Chapter 10

Discussion and Future Work

The final chapter of this thesis summarises and discusses the results reported. In particular, we emphasise on the impact of our findings on IR and IIR research, discuss the limitations of our work, and outline several potential future research directions.

10.1 Thesis Summary

In this thesis, we examined how stopping behaviours vary under different search contexts. In particular, we conducted and reported on two user studies under the domain of news search, examining how ❶ result summary lengths and ❷ a variation of search tasks, goals and retrieval systems affected search behaviours. A total of eight different interfaces and conditions were used to examine how behaviours vary – as summarised in Table 10.1.

From the first user study reported in Chapter 7, results showed that as result summary lengths increased (from T0 → T4), searchers became more confident in the decisions they took pertaining to the relevance of documents encountered. However, this was not reflected empirically; their accuracy in identifying relevant content did not improve with

Table 10.1 A summary table of the different experimental interfaces and conditions that were trialled. These are based upon the work reported in Chapters 7 and 8. In total, eight different experimental interfaces and conditioned were employed, considering different result summary lengths, systems and tasks.

		Summary Length	System	Task
Chapter 7	T0	Title only	ND (Non Div.)	AD (Ad-hoc)
	T1	Title + 1 snippet	ND (Non Div.)	AD (Ad-hoc)
	T2	Title + 2 snippets	ND (Non Div.)	AD (Ad-hoc)
	T4	Title + 4 snippets	ND (Non Div.)	AD (Ad-hoc)
Chapter 8	D-AS	Title + 2 snippets	D (Div.)	AS (Aspectual)
	ND-AS	Title + 2 snippets	ND (Non Div.)	AS (Aspectual)
	D-AD	Title + 2 snippets	D (Div.)	AD (Ad-hoc)
	ND-AD	Title + 2 snippets	ND (Non Div.)	AD (Ad-hoc)

longer summaries. In terms of stopping behaviours, a downward trend was observed. As the length of summaries increased, subjects examined to shallower depths per query – an intuitive result, given the increased examination times required for longer summaries.

Considering variations of tasks, goals and systems as reported in Chapter 8, we found that when using diversified system D (i.e. BM25 and XQuAD (Santos et al., 2010)), subjects issued more queries, and stopped at comparatively shallower depths per query. This was in comparison to the non-diversified system ND (i.e. BM25 baseline), where subjects reported feeling less confident about their decisions. Despite the significant differences we observed regarding how the two systems performed, few significant differences were observed when examining changes in searcher behaviours. Most subjects reported difficulty in identifying differences in performance between the two systems.

Analysis of interaction data from these user studies was then used to ground an extensive set of simulations of interaction. These simulations were designed to test a total of twelve individual stopping strategies, derived from six different stopping heuristics[1] and the RBP IR measure. Our approach to cataloguing these heuristics – together with the subsequent operationalisation of them into stopping strategies – provided an answer to **HL-RQ2**. We then tested the overall performance and how closely the simulations matched up to real-world searcher behaviours (across the eight experimental interfaces and conditions). In turn, this allowed us to provide answers to both **HL-RQ3a** and **HL-RQ3b**. The simulations were modelled with the *Complex Searcher Model (CSM)*, a high-level, conceptual model of the search process. By incorporating a new SERP level stopping decision point into the CSM, complete with subsequent empirical evaluation (as presented in Chapter 9), we could then provide an answer to **HL-RQ1**.

Results show that when enabled, the new SERP stopping decision point led to significant improvements over the baseline implementation, with consistent improvements in overall performance (measured in CG) reported across a range of experimental conditions, interfaces and stopping strategies. Improvements in approximations of real-world searcher stopping behaviours were also achieved. However, statistical significance for these improvements was not obtained. Overall, these results provide compelling evidence to address **HL-RQ1**. The results also demonstrate a promising direction for future research in developing our understanding of the search process.

With respect to our simulated analyses of individual stopping strategies, we found several stopping strategies offered high levels of mean CG, and good approximations toward actual searcher stopping behaviours. For example, we found that with increased result summary length, **SS11-COMB** consistently offered the best performance. **SS1-FIX** and **SS4-SAT** offered the best real-world searcher approximations. Furthermore, **SS5-COMB** offered the best level of CG within the second user study, while **SS1-FIX** offered the best level

[1] Stopping heuristics for example considered a searcher's tolerance to non-relevance, or their *frustration* with observing non-relevant content (Kraft and Lee, 1979).

of performance across condition ND AD. However, SS1-FIX and SS10-RELTIME yielded the lowest MSE values. Despite several strategies performing well, no single strategy clearly emerged as offering significantly improved levels of performance or approximations when acting alone. On the contrary, several more complex stopping strategies offered poorer performance, such as SS6-DT and SS7-DKL. This was a common theme in our results: simple and combination-based stopping strategies generally provided the highest levels of performance. This includes the fixed-depth stopping strategy, SS1-FIX, which, counter to our intuition, consistently performed well.

