Searching and Stopping: An Analysis of Stopping Rules and Strategies

David Maxwell and Leif Azzopardi
School of Computing Science
University of Glasgow
Glasgow, Scotland
d.maxwell.1@research.gla.ac.uk
Leif.Azzopardi@Glasgow.ac.uk

Kalervo Järvelin and Heikki Keskustalo
School of Information Sciences
University of Tampere
Tampere, Finland
kalervo.jarvelin@uta.fi
heikki.keskustalo@uta.fi

ABSTRACT

Searching naturally involves stopping points, both at a query level *(how far down the ranked list should I go?)* and at a session level *(how many queries should I issue?)*. Understanding when searchers stop has been of much interest to the community because it is fundamental to how we evaluate search behaviour and performance. Research has shown that searchers find it difficult to formalise stopping criteria, and typically resort to their intuition of what is *"good enough"*. While various heuristics and stopping criteria have been proposed, little work has investigated how well they perform, and whether searchers actually conform to any of these rules. In this paper, we undertake the first large scale study of stopping rules, investigating how they influence overall session performance, and which rules best match actual stopping behaviour. Our work is focused on stopping at the query level in the context of ad-hoc topic retrieval, where searchers undertake search tasks within a fixed time period. We show that stopping strategies based upon the *disgust* or *frustration point* rules - both of which capture a searcher's tolerance to non-relevance - typically result in *(i)* the best overall performance, and *(ii)* provide the closest approximation to actual searcher behaviour, although a fixed depth approach also performs remarkably well. Findings from this study have implications regarding how we build measures, and how we conduct simulations of search behaviours.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval: Search Process; H.3.4 [**Information Storage and Retrieval**]: Systems and Software: Performance Evaluation

General Terms

Theory, Experimentation, Simulation, Human Factors

Keywords

Retrieval Strategies, Search Behavior, Evaluation

1. INTRODUCTION

Interactive Information Retrieval (IIR) is a non-trivial process where searchers issue numerous queries and examine a varying number of snippets and documents per query [14]. During the search process, searchers need to decide when they should abandon the current query (and thus issue a new query after examining the current results list), and when to curtail their search (stopping the search session altogether). Knowing when to stop is considered a fundamental aspect of human behaviour [26]. Stop too early, and important information may be missed. Stop too late, and time and effort is wasted. Worse still, the examination of fruitless result lists will mean not having time to examine other lists which may potentially contain greater yields for the searcher.

While important, research into stopping behaviours when searching has been relatively sparse. Of the studies undertaken in this area, many have concluded that the decision to stop is based upon a searcher's intuition - or the feeling that what they have found is *"good enough"* [36]. However, past research has revealed that certain factors impact upon stopping behaviour. Examples include a searcher's prior knowledge of the domain being searched [33], their knowledge and search experience, how they interpret the cues/snippets presented on the *Search Engine Results Page (SERP)* [35], their understanding of the given information need, and their feeling of *enough*. To quantify this feeling of *"good enough"*, various researchers have proposed an array of stopping rules and heuristics that try to encode this intuition [6, 9, 19, 26]. These proposals are detailed in Section 2.

In this paper, we aim to catalogue and evaluate a range of proposed stopping rules. Specifically, we focus on rules regarding when a searcher should stop examining the results for the current query, and move onto the next query (i.e. query stopping, rather than session stopping). Once catalogued, we operationalise and implement these rules. We then perform a large scale simulated analysis over two ad-hoc *TREC* test collections, exploring:

RQ1 how performance varies due to each stopping rule; and

RQ2 which stopping rule delivers the best overall performance.

We then conduct an analysis comparing the search stopping behaviour of 48 searchers to the set of stopping rules defined, exploring:

RQ3 which stopping rule(s) best reflect actual stopping behaviour; and

RQ4 whether different stopping rule(s) would lead to improved performance.

This work represents one of the first attempts to enumerate and encode the various stopping rules, and empirically test their validity in the context of ad-hoc topic retrieval. Our findings show that stopping strategies based upon the *frustration point rule* [9] and *disgust rule* [19], where searchers stop after examining a given number of non-relevant documents, delivers the best overall performance during simulations. These stopping strategies also best approximated when actual searchers stopped. However, a fixed depth stopping strategy - where searchers stopped after examining 13 documents - also provided a good approximation of actual stopping behaviour. These findings suggest that simulations can be improved upon and made more realistic if they encode a stopping strategy based upon the frustration point and disgust rules. However, our findings also suggest that measures and models using a fixed depth stopping strategy provide a good approximation on average.

Next, we provide an overview of the background related to stopping rules before operationalising these rules into a number of strategies that we can implement and evaluate in Section 3. Our methodology, which consists of both a simulated comparison - and a comparison with data taken from a previous user study - is then described in Section 4. Results follow in Section 5. A summary of our findings and directions for future work then concludes this paper.

2. RELATED WORK

In *Information Retrieval (IR)*, research into stopping rules and stopping behaviour has been approached from two main perspectives. *User-sided research* focuses on how and why searchers stop, and often formulate motivations, heuristics and guidelines regarding stopping behaviour [11, 28, 31, 34, 35]. Additionally, *system-sided research* focuses on encoding stopping models into measures and models to evaluate IR system performance [9, 15, 24], and/or conduct simulations of user interaction [1, 4]. However, when people stop is considered a fundamental aspect of human behaviour [26], and as such has been considered more broadly from cognitive and decision-making perspectives [7, 19]. Determining when to stop is an inherently difficult task. This is due to the fact that the decision is usually instrumented by a series of internal factors of the decision maker's thinking [26]. This therefore makes the concept of stopping an extremely difficult phenomenon to model effectively.

Wu et al. [35] explain that a majority of studies examining stopping behaviours have been conducted with a series of interviews. This was to obtain an understanding of why people decide to stop. Many of these studies [11, 28, 31, 36] have often concluded that the decision to stop searching is based on intuition, or the *"feeling of good enough"* (termed *satisficing*) [36]. A small number of information-seeking and retrieval studies have also shown that stopping decisions are dependent upon the task type being undertaken, the time constraints imposed upon them, and a range of more specific, internally constructed stopping rules [28, 36]. For example, Prabha et al. [28] found that time constraints reduced how many documents were examined, whereas Zach [36] found that decision makers often stopped their searching earlier because they were satisfied with what they had found, even though they believed more information was available.

As highlighted by Marchionini [21], these internal stopping rules are dependent upon the searcher's task domain knowledge and information-seeking ability. Wu and Kelly [34] more recently undertook a study where subjects performed a series of search tasks. Subjects were then interviewed about their query stopping and task stopping behaviours. Results from this study showed that query stopping decisions were taken primarily on the face of search results, queries and search tasks. Task stopping decisions were determined by the subject's overall goal for each task, the content examined (and their subjective perceptions on the examined content), and the study constraints imposed upon them (e.g. imposed time constraints, and the search interface used).

Despite the difficulties associated with the modelling of stopping behaviours, researchers have proposed numerous stopping rules which are *believed* to quantify/explain when searchers stop searching. Proposed rules tend to be high-level descriptions of when a search should be stopped, and may apply to stopping at the query level, the session/task level, or both. Two of the earliest stopping rules proposed were devised by Cooper [9], who presented:

- the frustration point rule, where a searcher stops after examining a certain number of non-relevant documents; and
- the *satisfaction stopping rule*, where searchers would stop only when a certain number of relevant documents were found.

