A Study of Selection Noise in Collaborative Web Search*

Oisín Boydell and **Barry Smyth**
Adaptive Information Cluster
Smart Media Institute, Department of Computer Science,
University College Dublin, Belfield, Dublin 4, Ireland
{oisin.boydell, barry.smyth}@ucd.ie

Cathal Gurrin and **Alan F. Smeaton**
Adaptive Information Cluster, Centre for Digital Video Processing,
Dublin City University, Glasnevin, Dublin 9, Ireland
{cathal.gurrin, alan.smeaton}@computing.dcu.ie

Abstract

Collaborative Web search uses the past search behaviour (queries and selections) of a community of users to promote search results that are relevant to the community. The extent to which these promotions are likely to be relevant depends on how reliably past search behaviour can be captured. We consider this issue by analysing the results of collaborative Web search in circumstances where the behaviour of searchers is unreliable.

1 Introduction

Traditional approaches to Web search are term-based; they assume that relevance is best understood from the terms within documents. It has been shown that typical Web users are not experts when it comes to searching for documents - they routinely provide ambiguous queries [Lawrence and Giles, 1998], making it difficult for search engines to predict their needs. In such an environment, query and document terms are not always a good indication of relevance and so content-based approaches often fail [Furnas *et al.*, 1987]. Recent innovations have seen the introduction of new sources of relevance knowledge. For example, the work of [Brin and Page, 1998] has famously demonstrated the usefulness of inter-document relationships by exploiting link-connectivity information. More recently researchers have also focused on ways to exploit context in Web search as a way to disambiguate vague queries, either by explicitly establishing the search context up-front or by implicitly inferring the context as part of the search process (see [Lawrence, 2000]). Yet others have focused on leveraging knowledge about the query-space to directly address the correspondence problem, by mining query logs in order to identify useful past queries that may help the current searcher (e.g., [Glance, 2001; Raghavan and Sever, 1995; Wen *et al.*, 2002]). The work of [Smyth *et al.*, In Pressb] has suggested another source of relevance knowledge: in the *Collaborative Web Search* (CWS) approach patterns of queries and result selections from a community of users are used to re-rank the results of an underlying search engine. A community can be a well-defined group of users, such as the employees of a given company, or it can be an ad-hoc group of individuals, such as the users of some topical Web site, for example a motoring Web site. The point is that many searches can be traced back to communities of users and our research indicates that these communities are likely to submit similar queries and select the same types of results [Smyth *et al.*, In Pressb]. In a recent analysis of search logs from a variety of sources we have found that within communities of searchers up to 70% or more of queries share at least 50% of their query terms with past queries from the community; see [Smyth *et al.*, In Pressb]. For example, in [Smyth *et al.*, In Pressa] we describe the results of a recent trial of CWS within a corporate context and a community of about 50 searchers (the employees of an Irish software company). The data collected as part of this trial indicates that 57% of queries share half of their query terms with previous queries. The results of this trial go on to show that CWS can take good advantage of this type of repetition and regularity to significantly enhance search performance.

CWS relies on a relevance assumption—*that the selection of a page for some query is a reliable indicator of page relevance*—which cannot be fully relied upon. Search engines often return pages that turn out not to be relevant, and users often mistakenly select them. The question is: how sensitive are the results of CWS to selection noise? We will answer this question in what follows by analysing the impact of selection noise on result precision.

2 Collaborative Web Search

CWS is a form of meta-search, relying on the search services of a set of underlying search engines, but manipulating their results in response to the learned preferences of a given community of users. A key data structure in CWS is the *hit ma-*

*This material is based on works supported by Science Foundation Ireland under Grant No. 03/IN.3/I361.

trix, H. It represents the search behaviour of a given community of users and each time a member of a community selects a result p_j in response to some query q_i the entry in cell H_{ij} is incremented. In turn, the *relevance* of a page p_j to q_i can be estimated as the relative number of selections p_j has received in the past for q_i; see Equation 1.

$$Relevance(p_j, q_i) = \frac{H_{ij}}{\sum_{\forall j} H_{ij}} \quad (1)$$

$$WRel(p_j, q_T, q_1, ..., q_n) = \frac{\sum_{i=1...n} Relevance(p_j, q_i) \bullet Sim(q_T, q_i)}{\sum_{i=1...n} Exists(p_j, q_i) \bullet Sim(q_T, q_i)} \quad (2)$$

More generally, the relevance of a page p_j to some query q_T can be calculated as the weighted sum of its relevance to a set of queries that are similar to q_T, namely $q_1, ..., q_n$, with each individual relevance value discounted by the similarity between q_T and the query in question; see Equation 2. The similarity between q_T and the query in question, q_i is calculated as in Equation 3; it is worth noting that we have recently evaluated a variety of alternative and more sophisticated similarity metrics, the results of which are presented in [Balfe and Smyth, 2005; In Press].