10.2 Discussion

From the analysis of our simulations of interaction, a number of novel, interesting areas of discussion were revealed. In this section, we discuss our findings with an emphasis on examining the result summary level stopping strategies. In particular, our discussion is guided by our four high-level research questions. We repeat these below.

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

With these research questions pertaining to the simulations of interaction (along with the implemented stopping strategies), in-depth discussion of our user studies can be found in Sections 7.2.3 on page 214 and 8.2.3 on page 278. However, we do briefly touch on summarising statements relating to searcher behaviours in Section 10.2.3.

10.2.1 Searcher Models and Realism

Work in this thesis has reported advancements to modelling the IIR process – particularly with the inclusion of the new SERP level stopping decision point. The inclusion of the new stopping decision point led to significant improvements in terms of the level of CG that could be attained, together with improved approximations of real-world behaviours.

However, these significant improvements from the SERP Always baseline (as reported in Chapter 9) were only achieved with the SERP Perfect implementation. This is a limitation, as the implementation relied upon access to the TREC QRELs in order for the impression to be determined – although this implementation acted as a good upper bound. While improvements in performance and approximations were noted with the SERP Average implementation, these changes did not achieve a significant level of improved performance. We discuss this limitation of our simulations later in Section 10.2.4.

Of course, attaining access to the *gold standard* is wholly unrealistic. However, the present study demonstrates the *maximum performance* that can be reached with the inclusion of this new stopping decision point. The observed improvements demonstrate that more realistic simulations of interaction may be produced. With further work examining the proximal cues that searchers observe when forming an initial impression of the SERP, incorporating these findings into future models and simulations of the search process would arguably make them even more realistic.

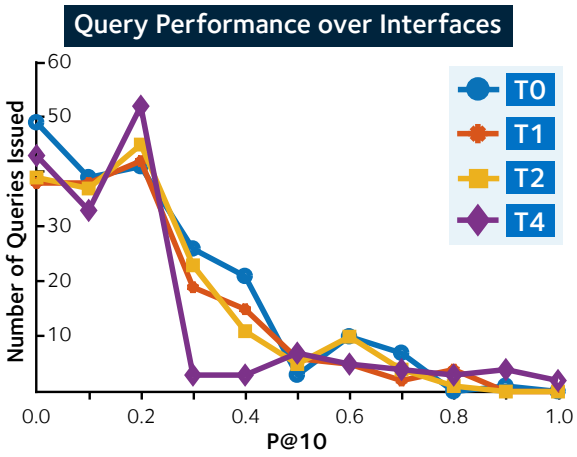
10.2.2 Stopping Strategy Operationalisation

In general, findings across all interfaces and conditions demonstrated that simple stopping strategies tended to yield better performance, and matched better with real-world searcher stopping behaviours. Stopping strategies SS2-NT, SS3-NC, SS4-SAT, SS9-TIME and SS10-RELTIME for example performed and approximated well. We consider these to be simple in the sense that the stopping criterion that they each encoded was straightforward to implement and subsequently measure. Examples included the consideration of aspects such as the number of non-relevant documents encountered, or the elapsed time spent searching since a query was issued.

In contrast, findings also demonstrated that the more complex stopping strategies tended to perform worse. They consistently offered poorer performance and approximations. Complexity was again denoted by the criterion/criteria that were considered by each of the stopping strategies, with more complex computations required in order to determine when the simulated searcher should stop. Given these general findings, *why did the more complex stopping strategies perform and approximate worse on average?* The present section of the discussion focuses primarily on this question, considering the difference-based strategies SS6-DT and SS7-DKL, the IFT-based strategy SS8-IFT, and the RBP-based strategy SS12-RBP. We also discuss the importance of more performant stopping strategies, such as SS5-COMB and SS11-COMB.

Difference Stopping Strategies Considering SS6-DT and SS7-DKL, we hypothesise that the performance of issued queries may be having an effect on the way in which these strategies perform. In other words, the stopping strategies *may not be very robust* to varying levels of query performance. Recall that for our *what-if* performance runs, we employed an interleaved querying strategy QS13, where single and three term queries were interleaved.[2] From empirical evidence, it was shown that single term queries offered poor

[2]Refer to Section 6.4.2.2 on page 164 for additional information on how the querying strategy was implemented.



Query Performance vs. Lengths (All Interfaces)

#Terms	#Queries	Mean P@10
1	12	0.16 ± 0.13
2	162	0.23 ± 0.18
3	287	0.24 ± 0.21

On average, as query length increases, P@10 scores also increase. Values reported represent the mean performance ± standard deviations.

