An Initial Investigation into Fixed and Adaptive Stopping 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

Most models, measures and simulations often assume that a searcher will stop at a predetermined place in a ranked list of results. However, during the course of a search session, real-world searchers will vary and adapt their interactions with a ranked list. These interactions depend upon a variety of factors, including the content and quality of the results returned, and the searcher's information need. In this paper, we perform a preliminary simulated analysis into the influence of stopping strategies when query quality varies. Placed in the context of ad-hoc topic retrieval during a multi-query search session, we examine the influence of fixed and adaptive stopping strategies on overall performance. Surprisingly, we find that a fixed strategy can perform as well as the examined adaptive strategies, but the fixed depth needs to be adjusted depending on the querying strategy used. Further work is required to explore how well the stopping strategies reflect actual search behaviour, and to determine whether one stopping strategy is dominant.

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

Keywords Search Strategies; Search Behaviour; Stopping Strategies; Evaluation

1. INTRODUCTION

Most models, simulations and measures that examine or evaluate searcher interaction typically rely on the assumption that searchers will reach a fixed depth, with *precision-at-n (P@n)* being a prime example. In practice, this assumption is unlikely to hold. Indeed, searchers are likely to vary their interaction and the depth to which they inspect snippets and documents, depending on the performance of the system, their information need/task, and the amount of time they have available [4, 17, 18, 20, 22]. For example, a searcher may issue a query that does not return any relevant documents (i.e. a *'dud'* query). It is then likely that once the searcher inspects a few snippets and/or documents, (s)he will conclude that the issued query was unsuccessful. In this case, the searcher is then likely to issue a new query rather than continue down the current results list. This intuition has been confirmed by empirical analysis, where research shows that searchers examined significantly fewer documents when the search system failed to retrieve any relevant material in the top ten results, in contrast to when it did [1]. Thus, real searchers are inherently adaptive, and their behaviour is conditioned on the quality of the ranked lists that they interact with. In this paper, we examine the aforementioned fixed depth assumption, and perform a preliminary analysis to determine the impact of assuming a fixed depth when interacting with a search system over the course of a session. To this end, we will propose two adaptive stopping strategies - motivated by various stopping rules - and compare them to the fixed depth stopping strategy.

While we assume that adaptive approaches may perform better, we hypothesise that this depends on the quality of the queries issued. For example, if all queries issued by a searcher are 'duds', then the stopping strategy may be irrelevant. However, what if all their queries are successful, or if their queries are of varying quality? What then is the influence of the stopping strategy on overall performance?

2. RELATED WORK

Knowing when to stop is considered a fundamental aspect of human thinking [15]. Consequently, IR researchers have examined stopping behaviours in a bid to understand why, and when, searchers stop. Many studies investigating when people stop searching have concluded that the decision was mainly based on intuition, or the subjective notion of *"feeling good enough"* [22] - often termed *satisficing* [9, 21, 22]. Furthermore, decisions have also been shown to be highly dependent upon the task type being undertaken, time constraints, and various actions that the searcher performs [4, 17, 20, 22]. However, in this work, we are interested in the stopping rules and heuristics that have been proposed [5, 6, 7, 11, 15] and how to operationalise them.

When formulating such rules, researchers have considered stopping behaviours with respect to the overall search task (e.g. ceasing the search when enough information has been acquired to meet some threshold, or satisfy some task or goal) [5, 6, 15]. For example, Nickles et al. [15] proposed a

number of rules investigating the sufficiency of information: the *mental list rule*, where searchers construct a mental list of criteria about a given item (such as 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 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 reached before he or she stops.

Other work focuses on stopping behaviour at the query level (e.g. whether the searcher continues to examine documents, or decides to submit a new query) [7, 11]. For example, Cooper [7] devised two stopping rules for examining a list of ranked results: the *frustration point rule*, where a searcher stops after a certain number of non-relevant documents are encountered; and the *satisfaction stopping rule*, where searchers would stop when a certain number of relevant documents were obtained. Later, Cooper [8] developed rules using *utility theory*, positing that searchers stop examining documents once the effort of examining another document outweighs the benefit of moving to a new results list. Similar rules can be obtained from *Search Economic Theory* [1] and *Information Foraging Theory* [16].

In this work, we focus on query level stopping rules and implement two variations of the frustration point rule to explore the relationship between stopping strategies and query performance. This rule was also implemented by Lin and Smucker [12]. When navigating similar documents, their simulated searchers stopped after seeing two contiguous non-relevant documents. Our work however considers different implementations, and explores a range of stopping thresholds for ad-hoc topic retrieval.