Kraft and Lee [19] also presented three stopping rules which were devised for measuring the expected *search length* (the number of snippets scanned in a ranked list). The *satiation rule* borrowed the underlying theory of the satisfaction stopping rule [9], where searchers would stop after being *satiated* by finding a desired number of relevant documents. In addition, the disgust rule borrowed the underlying theory from the frustration point rule of Cooper [9]. The *combination rule* is the final rule introduced by Kraft and Lee [19]. This rule combines both the satiation and disgust rules into one. Here, searchers would stop if they are satisfied with what has been found, or disgusted by having examined too many non-relevant documents - whichever condition is met first. Kraft and Lee [19] further demonstrated that the expected search length could be approximated using each of their stopping rules by considering the size of the retrieval set, the number of relevant documents a searcher wished to obtain, and the number of non-relevant documents a searcher would be willing to tolerate. The number of documents required to consider a search successful is dependent upon whether the search task is high-precision (stopping comparatively early) or high-recall (stopping comparatively later), as hypothesised by Bates [5].

In addition to the rules posited by Cooper [9] and Kraft and Lee [19], Nickles [26] proposed four cognitive rules investigating the sufficiency of information:

- the *mental list rule*, where searchers construct a mental list of criteria about a given item (such as the number of seats and engine displacement in a car) that must be satisfied before stopping;
- the *representational stability rule*, where a searcher continues to examine information until the underlying mental model that they possess of the topic in question begins to stabilise;

- the *difference threshold rule*, where a searcher sets an *a priori* difference level to gauge when he or she is not learning anything new; and
- the *magnitude threshold rule*, where a searcher has a cumulative amount of information that must be found before he or she stops searching. In this rule, the focus is attaining 'enough' information.

A further fifth rule, the *single criterion rule*, was proposed by Browne et al. [6]. Here, searchers would examine documents for information on a single criterion, stopping when enough information had been collected about said criterion.

A decision-theoretic approach using utility theory was proposed by Cooper [10]. This approach posited that searchers will stop examining documents once the effort of conducting a new search outweighs the benefit of any new information that may be obtained. Similarly, the economic model of search proposed by Azzopardi [1] suggests that a searcher should stop when the marginal gain equals the marginal cost. Related to these approaches is the *patch model* from *Information Foraging Theory (IFT)* [27] to predict when a searcher should stop examining a ranked list. Applying *Charnov's Marginal Value Theorem* [8] to search, Pirolli and Card's theory suggests that a searcher should stop when the rate of gain from a given result falls below the average rate of gain experienced by the searcher [27].

In previous work, we performed an initial analysis comparing a fixed depth stopping strategy against two implementations of the frustration point/disgust rules [23]. The fixed depth stopping strategy assumed that the searcher would stop at a particular document n in the ranked list, regardless of whether previous documents were considered relevant or non-relevant. This strategy assumed the same stopping model as encoded in many IR evaluation measures, such as *Precision-at-k (P@k)* and *Normalised Discounted Cumulative Gain (NDCG)* [15]. Our implementation of the frustration point rule consisted of counting the number of non-relevant documents seen in the ranked list at position k. If the total number of non-relevant documents exceeded a given threshold, the searcher would then stop. A similar implementation of this rule was also examined by Lin and Smucker [20], but in the context of document browsing. Our second implementation of the frustration point rule considered that if the total number of non-relevant documents observed *contiguously* exceeded some threshold, then the searcher would stop. The initial findings from the simulation we conducted showed that our implementations of the frustration point rule resulted in higher levels of gain per second across a number of different querying strategies. In this paper, we go beyond our initial study. We implement more sophisticated stopping rules and evaluate them across two TREC test collections, before comparing the rules to actual observed searcher behaviour.

As previously mentioned, retrieval measures also encode stopping models that dictate the likelihood of a searcher stopping at a particular rank [25]. For example, $P@k$ assumes that a searcher will stop at rank k. However, more sophisticated measures exist, such as *Rank-Biased Precision* [24]. This measure has a patience/persistence parameter that determines whether a searcher is more likely to go further down a ranked list (if they are patient), or less likely (if they are impatient). Indeed, Moffat et al. [25] argue that all measures have an implicit stopping model, and show how the stopping model used can be derived from a number of common measures. Consequently, forming a better understanding of how and why searchers stop is quintessential to the development of IR measures.

In terms of modelling search and stopping behaviours, numerous studies have been performed that simulate session-based IIR. Studies examining click models implicitly employ the use of stopping probabilities which determine the likelihood of how far down a searcher will examine a ranked list of results [12, 13, 32]. However, in many other studies, it has been assumed that for a given query, a searcher will go to a prescribed and fixed depth for each query issued during the session [1, 3, 17, 18]. This is clearly a major limitation as actual searcher behaviour varies according to a number of factors. Consequently, finding and selecting a stopping rule that is more in line with the actual stopping behaviour of searchers will result in more realistic simulations.

3. STOPPING RULES

In the previous section, we outlined a number of stopping rules defined in the literature. In this section, we take a subset of these rules and propose a number of ways in which we can operationalise them, such that we can subsequently implement them. The implemented stopping strategies fit into the wider simulated searcher model, illustrated in Figure 1 (where key decision points are denoted with a `?` marker). The model is explained in detail in Section 4.2.

The first rule (and baseline) for this study is the fixed depth stopping rule.

SS1 ***(Fixed Depth)*** Under this stopping strategy, the simulated searcher will stop once they have observed x_1 snippets, regardless of their relevance to the given topic.

This represents the typical approach taken in most simulations, and represents the basic stopping searcher model in $P@k$ measures. On average, this rule makes sense, but when result lists are of varying quality, it is not a very sensible approach. For example, consider the situation where a searcher submits a poor query that returns no relevant information. Naïvely examining ten snippets (where $x_1 = 10$) - and potentially some documents along the way - is by and large a waste of the searcher's time.

Thus, we derive two further stopping strategies and their subsequent variants - ***SS2*** and ***SS3*** - based upon the frustration point and disgust stopping rules, as defined by Cooper [9] and Kraft and Lee [19].

SS2 ***(Total Non-Relevant)*** Under this stopping strategy, the searcher will stop once they have observed x_2 non-relevant snippets. If a snippet has been previously seen in the search session and was considered non-relevant, it is included in the count.

SS3 ***(Contiguous Non-Relevant)*** Similar to ***SS2*** above, the searcher will stop once they have observed x_3 non-relevant snippets in a row (contiguously). As with ***SS2***, previously seen non-relevant snippets within the search session are included in the count.