$$Sim(q_T, q_i) = \frac{|q_T \cap q_i|}{|q_T \cup q_i|} \quad (3)$$

Thus on receipt of some target query, q_T, CWS dispatches it to a set of underlying search engines and their results are combined to form a meta-search result-list, R_M. At the same time, q_T is compared to other queries in the hit-matrix to select a set of similar queries, $q_1, ..., q_n$, that are selected if their similarity to q_T exceeds some set threshold. The results selected for these queries in the past are ranked by their weighted relevance to q_T, according to Equation 2, to produce a new *promoted* result-list, R_p. R_P is combined with R_M to produce R_T, which is then returned to the user; in our implementation $R_T = R_P \cup R_M$. For further detail on the technical details of the CWS architecture and operation see [Smyth *et al.*, In Pressb].

3 Evaluation

The success of CWS depends on the quality of its promoted results, R_P, and their quality in turn depends on the reliability of the user selections that underpin the core relevancy calculations. In this evaluation we will assess the quality of these promotions under different degrees of selection noise with an experiment using the TREC Terabyte collection (see *http://www-nlpir.nist.gov/projects/terabyte*) and the Físréal search engine [Blott *et al.*, In Press].

3.1 The TREC Terabyte Collection

This collection consists of over 25 million Web pages (approx, 426GB) crawled from the $.gov$ domain during early 2004. The TREC Terabyte track included 50 random topics as target search topics. Each topic included a short textual description and during the evaluation competing search engines were evaluated with respect to their ability to locate pages relevant to these topics. After the evaluation a *relevance engine* was made available to help with the evaluation of new search techniques. This engine provides ground-truth relevance for the topics and allows for the detailed analysis of result-lists in relation to the test queries.

3.2 Training the Hit-Matrix

Normally in CWS the hit-matrix is trained by the searches of a given community of users, but we need a mechanism for manipulating selection noise and so we need a more controlled experimental framework. For this we require a set of queries and a method for judging the relevance of the results that are returned and that might be selected by users. We generate queries by extracting subsets of terms from the TREC topic descriptions, after first removing commonly occurring stop words; for each topic we generate 50 queries with between 2 and 8 terms each. To simulate the action of a searcher we use the official TREC relevance engine, which is capable of identifying all result pages that are relevant for a given topic. Thus, for a training query q generated from topic t, we can identify the k pages returned by a baseline search engine (see [Blott *et al.*, In Press]) that are relevant to t as the basis for updating our hit-matrix for topic t. During the update we use two types of noise. $TypeA$ noise occurs when additional non-relevant results are selected and thus during training we assume that the user selects all k retrieved documents that are relevant and an additional $n\%$ of k non-relevant pages. $TypeB$ noise occurs when users fail to select all of the relevant results returned and thus during training we assume that the user only selects $(100 - n)\%$ of the k relevant results and $n\%$ of k irrelevant results returned for each query.

3.3 Precision Analysis

To assess the precision of CWS we use the official TREC test queries that formed the basis of the TREC 2004 evaluation; none of these queries were used in training. Two different versions of CWS are used for query similarity thresholds of > 0 and $>= 0.5$, with different types and levels of selection noise introduced. Each test query is replayed and the percentage of relevant results for different result-list sizes is computed. The average of these precision values is computed to produce an overall *mean average precision* (MAP) for each version of CWS. We also compute a baseline MAP for the TREC Benchmark search engine used in [Blott *et al.*, In Press], which serves as the underlying engine for CWS.

The results are presented in Figures 1 and 2. Each graph shows the MAP for the baseline and for CWS with Type A and B noise. They show that CWS offers significant precision increases over the baseline. For example, for the > 0 threshold, at the 0% noise level, CWS delivers over twice the precision of the baseline (0.34 vs. 0.15). We also see that these precision benefits degrade gracefully as noise is increased, and the benefits are preserved even for high noise levels. For instance, even with 100% Type A noise we see that CWS delivers a MAP of 0.17, which is greater than the corresponding baseline MAP (0.14). We see that CWS is more sensitive to Type B noise, especially for the $>= 0.5$ threshold. Here, CWS MAP falls below the baseline for 80% noise;