3. EXPERIMENTAL METHOD

To explore the influence of the stopping strategy on overall performance, we conducted a number of simulations where we varied the stopping strategy and querying strategy. This was to determine which stopping strategy achieved the best performance when the quality of queries was varied. Our simulations consisted of 'searchers' performing ad-hoc topic retrieval over a series of topics on two TREC collections.

Corpora, Topics and System: We used two test collections: *TREC AQUAINT* with the *TREC 2005 Robust Track* topics, and *TREC WT2g* with the *TREC Ad-Hoc and Small Web* topics. Both topic sets were comprised of 50 topics. Each collection was indexed using the Whoosh IR toolkit[1], where stopwords[2] were removed and Porter stemming applied. The retrieval model used was PL2 ($c = 10.0$).

3.1 Simulations

While various methods have been proposed to model or simulate session-based retrieval [2, 3, 14, 19], we utilise an adaption of the method proposed by Baskaya et al. [3] as follows. A searcher ***(1)*** issues a query to the system, and then ***(2)*** proceeds to examine the first/next result snippet, or decides to issue a new query (back to *(1)*). If a given snippet is considered relevant, ***(3)*** the document is examined. If said document is considered relevant, ***(4)*** it is marked as such, and the searcher returns to *(2)*. If either a snippet or document are considered non-relevant, the searcher returns to *(2)* - with the document remaining unmarked.

[1] `https://pypi.python.org/pypi/Whoosh/`

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

Table 1: Summary of interaction times (in seconds) used for the simulations in this study.

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

Experimental Setup: At *(2)*, Baskaya et al. [3] assumed in one of their baselines a fixed depth of ten. In this paper, we consider a range of fixed depths n. In addition, we also include two adaptive stopping strategies (see Section 3.2). To generate queries of varying quality, we will employ a range of strategies as suggested by Keskustalo et al. [10] (see Section 3.3). To determine the relevance of a document, the corresponding TREC relevance assessments are typically used. Here, the action/decision of clicking on a relevant snippet or marking a relevant document is determined in a probabilistic manner. In this work, we used the probabilities of clicking on a (non)relevant snippet and the probabilities of marking a document as (non)relevant from the study performed by Smucker and Clarke [18]. In previous work, for each run, whether a document is examined or marked relevant is determined on the fly. This means that for the same query (or even a different query), the same snippet/document can be considered relevant and then non-relevant, or vice versa. In this paper, we pre-compute whether a document is considered relevant or not *a priori*, so that for different thresholds, depths and other factors, the same judgements are made for each run. This means we can perform a pairwise comparison, thus reducing the total number of simulations required.

Finally, the goal of the search task is to find as many relevant documents in a fixed time period of 1200 seconds (20 minutes). For each action performed during the simulation, the times in Table 1 were used. The estimates for each action were obtained from a user study we performed with 48 subjects over the TREC 2005 Robust Track [13].

3.2 Stopping Strategies

We considered three stopping strategies - the default fixed depth strategy (***SS1***), and two other strategies based on the frustration point rule (***SS2*** & ***SS3***) [7].

SS1 (Fixed Depth): This fixed stopping strategy encodes the heuristic that a searcher will stop examining a results list after they have viewed x_1 snippets, regardless of their relevance to the given topic.

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 and was considered non-relevant, it is included in the count.

SS3 (Contiguous Non-Relevant): Similar to ***SS2*** above, the searcher will stop using this strategy when they observe x_3 non-relevant documents *in a row*. As above, previously seen non-relevant snippets are included in the count.

For this analysis, we set the thresholds (x_1, x_2 & x_3) to be 1-20 in steps of 1, and 25-50 in steps of 5. The final value of 50 was sufficiently deep enough such that if a simulated

searcher only issued one query and examined all documents, they would run out of time. 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. In our results, we will report the average depth per query for each x_i. This will therefore allow us to compare across the three implemented stopping strategies.

3.3 Querying Strategies and Selection

Keskustalo et al. [10] define and analyse a number of different querying strategies. For the purposes of this paper, we have selected the best performing querying strategy (referred to as ***QS3***, consisting of two *pivot terms* and one other term), and the worst performing querying strategy (referred to as ***QS1***, consisting of a series of single terms) [3]. These querying strategies were used to determine how the different stopping strategies performed when combined with differing querying quality. We also implemented a blended querying strategy ***QS1+3***, which interleaved the queries generated by ***QS1*** and ***QS3***. ***QS1+3*** was implemented to determine how robust the different stopping strategies were against poor performing (or 'dud') queries.

Queries were generated as follows. For each topic, the title and description were used to create a *Maximum Likelihood Estimate (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 term combinations of the title terms, and selected the pair with the highest joint probability as the pivot. A list of three term candidate queries q was then constructed by appending another term from the topic to the pivot. These were then ranked according to $p(q|topic)$.

