Information Scent, Searching and Stopping
Modelling SERP Level Stopping Behaviour

David Maxwell[1] and Leif Azzopardi[2]

[1] School of Computing Science
University of Glasgow, Scotland
[2] Computer and Information Sciences
University of Strathclyde, Scotland
d.maxwell.1@research.gla.ac.uk, leif.azzopardi@strath.ac.uk

Abstract. Current models and measures of the *Interactive Information Retrieval (IIR)* process typically assume that a searcher will always examine the first snippet in a given *Search Engine Results Page (SERP)*, and then with some probability or cutoff, he or she will stop examining snippets and/or documents in the ranked list (snippet level stopping). Prior work has however shown that searchers will form an initial impression of the SERP, and will often abandon a page without clicking on or inspecting in detail any snippets or documents. That is, the *information scent* affects their decision to continue. In this work, we examine whether considering the information scent of a page leads to better predictions of stopping behaviour. In a simulated analysis, grounded with data from a prior user study, we show that introducing a SERP level stopping strategy can improve the performance attained by simulated users, resulting in an increase in gain across most snippet level stopping strategies. When compared to actual search and stopping behaviour, incorporating SERP level stopping offers a closer approximation than without. These findings show that models and measures that naïvely assume snippets and documents in a ranked list are actually examined in detail are less accurate, and that modelling SERP level stopping is required to create more realistic models of the search process.

1 Introduction

Interactive Information Retrieval (IIR) is a complex, non-trivial process in which during a search session, searchers may issue multiple queries and examine a varying number of snippets and documents per query [12]. One particularly important part of this process is knowing *when to stop* [25]. Stop too early, and you could miss useful information; stop too late, and you could be wasting valuable time examining non-relevant material. Research into examining stopping behaviour has been until recently relatively sparse, with a series of studies finding that people stop based upon their intuition, or what is simply *"good enough"* [38]. Formally, stopping behaviour has typically been considered at two levels: *(i)* the query (or snippet) level; and *(ii)* the session level. As such, researchers have attempted to quantify the sense of *"enough"* at both levels by proposing a series of *stopping rules* and heuristics that attempt to encode this intuition (e.g. [3,

6, 15, 25]). Models of stopping behaviour have also been encoded with measures used to evaluate the quality of ranked lists, and within simulations of interaction. However, the majority of work in this area currently assumes that a searcher will always examine the first snippet, and will either examine to a fixed depth, or stop based upon some probability on continuing. Yet the *Search Engine Results Page (SERP)* provides various cues which searchers use to decide when to stop, or even whether to begin examining the SERP in detail at all. Thus, current stopping models tend to be agnostic of the *information scent* [5, 30], which has been previously shown to affect a searcher's (stopping) behaviours [4, 37]. This scent can be used to determine whether a given SERP *smells good enough* to *enter* and examine individual snippets within the SERP in more detail, as per the *Patch Model* in *Foraging Theory* [32].

To this end, this paper: *(i)* introduces a new SERP level decision point in an established interaction model, allowing searchers following the model to obtain an initial impression (or *'overview'*) of the SERP before deciding to enter or abandon it; and *(ii)* enumerates a series of simple SERP level stopping strategies, implementing the new decision point in several ways. These strategies are grounded using analysis from a prior user study [21] examining information scent. We report on a large-scale simulation, allowing us to address our two main research questions. Does incorporating a SERP level stopping decision point, motivated by information scent, lead to: **RQ1** higher overall performance, and **RQ2** better approximations of searcher stopping behaviour?

2 Background

A user is said to *abandon* a SERP when he or she fails to click on any of the results returned for the given query [7, 10]. This may be for a variety of reasons, the primary reason being user satisfaction (or lack of) [10]. Satisfaction from simply examining snippets may lead to *good abandonment* [16, 37]. Alternatively, if the presented SERP looks poor, dissatisfaction occurs. This phenomenon has been shown to lead to a difference in information seeking behaviour, which have been analysed and subsequently modelled [14]. We consider in this study *Information Foraging Theory (IFT)* [30] as a means for attempting to model such a process, where a user abandons a SERP through dissatisfaction with the presented SERP – good abandonment in this study is not considered.

