisfies his or her information need by examining the SERP, requiring no further interactions with it. This is more prevalent on small-screen devices, where an information card presented to the searcher on the SERP may provide all the information required to satisfy the searcher on a simple lookup task, for example.

The CSM does not explicitly consider the notion of good or bad SERP abandonment; the provision exists however for both to be modelled effectively within the scope of the new stopping decision point. Good abandonment can be for example catered for with the inclusion of an additional decision point after determining a result summary to be attractive; the searcher could then make the decision to abandon the SERP if they feel satisfied with the result obtained. This is illustrated as an excerpt of a searcher model flowchart in Figure 4.2. The excerpt demonstrates the result summary **Attractiveness** decision point, the additional decision point determining **Satisfaction** with the result, and the final decision point that determines whether the searcher should abandon the SERP.

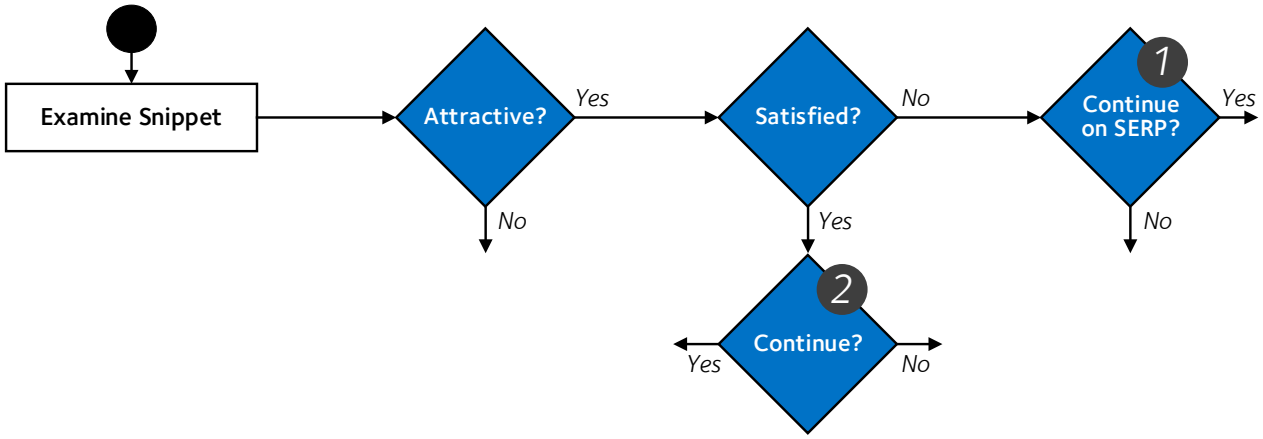


Figure 4.2 The interaction processes that can provide for incorporating *good* SERP *abandonment*, where a searcher satisfies his or her information need by simply examining a presented result summary. This is opposed to bad SERP abandonment, where the searcher will abandon a SERP if he or she feels the presented results are not of good quality. Upon examining a result summary **(Attractive?)**, a searcher will then determine if the summary addresses their information need **(Satisfied?)**. If so, they reach the session level stopping decision point ❷. Otherwise, they reach the result summary level stopping decision point ❶.

However, for the work in this thesis, we assume a simple SERP consisting only of a ranked list of results. We also assume that searchers subscribing to the CSM will have complex information needs, as discussed in Section 4.4.1 above. As such, we assume that the elements provided as part of the simplistic SERP are unlikely to fully satisfy their information need, and thus we consider SERP abandonment in this thesis exclusively from the perspective of **bad abandonment.** This is further discussed in Section 4.5.

4.4.5 External Factors

Given the flowchart of the CSM in Figure 4.1, it is clear that the model is completely agnostic of *external factors* that could influence how an individual behaves when searching. Kelly (2009), for example, cited that:

"searcher behavior [sic] can be governed by a number of external factors. For instance,

> *the occurrences of a holiday or a project deadline will likely change the kinds of behaviors users exhibit and these behaviors may not represent their typical behaviors."*
>
> Kelly (2009)

These examples allude to time pressures, but there are a virtually unlimited number of other external reasons that may influence a searcher's behaviour. Even simple everyday occurrences such as a phone call or an incoming e-mail can sufficiently distract the searcher to the point that their behaviours are altered. Our assumption is that an individual searches in a more controlled environment, where they are exclusively tasked to search.

4.5 Evaluating the CSM

The CSM is presented as a generalised, conceptual model of the search process. It captures the key activities and decisions that a searcher must undertake. Given the current searcher models presented in Section 2.3.5, the CSM introduces further levels of complexity and realism into searcher models. Given our choice of search tasks, types, and assumptions, four key assumptions are made for the evaluation of the model.

- **Costs** We assume that a searcher will incur some cost when performing an individual activity within the CSM. For example, a document examination cost will be incurred when a searcher decides to examine a document for relevance.
- **Bad Abandonment** As described previously, a searcher subscribing to the CSM will only abandon a SERP if they consider it to be of poor quality. Given the complex information needs we consider in this work, this is a reasonable assumption to make.
- **Saving Documents** Documents that a searcher subscribing to the CSM will be saved to a list. This provides us with a mechanism of identifying content the searcher deems relevant, which can be used in calculating performance measures (see below).

- **Accruing Gain** Following on from the above, we assume that searchers only gain from documents they examine and save. We do not assume that a searcher will be able to gain from the examination of result summaries, for instance – the information need is complex, and short result summaries would be unlikely to provide an answer to their information need.

4.6 Chapter Summary

This chapter has proposed the *Complex Searcher Model (CSM)*, our solution to partially addressing **HL-RQ1**. Building on prior searcher models, the CSM proposes different advancements to modelling a searcher's interactions, the main development being the inclusion of a new, SERP level stopping decision point. We have outlined a number of different assumptions that we make in the CSM, and also discussed some evaluation considerations related to the work in this thesis. Empirical work presented later in Chapter 9 tests the CSM, providing evidence to support **HL-RQ1** in that the CSM does provide improvements over current searcher models.

In the next chapter, we turn our attention to the twelve stopping strategies that we operationalise and subsequently implement. These then allow us to operationalise the stopping decision points of the CSM.