4. RESULTS

Figure 1 plots the mean depth per query[3] versus the rate of gain per second (averaged over all sessions and topics), given each threshold value for the AQUAINT collection[4]. All nine combinations of the aforementioned querying strategies (***QS1***, ***QS1+3*** & ***QS3***) and stopping strategies (***SS1***, ***SS2*** & ***SS3***) are shown. From the plots, we can see that the adaptive stopping strategies (***SS2*** & ***SS3***) generally outperformed the fixed depth strategy (***SS1***), regardless of the querying strategy employed, or the depth attained.

Table 2 reports the maximum gain attained across each of the stopping and querying strategies for both AQUAINT and WT2g, together with the thresholds (x_i) and mean depth per query (d). To determine whether one stopping strategy outperformed another, we performed a series of paired t-tests comparing each stopping strategy with a given querying strategy (e.g. ***SS1*** & ***QSx*** vs. ***SS2*** & ***QSx***). We found that there were no significant differences at $p = 0.05$.

To examine why this was the case, given that adaptive strategies should be intuitively more successful, we evaluated the retrieval performance of the queries that were issued during the simulation. Table 3 reports the mean retrieval performance metrics (P@5, P@10 & P@20) for

[3]The depth for a query is the lowest item in the results list for which a simulated searcher examined the associated snippet of. Mean depth is averaged for all simulated searchers over each query issued.

[4]Similar plots were obtained for the WT2g collection.

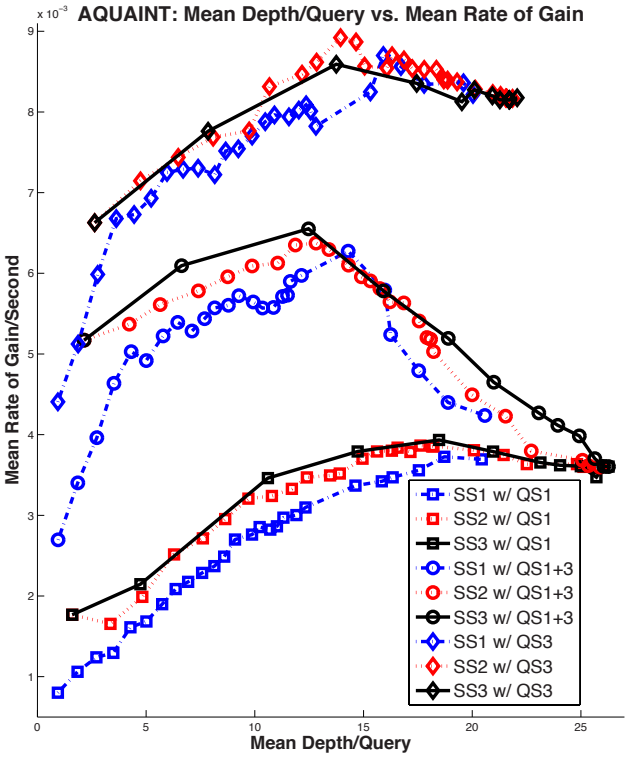

Figure 1: The mean depth per query versus the mean rate of gain per second for the AQUAINT collection, for each query and stopping strategy. Each point represents a particular threshold value, i.e. x_i.

both collections and for each of the three querying strategies used (***QS1***, ***QS1+3*** & ***QS3***). In line with previous work, the three-term querying strategy (***QS3***) outperformed the single-term strategy (***QS1***) [10], while blended ***QS1+3*** queries performed better than those generated by ***QS1***, but worse than ***QS3*** queries. Notably, each querying strategy had a high variance. This resulted in 35-40% of queries issued under ***QS1*** achieving P@10 = 0 ('dud' queries), while approximately 25% of queries under ***QS1+3*** and ***QS3*** were 'duds'. This suggests that our manipulation provided a mixture of highly performing and underperforming queries, yet the performance of the best cases for the fixed strategy were similar in overall gain to the adaptive strategies. This may be an artefact of the simulation as fixed interaction probabilities were used, whereas searchers may be more or less likely to click depending on the quality of the list (and the information scent [21]). In future work, we will examine how the behaviour of a searcher changes given the quality of the ranked list to provide a more grounded simulation.

When we examined the threshold for the fixed depth stopping strategy (***SS1***), we observed that it was quite high, ranging from 25-50. This range is much deeper than inspecting ten results per query, which is typically assumed in simulations [3]. Furthermore, real searchers are unlikely to know in advance the average performance of their queries or adopt only one particular querying strategy, so it is unlikely that a real searcher would subscribe to a fixed depth strategy. Similarly to ***SS1***, ***SS2*** also requires a range of thresholds in order to achieve maximum gain. In contrast, ***SS3*** is more consistent, where thresholds range from three to five over both collections, depending on the querying strategy used.