IFT is primarily composed of three models: the *Information Scent model*, the *Information Patch model*, and the *Information Diet model* [30]. Of particular relevance to this work are the related studies on *scent* and *patches* (as discussed in Section 3). Pirolli and Card [30] argue that information seekers are like animals foraging in the wild, and as such will follow a scent to find food. Similarly, information seekers follow *proximal cues* provided by hypertext links, titles, snippets and thumbnails to help locate relevant information [5, 26, 28–30]. In the context of news search, cues were examined by Sundar et al. [33]. Cues such as an article's source were shown to have a powerful effect on the perception of said article. If cues can provide a rationale as to what leads to a promising scent trail, it follows that scent also provides a rationale as to when a searcher will stop examining a set of results [30, 36, 37]. The distribution of relevant search

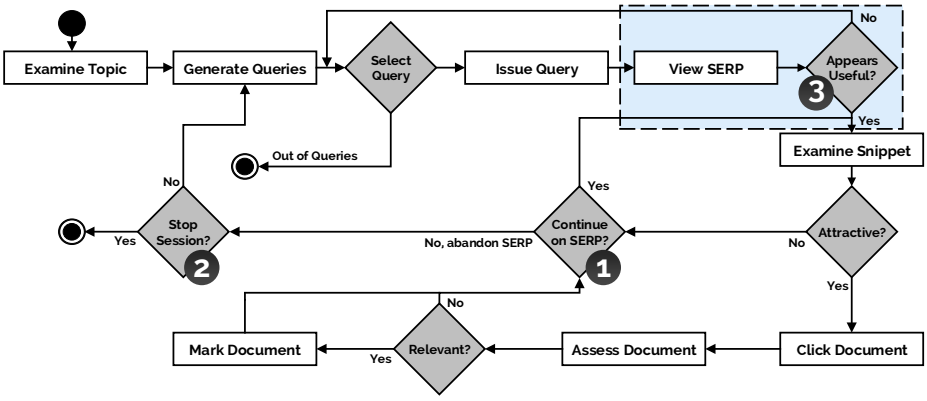

Fig. 1. The updated *Complex Searcher Model (CSM)* with the key decisions (shown in grey) and actions (shown in white). Stopping level decision points are numbered *1–3*. The new SERP level stopping decision point is highlighted in the dashed box.

items also matters: a searcher may continue to forage to greater depths if the SERP appears to contain many relevant items [37]. A similar trend was also observed by Card et al. [4], who found that when navigating through webpages, searchers were more likely to leave when the information scent began to decline.

Examining searcher behaviours when considering scent has been examined by several researchers (e.g. [8, 21, 27, 33, 37]). Wu et al. [37] conducted a user study where the scent of the first SERP was manipulated. They created low, medium and high scent SERPs by changing the number and distribution of relevant items on the page. Subjects interacting with SERPs with a higher scent examined more content and clicked to greater depths, while subjects on low scent pages examined less, and were more likely to abandon the page altogether. A study by Ong et al. [27] replicated the same experimental setup as used by Wu et al. [37], but for both desktop and mobile environments, where similar findings were observed – subjects using the desktop interface however tended to perform better. Maxwell et al. [21] conducted a user study where information scent was varied by manipulating the length of snippets (changing proximal cues) as opposed to manipulating performance of SERPs as done before [27, 37]. It was found that as result snippets increased in length from title only to title plus four line summaries, subjects examined fewer snippets – and were more likely to click on documents, but with lower accuracy. Taken together, these studies suggest that the information scent does indeed influence stopping behaviour. In this study, we operationalise scent as the performance of a SERP (as done in [27, 37]), examining how scent affects search, stopping and overall performance.

3 Updating the *Complex Searcher Model*

We propose the introduction of a SERP level decision point within the *Complex Searcher Model (CSM)*. The CSM combines several frameworks previously proposed in the literature [2, 18, 20, 34] that model the search process for the simulation and evaluation of user behaviour and performance. The process models

ad-hoc topic retrieval based tasks, where the searcher has to identify documents that are relevant to a given information need. Represented as a flow diagram, the CSM is illustrated in Figure 1.[3] The key stopping decision points in the CSM are highlighted in the figure as: *(1)* the *snippet level* decision point (referred to as the *query level* in the literature); *(2) session level* stopping; and *(3)* the proposed *SERP level* stopping decision point.