Intuitively, these strategies will mean that the simulated searcher adapts their interaction with a ranked list of results depending upon the performance of the underlying query. For example, a ranked list is judged by a simulated searcher

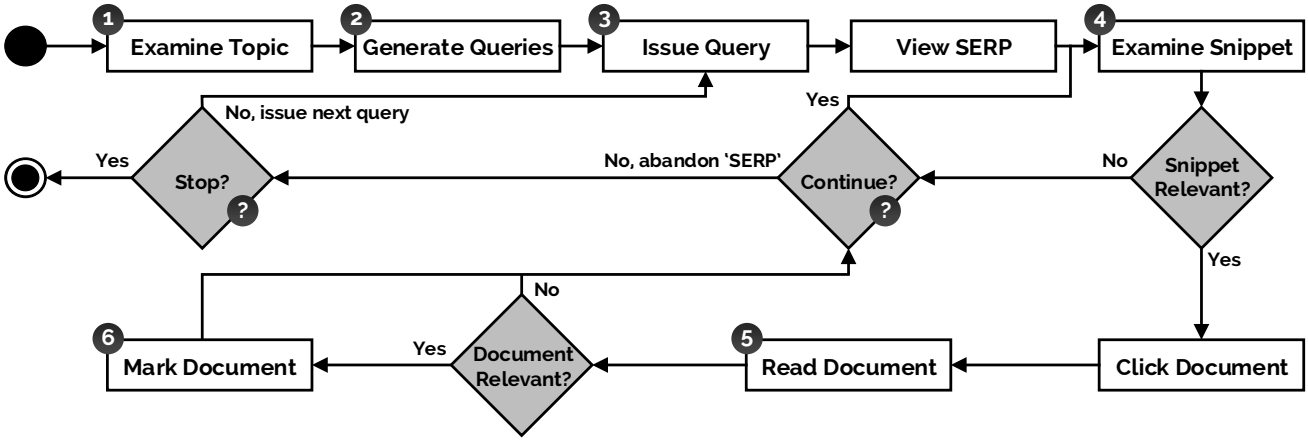

Figure 1: A flowchart of decisions (shown in grey) and tasks (shown in white) undertaken by simulated searchers 'participating' in this study. The model is adapted from Baskaya et al. [4] and Thomas et al. [30]. Numbers associated with tasks correspond to the steps detailed in Section 4.2. The two *?* markers indicate the decisions that are encoded within each implemented stopping strategy (refer to Section 4.5).

as `[R,N,N,R,N,N,N]`, where `R` and `N`[1] denote relevant and non-relevant items respectively. Under ***SS2***, with $x_2 = x_3 = 3$, the simulated searcher would stop at rank five. Under ***SS3***, the simulated searcher would stop at rank seven.

Since documents are examined when the associated snippet is considered relevant (see Figure 1), a simulated searcher may revise their opinion of the amount of relevant information observed once they see the document. A simulated searcher may for example inspect the first snippet thinking it is relevant, and then examine the document. The document is subsequently considered non-relevant, meaning the initial `R` is changed to a `N`. This updates the count, so that under a stopping strategy with *revised relevance*, simulated searchers would stop at rank three in the list shown above using ***SS2***, and similarly for ***SS3***. This therefore introduces two variants of ***SS2*** and ***SS3***, which were found in a pilot study to perform slightly better. Therefore, we only report the revised relevance variants of ***SS2*** and ***SS3*** in this paper.

The next two stopping strategies are based upon the difference threshold rule [26]. To operationalise this rule, we consider the difference between the text of the current snippet and the text of previously examined snippets. Here, the idea is that as simulated searchers examine snippets, they may encounter a snippet that is not sufficiently different from what they already have observed, meaning that they are unlikely to find new information. The searcher therefore stops and issues a new query. From this rule, we devised two separate stopping strategies where we computed the difference based upon *term overlap* and *KL-Divergence* scores.

SS4 (Term Overlap Difference) This stopping strategy compares the occurrences of terms in a given snippet against all terms in previously examined snippets. The more terms that overlap, the greater the chance that the new snippet does not contain any new information. If $\frac{|s_{curr} \cap s_{prev}|}{|s_{curr}|} > x_4$, the new snippet is considered too similar to previously examined content. The simulated searcher then moves to the next query. Here, s_{curr} denotes the terms of the current snippet, s_{prev} denotes terms from all previously observed snippets, and x_4 is the threshold at which the simulated searcher stops.

SS5 (KL-Divergence Difference) This stopping strategy compares a given snippet against previously seen snippets using the KL-Divergence measure. If the resulting value is less than threshold x_5, then the snippet is considered too similar to previously seen content, and the simulated searcher stops, moving to the next query.

When implementing ***SS4*** and ***SS5***, we considered the *per-query difference* and the *per-session difference*. For the per-query variant, previously observed text consisted of the first snippet, thus meaning that the simulated searcher always considers at least two snippets before stopping. For the per-session variant, all previously seen snippets over the simulated search session are used. In this paper, we will only report the per-query variants of ***SS4*** and ***SS5***, as both performed somewhat better than their per-session variants in a pilot study. A number of other variants were also considered but not explored, such as using the document and snippet text, and using only text from snippets considered relevant. To compute the KL-Divergence, we used a *Maximum Likelihood Estimate (MLE)* of the term distribution given the new snippet, and all the previously examined snippets. We also explored smoothing the distribution with the probabilities of each collection used. However, this approach was not used; performance was not increased, only complexity.

The final stopping strategy that we implemented was based upon IFT to represent utility-based approaches.

SS6 (IFT) With this stopping strategy, a searcher is assumed to have some idea of the average rate of gain (denoted as x_6). If the rate of gain from the observed documents thus far does not exceed x_6, the searcher then stops and proceeds to issue the next query.

To determine the rate of gain at the current snippet, we first computed the *Discounted Cumulative Gain (DCG)* g received from the observed documents up to that point in the ranked list at position i. We then divided g by the total time taken, i.e. $i * t_d + t_q$, where i represents the rank,

[1]NB: Judgements `R` and `N` do not (necessarily) represent TREC relevance judgements. Instead, they represent the decision taken by the simulated searcher as to whether snippets are relevant (or non-relevant) to the given topic.

t_d is the time taken to examine a document, and t_q is the time taken to issue a query. This estimate is very dependent upon the first document. For example, if the first document is non-relevant, then the gain is zero, and thus the simulated searcher would immediately stop when $x_6 > 0$. We also included another parameter which specifies how many snippets they should first consider before making their decision based on the rate of gain[2]. This would essentially mean that the simulated searcher would look at y_6 snippets/documents, and then decide to continue with the current query.

Stopping strategies ***SS1***-***SS6*** represent a subset of the rules that we have catalogued in Section 2. It should be noted that we have not selected rules which are based upon satisfaction or satiation because the task we are investigating is ad-hoc topic retrieval, where the goal is to find as many relevant documents as possible in a given period of time. The satisfaction/satiation rules therefore do not seem particularly applicable in this context. Furthermore, rules such as the mental list rule seem to be more topic specific, requiring a searcher to know in advance all the criteria that they need to check off in their head *a priori*. However, these criteria are largely unknown in ad-hoc circumstances.

4. EXPERIMENTAL METHOD

To address research questions **RQ1**-**RQ4** as posed in Section 1, we conducted a large scale simulation to compare a total of six stopping strategies based upon four stopping rules. Each stopping strategy was then compared against the actual stopping behaviour of 48 subjects, all of whom took part in our previously conducted user study [22].

4.1 Corpora, Topics, System and Users

We used two test collections: TREC *AQUAINT* with the TREC *2005 Robust Track* topic set, and TREC *WT2g* with the TREC *Ad-Hoc and Small Web* topic set. The topic sets contain a total of 50 topics each, all of which were used for our stopping strategy comparisons. Both collections were indexed using the *Whoosh* IR toolkit[3], where stopwords[4] were removed with Porter stemming applied. The retrieval model used for all simulations was PL2 ($c = 10.0$).

