Evaluating the Impact of Selection Noise in Community-Based Web Search*

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

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

ABSTRACT

The I-SPY meta-search engine uses a technique called *collaborative Web search* to leverage the past search behaviour (queries and selections) of a community of users in order to promote search results that are relevant to the community. In this paper we describe recent studies to clarify the benefits of this approach in situations when the behaviour of users cannot be relied upon in terms of their ability to consistently select relevant results during search sessions.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*relevance feedback, retrieval models*

General Terms

Experimentation, Measurement, Human Factors

Keywords

Collaborative Web search, personalisation, relevance, noise

1. INTRODUCTION

Collaborative Web search (CWS) [4] is a form of meta-search that manipulates the results of underlying Web search engines, such as Google and HotBot, in response to the learned preferences of a given community of users. A community shares similar information needs, and may be defined implicitly, for example an ad-hoc group of searchers using a search box located on a particular themed Web site, or explicitly where a particular information need is defined for a community. Within a community, results that have been preferred in the past for similar queries are actively promoted in a new result-list. To do this CWS maintains a data structure called the *hit matrix*, H to represent the search behaviour of a given community of users. Each time a community member 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 is estimated as the relative number of selections p_j has received in the past for q_i; see Equation 1.

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

$$\text{WeightedRelevance}(p_j, q_T, q_1, \ldots, q_n) = \frac{\sum_{i=1}^{n} \text{Relevance}(p_j, q_i) \cdot \text{Sim}(q_T, q_i)}{\sum_{i=1}^{n} \exists(p_j, q_i) \cdot \text{Sim}(q_T, q_i)} \quad (2)$$

More generally, the relevance of p_j to some q_T is calculated as the weighted sum of its relevance to a set of queries, $q_1, \ldots, q_n$, which are similar to q_T; each individual relevance value is discounted by the similarity between q_T and the query in question (Equation 2). We assume query similarity is measured using a suitable metric (e.g., weighted term overlap).

On receipt of some target query, q_T, CWS dispatches it to its 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 the hit-matrix to choose a set of similar queries, $q_1, \ldots, q_n$ whose 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 $R_T = R_P \cup R_M$.

In theory, CWS can be used to overlay a new relevance model on top of an existing search engine. Ordinarily this relevance model corresponds to the search preferences of communities, by utilizing past selection behaviours, but it could just as easily use different types of relevance knowledge. In this paper we describe experiments as part of the TREC terabyte track [2] which test the CWS approach to implement a relevance model based on link connectivity information over a benchmark search engine that relied on page content alone. While the results of this initial experiment met with only limited success, the experience enabled us to gain new insights into the key factors that influence the success of CWS. In particular these results helped to evaluate the impact of selection noise on CWS performance, using TREC relevance results.

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

SIGIR'05, August 15–19, 2005, Salvador, Brazil.
ACM 1-59593-034-5/05/0008.

2. AN INITIAL TREC EXPERIMENT

The TREC Terabyte collection consists of 25 million Web pages (approx. 426GB) from the *.gov* domain. The TREC Terabyte track of 2004 includes 50 topics as target search topics. Each includes a short text description and during the evaluation, competing search engines are compared by their ability to retrieve documents relevant to these topics.

Normally in CWS the hit-matrix is trained by the searches of a given community of users, but in our TREC experiment we were interested in whether CWS could be used to implement a relevance model that was based on link connectivity and anchor text; each document d_i was represented by a new document d_i', made up of the anchor text entries for all links to d_i within the collection. The CWS hit-matrix was then trained using a version of the Físréal benchmark search engine[2] that relied on an anchor text index produced from the d_i''s. A set of training queries was generated by extracting subsets of terms from the TREC topic descriptions and narratives; for each topic we generated 250 queries with between 2 and 8 terms each. Each of these queries was submitted to the Físréal engine and the hit-matrix was updated with the top 20 results. During testing, the TREC test queries were submitted to a version of CWS that used the above hit-matrix and the Físréal benchmark search engine using a standard document index, so that documents that tended to match on anchor text terms were promoted within the final result-list.

The results were mixed, with our TREC Terabyte track run ranking only 56th out of 71 submitted runs [2, 3]. It quickly became apparent that our approach to training was unlikely to produce high-quality (relevant) result promotions. Specifically, naively updating the hit-matrix with the top 20 results led to a significant degree of noise (non-relevant results) being added to the hit-matrix. And, of course, this noise was being expressed during testing through the promoted results.