The new decision point considers the SERP as a whole, and the impression that the searcher obtains from the cues and information visible to them within the browser's viewport. Searchers could, for example, *skim* the SERP, examining the titles and/or URLs of visible results, before determining – through the process of *information triage* [17] – whether enough relevant content is present to examine them in more detail. By considering the SERP as a whole, this provides a way to model abandonment within the search process, rather than assuming that a searcher will assess the first snippet specifically. This therefore marks a departure from assumptions encoded within many *Information Retrieval (IR)* models and measures, such as $P@k$, *RBP* [24], and *INST* [1, 23, 31]. The motivation for including this additional decision point stems from empirical research (i.e. query abandonment [9]) and theory. In Foraging Theory, as mentioned in Section 2, a forager, when presented with a patch, will survey said patch to assess its potential gain before making a decision as to whether it would be worth their while entering it [32]. For example, McNair [22] showed that foragers assess patches – and even select different strategies – based upon this initial patch impression. When considering a SERP as being analogous to a patch, we posit that, given the opportunity to judge a SERP for potential usefulness, a searcher will be able to save time by abandoning SERPs that appear to offer poor yields, and thus search more efficiently. In this work, we will compare the stopping behaviour and overall performance with and without the SERP level stopping decision point across established snippet level stopping strategies, along with an examination as to which approach best approximates actual searcher behaviour.

4 Experimental Method

To address **RQ1** and **RQ2**, we first conducted a large scale simulation to assess the performance when different SERP level stopping strategies are employed (***performance runs***). Then we conducted a simulated analysis replaying actual user queries to determine which SERP level stopping strategy offers the best approximation to actual searcher stopping behaviour (***comparison runs***).

4.1 Corpus, Topic and System

For this study, we used the *TREC AQUAINT* newswire collection, complete with the *TREC 2005 Robust Track* topic set. The set consists of a total of 50 topics, all of which were used for our performance runs. The AQUAINT collection was indexed with the *Whoosh IR toolkit*[4] (version 2.7.4), where stopwords

[3] For a more detailed description of the flow of interaction and various processes represented within the CSM, refer to Maxwell et al. [18].

[4] Whoosh is available on *PyPi* at `https://pypi.python.org/pypi/Whoosh/`.

Table 1. Interaction costs and probabilities (as observed from the user study by Maxwell et al. [21]) that are used to ground our simulated analysis. Refer to Section 4.2 for an explanation of each of the probabilities listed.

Time required to...	Seconds
...issue a query	9.42
...examine a SERP	3.93
...examine a snippet	2.35
...examine a document	17.19
...mark a document	1.26
Session Time	**360**

(a) Interaction costs

Probability		Avg.	Savvy	Naïve
SERP	$P(E\|LS)$	0.34	0.00	0.74
	$P(E\|HS)$	0.77	0.80	0.82
Snip.	$P(C\|R)$	0.35		
	$P(C\|N)$	0.25		
Doc.	$P(M\|R)$	0.67		
	$P(M\|N)$	0.58		

(b) Interaction probabilities

were removed with Porter stemming applied. The retrieval model used for all simulations was BM25 ($b = 0.75$). The simulation framework *SimIIR*[5] was used, where we added the proposed SERP level component to the framework.

4.2 User Study, Subjects, Costs and Probabilities

Log interaction data was obtained from a within-subjects user study by Maxwell et al. [21], using the same collection and retrieval model as above. In the study, 53 subjects undertook ad-hoc topic retrieval using the same configuration of search engine and corpus as described above. Subjects were asked to identify (mark) as many relevant documents as they could over four topics, with each subject allocated a total of 10 minutes per topic[6]. For each topic, the search system was configured to present *query biased snippets* [35] of different lengths. For this study however, we consider only one of those interfaces – where two fragments were presented. This decision was taken: *(i)* to simplify the reporting of our results; and *(ii)* because the interfaces all yielded similar interaction probabilities. Two snippet fragments (roughly equivalent to two lines of surrogate text) is considered to provide a good tradeoff between length and examination cost [11].

Given the log data for the interface, we were able to estimate the interaction probabilities and costs to ground our simulations for this study. Table 1(b) presents: the probability of clicking on a result summary, given it is TREC relevant or not ($P(C|R)$ and $P(C|N)$, respectively); and the probability of marking a document relevant, given it has been clicked on and is TREC relevant or not ($P(M|R)$ and $P(M|N)$, respectively). The table also includes the probabilities of *examining* a SERP yielding a *high* information scent (good results), represented by $P(E|HS)$ – with the converse for *low* information scent (poor results)

[5] SimIIR is available at `https://github.com/leifos/simiir`.