In the user study we performed [22], 48 undergraduate subjects were recruited to undertake ad-hoc topic retrieval over two 2005 Robust Track topics, №. 347 and №. 435. For each topic, subjects had a total of 1200 seconds (20 minutes) to complete each task using the same system and setup as defined above. For each query submitted, we recorded various interactions that subjects undertook. Of particular interest for this study is the data that report the documents that each subject clicked on, what documents they marked as relevant, the depth to which they hovered on each SERP (inferred from mouseover events), and the depth of the last document that they clicked. This data were used to first ground our simulations (refer to Section 4.3), and then to evaluate the different stopping strategies against each other.

4.2 Implemented Searcher Model

Figure 1 presents a flowchart of the operations that simulated searchers undertook during a simulated search session. The processes are largely based upon the simulation framework used by Baskaya et al. [4], but includes additional decision points as suggested by Thomas et al. [30].

Essentially, the simulated searcher begins by ***(1)*** examining the given topic and title description. From the title and description, the simulated searcher then ***(2)*** generates a series of queries[5] which are issued to the underlying search engine. The simulated searcher then ***(3)*** issues a query from the generated list, and then ***(4)*** proceeds to examine the first/next snippet in the ranked list provided. The simulated searcher can also decide to issue a new query, thus returning to *(3)*. If the snippet is considered relevant by the simulated searcher, ***(5)*** the document is then examined in full. If the document is also considered relevant, ***(6)*** the document is then marked relevant. If either the snippet or document are considered non-relevant, the simulated searcher then returns to *(4)* with the document unmarked. The gain for each marked document is determined from TREC relevance judgements. A marked document can score either *0* (non-relevant), *1* (somewhat relevant), or *2* (very relevant).

For the purposes of this study, we assume that *(i)* there is only one 'SERP', consisting of a maximum of 75 results (thus meaning no pagination and associated interaction costs), and *(ii)* the results in the 'SERP' are examined using the *cascade* assumption [16], meaning examination of a results list top-down in sequential order.

The searcher model detailed above is encoded within our simulator. It was developed as a highly modularised framework to allow for the easy experimentation of different simulated searchers, test collections, querying strategies, stopping strategies, and more[6]. Below, we describe how the components in the framework were instantiated.

4.3 Interaction Times and Probabilities

To ground our simulations, we computed the interaction probabilities and the times it took subjects to perform various actions when undertaking our user study [22]. Average times for each action simulated are reported in Table 1. For each action that a simulated searcher performed during a simulated run, the time spent was accumulated given the times reported in Table 1 until the time limit of 1200 seconds was exceeded. After reaching this limit, the simulated search session ended. The interaction probabilities, associated with clicking snippets and marking documents when TREC relevant/non-relevant, are reported in Table 2.

Since the simulation is stochastic based upon the interaction probabilities, this means that for different runs (i.e. stopping strategies and different thresholds), the same document can be considered relevant in one run and non-relevant in another. To ensure a fair comparison, we used a set of pre-rolled outcomes. For each snippet/document, we pre-computed the probabilities, and stored these in *action judgement files* - one related to the action of clicking on a snippet, another related to the action of marking a document. If a document was judged as relevant in one run, then by pre-computing these actions in advance, the same document would be considered relevant in other runs. We repeated this process ten times, so that for each particular stopping strategy and threshold value, we performed ten trials (in which documents could be marked as either relevant or non-

[2]This parameter was set to 2 for this study - refer to Section 4.5.

[3]Whoosh is available at `https://pypi.python.org/pypi/Whoosh`.

[4]Fox's classical stopword list was used. Refer to `http://git.io/vT3So` for the complete list.

[5]Refer to Section 4.4 for more information on the query generation strategy used for this study.

[6]The simulator is open-source, and freely available for download from *GitHub* at `http://git.io/vOBLz`.

Table 1: A summary of the interaction times (in seconds) used for all simulations in this study. Values are obtained from the user study conducted by Maxwell and Azzopardi [22].

Time Required to...	Seconds
...issue a query	15.1
...undertake an initial SERP examination	1.1
...examine an individual snippet	1.3
...examine a document	21.45
...mark a document as relevant	2.57

Table 2: The four probabilities used in all simulations run as part of this study. Probabilities are derived from the user study conducted by Maxwell and Azzopardi [22].

Probability	Value
$P(C\|R_s)$ (Clicking a relevant snippet)	0.36
$P(C\|N_s)$ (Clicking a non-relevant snippet)	0.21
$P(M\|R_d)$ (Marking a relevant document)	0.71
$P(M\|N_d)$ (Marking a non-relevant document)	0.53

relevant). This meant a fairer and paired comparison could be undertaken within runs. The action judgement files for each trial were produced with different seeded probabilities, allowing for reproducible results[7]. For each run, we report the average over all trials.

4.4 Query Generation Strategy

To successfully run our simulations, we required a component to generate a series of queries to be issued to the underlying search engine. We followed two approaches to query selection: automatically generating queries with a given, *idealised querying strategy*, and using the queries from our user study [22] to provide a direct comparison between real-world and simulated searcher behaviours. This section provides details of the automatic query generation component.

Keskustalo et al. [18] and Baskaya [2] defined and analysed five prototypical querying strategies, each of which represented an idealised subset of searcher behaviours [18]. Of particular interest to us were the two querying strategies which yielded the worst (***QS1***) and best performance (***QS3***). These are briefly explained below. Our explanations use the following notation, where Q_n represents query n within a search session, with t_n representing term n from a list of terms available to formulate queries.

QS1 ***(Single Term)*** This strategy generated a series of single term queries, e.g.:

$$\mathbf{Q_1}\ t_1 \rightarrow \mathbf{Q_2}\ t_2 \rightarrow \mathbf{Q_3}\ t_3$$

QS3 ***(Three-Term)*** Here, queries are produced with two *pivot* terms and one other term. The first two terms therefore remain constant, with the third term changing for each subsequent query, e.g.:

$$\mathbf{Q_1}\ t_1t_2t_3 \rightarrow \mathbf{Q_2}\ t_1t_2t_4 \rightarrow \mathbf{Q_3}\ t_1t_2t_5$$

For this study, we implemented a blended querying strategy, ***QS1+3***, where a single term from ***QS1*** is issued, followed by a three-term query from ***QS3***. This is repeated until the generated queries are exhausted. All stopping strategy comparison simulations in this study use querying strategy ***QS1+3***. The blended querying strategy provided a balance between good queries (generated by ***QS3***) and poor queries (generated by ***QS1***) in terms of performance (see Table 3). This would allow us to determine the robustness of each stopping strategy when faced with a poor set of results. Interleaving good and bad queries poses a key challenge to stopping strategies: to spot a poor performing query and subsequently stop early, thus saving time.

Queries for the two underlying strategies ***QS1*** and ***QS3*** were generated as follows. For each topic used, the title and description were used to create a MLE language model, i.e. $p(term|topic)$. For ***QS1***, we then extracted a list of all single terms, ranking them according to this probability. For ***QS3***, we took all two-word combinations of the title terms, and selected the pair with the highest joint probability to act as the two pivot terms as described above. A list of three-term candidate queries, q, was then constructed by appending another term from the topic to the pivot terms. These three-term queries were then ranked according to $p(q|topic)$.