Figure 10.1 Plot demonstrating the performance of queries (**P@10**) across the four experimental interfaces trialled in the user study we report on in Chapter 7. On the right, a table highlights the varying levels of performance (averaged over all four experimental interfaces) in relation to query term lengths. As query term length increases, so too does the mean P@10 score. Similar findings were observed for the study reported in Chapter 8.

performance (in terms of *P@k*) when compared to three term queries – single term queries offered higher levels of query ambiguity compared to the three term queries.[3] As such, using a fixed threshold across queries of varying performance would not necessarily make sense. A low threshold for **SS6-DT** and **SS7-DKL** would mean that searchers would stop too early for single term queries, and examine to excessive depths for three term queries. A higher threshold would mean that searchers would examine to excessive depths generally. This means for example that a low threshold would be too stringent for single term queries, and suggests that poor levels of gain would be achieved.

As such, we hypothesise that for stopping strategies based upon the difference-based heuristic, thresholds should likely be query specific – perhaps dependent upon the length of the query issued. Given queries issued by the real-world subjects in Chapter 7, we also observed a large variation in performance for the queries that were issued. We report this in Figure 10.1, with a plot showing the number of queries issued across each of the four in-

[3] For example, consider the queries `'piracy'` and `'piracy china sea'`. The first single term query returned a majority of documents pertaining to software piracy, along with piracy at sea. In contrast, the three term query returned a majority of its matched documents to instances of piracy on the South China Sea, relevant to the TREC piracy topic.

terfaces, plotted against the performance of the queries. A table also provides evidence to support our hypothesis, showing that as the number of terms in the queries increased, so too did the mean level of query performance. Similar findings were observed in the user study reported in Chapter 8.

IFT Stopping Strategy Next, we consider the poor performance and approximations afforded by **SS8-IFT**. Evidence has shown that IFT has been proven to be good at predicting search behaviours (Ong et al., 2017; Azzopardi et al., 2018). In Section 8.2.2.5 on page 274, we demonstrated that our IFT-based hypotheses matched closely to empirical evidence. So, why did **SS8-IFT** consistently offer poorer performance and approximations when compared to more simplistic stopping strategies? We hypothesise that this comparative lack of performance can be attributed to how the *rate of gain* was operationalised, which serves as the stopping criterion for **SS8-IFT**. This is an inherently difficult value to compute, with limitations relating to the rate of gain considered from two angles:

1. the *per-topic* rate of gain; and
2. *how the rate of gain is estimated* by searchers in the first instance.

Considering point 1 first, we note that the same gain stopping threshold values (for x_8) were trialled over all five topics in the reported simulations of interaction. Table 6.1 on page 140 demonstrated that the number of TREC relevant documents for each of the five topics varies considerably. As such, one would expect that the computed rate of gain would also vary considerably on a per-topic basis. This way, expectations of gain can be kept in check – a rate of gain threshold computed over a performant TREC topic with many relevant documents would perform much worse under a topic for which it is much harder to find relevant documents for (i.e. a comparatively smaller number of TREC relevant documents). This variation in the number of relevant documents over topics (amongst other factors, such as the retrieval system used) is illustrated in Figure 10.2. Using interface **T2** (left) and condition **ND** **AD** (right) over **SS3-NC**, the two plots illustrate how performance varies