[6] Despite the allocation of 10 minutes per topic, only the first six minutes (360 seconds) of interaction data were considered in the results of Maxwell et al. [21]. As such, we use this as our simulated search session time limit. Refer to Maxwell et al. [21] for the rationale behind this decision.

defined as $P(E|LS)$. To compute the latter two probabilities, we first categorised queries issued by each subject according to scent, such that if $P@10 = 0.0$, the scent level would be considered to be low. This definition follows from work by Wu et al. [37], who state that a page that returns little or no relevant content can be considered to offer a low information scent. We then counted the number of SERPs that recorded no clicks as abandoned SERPs (as per Hassan and White [9]), and divided this value by the number of queries issued. From Table 1(b), we observe that the probability of continuing after observing a SERP of high scent ($P(E|HS)$) is greater than the probability of continuing to examine a low scent SERP ($P(E|LS)$). This provides evidence that searchers do indeed attempt to avoid low quality SERPs. In this study, we used three sets of SERP interaction probabilities to examine the effect of information scent on search behaviour. These are subsequently detailed in Section 4.3.

4.3 User Simulations Setup

To run a given simulation, we instantiated each component of the CSM (as illustrated in Figure 1). Since we employed various interaction probabilities, the stochastic components using these were each trialled ten times. Each run's pseudo-random number generator was seeded to ensure reproducible results, with the same seed used across SERP conditions. This allowed us to perform a pairwise comparison of performance. Considering 50 topics, each component, and the numerous parameter settings trialled, a total of approximately $356,000$ runs were executed to produce the required results. Each of the different simulation components and their instantiated configurations are described below.

Query Generation Keskustalo et al. [13] proposed a series of *idealised, prototypical* approaches to generating queries, as identified from a user study. In particular, strategy **QS3** identified by the authors offered reasonably good performance. The strategy generates queries of three terms in length.[7] Two *pivot terms* were selected, with an additional third term added. However, users have been shown to steadily build up their queries as they acquire more information, first issuing short queries then increasing their length [13]. To this end, we created a modified querying strategy, taking the pivot terms, issuing these as individual queries first, and then combining them as the pivot. This approach: *(i)* makes the querying strategy more realistic [13]; and *(ii)* allows us to test the robustness of both the SERP level and snippet level stopping strategies when faced with both good and poor performing queries.[8]

SERP Decision Making The new SERP level decision point allows for a searcher to begin examining individual snippets for attractiveness, or abandon the SERP completely. To examine this component in detail, we report on three different implementations.

[7] Human subjects issued queries of 3.31 terms on average. This means that the three term queries generated by QS3 can be considered as a reasonable approximation.

[8] For example, a robust snippet level stopping strategy would ideally stop early in the ranked list for poor performing queries, and later for good queries – good queries will return more relevant documents in the ranked list of results.

– **Always *(No SERP Judgements – Always Examine):*** With this strategy, a user will always *enter* the SERP and examine a number of snippets, determined by the snippet level stopping strategy. This is the current state of the art that we consider as our baseline approach.
– **Perfect *(Perfect SERP Judgements):*** Here, a simulated user will only begin to examine a SERP in detail if $P@k > 0$ (the patch threshold). If $P@k = 0$, the user will abandon the SERP and proceed to the next action as dictated by the CSM. This is the upper bound for our simulations – analogous to, as an example, the *ideal user* of Hagen et al. [8].
– **Stochastic SERP Judgements:** This strategy uses a stochastic element to determine whether the simulated user should enter the SERP or not. Like above, the $P@k$ of the SERP is computed. If the SERP is of high scent, $P(E|HS)$ is used to determine whether the user enters the SERP. If the SERP is considered to be of low scent, $P(E|LS)$ is used to determine the likelihood of abandonment. The three sets of probabilities we considered in this study are detailed below.

  - **Average:** $P(E|HS)$ and $P(E|LS)$ are estimated over all users.
  - **Savvy:** $P(E|HS)$ and $P(E|LS)$ are estimated based on the top 15 users with the lowest $P(E|LS)$.
  - **Naïve:** $P(E|HS)$ and $P(E|LS)$ are estimated based on the top 15 users with the highest $P(E|LS)$.

Table 1(b) shows the probabilities used for **Average**, **Savvy** and **Naïve**. The information scent of the SERP was estimated using the associated TREC QRELs for the given topic and based on the top seven snippets returned ($k = 7$). This value was selected as the interface in the user study displayed, on average, seven snippets in the browser's viewport. We also considered additional ways to estimate the scent of page, such as considering the uniqueness of relevant documents within a SERP (i.e. stop if lots of *purple links* – visited links – were on the SERP page). However, the findings were similar to those reported here – and so were not included due to space constraints.