4.5 Stopping Strategies

For stopping strategies ***SS1***, ***SS2*** and ***SS3***, we set the associated thresholds (x_1, x_2 and x_3 respectively) to explore a range of values, from 1-20 in steps of 1, and 25-50 in steps of 5. The final threshold value of 50 was sufficiently deep such that if a simulated searcher only issued one query and examined all documents, they would reach their 1200-second time limit. Note that for ***SS1***, x_1 corresponds to the *maximum depth per query*, whereas for ***SS2*** and ***SS3***, x_2 and x_3 represent the *minimum depth per query*. For example, when $x_2 = 3$, a searcher is willing to tolerate three non-relevant snippets. However, they may see two relevant snippets in the process, and thus stop at a depth of five. Section 3 also provides a further example of this scenario.

For our difference rule-based stopping strategies ***SS4*** and ***SS5***, we explored a different threshold range for both. For ***SS4***, which considered term overlap differences, we explored a threshold range where $x_4 = 0.0$ to $x_4 = 1.0$ in steps of 0.05. The lower the threshold, the less similar the content. For ***SS5*** which utilised KL-Divergence, the threshold range varied between $x_5 = 3.0$ to $x_5 = 8.0$ in steps of 0.5. A small scale pilot study revealed that the majority of KL-Divergence scores for both the AQUAINT and WT2g collections fell within this range.

Finally, for IFT-based ***SS6***, we experimented with manipulating the gain threshold parameter and the number of documents first viewed before estimating the gain. From this pilot study, we observed that ***SS6*** was very sensitive when the number of documents first viewed was set to one. This was due to the fact that if the first document was non-relevant, the searcher would always stop. We found that basing the estimate on two snippets was much less sensitive and resulted in better performance. Thus, we report findings based upon that value. For x_6 (i.e. the average rate of gain that needed to be exceeded to continue searching), we explored the range 0.002 to 0.03 in steps of 0.002.

4.6 Searcher Comparisons

To determine which stopping strategy best approximated when people stopped searching, we evaluated each strategy against actual searcher stopping behaviour. For each

[7]Configuration files required to run the experiments reported in this paper can be obtained by contacting the lead author.

of the 48 subjects who undertook our previously discussed user study [22], we extracted from the corresponding log file all of the queries that the subjects issued across the two AQUAINT topics (№. 347 and №. 435). We issued the queries to the same search engine that the subjects used during the user study, and then computed when each stopping strategy said they would stop (for each threshold). To determine which stopping strategy/threshold best approximated the subjects of the user study, we calculated the *mean squared error (MSE)* between the actual observed subject/searcher behaviour and the simulated searcher's behaviours (as defined by the stopping strategy). This was calculated for both click depth (the last document clicked) and hover depth (the last document hovered over). The mean ± standard deviation for the click depth of the last document was 9.6 ± 15.8, while the mean ± standard deviation of the last snippet hovered over was 13.5 ± 17.5.

5. RESULTS

We present our results over two main sections. Section 5.1 provides results for our stopping strategy comparison simulations. Section 5.2 provides the key findings for comparing the stopping behaviour of 48 subjects against the six implemented stopping strategies (***SS1***-***SS6***).

5.1 Stopping Strategy Comparison

Figure 2 provides an overview of the simulation results over both the AQUAINT and WT2g collections. The two plots show how the mean *Cumulative Gain (CG)* varies across each of the six stopping strategies (***SS1***-***SS6***) versus the mean depth per query, given the threshold values used. All results presented in this section are averaged out over the 50 topics used for the associated collection, and over the ten times each simulation was run (refer to Section 4.3). The plots in Figure 2 show how the performance of the six stopping strategies ***SS1***-***SS6*** vary, thus answering **RQ1**. Table 4 reports the best performing threshold - complete with corresponding CG and depth per query values - across the two collections used and all six stopping strategies. This therefore provides an answer to **RQ2**.

By inspecting the plots in Figure 2, we first note that the two stopping strategies ***SS2*** and ***SS3*** - both of which are based upon the frustration point/disgust rules - both perform very well, especially at lower depths per query. Indeed, ***SS3*** generally outperforms all of the other stopping strategies over both the AQUAINT and WT2g collections. ***SS3*** also achieves the highest CG (8.6 ± 1.1) over AQUAINT, at a depth per query of 14.2, where $x_3 = 5$. The counterpart stopping strategy to ***SS3***, ***SS2***, also performs very well, close to the baseline fixed depth stopping strategy ***SS1***. Interestingly, ***SS1*** marginally attained the highest level of CG over the WT2g collection, with a value of 5.6 ± 1.2, at a depth per query of 22.2, with $x_1 = 40$. ***SS2*** and ***SS3*** attained CG values of 5.55 ± 0.9 and 5.48 ± 0.7 respectively. Recall the difference between these two stopping strategies: ***SS2*** considers the total number of non-relevant documents observed, while ***SS3*** considers a certain number of non-relevant documents observed contiguously. The fixed depth stopping strategy ***SS1*** results in good performance when a sufficiently high threshold is selected, resulting in a depth of approximately 21-22 documents per query. The utility-based stopping strategy ***SS6*** also performed reasonably well. Simulated searchers examined to comparatively lower depths per query (17.3 and 19.7 for AQUAINT and WT2g respectively), yet still attained a respectable level of CG. The general trend of lower CG over the WT2g collection confirms an underlying concept of IFT. If the rate of gain experienced is lower, searchers will examine result lists to greater depths. This suggests that ***SS6*** is somewhat less adaptive when experienced performance is low (on average). However, further simulations are required to confirm this and are left for future work. Finally, difference-based stopping strategies ***SS4*** and ***SS5*** consistently performed poorly. As thresholds x_4 and x_5 decreased (i.e. less overlap, more difference), simulated searchers went deeper. This did not however translate into substantially higher returns in gain.

The baseline stopping strategy ***SS1*** - which is commonly used in IR simulations and is employed in many IR measures - perhaps surprisingly performed well. To determine the difference between the baseline and other stopping strategies, we performed a series of paired t-tests. For each stopping strategy, we took the best performing threshold and assumed that there was a significant difference in total CG when compared against that of ***SS1*** when $p < 0.05$. Over AQUAINT, the frustration point/disgust-based stopping strategies ***SS2*** ($p = 0.1$) and ***SS3*** ($p = 0.8$) were not statistically significant from the CG attained by ***SS1***. However, ***SS4*** ($p = 0.003$), ***SS5*** ($p = 0.0004$) and ***SS6*** ($p = 0.0008$) all attained significantly lower levels of CG. A similar trend was found over WT2g, where the frustration point/disgust-based stopping strategies ***SS2*** ($p = 0.07$) and ***SS3*** ($p = 0.2$) were not statistically significant when compared to the CG value of ***SS1***.

Table 3: Mean (±) standard deviations of the performance of the three simulated querying strategies used (*QS1*, *QS1+3* and *QS3*) over both the AQUAINT (AQ.) and WT2g (WT.) collections.

		QS1	*QS1+3*	*QS3*
AQ.	***P@10***	0.02 ± 0.09	0.14 ± 0.24	0.25 ± 0.28
	P@20	0.02 ± 0.08	0.13 ± 0.22	0.24 ± 0.26
WT.	***P@10***	0.03 ± 0.12	0.17 ± 0.25	0.32 ± 0.27
	P@20	0.03 ± 0.10	0.16 ± 0.22	0.28 ± 0.24

Table 4: Summary of the best performing thresholds (x_n) for the six stopping strategies *SS1*-*SS6*, along with the corresponding CG values (± standard deviations) and depths per query. Results are presented for both the AQUAINT and WT2g collections.