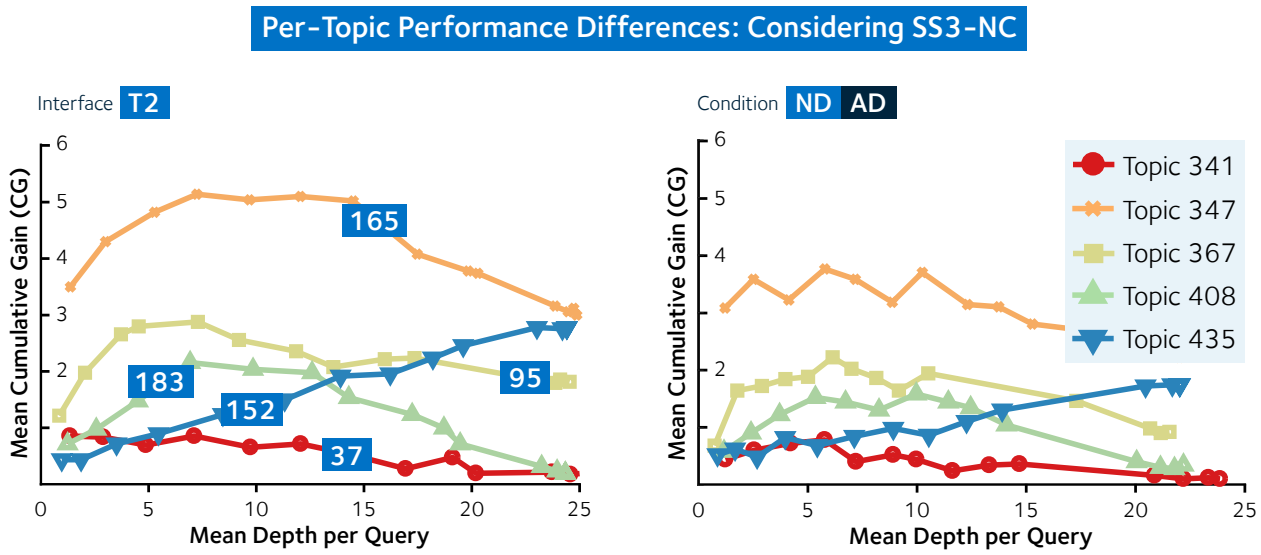


Figure 10.2 Plots demonstrating the wide per-topic variance over the *what-if* performance simulations. On the left, performance over interface **T2** is shown — **ND** **AD** is shown on the right. Stopping strategy **SS3-NC** is used for this demonstration. Similar observations were observed across other interfaces, conditions and stopping strategies. Also **highlighted** on the left plot is the number of TREC relevant documents for each topic. Note the general performance improvement as the number of TREC relevant documents increases for a topic.

across the five topics. We also note a general trend of higher performance for a topic in the plots if a greater number of TREC relevant documents are present.

We also consider how the rate of gain is computed, as per point ❷. *How do searchers estimate a rate of gain threshold?* This is a difficult question to answer, with further study required to address this. However, one would be pressed to believe that from an initial impression of a SERP, a searcher would undertake a series of computations in their head to reach an estimation for a rate of gain threshold value. It is much easier to believe that searchers would rather employ a simpler stopping criterion in this instance, such as *stopping after observing k non-relevant result summaries* (i.e. the frustration-based heuristic). This can be simplified with the trivial example of an individual throwing a ball in the air, as illustrated in Figure 10.3. It would be easier to believe that the ball thrower would think of how to catch the ball in relation to how it is falling through the air, with feedback from their visual

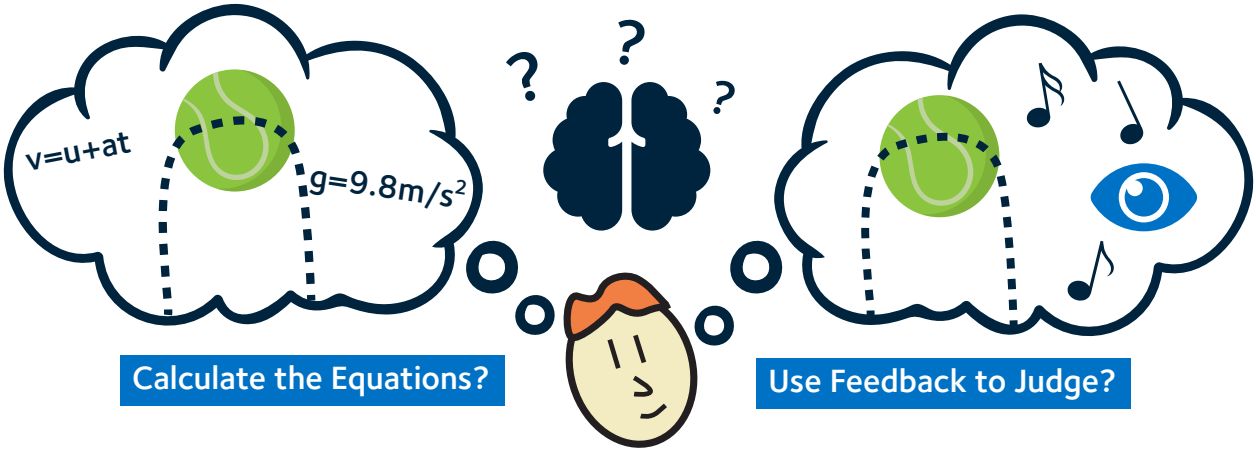


Figure 10.3 *How would you catch a ball in the air?* Would you consider all of the equations required to work out when and where in space the ball will be for you to catch it, or would you rely on your visual/proprioception systems to guide you? If you are human, it'll be the latter.

and proprioception systems. This is opposed to believing that the thrower of the ball may catch it by calculating the equations relating to the physics of the falling ball to predict the optimal point in space at which to intercept it.

However, even if we were able to provide better values for the rate of gain, would we see improvements in real-world approximations? While IFT says that we will, individuals may be behaving in a suboptimal way. A body of literature in ecology suggests that when foraging for food in the wild, animals *do* behave in a suboptimal way. Janetos and Cole (1981) and Krebs et al. (1983) state that animals may employ some *rule of thumb* that is less than perfect, with an example cited as *'take the largest thing you can eat'*. This is some ways analogous to the more simplistic stopping strategies we trialled. Krebs et al. (1983) also argue that these simplistic approaches are actually an optimisation within a wide number of constraints, such as sensory limitations. This may be true of searchers, too – with limited working memory, a more simplistic approach may, in reality, be the better, optimal choice even though the theory may suggest otherwise.