Snippet and Document Decision Making As done in prior simulations [2, 19], the decision to click on a snippet – and the subsequent decision to mark a document relevant – are based upon interaction probabilities. The clicking ($P(C)$) and marking ($P(M)$) probabilities used here are reported in Table 1(b).

Snippet Level Stopping Strategies If the scent of a result page appears to be good enough to examine in more detail, a simulated user will employ the use of one of the following established SERP level stopping strategies.

– **SS1** ***(Fixed Depth):*** A simulated user will stop examining the ranked list once they have observed x_1 snippets, regardless of their relevance.
– **SS2** ***(Adaptive):*** A simulated user will stop once they have observed x_2 non-relevant snippets on the provided SERP.
– **SS3** ***(Adaptive):*** A simulated user will stop once they have observed x_3 non-relevant snippets in a row *(contiguously)*.

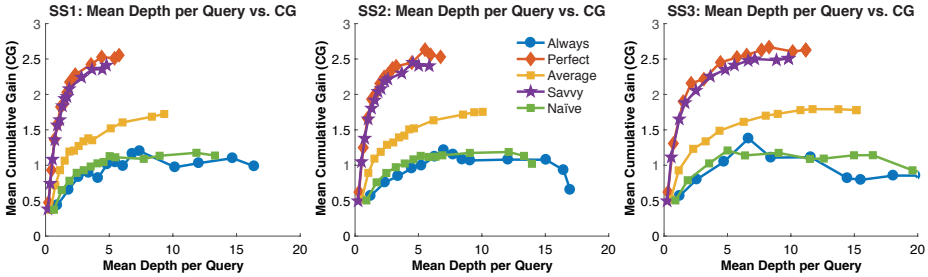

Fig. 2. Plots illustrating the results of our performance runs over each SERP stopping strategy and snippet level stopping strategies **SS1** (L), **SS2** (C) and **SS3** (R).

All three strategies have been used in prior simulations [18, 20]. **SS1** can be considered to be the *de facto* approach used by many models and measures we use in the field (e.g. $P@k$). The adaptive strategies **SS2** and **SS3** that consider a user's tolerance to non-relevance are based upon the *frustration point* and *disgust* rules, proposed by Cooper [6] and Kraft and Lee [15] respectively. These strategies were selected as they had previously been shown to provide good approximations of actual searcher behaviours [20]. We explored a range of values for x_1, x_2 and x_3, trialling 1–10 in steps of 1, and 15–30 in steps of 5.

4.4 User Comparisons

Comparison runs used the same configurations as described previously, save for the querying strategy. Here, we *replayed* each of the queries issued by the real-world subjects from the associated user study.[9] Each query was issued over the different configurations, allowing us to then calculate the simulated click depths per query. We then compared the actual click depths against the simulated click depths for each query, calculating the *Mean Squared Error (MSE)* between the two. For this analysis, we used a total of 175 user queries.

5 Results

Here, the performance of simulated users is reported both in terms of *Cumulative Gain (CG)* and the click depth reached per query *(D/Q)*. CG is measured by summing the TREC QRELs judgement scores of all documents marked as relevant over the course of a search session.

RQ1 ***(Examining Performance)*** Figure 2 shows three plots, each of which illustrates the maximum levels of CG attained by simulated users at varying *D/Q* values. These are shown over the different SERP level stopping strategies: **Always** (baseline); **Perfect**; **Average**; **Savvy**; and **Naïve**, over the three snippet level stopping strategies **SS1** (left), **SS2** (centre) and **SS3** (right). From the plots, we can immediately see that compared to our baseline approach **Always**, **Perfect** attained a much higher level of CG (e.g. 2.55 for **Perfect** at a D/Q

[9] This also meant that for our comparison runs, only the four topics selected by Maxwell et al. [21] were trialled, rather than the full set of 50 topics from the TREC 2005 Robust Track as used in our performance runs.

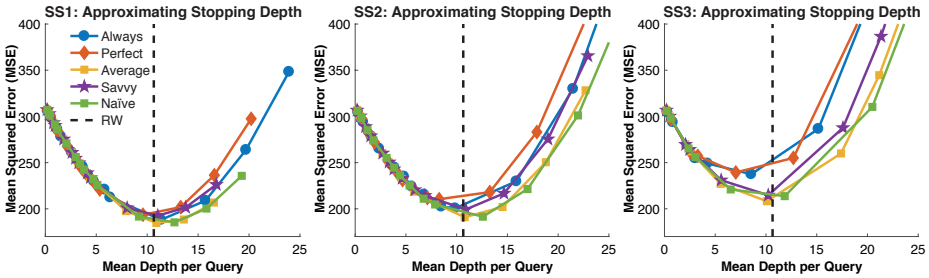

Fig. 3. Plots illustrating the results of our comparison runs over each SERP stopping strategy and snippet level stopping strategies **SS1** (L), **SS2** (C) and **SS3** (R). Also included for comparison is the mean click depth exhibited by the real world subjects.