		Threshold	CG	Depth/Query
AQUAINT	***SS1***	$x_1 = 35$	8.3 ± 1.5	20.9
	SS2	$x_2 = 25$	8.3 ± 1.3	21.4
	SS3	$\mathbf{x_3 = 5}$	$\mathbf{8.6 \pm 1.1}$	**14.2**
	SS4	$x_4 = 0.95$	6.9 ± 1.3	27.4
	SS5	$x_5 = 3$	6.3 ± 1.2	17.7
	SS6	$x_6 = 0.004$	7.3 ± 1.5	17.3
WT2g	***SS1***	$\mathbf{x_1 = 40}$	$\mathbf{5.6 \pm 1.2}$	**22.2**
	SS2	$x_2 = 19$	5.6 ± 0.9	18.9
	SS3	$x_3 = 6$	5.5 ± 0.7	19.7
	SS4	$x_4 = 1$	4.3 ± 0.8	33.2
	SS5	$x_5 = 3$	3.9 ± 0.8	9.6
	SS6	$x_6 = 0.004$	4.8 ± 1.1	19.7

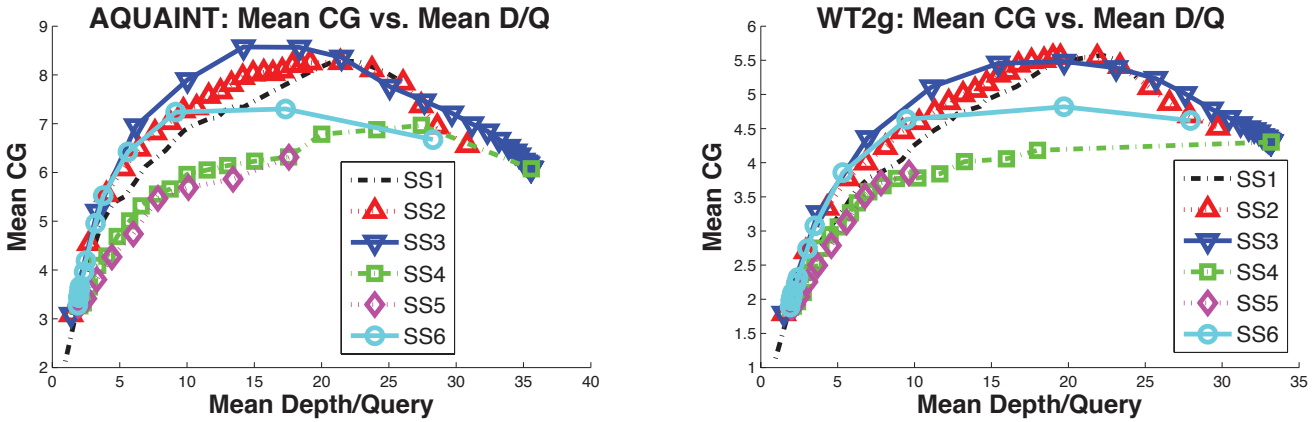

Figure 2: Plots showing the mean CG attained versus the mean depth per query, for both the AQUAINT (left) and WT2g (right) collections. Note the particularly strong performance of our baseline, *SS1*.

In addition, ***SS4*** ($p = 0.001$), ***SS5*** ($p = 0.0005$) and ***SS6*** ($p = 0.001$) were all significantly lower when compared to ***SS1***.

Table 3 reports the mean retrieval performance metrics (P@10 and P@20) over both the AQUAINT and WT2g collections, examining each individual querying strategy (***QS1***, ***QS1+3*** and ***QS3***) as detailed in Section 4.4. As expected, we find that queries generated by ***QS1*** generally performed poorly, while queries generated by ***QS3*** performed substantially better. The performance of blended querying strategy ***QS1+3*** was roughly the average of querying strategies ***QS1*** and ***QS3***. However, this hides the bimodal distribution of performance that would be experienced during searching. We posit that this may have been very detrimental to the difference-based strategies ***SS4*** and ***SS5***. This is due to the fact that for single term queries, the diversity of the results are likely to be much greater than for three term queries. Consequently, for single term queries, which provide little gain, the simulated searcher is likely to go far deeper. This means that the searcher wastes time, getting little gain in return. The plots in Figure 2 confirm this, showing that performance barely increases past a depth of 10. As a result, this suggests that such difference methods would need to consider the length of the query and expected diversity of the results, with the stopping threshold set accordingly. Similarly for the utility-based strategy ***SS6***, between-topic variation would mean that the rate of gain experienced from topic to topic will be quite different. Ultimately, this means that a more adaptive threshold for ***SS6*** may be more appropriate. For example, the threshold set could be updated based upon the performance of the queries issued by searchers. Therefore, if the issued queries are delivering a high rate of gain, the threshold could then be increased - and conversely for a low rate of gain. These directions are left for future work.

5.2 Real-World Searcher Analysis

As previously explained in Section 4.6, we used the MSE to measure the difference between the search behaviours of subjects and their simulated counterparts. This was realised for each implemented stopping strategy (***SS1***-***SS6***). For our analysis, we used two behavioural aspects of the real-world searchers - their click depths and their hover depths. Regarding our simulated searchers, the click depth was the last document observed by the searcher, irrespective of whether it was considered relevant or non-relevant. The hover depth was operationalised as the last snippet that was inspected before the stopping threshold criteria was met or exceeded.

Figure 3 plots the MSE for the click depths (top) and hover depths (bottom) for all six stopping strategies (***SS1***-***SS6***) across the depth per query (given the thresholds used; refer to Section 4.5). Table 5 reports the MSE for the best threshold over each stopping strategy, both for click depths and hover depths, broken down across topics (complete with a mean over the two topics used). On inspecting the MSE plots, we can see that the frustration point/disgust-based stopping strategies ***SS2*** and ***SS3*** both tend to result in lower MSE values across all depths (thresholds), with the baseline fixed depth stopping strategy ***SS1*** close behind. On average, the click and hover depths were best approximated when $x_2 = 9$. These findings therefore provide an answer to **RQ3**. Utility-based stopping strategy ***SS6*** then followed, performing consistently well. As with the findings reported in Section 5.1, lower thresholds for ***SS6*** yielded a closer approximation to actual searcher behaviour. This suggests real-world searchers may be attuned to obtaining small increases of gain between snippets and documents. However, it is likely that the way in which we calculated the gain (approximated as the rate of gain) and the lack of adaptive thresholds impacted the accuracy of results for ***SS6***. Furthermore, difference-based stopping strategies ***SS4*** and ***SS5*** consistently performed poorly. Despite comparatively poor approximations, we posit that the difference methods would improve with longer queries. The queries posed by the 48 real-world subjects typically were around 2-4 terms in length. It should also be noted that ***SS2*** generally outperformed its counterpart ***SS3*** in this scenario. This suggests that real-world searchers are more sensitive to the total amount of non-relevant information encountered per query. In addition, note that ***SS1***-***SS6*** all provided better approximations for click depths than hover depths, with click depth approximations yielding a lower MSE.