RBP Stopping Strategy We also observed that **SS12-RBP**, the RBP-based stopping strategy, also generally failed to provide a good approximation. Performance in the *what-if* simulations was generally significantly different from the best performing stopping strategies,

although instances such as T2 did not demonstrate any significant difference. While performance matchups might have been relatively good, the depths to which simulated searchers examined content using this stopping strategy were weak to a considerable degree. Refer to plot SS12-RBP in Figure 7.8 on page 232 for an example. Recall the patience parameter p of RBP, that dictated how deep down a list of ranked results a searcher would be prepared to go. The point at which the searcher would decide to stop was modelled stochastically. In the real-world, searchers do not roll a dice to determine when to stop, but rather rely upon some form of an intuitive informational cue, as have been previously shown to affect search behaviours. However, it may also be the case that this way of representing human behaviour is also correct at times – humans can often behave irrationally.

Considering more Performant Stopping Strategies Both the combination-based stopping strategies SS5-COMB and SS11-COMB performed and approximated searcher stopping behaviours well. Formed of more simplistic stopping strategies (e.g. SS2-NT), results seem to suggest that searchers do not consider a single criterion when determining the point at which they should stop examining results – an interesting finding.

This interesting conclusion can be corroborated by other recent studies. Work by Zhang et al. (2017a) used the *Bejeweled Player Model (BPM)* to model a searcher's stopping behaviours, where they would stop when:

> *"he/she either has found sufficient useful information, or no more patience to continue."*
>
> Zhang et al. (2017a)

Findings from this study demonstrated improvements in the correlations between searcher satisfaction and existing IR evaluation measures. This was also corroborated in a recent study by Azzopardi et al. (2018). Central to this argument is the similarity of the BPM to SS5-COMB, that considered a combination of the satiation (SS4-SAT) and frustration (SS2-NT or SS3-NC) stopping strategies. This provides evidence that empirically validates the inclusion of both satiation and frustration-based stopping heuristics within the

searcher model. The evidence is clearly showing that multiple criteria are being considered when a stopping point is decided, and future work should consider the development of measures that support both criteria.

The Fixed Depth Fallacy Overall, a majority of stopping strategies performed well and produced approximations that were very close to one another, with few significant differences. One particularly surprising result was that of **SS1-FIX**. The fixed-depth, non-adaptive baseline approach consistently offered good performance and approximations. This is counter-intuitive, as it would make sense for more adaptive strategies to offer improved approximations. It is likely that different subjects would have employed different stopping strategies, or a variety of different strategies depending upon the situation (i.e. as demonstrated by **SS11-COMB**). In this regard, next steps should consider stopping behaviours on an individual level. However, from the perspective of averaging over a population, many of the stopping strategies trialled, and when tuned appropriately (i.e. would **SS1-FIX @24** really be considered as realistic? It is unlikely!), offer good approximations and performance. This provides a rationale as to how the fixed depth strategies consistently performed and approximated so well across our results.

10.2.3 Searcher Behaviours

From the reported user studies, it is clear that the interfaces and conditions that we trialled do affect the behaviours of searchers. In terms of stopping behaviours, we did observe differences, but differences often were not significant. We hypothesise that due to the high levels of variance that we observed, larger sample sizes over each study would be required in order to tease out significant differences and to provide data for further examination.

Understanding stopping behaviours is difficult. What findings from our studies do suggest is that variations in interfaces, tasks, goals and systems do impact upon performance. For example, as we increased result summary lengths, stopping depths became shallower

(i.e. from T0 → T4). More extreme interfaces and conditions would likely amplify the effect. Factors such as how the prior topic knowledge that a subject possesses were also not considered, and would likely play a role in stopping behaviours.

10.2.4 Simulations of Interaction

In this thesis, we have presented significant advancements in terms of modelling and understanding the IIR process. We developed an extensive framework that allowed us to change components of the underlying CSM. Given this framework, we could then formulate the search problem more precisely, and explore the impact that each of the component variations had on the wider search process. As components were changed, we were able to demonstrate improvements in the approximations of human searcher behaviours. Given the limitations of our user studies with the risk of an insufficient amount of interaction data, simulations of interaction allowed us to generate more data at a much lower cost.

One particularly novel contribution concerning the simulations of interaction was addressing the issue of comparing results across different configurations. Being stochastic in nature, the simulations relied upon the roll of a dice to determine whether a simulated searcher would click on a result summary link (if deemed sufficiently attractive to warrant further examination), or save a document as relevant (if deemed relevant to the given information need). These were grounded on the TREC relevance judgements and interaction probabilities extracted from the user studies. Across different configurations however, outcomes of the dice roll would have resulted in different decisions being taken – which in turn ensured that when examining two configurations, their outcomes would not be comparable.

Section 6.4.2.3 on page 167 outlined a *pre-rolled judgements* technique that rolled the dice *a priori* 50 times, with 50 being the number of trials that were run per configuration. This then meant that during the simulations, the decision maker components of the **SimIIR** framework essentially became deterministic, extracting the judgement for a particular trial from

a pre-rolled *action judgement file*. In turn, this addressed the issue of comparability between different configurations. With the same judgements, comparisons became fairer. Of course, a larger number of trials would always be more desirable as a means of teasing out further differences that perhaps would otherwise not have been observed.