3. EVALUATING SELECTION NOISE

After the TREC 2004 Terabyte track, *relevance results* were made available to participants to help with the evaluation of new search techniques. These provide ground-truth relevance assessments for the topics and allow for a more detailed and principled evaluation of the factors that contribute to the success of CWS, especially in relation to the presence of selection noise in the hit-matrix data.

To do this we configured our CWS engine to work with the Físréal benchmark search engine and the same training queries were used during hit-matrix training; we also set a query similarity threshold of > 0 so that query reuse is triggered once the queries share at least one common term. This query similarity method is evaluated with respect to CWS in [1]. However, this time we use the TREC relevance judgments to simulate the selections of a live-user under different noise conditions. We control two basic parameters: k refers to the number of selections made by this user during a search sessions; and n refers to the percentage of these selections that are noisy (non-relevant). So, for example, $k = 10$ and $n = 0.4$ indicates that during training 10 results were selected per search but that only 60% of these results were actually relevant according to TREC relevance assessments. We trained different hit-matrices for a range of different combinations of k and n, and for each we calculated the mean average precision (MAP) of CWS for the official TREC 2004 Terabyte Track test queries; of course, none of these test queries were used in training. We also computed a baseline MAP for the Físréal benchmark search engine, which serves as the underlying engine for CWS.

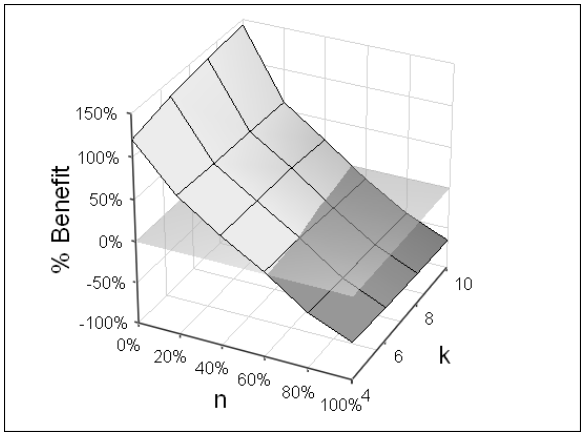

Figure 1: CWS Benefit vs. k & n.

The results are presented in Figure 1 as a surface plot of percentage increase in CWS MAP relative to the baseline versus k and n for CWS. The results clearly indicate a significant benefit for CWS for noise levels up to approximately 50%. In other words, as long as user selections are more reliable than they are unreliable we can expect CWS to deliver a precision benefit; for example, for low levels of noise ($< 20\%$) we see a $> 50\%$ increase in result precision for CWS. We also see that precision is less sensitive to the number of selections per session and that even when a user is only selecting a few results per session, there is an improvement in precision. For example, for $k = 4$ and $n = 0.4$ there is a 33% relative improvement for CWS even though only 2-3 relevant results are selected per session against 1-2 irrelevant results.

4. CONCLUSIONS

Collaborative Web Search is a way of personalizing search results for a community of like-minded searchers based on their prior search histories. We have evaluated the sensitivity of CWS to noisy relevance judgments within these search histories. Our results indicate that CWS is robust to reasonable levels of noise with significant precision benefits available even for relatively high levels of selection noise.

5. REFERENCES

[1] E. Balfe and B. Smyth. An Analysis of Query Similarity in Collaborative Web Search. In *Proceedings of the European Conference on Information Retrieval*. Springer-Verlag, 2005.

[2] S. Blott, O. Boydell, F. Camous, P. Ferguson, G. Gaughan, C. Gurrin, N. Murphy, N. O'Connor, A. F. Smeaton, B. Smyth, and P. Wilkins. Experiments In Terabyte Searching, Genomic Retrieval And Novelty Detection For TREC-2004. In *Proceedings of the Thirteenth Text REtrieval Conference*, (In Press).

[3] O. Boydell, C. Gurrin, A. F. Smeaton, and B. Smyth. Manipulating the Relevance Models of Existing Search Engines. In *Proceedings of the European Conference on Information Retrieval*. Springer-Verlag, 2005.

[4] B. Smyth, E. Balfe, J. Freyne, P. Briggs, M. Coyle, and O. Boydell. Exploiting query repetition and regularity in an adaptive community-based web search engine. *User Modeling and User-Adapted Interaction*, 14(5):383–423, 2004.