The thresholds yielding the lowest MSE for click depth for ***SS1***-***SS6*** were then used as the basis for Table 6. This is because those thresholds best approximated actual searcher stopping behaviour. Here, we report the mean CG that would have resulted from the queries issued by subjects if they had followed exactly each of the six stopping strategies - and the actual mean CG attained by the subjects (denoted as **RW**). Both frustration point/disgust-based stopping strategies ***SS2*** and ***SS3*** again performed well, with

Table 5: The lowest observed MSE values for each stopping strategy (*SS1-SS6*) along with the threshold at which the value was obtained (x_n). Values are reported for both topics examined, and the mean.

		Topic №. 347		Topic №. 435		Mean	
		MSE	x_n	MSE	x_n	MSE	x_n
Click Depth	*SS1*	335.4	13	166.4	13	250.2	13
	SS2	328.9	10	**159.7**	**9**	**244.4**	**9**
	SS3	**319.7**	**5**	164.7	4	253.3	5
	SS4	366.3	0.5	193.6	0.45	279.3	0.5
	SS5	392.6	5	211.5	5	301.3	5
	SS6	357.1	0.01	198.4	0.01	277.1	0.01
Hover Depth	*SS1*	397.3	14	219.3	13	307.9	13
	SS2	386.2	10	**210.3**	**9**	**298.4**	**9**
	SS3	**383.2**	**5**	213.7	4	302.4	4
	SS4	444.1	0.5	259.2	0.6	351.1	0.55
	SS5	480.6	4.5	301.3	4.5	390.2	4.5
	SS6	436.7	0.006	281.8	0.006	358.6	0.006

Table 6: The mean CG ± standard deviations for each implemented stopping strategy (*SS1-SS6*). CG values reported are attained at threshold x_n which attained the lowest MSE for click depths, as reported in Table 5. Also included is the mean CG actually attained by the real-world (RW) searchers.

	Topic №. 347		Topic №. 435		Mean	
	CG	x_n	CG	x_n	CG	x_n
RW	11.7 ± 8.9	-	13.5 ± 7.9	-	12.6 ± 8.5	-
SS1	13.1 ± 8.2	13	17 ± 7.9	13	15 ± 8.3	13
SS2	14.4 ± 8.9	10	18.2 ± 8.8	9	15.8 ± 9	9
SS3	15.9 ± 10.8	5	17.8 ± 10.2	4	19.6 ± 11.8	5
SS4	9.7 ± 6.7	0.5	9.6 ± 4.7	0.45	10.6 ± 6.1	0.5
SS5	6.7 ± 4.9	5	6.9 ± 3.9	5	6.8 ± 4.4	5
SS6	5.4 ± 5.1	0.01	8.2 ± 5.4	0.01	6.8 ± 5.4	0.01

SS3 attaining the highest mean CG (19.6). ***SS2*** followed with 15.8, with ***SS1*** attaining 15. Interestingly, ***SS1-SS3*** all attained a higher mean CG (15-19.6) than the subjects of the user study (at 12.6). This answers **RQ4**. Of note, ***SS1*** performed very well with a fixed depth ($x_1 = 13$) slightly deeper than what is typically assumed (i.e. $x_1 = 10$ for P@10). However, the high levels of variance observed will motivate further work to examine whether these findings hold at an individual searcher level.

6. DISCUSSION AND CONCLUSION

In this paper, we have examined a series of stopping rules and transformed them into stopping strategies that can be implemented and evaluated. In the context of ad-hoc topic retrieval, we found that stopping strategies ***SS1*** and ***SS3*** delivered the best overall performance. ***SS3*** was not however significantly better than a well chosen fixed depth threshold. Intuitively, an adaptive strategy appears much more realistic, and more in line with what searchers are likely to adopt. Indeed, our simulations confirmed this hypothesis, with frustration point/disgust-based ***SS2*** consistently providing the best approximation for actual searcher behaviour. In all cases however, the fixed depth baseline stopping strategy ***SS1*** performed remarkably well.

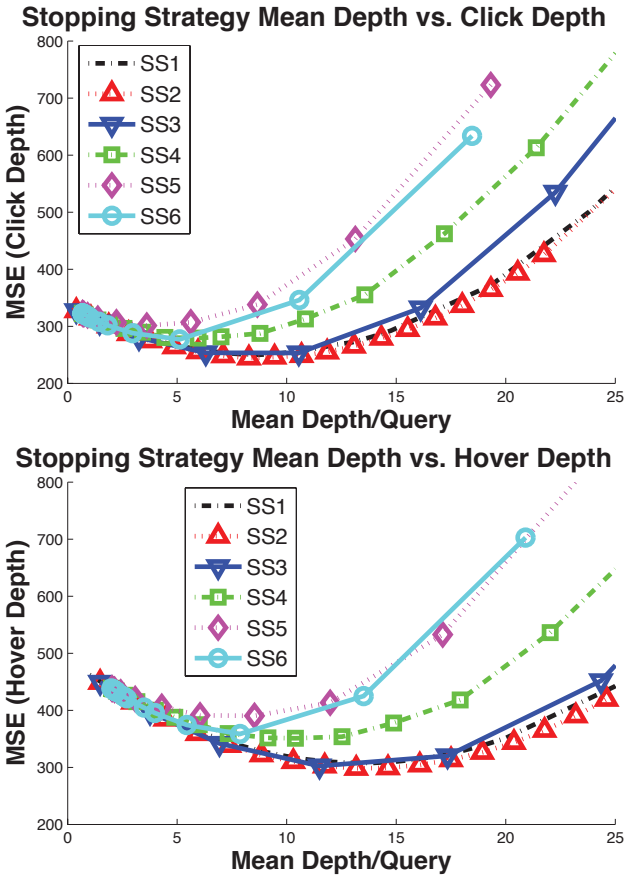

Figure 3: The MSE plots for each stopping strategy (*SS1-SS6*) across the mean depth per query. The plot at the top compares click depths, with hover depths at the bottom. Smaller MSE values indicate a closer approximation to actual searcher behaviour. Both plots indicate that *SS1-SS3* are closer to actual searcher behaviour at reasonable depths.

Our findings are both reassuring and promising. Firstly, they are reassuring in the sense that past simulations and various measures that are employed which assume a fixed depth (i.e. ***SS1***) are reasonable approximations for actual observed stopping behaviour. However, this is only in terms of the *average* stopping distance, and not the variance. A caveat here is that we have selected the best fixed depth. Therefore, an open question is how to determine this value in advance. Using the average stopping distance provides a very strong baseline to which we can improve upon. The results we have obtained from this study are promising as we have examined only a subset of possible stopping rules, and explored only a small number of possible implementations. An obvious direction for future work would be to explore this area further and consider how more adaptive variants of the proposed stopping strategies would fare. This would allow us to provide more accurate estimates of when searchers stop examining search results.

Aspects that we did not vary were the interaction probabilities and times which vary according to the veracity and

expertise of the searchers. It is likely that these probabilities and times will influence which stopping strategy is employed and which stopping strategy yields the highest gains. For example, assessment strategies such as *"fast and liberal"* or *"slow and neutral"* [29] which characterise different types of searchers when assessing are likely to impact upon the depth to which they search. On the other hand, the utility-based stopping strategy ***SS6*** is sensitive to the time spent examining snippets and documents. Accounting for and encoding these values on a group or per-searcher basis would result in a deeper understanding of how such factors influence stopping behaviour. We leave these directions for future work.