of 5.77 vs. 1.2 for **Always** at a D/Q of 7.39 over **SS1**). Turning to our three stochastic variants, the **Savvy** searcher always abandoned a low scent SERP and examined a high scent SERP about 80% of the time. This led to a general trend similar to that of **Perfect**, yet with a slightly lower maximum level of CG (e.g. 2.41 at a D/Q of 4.78 over **SS1**). This is in line with intuition, as 20% of the time, the **Savvy** user would have abandoned a high scent SERP, accounting for the slightly lower levels of performance. On the other hand, the **Naïve** searcher followed a similar trend to our baseline approach, **Always**. This is again in line with expectations, as the probabilities used by **Naïve** led to a high probability of examining high *and* low scent SERPs. In turn, this led to an inefficient search strategy (like **Always**) – one where searchers would by and large waste time examining low scent SERPs. The final **Average** searcher however fell between the extremes of **Savvy** and **Naïve**, and attained a maximum CG of 1.72 at a D/Q of 9.32 over **SS1**. Similar trends as discussed previously can be seen across all three snippet level stopping strategies **SS1**, **SS2**, and **SS3** – although for **SS3**, simulated users on average examined to slightly greater depths per query. Overall, the highest CG was attained by **Perfect** over **SS3**, with the lowest CG of 1.18 reached by **Naïve** – baseline **Always** was close with a CG of 1.2. Overall, the **Savvy**, **Average** and **Naïve** searchers tended to outperform the **Always** baseline, and suggests that performance improvements can be made to varying degrees depending upon how well the searcher can identify good quality SERPs. Interestingly, searchers need not be **Perfect**, with **Average** searchers still performing much better than **Always**. These findings show that including the SERP level decision point does indeed lead to improvements in performance.

RQ2 ***(Approximating Stopping Depths)*** To determine whether including the SERP decision point could lead to better approximations of stopping behaviour, we calculated and plotted the MSE for each SERP and snippet level stopping strategy (see Figure 3). Again, **SS1** is shown on the left, with **SS2** in the centre and **SS3** on the right. Also included in each of the plots – denoted by the black dashed line – is the actual mean click depth that the 53 subjects of the user study examined to – a depth computed across all issued queries as 10.65. From the plots, we can immediately observe that the lowest (and there-

Table 2. Tables showing the lowest MSE (**MSE**) approximations attained over each SERP stopping strategy (***Strat.***) and snippet level stopping strategy **SS1** (L), **SS2** (C) and **SS3** (R). Also included are the associated threshold values ($\boldsymbol{x_n}$) and mean depths per query (**D/Q**) at which the lowest MSE values were reached at.

Strat.	$\boldsymbol{x_1}$	**D/Q**	**MSE**
Always	15	11.37	188.12
Perfect	15	9.59	193.57
Avg.	**20**	**10.90**	**184.66**
Savvy	20	11.05	192.53
Naïve	20	12.68	185.36

(a) **SS1**

Strat.	$\boldsymbol{x_2}$	**D/Q**	**MSE**
Always	10	9.85	200.62
Perfect	10	8.28	210.39
Avg.	**15**	**10.79**	**190.91**
Savvy	15	10.90	199.32
Naïve	15	12.58	191.47

(b) **SS2**

Strat.	$\boldsymbol{x_3}$	**D/Q**	**MSE**
Always	5	8.55	237.45
Perfect	5	7.03	239.36
Avg.	**6**	**10.09**	**208.23**
Savvy	6	10.22	214.67
Naïve	6	11.85	213.71

(c) **SS3**

fore best) MSE values were found to be close to the real mean click depth for both **SS1** and **SS2**, but the approximations offered by **SS3** were slightly further away, with the best approximation for **SS3** yielding a D/Q of 10.09. The best MSE approximations – and the corresponding x_n threshold and D/Q that it was attained at – can be seen in Tables 2(a), (b), (c) for **SS1**, **SS2** and **SS3** respectively. Closer examination of the tables show that the best approximation over **SS1** was achieved at a D/Q of 10.90 ($x_1 = 20$) for **Naïve**, with a D/Q of 10.79 ($x_2 = 15$) for **Average**. Indeed, the stochastic users gave the best approximations over all three snippet level stopping strategies. This finding is intuitive as nobody from the user study correctly identified high and low scent SERPs 100% of the time, making **Perfect** an unrealistic strategy to use. Interestingly, stopping behaviour was best approximated by **Average** searchers.