We also note limitations of the approach taken in conducting our simulations of interaction. Most notably, we considered the *most optimistic outcome* at several points in our simulations, mainly pertaining to the perceived quality of a SERP. A prime example of this was highlighted in Section 10.2.1, with the SERP Perfect SERP level stopping decision point implementation highlighted as the implementation yielding significant improvements in performance, yet attaining this with access to TREC QREL judgements.

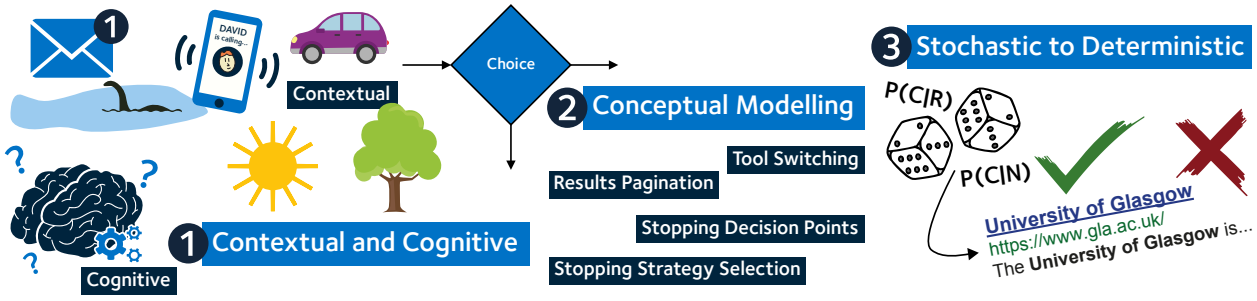
A further example of this approach was demonstrated with combination stopping strategy SS11-COMB. Similar to the SERP level stopping decision point, this strategy took an initial impression of a presented SERP, and used this impression to select an appropriate constituent stopping strategy – with either SS4-SAT for a SERP yielding relevant documents early in the rankings, or SS10-RELTIME for a SERP of dubious quality. Under these conditions, such strategies do intuitively make sense. However, the decision was again made with access to TREC QRELs – *P*@1 was used to determine if the SERP yielded relevant content at shallow ranks. If, for example, a stochastic approach were to be implemented in determining what stopping strategy to employ, it may mean that even better approximations of real-world searcher behaviours could be achieved.

10.3 Future Research Directions

From the summary and discussion of our empirical results, a number of potential avenues for future work may be considered. In this section, we consider: how to improve the realism of simulations of interaction further; stopping heuristics and strategies; simulation trials and topics; and the modelling of stopping from the level of individual searchers.

10.3.1 Improving Simulation Realism

In this thesis, we presented the CSM, a high-level, conceptual searcher model. It encapsulates many of the different activities and decision points that searchers would contend with across informational search tasks. With the inclusion of the new SERP level stopping decision point, improvements were made to the realism of the simulations that were executed with the CSM. However, *what changes could we subsequently make to the CSM and related infrastructure in future work that would aid in advancing the realism of these simulations further?* As illustrated below, we consider this open question from three main research strands.



Contextual and Cognitive Our first strand considers **contextual** and **cognitive** factors. All experimentation in this thesis was conducted under the domain of news search, with subjects of the user studies asked to imagine that they were newspaper reporters, having being given a task to find documents that they thought were relevant to a particular topic. However, this scenario is very specific. If we performed studies with the same methodology, but under a different search context, would we find similar results? Arguably, behaviours will change – general web search and a detailed examination of content under the context we employed will result in different outcomes, for example. Different tasks can also be considered. Aspectual and ad-hoc tasks were considered as we believed they would offer the greatest difference in terms of stopping behaviours. Would other retrieval tasks offer even bigger differences in terms of searcher behaviours?

Other factors such as the location at which the search is undertaken, the device upon which the search is undertaken and other external pressures will also undoubtedly influence the

outcome of the results obtained. Crowdsourced subjects whose behaviours are reported in this thesis conducted our experiments on a desktop or laptop computer. They were instructed to be in a comfortable, quiet location, free from major distractions. In reality, individuals are less likely to search in such conditions. Perhaps time pressures would influence their behaviours – a student under pressure to finish a draft of her paper will behave differently to one who is not under the same pressure. With the proliferation of mobile devices such as smartphones, searching on such devices must also be considered. A recent study by Ong et al. (2017) demonstrated that search behaviours, for example, do differ between individuals using desktop computers and smartphones.