In conclusion, we have examined a range of stopping strategies and have found that strategies based upon the frustration point/disgust rules generally provide the best overall performance under simulated conditions. Furthermore, a frustration point/disgust-based strategy also closely approximated actual stopping behaviour. However, a well chosen depth with a fixed depth strategy also provided a very close approximation to actual stopping behaviour. Nevertheless, searchers can obtain markedly better performance by adopting an adaptive strategy based upon the frustration point/disgust rules. This work opens up a number of interesting lines of investigation regarding stopping rules and strategies which will be the subject of future work.

Acknowledgments: Funding for this project was provided by the MUMIA Cost Action, ref `ECOST-STSM-IC1002-080914-049840`. Funding was also provided by the ACM SIGIR for the lead author to attend the conference. We'd also like to thank the three anonymous reviewers and Alastair Maxwell for their feedback. Their comments greatly improved the clarity of this work.

References

[1] L. Azzopardi. The economics in interactive information retrieval. In *Proc. 34th ACM SIGIR*, pages 15–24, 2011.

[2] F. Baskaya. *Simulating Search Sessions in Interactive Information Retrieval Evaluation*. PhD thesis, University of Tampere, School of Information Sciences, Finland, 2014.

[3] F. Baskaya, H. Keskustalo, and K. Järvelin. Time drives interaction: Simulating sessions in diverse searching environments. In *Proc. 35th ACM SIGIR*, pages 105–114, 2012.

[4] F. Baskaya, H. Keskustalo, and K. Järvelin. Modeling behavioral factors in interactive information retrieval. In *Proc. 22nd ACM CIKM*, pages 2297–2302, 2013.

[5] M.J. Bates. The fallacy of the perfect thirty-item online search. *RQ*, 24(1):pp. 43–50, 1984.

[6] G.J. Browne, M.G. Pitts, and J.C. Wetherbe. Stopping rule use during web-based search. In *Proc. HICSS-38*, page 271b, 2005.

[7] G.J. Browne, M.G. Pitts, and J.C. Wetherbe. Cognitive stopping rules for terminating information search in online tasks. *MIS Quarterly*, 31(1):89–104, 2007.

[8] E.L. Charnov. Optimal foraging, the Marginal Value Theorem. *Theoretical Population Biology*, 9(2):129–136, 1976.

[9] W.S. Cooper. 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.

[10] W.S. Cooper. The paradoxical role of unexamined documents in the evaluation of retrieval effectiveness. *Info. Processing and Management*, 12(6):367 – 375, 1976.

[11] M. Dostert and D. Kelly. Users' stopping behaviors and estimates of recall. In *Proc. 32nd ACM SIGIR*, pages 820–821, 2009.

[12] F. Guo, C. Liu, A. Kannan, T. Minka, M. Taylor, Y. Wang, and C. Faloutsos. Click chain model in web search. In *Proc. 18th WWW*, pages 11–20, 2009.

[13] K. Hofmann. Fast and reliable online learning to rank for information retrieval. *SIGIR Forum*, 47(2):140–140, 2013.

[14] P. Ingwersen and K. Järvelin. *The Turn: Integration of Information Seeking and Retrieval in Context (The Information Retrieval Series)*. 2005.

[15] K. Järvelin and J. Kekäläinen. Cumulative gain-based evaluation of IR techniques. *ACM TOIS*, 20(4):422–446, 2002.

[16] T. Joachims, L. Granka, B. Pan, H. Hembrooke, F. Radlinski, and G. Gay. Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search. *ACM Trans. on Info. Systems*, 25(2), 2007.

[17] H. Keskustalo, K. Järvelin, and A. Pirkola. Evaluating the effectiveness of relevance feedback based on a user simulation model: Effects of a user scenario on cumulated gain value. *Information Retrieval*, 11(3):209–228, 2008.

[18] H. Keskustalo, K. Järvelin, A. Pirkola, T. Sharma, and M. Lykke. Test collection-based ir evaluation needs extension toward sessions — a case of extremely short queries. In *Proc. 5th AIRS*, pages 63–74, 2009.

[19] D.H. Kraft and T. Lee. Stopping rules and their effect on expected search length. *IPM*, 15(1):47 – 58, 1979.

[20] J. Lin and M.D. Smucker. How do users find things with pubmed?: Towards automatic utility evaluation with user simulations. In *Proc. 31st ACM SIGIR*, pages 19–26, 2008.

[21] G. Marchionini. *Information Seeking in Electronic Environments*. Cambridge University Press, 1995.

[22] D. Maxwell and L. Azzopardi. Stuck in traffic: How temporal delays affect search behaviour. In *Proc. 5th IIiX*, pages 155–164, 2014.

[23] D. Maxwell, L. Azzopardi, K. Järvelin, and H. Keskustalo. An initial investigation into fixed and adaptive stopping strategies. In *Proc. 38th ACM SIGIR*, pages 903–906, 2015.

[24] A. Moffat and J. Zobel. Rank-biased precision for measurement of retrieval effectiveness. *ACM Trans. on Info. Systems*, 27(1):2:1–2:27, 2008.

[25] A. Moffat, P. Thomas, and F. Scholer. Users versus models: What observation tells us about effectiveness metrics. In *Proc. 22nd ACM CIKM*, pages 659–668, 2013.

[26] K.R. Nickles. *Judgment-based and reasoning-based stopping rules in decision making under uncertainty*. PhD thesis, University of Minnesota, 1995.

[27] P. Pirolli and S.K. Card. Information foraging. *Psychological Review*, 106:643–675, 1999.

[28] C. Prabha, L.S. Connaway, L. Olszewski, and L.R. Jenkins. What is enough? Satisficing information needs. *J. of Documentation*, 63(1):74–89, 2007.

[29] M.D. Smucker. An analysis of user strategies for examining and processing ranked lists of documents. In *Proc. of 5th HCIR*, 2011.

[30] P. Thomas, A. Moffat, P. Bailey, and F. Scholer. Modeling decision points in user search behavior. In *Proc. 5th IIiX*, pages 239–242, 2014.

[31] E.G. Toms and L. Freund. Predicting stopping behaviour: A preliminary analysis. In *Proc. 32nd ACM SIGIR*, pages 750–751, 2009.

[32] S. Verberne, M. Sappelli, K. Järvelin, and W. Kraaij. User simulations for interactive search: Evaluating personalized query suggestion. In *Advances in Information Retrieval*, volume 9022 of *LNCS*. 2015.

[33] R.W. White, S.T. Dumais, and J. Teevan. Characterizing the influence of domain expertise on web search behavior. In *Proc 2nd ACM WSDM*, pages 132–141, 2009.

[34] W. Wu and D. Kelly. Online Search Stopping Behaviors: An Investigation of Query Abandonment and Task Stopping. In *77th ASIS&T Annual Meeting*, 2014.

[35] W. Wu, D. Kelly, and A. Sud. Using information scent and need for cognition to understand online search behavior. In *Proc 37th ACM SIGIR*, pages 557–566, 2014.

[36] L. Zach. When is "enough" enough? modeling the information-seeking and stopping behavior of senior arts administrators: Research articles. *J. American Soc. for Info. Sci. and Tech.*, 56(1):23–35, 2005.