Closer inspection of the results for **SS3** shows that this snippet level stopping strategy consistently yielded higher (and thus poorer) MSE values, although D/Q approximations remained close to the actual mean click depth – at least for the stochastic users, **Average**, **Perfect** and **Naïve**. This finding is interesting because the same strategy yielded the best approximations for searcher behaviour in previous work [20], and suggests the strategy may not be robust when applied in other contexts. Overall, the actual stopping behaviour of searchers *is* better approximated when incorporating a SERP level decision point.

6 Discussion and Future Work

In this paper, we have considered how information scent affects search and stopping behaviour, and have encoded this within the *Complex Searcher Model (CSM)* to provide a more realistic model of the search process. This was operationalised by the inclusion of a new SERP level decision point, where the *scent* of a SERP is attained from an initial impression of the page. This information is then used to decide if individual snippets should be examined, or whether to simply abandon the SERP. We found that the inclusion of this additional decision point can lead to more effective searching, but only if the searcher is able to discern between SERPs of high and low scent. Our study shows that **Savvy** users

can easily avoid poor quality SERPs, while **Naïve** users find it hard to recognise the quality of SERPs. This suggests that work should be directed towards improving how SERPs are rendered to increase how well people can identify good SERPs from the bad, as well as research into what cues searchers look for in a good SERP. Furthermore, we found that including the SERP level decision point led to more accurate modelling of actual stopping behaviour. This represents a major shift in modelling interaction – and has ramifications for how IR systems are measured, which typically assumes people examine ranked lists. These results suggest that future work needs to be directed towards measures that consider abandonment, and should also include how the sequence and quality of queries affects interactions taking place with ranked lists.

Acknowledgements Our thanks to Horaţiu Bota and Alastair Maxwell for their feedback – including Horaţiu's helpful comments on our results. We would also like to thank the anonymous reviewers for their comments and feedback. Finally, the lead author is funded by the UK Government though the EPSRC, grant number `1367507`.

References

1. Bailey, P., Moffat, A., Scholer, F., Thomas, P.: User variability and IR system evaluation. In: Proc. 38th ACM SIGIR. pp. 625–634 (2015)
2. Baskaya, F., Keskustalo, H., Järvelin, K.: Modeling behavioral factors in interactive info. retrieval. In: Proc. 22nd ACM CIKM. pp. 2297–2302 (2013)
3. Browne, G., Pitts, M., Wetherbe, J.: Stopping rule use during web-based search. In: Proc. HICSS-38. p. 271b (2005)
4. Card, S., Pirolli, P., Van Der Wege, M., Morrison, J., Reeder, R., Schraedley, P., Boshart, J.: Info. scent as a driver of web behavior graphs: Results of a protocol analysis method for web usability. In: Proc. 19th ACM CHI. pp. 498–505 (2001)
5. Chi, E., Pirolli, P., Chen, K., Pitkow, J.: Using info. scent to model user info. needs and actions and the web. In: Proc. 19th ACM CHI. pp. 490–497 (2001)
6. Cooper, W.: On selecting a measure of retrieval effectiveness part II. Implementation of the philosophy. J. of the American Soc. for Info. Sci. 24(6), 413–424 (1973)
7. Diriye, A., White, R., Buscher, G., Dumais, S.: Leaving so soon?: Understanding and predicting web search abandonment rationales. In: Proc. 21st ACM CIKM. pp. 1025–1034 (2012)
8. Hagen, M., Michel, M., Stein, B.: Simulating ideal and average users. In: Proc. 12th AIRS. pp. 138–154 (2016)
9. Hassan, A., White, R.: Personalized models of search satisfaction. In: Proc. 22nd ACM CIKM. pp. 2009–2018 (2013)
10. Hassan, A., Shi, X., Craswell, N., Ramsey, B.: Beyond clicks: Query reformulation as a predictor of search satisfaction. In: Proc. 22nd CIKM. pp. 2019–2028 (2013)
11. Hearst, M.: Search user interfaces. Cambridge University Press (2009)
12. Ingwersen, P., Järvelin, K.: The Turn: Integration of Info. Seeking and Retrieval in Context. Springer (2005)
13. Keskustalo, H., Järvelin, K., Pirkola, A., Sharma, T., Lykke, M.: Test collection-based IR evaluation needs extension toward sessions — A case of extremely short queries. In: Proc. 5th AIRS. pp. 63–74 (2009)
14. Kiseleva, J., Kamps, J., Nikulin, V., Makarov, N.: Behavioral dynamics from the SERP's perspective: What are failed SERPs and how to fix them? In: Proc. 24th ACM CIKM. pp. 1561–1570 (2015)