Much work remains to determine how we can try to understand and subsequently model the cognitive processes and factors that influence how individuals behave when searching. Individuals are products of their prior experiences, and are therefore unique; behaviours will undoubtedly differ from person to person. Within the modelling process, novel techniques can be applied that could possibly improve the realism of simulations. For example, within the **SimIIR** framework, the search context component tracks a list of queries issued, documents examined (and saved), along with other measures. Could this component be manipulated in such a way as to better mimic the behaviours of a human? Rather than maintaining a perfect list of everything that has been examined, a simulated searcher could be programmed to become 'forgetful' in remembering what they have examined, with cues within a document reminding them that they previously examined it. Other factors, such as prior topic knowledge (as alluded to in Section 10.2.3) ought to be considered, as such aspects would likely impact upon the stopping behaviours of searchers.

As alluded to in Chapter 8, further work could also be undertaken in relation to the decision making components of the **SimIIR** framework. This work would consider how simulated searchers would judge the attractiveness of result summaries and relevance of documents to a given topic. Decision makers were implemented primarily with ad-hoc retrieval in mind, considering only the probability of clicking or saving with respect to the TREC *Query Relevance Judgement (QREL)* judgement. For aspectual retrieval tasks, further work would

consider whether the result summary or document contains a discussion of new entities for the topic, such as a previously unseen species of animal.

Conceptual Modelling We next consider a number of further enhancements to the CSM that could improve the realism of simulations further. Examples in the illustration above consider potential areas for future improvement. One such example, **tool switching**, (demonstrated by Thomas et al. (2014)) would be considered at the beginning of the search process. It would enable a searcher to determine what tool (or retrieval system) would be better suited to help them satisfy their information need. This is opposed to the current CSM as presented in this thesis that assumes a retrieval system has been selected *a priori*. A study by White and Dumais (2009) has shown that predicting tool switching is feasible. They reported that sufficiently consistent behaviours exhibited by searchers in relation to this phenomenon led to accurate predictions of tool switching events.

Results pagination is also listed in the illustration above. Here, a simulated searcher will be presented with SERPs that are split across a number of different pages, rather than examining a continuous ranked list of results. This would involve the notion of extracting additional grounding data from interaction logs, perhaps such as the likelihood of a searcher continuing to the next SERP page. This would likely impact upon the realism of simulations, as a study by Jansen and Spink (2005) showed a sharp decrease in content examined after the first page of results. Further examination of modelling stopping behaviours within the CSM is also considered; refer to Section 10.3.2 for further details.

Stochastic to Deterministic Decisions pertaining to the attractiveness of result summaries and the relevance of documents within our simulations of interaction were determined *stochastically*, or by a roll of the dice. While a simplifying assumption that has been used in many other studies employing simulations of interaction, this is an unrealistic approach. If implemented correctly, a more *deterministic* solution would offer more realistic simulations, where simulated searchers would be able to *learn* as they traverse through content, improving their decision making abilities based upon the content observed, rather than the

outcome of a roll of a dice. Advancements in understanding the *information triage* process would undoubtedly lead to improved realism. In addition, the inclusion of *variable interaction costs* would also benefit the realism of future simulations.[4]

10.3.2 Stopping Heuristics and Strategies

In this thesis, we considered a total of twelve different stopping strategies, operationalised from a total of seven different stopping heuristics. We showed how each of the different strategies perform over a number of different experimental interfaces and conditions. During the methodological design stage, it became apparent that approaches taken for the operationalisation of our stopping strategies were just one of many. *What if we implemented our stopping strategies in different ways? Why did we select these strategies?* Here, we consider these questions with insight into what might happen if they were to be addressed.

Stopping Decision Points Following on with the theme of improving the underlying CSM, additional stopping decision points could be included. These would provide searchers subscribing to the CSM with greater flexibility regarding when they stop examining content. Additional stopping decision points could, for example, include one for tool switching. In this example, as we discussed earlier, a searcher, after spending some period of time on one retrieval system, could decide to stop using it after certain criteria are met. After this point has been reached, they will then switch to a different retrieval system. A further interesting research question would be whether the result summary level stopping strategies trialled in this thesis would work at different stopping decision points. For example, at a session level, would these strategies make sense? Would using them at that decision point lead to a better matchup with real-world stopping behaviours?

Stopping Strategy Selection From here, we can also consider a further decision point that could be encoded within the CSM. Inspired by **SS11-COMB**, consideration must be

[4] As discussed previously in this thesis, *time-biased gain* (Smucker and Clarke, 2012) is an example of such an approach.

taken into deciding *why* and *when* a particular stopping strategy could be employed. As we demonstrated in Figure 5.3 on page 132, SS11-COMB employs both the frustration and give-up time-based stopping heuristics – but not at the same time. Rather, a decision is made pertaining to the quality of the presented SERP (much like the SERP level stopping decision point). The outcome of this decision then dictates what stopping strategy is employed for the remainder of the query. Further refinements to this approach could, for example, include additional stopping strategies and a wider range of conditions for employing them. Empirical evidence could be extracted from interaction logs to determine if, under certain circumstances, searchers would favour one approach over another.