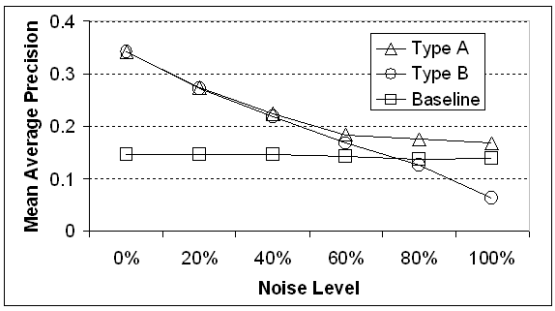

Figure 1: Similarity Threshold > 0

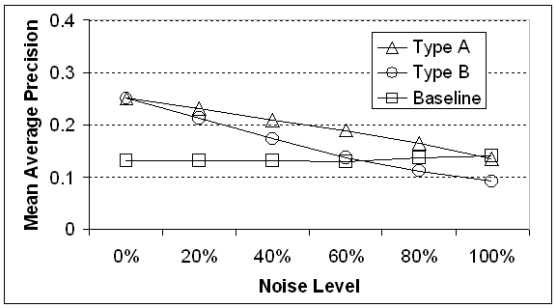

Figure 2: Similarity Threshold ≥ 0.5

this corresponds to a situation where in many search sessions no relevant results are selected by users.

4 Conclusions

CWS personalizes search results for a community of like-minded searchers based on their prior search histories. It is appropriate in a wide variety of search scenarios because many Web searches are initiated within a community context; for example, the searches initiated from a search box on some themed Web site or the searches of some online social group. When responding to a new search query, CWS exploits the results of searches for similar queries that have taken place in the past (for a given community) and actively promotes those results. Thus, results that are consistently selected for queries will tend to be promoted, although care must be taken to limit the influence of promotional biases as a result of selection noise or malicious users.

In this paper we have evaluated the performance of CWS when search histories contain result selections that are unreliable. Obviously, if users select results that are not relevant to their query then there is the risk that these results will be promoted in the future. The question that we have attempted to answer is the degree to which CWS is sensitive to such noisy selections. Our results indicate that CWS is relatively robust to noise with significant precision benefits available even for very high levels of selection noise. This bodes well for CWS and is in agreement with the performance benefits witnessed in live-user trials, in which noisy selections are likely to occur. In finishing, it is worth highlighting that the CWS approach has been fully implemented in the I-SPY system which can be accessed at *http://ispy.ucd.ie*.

References

[Balfe and Smyth, 2005] Evelyn Balfe and Barry Smyth. An analysis of query similarity in collaborative web search. In *Proceedings of the 27th European Conference on Information Retrieval*, pages 330–344, 2005.

[Balfe and Smyth, In Press] Evelyn Balfe and Barry Smyth. A Comparative Analysis of Query Similarity Metrics for Community-Based Web Search. In *Proceedings of the 6th International Conference on Case-Based Reasoning*, (In Press).

[Blott *et al.*, In Press] Stephen Blott, Oisín Boydell, Fabrice Camous, Paul Ferguson, Georgina Gaughan, Cathal Gurrin, Noel Murphy, Noel O'Connor, Alan F. Smeaton, Barry Smyth, and Peter Wilkins. Experiments In Terabyte Searching, Genomic Retrieval And Novelty Detection For TREC-2004. In *Proceedings of the Thirteenth Text REtrieval Conference*, (In Press).

[Brin and Page, 1998] Sergey Brin and Lawrence Page. The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117, 1998.

[Furnas *et al.*, 1987] George W. Furnas, Thomas K. Landauer, Louis M. Gomez, and Susan T. Dumais. The vocabulary problem in human-system communication. *Communications of the ACM*, 30(11):964–971, 1987.

[Glance, 2001] Natalie S. Glance. Community Search Assistant. In *Proceedings of the International Conference on Intelligent User Interfaces*, pages 91–96. ACM Press, 2001.

[Lawrence and Giles, 1998] Steve Lawrence and C. Lee Giles. Context and Page Analysis for Improved Web Search. *IEEE Internet Computing*, July-August:38–46, 1998.

[Lawrence, 2000] Steve Lawrence. Context in Web Search. *IEEE Data Engineering Bulletin*, 23(3):25–32, 2000.

[Raghavan and Sever, 1995] Vijay V. Raghavan and Hayri Sever. On the reuse of past optimal queries. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 344–350. ACM Press, 1995.

[Smyth *et al.*, In Pressa] Barry Smyth, Evelyn Balfe, Oisín Boydell, Keith Bradley, Peter Briggs, Maurice Coyle, and Jill Freyne. A Live-User Evaluation of Collaborative Web Search. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence, IJCAI-05*, (In Press).

[Smyth *et al.*, In Pressb] Barry Smyth, Evelyn Balfe, Jill Freyne, Peter Briggs, Maurice Coyle, and Oisín Boydell. Exploiting Query Repetition & Regularity in an Adaptive Community-based Web Search Engine. *User Modeling and User-Adapted Interaction: The Journal of Personalization Research*, (In Press).

[Wen *et al.*, 2002] Ji-Rong Wen, Jian-Yun Nie, and Hong-Jiang Zhang. Query clustering using user logs. *ACM Transactions on Information Systems*, 20(1):59–81, 2002.