Table 2: Maximum cumulative gain values with corresponding thresholds and depths for each stopping and querying strategy for AQUAINT and WT2g. Significance tests indicate that there are no differences between stopping strategies.

		QS1	*QS1+QS3*	*QS3*
AQUAINT	*SS1*	4.5 ± 1.1 $x_1 = 45, d = 18.7$	7.6 ± 1.1 $x_1 = 25, d = 14.3$	10.6 ± 1.8 $x_1 = 30, d = 15.9$
	SS2	4.7 ± 0.9 $x_2 = 19, d = 17.9$	7.8 ± 1.1 $x_2 = 9, d = 12.8$	10.9 ± 0.9 $x_2 = 9, d = 13.9$
	SS3	4.7 ± 0.7 $x_3 = 5, d = 18.5$	8 ± 1.1 $x_3 = 3, d = 12.5$	10.5 ± 0.7 $x_3 = 3, d = 13.7$
WT2g	*SS1*	3.9 ± 1 $x_1 = 45, d = 18.2$	4.9 ± 0.9 $x_1 = 25, d = 13.6$	6.6 ± 1.4 $x_1 = 50, d = 18.9$
	SS2	3.9 ± 0.7 $x_2 = 17, d = 17$	5 ± 0.7 $x_2 = 9, d = 12.8$	6.7 ± 0.9 $x_2 = 30, d = 19.6$
	SS3	3.8 ± 0.6 $x_3 = 5, d = 19$	4.9 ± 0.6 $x_3 = 3, d = 12.2$	6.6 ± 0.8 $x_3 = 4, d = 16.9$

Table 3: Means (and standard deviations) of the queries issued during the simulation for each querying strategy, both for AQUAINT (AQ.) and WT2g.

		QS1	*QS1+3*	*QS3*
AQ.	*P@5*	0.2 ± 0.27	0.32 ± 0.31	0.41 ± 0.31
	P@10	0.21 ± 0.26	0.31 ± 0.29	0.41 ± 0.28
	P@20	0.13 ± 0.18	0.2 ± 0.2	0.32 ± 0.25
WT2g	*P@5*	0.19 ± 0.26	0.25 ± 0.29	0.32 ± 0.33
	P@10	0.19 ± 0.22	0.26 ± 0.25	0.35 ± 0.29
	P@20	0.15 ± 0.21	0.18 ± 0.21	0.29 ± 0.25

This strategy is also more in line with intuition. Here, the searcher moves to the next query after encountering three to five contiguous non-relevant documents. This suggests that ***SS3*** is more robust across query performance, but further research is required to confirm this.

5. SUMMARY AND FUTURE WORK

In this paper, we used simulations to examine different stopping strategies. Overall, the adaptive stopping strategies tend to outperform the fixed stopping strategy. However, to our surprise, this was not significantly so. The caveat being that for the fixed stopping strategy to provide similar performance, the right threshold needs to be chosen. In practice, this would require a different type of adaptive behaviour, where the searcher changes the depth they are willing to go to based on their querying strategy. This seems unlikely. The most robust stopping strategy appeared to be ***SS3*** (contiguous non-relevant), with a threshold of around three to five non-relevant documents. This stopping strategy seems to match better with intuition, but whether real searchers adopt such a strategy is an open question. In future work, we will examine a greater variety of querying strategies/selection methods (such as lower and higher precision) to determine whether the adaptive strategies result in greater gains. We shall also explore which strategy, if any, best characterises real searcher stopping behaviour, and explore whether there is a relationship between results list quality and interaction probabilities. Other adaptive stopping strategies will be examined, as well as comparing proposed strategies against observed stopping behaviours.

Acknowledgments: We would like to thank the ESF-funded MUMIA COST Action (ref. `ECOST-STSM-IC1002-080914-049840`). We would also like to thank Horaţiu Bota, Sean McKeown, Alastair Maxwell and Paul Thomas for their input.

References

[1] L. Azzopardi. Modelling interaction with economic models of search. In *Proc. 37th ACM SIGIR*, pages 3–12, 2014.

[2] 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.

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

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

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

[6] 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.

[7] W. S. Cooper. On selecting a measure of retrieval effectiveness part ii. implementation of the philosophy. *J. of the American Society for Info. Sci.*, 24(6):413–424, 1973.

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

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

[10] 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.

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

[12] 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.

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

[14] 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.

[15] K. R. Nickles, S. P. Curley, and P. G. Benson. Judgment-based and reasoning-based stopping rules in decision making under uncertainty. Technical report, U. of Minnesota, 1995.

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

[17] 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.

[18] M. D. Smucker and C. L. Clarke. Time-based calibration of effectiveness measures. In *Proc. 35th ACM SIGIR*, pages 95–104, 2012.

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

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

[21] W. C. 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.

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