Stopping Strategy Operationalisation An open question arising from the work in this thesis considers: *how do you operationalise the stopping heuristics?* Clearly, from the outline of the twelve stopping strategies in Chapter 5 on page 121 (and implementation methodology in Section 6.4.2.6 on page 173), there are a large number of different ways in which the stopping strategies can be implemented. While we provided a means and justification for the approaches that we took in this thesis, we have reason to believe that some of the stopping strategies – especially SS6-DT, SS7-DKL and SS8-IFT – performed poorly, perhaps because of our implementations (refer to Section 10.2.2). For example, the rate of gain for SS8-IFT could have been computed on a per topic basis. Further work will be required in order to determine if different implementations would lead to performance improvements.

Considering Additional Stopping Heuristics Of course, the seven stopping heuristics that we considered do not constitute the entirety of the heuristics defined in the literature. We selected these heuristics as they offered interesting differences between one another, were *relatively* straightforward to implement, and would likely be discernible across complex informational search tasks (though this was not necessarily proved). Unused heuristics such as the mental list heuristic (considering different criteria that must be met, as outlined by Nickles (1995) and detailed in Section 3.2.2.3 on page 87) would have been much more challenging to operationalise and implement – and even so, would such a heuristic be suitable for the task at hand? The ability for a searcher to create a series of bullet points about a

topic would imply he or she has some sound idea of their objective. The searcher's knowledge of a topic may be so limited that such a heuristic would be unsuitable. Linking back to contextual factors above, considering additional search contexts (perhaps with searchers of astute and limited knowledge of a topic) would be interesting to examine.

Towards Future IR Measures Given the above, findings from this research provide motivation for further work considering the inclusion of stopping heuristics within the measures that are used within IR research. For example, stopping strategy **SS5-COMB** demonstrated good overall and performance considering a searcher's satisfaction and tolerance toward non-relevant material. This has also been shown in the BPM (Zhang et al., 2017a).

10.3.3 Simulation Trials and Topics

We also consider future work in terms of *how* the simulations of interaction could be run. While 50 trials were selected because of the fact that approximately 50 subjects partook in each user study, there are likely trends and significant differences that exist that we simply did not observe because of a lack of experimental power. This limitation was also imposed with an insufficient amount of processing power to complete the experiments in a reasonable timeframe.[5] With more powerful computer hardware, scaling up the experiments with more trials would have become a more realistic prospect.

We also consider using five topics for our performance *(what-if)* experiments to be a limiting factor. While the decision to use five topics was justified due to a lack of data (considering entities across the remaining 45 topics in Chapter 8) – and to ensure that comparisons between interfaces and conditions were fair – 50 topics would have been preferred (refer to Figure 10.2). If (aspectual) data were available for the remaining 45 topics, we could then trial additional performance runs, which may also lead to the observation of other trends and potential significant differences.

[5] Using the experimental setup detailed in this thesis, all simulations of interaction took approximately 38 days of processing time.

10.3.4 Individual Searcher Stopping Behaviours

Our final consideration for future work revolves around the notion of *individual searcher stopping behaviours*. In this thesis, we considered searcher stopping behaviours, reported across ≈ 50 subjects, over each interface and condition that was trialled. This provided us with a rough approximation as to what strategies work best, with similar findings reported across interfaces and conditions. However, research has shown that individual searcher behaviours may differ to a significant degree. If we considered individual searchers, what trends would we then observe? We may see a decrease in how well the fixed depth stopping strategy SS1-FIX fares, given that we hypothesised in Section 5.1 on page 123 that such an approach would work well on average. If we examined behaviours on a per-searcher basis (or even at a session level), how would the strategy then fare?

Examining behaviours on a per-searcher level will avoid watering down results through averaging over a particular cohort, exposing more interesting results. For example, could we perform a classification of searcher stopping behaviours? Such an approach was followed, for example, by Smucker (2011), who devised a classification of searchers when examining documents – with searchers being categorised into one of either *fast and liberal* or *slow and neutral*. This is undoubtedly one key area of future work that we must consider in order to develop a deeper understanding of the stopping behaviours that searchers employ.

10.4 Final Remarks

Stopping during the search process is a difficult phenomenon to understand and model effectively. A wide range of different factors influence the internal decision-making process of searchers. We have shown in this thesis that a number of simple stopping strategies can offer improved performance and approximations of real-world searcher behaviours. We also provide novel evidence to motivate the fact that multiple stopping criteria need to be

considered in the development of future IR evaluation measures, along with the inclusion of additional stopping decision points to improve the realism of future searcher models. The development of the CSM has also been positive, with a solid baseline provided for future work in developing ever more realistic simulations of interaction.

Despite the inherently difficult task that understanding and modelling stopping behaviours represent, we believe that the potential benefits of further exploration in this area will undoubtedly aid the searchers and researchers of future retrieval systems.