15. Kraft, D., Lee, T.: Stopping rules and their effect on expected search length. IPM 15(1), 47 – 58 (1979)
16. Loumakis, F., Stumpf, S., Grayson, D.: This image smells good: Effects of image info. scent in search engine results pages. In: Proc. 20th ACM CIKM. pp. 475–484 (2011)
17. Marshall, C.C., Shipman (III), F.M.: Spatial hypertext and the practice of information triage. In: Proc. 8th ACM HYPERTEXT. pp. 124–133 (1997)
18. Maxwell, D., Azzopardi, L.: Agents, simulated users and humans: An analysis of performance and behaviour. In: Proc. 25th ACM CIKM. pp. 731–740 (2016)
19. Maxwell, D., Azzopardi, L., Järvelin, K., Keskustalo, H.: An initial investigation into fixed and adaptive stopping strategies. In: Proc. 38th ACM SIGIR. pp. 903–906 (2015)
20. Maxwell, D., Azzopardi, L., Järvelin, K., Keskustalo, H.: Searching and stopping: An analysis of stopping rules and strategies. In: Proc. 24th ACM CIKM. pp. 313–322 (2015)
21. Maxwell, D., Azzopardi, L., Moshfeghi, Y.: A study of snippet length and informativeness: Behaviour, performance and UX. In: Proc. 40th ACM SIGIR (2017)
22. McNair, J.N.: Optimal giving-up times and the marginal value theorem. The American Naturalist 119(4), 511–529 (1982)
23. Moffat, A., Bailey, P., Scholer, F., Thomas, P.: INST: An adaptive metric for IR evaluation. In: Proc. 20th ADCS. pp. 5:1–5:4 (2015)
24. Moffat, A., Zobel, J.: Rank-biased precision for measurement of retrieval effectiveness. ACM Trans. on Info. Systems 27(1), 2:1–2:27 (2008)
25. Nickles, K.: Judgment-based and reasoning-based stopping rules in decision making under uncertainty. Ph.D. thesis, University of Minnesota (1995)
26. Olston, C., Chi, E.: Scenttrails: Integrating browsing and searching on the web. ACM Trans. Comput.-Hum. Interact. 10(3) (2003)
27. Ong, K., Järvelin, K., Sanderson, M., Scholer, F.: Using info. scent to understand mobile and desktop web search behavior. In: Proc. 40th ACM SIGIR (2017)
28. Pirolli, P.: Info. Foraging Theory: Adaptive interaction with info. 1 edn. (2007)
29. Pirolli, P., Card, S.: Info. foraging in information access environments. In: Proc. 13th ACM SIGCHI. pp. 51–58 (1995)
30. Pirolli, P., Card, S.: Info. foraging. Psychological Review 106, 643–675 (1999)
31. Smucker, M., Clarke, C.: Modeling optimal switching behavior. In: Proc. 1st ACM CHIIR. pp. 317–320 (2016)
32. Stephens, D., Krebs, J.: Foraging Theory (1986)
33. Sundar, S., Knobloch-Westerwick, S., Hastall, M.: News cues: Info. scent and cognitive heuristics. J. Am. Soc. Inf. Sci. Technol. 58(3), 366–378 (2007)
34. Thomas, P., Moffat, A., Bailey, P., Scholer, F.: Modeling decision points in user search behavior. In: Proc. 5th IIiX. pp. 239–242 (2014)
35. Tombros, A., Sanderson, M.: Advantages of query biased summaries in info. retrieval. In: Proc. 21st ACM SIGIR. pp. 2–10 (1998)
36. Wu, W.: How far will you go?: Using need for closure and information scent to model search stopping behavior. In: Proc. 4th IIiX. pp. 328–328 (2012)
37. Wu, W., Kelly, D., Sud, A.: Using info. scent and need for cognition to understand online search behavior. In: Proc. 37th ACM SIGIR. pp. 557–566 (2014)
38. Zach, L.: When is "enough" enough? modeling the info.-seeking and stopping behavior of senior arts administrators. J. American Soc. for Info. Sci. and Tech. 56(1), 23–35 (2005)