Community-Based Snippet-Indexes for Pseudo-Anonymous Personalization in Web Search*

Oisín Boydell
Adaptive Information Cluster
School of Computer Science and Informatics
University College Dublin, Dublin 4, Ireland
oisin.boydell@ucd.ie

Barry Smyth
Adaptive Information Cluster
School of Computer Science and Informatics
University College Dublin, Dublin 4, Ireland
barry.smyth@ucd.ie

ABSTRACT

We describe and evaluate an approach to personalizing Web search that involves post-processing the results returned by some underlying search engine so that they reflect the interests of a community of like-minded searchers. To do this we leverage the search experiences of the community by mining the title and snippet texts of results that have been selected by community members in response to their queries. Our approach seeks to build a community-based snippet index that reflects the evolving interests of a group of searchers. This index is then used to re-rank the results returned by the underlying search engine by boosting the ranking of key results that have been frequently selected for similar queries by community members in the past.

Categories and Subject Descriptors: H.3.3 Information Search and Retrieval: search process, selection process

General Terms: Human Factors, Design

Keywords: personalization, Web search, community

1. INTRODUCTION

Dealing with the type of vague queries that are commonplace in Web search is an important and challenging problem. It has been well documented that typical Web queries contain an average of only 2-3 terms [1], for example, and queries like *"jordan pictures"* offer no clues about whether the searcher is likely to be looking for images of the racing team, the middle eastern state, the basketball star, or the celebrity. Approaches which attempt to personalize the selection and ranking of search results offer a solution [2]. By learning about the personal preferences of the searcher and/or the context of their search it may be possible to prioritise certain results that are more likely to be relevant.

The work described in this paper has been inspired by previous research on *Collaborative Web Search* (CWS) [3] which highlighted the high degree of query repetition and result selection regularity that naturally exists within community-based search scenarios. For reasons of space it is not possible to discuss in detail the origins of such communities of searchers although the interested reader is referred to the work of [3] for a more complete treatment of this issue. Suffice to say that such communities can be readily identified, whether they are a formal community of users (e.g., the employees of a company operating in a specific business sector) or an ad-hoc group of searchers (e.g., the visitors to a Web site specialising in wildlife and endangered species). CWS maintains a community search profile by recording the queries submitted and the results selected by community members. When faced with a new target query, CWS promotes results that have consistently been selected for this and similar queries by the community in the past. However, a major limitation of using community profiles based solely on queries and previously-selected results is that results in a community's history which are relevant to a new target query can only be identified as such if there are overlapping terms between the target and previous relevant queries.

The work presented here proposes a more elaborate community model to improve promotion quality by maintaining a community-based *snippet index* as a way to drive promotions. Thus, instead of simply storing the queries and the URLs of the pages selected, we produce a local index based on the terms that are contained in the title and (query-sensitive) snippets for selected results.

The use of snippets for document indexing in IR was suggested as early as 1958 [4], and more recently work by [5] has looked at generating an alternative index using generic document summaries which can then be queried in parallel to a full content index or used as a source for pseudo-relevance feedback. Our approach is different in that we use query-sensitive document snippets which summarise a document for a particular community of searchers.

The work of [6] on *document transformation* suggests modifying indexed document content according to previous selection behaviour in order to bring documents closer to the queries that led to their selection. Again our approach is different in that we create a new personalized index for a specific community without altering the existing full text index, and this enables our approach to be applied as a personalized meta search engine on top of existing Web search engines. We also use query-sensitive snippets which provide a far richer set of terms than the query terms alone.

2. COMMUNITY-BASED SNIPPET INDEXING

Consider some user u, a member of some community C. A new target query q_T from u is initially answered by a traditional meta-search engine to produce a result-list, R_M. In parallel, q_T is used to query a local document index that

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

SIGIR'06, August 6–11, 2006, Seattle, Washington, USA.
ACM 1-59593-369-7/06/0008.

has been constructed from the title and snippet texts of results that have been selected by the community in the past. This produces a new list of results, R_C, that are more closely aligned with community interests and R_M and R_C are combined and returned to the user as R_T.

We use (C, u, q_T) to denote a search for query q_T by user u in community C. Consider a result r selected in response to such a search. We can reasonably assume that the snippet for this result $s(r)$ must contain terms which are of special interest to the user in relation to their query. Therefore, $s(r, q_T)$ can be used to represent the document corresponding to r for (C, u, q_T). In this sense $s(r, q_T)$ is a *surrogate* for r in the context of (C, u, q_T) and thus we propose that r can be indexed by using the terms contained within $s(r)$.

Accordingly, our approach to collaborative Web search involves constructing a community-based index by indexing each selected result document by its snippet terms. In general then, given that a result r might actually be selected for a number of different queries, $q_1, ..., q_n$, it will come to be indexed under a number of different snippets, $s(r, q_1), \ldots, s(r, q_n)$. Thus, for a given community of searchers each document will come to be represented by its surrogate, $S^C(r)$ as shown in Equation 1

$$S^C(r) = \bigcup_{\forall i} s(r, q_i) \tag{1}$$

Documents that are broadly relevant to a community's interests are likely to be retrieved for a wide variety of queries and are likely to be selected for many of these queries. As a result the document surrogate will cover a significant portion of the document's contents and the snippet index will reflect this by associating the document with a broad set of index terms. In contrast, we might consider other documents that are only relevant to a community through some small part of their contents. These are more likely to be retrieved for a much more restricted set of query terms and their snippets will also be drawn from a limited subset of their content, and so their index terms will also be very limited.

2.1 Community-Based Promotion

In the current implementation we use Lucene [1] to perform the basic indexing and retrieval on the community-based snippet index. At retrieval time, Lucene queries the snippet index for C using q_T to produce R_C, which is ranked using a function based on each result's *TF-IDF* score boosted by relative hits count and query similarity according to Equation 2. r_j is a result in R_C and $q_1, \ldots, q_n$ are the queries for which r_j was previously selected. $Rel(r_j, q_i)$ is the relative hits count, which is the number of times r_j was previously selected for q_i compared to the total hits for q_i. $QuerySim(q_T, q_i))$ is a simple term-based query similarity metric based on Jaccard's coefficient.

$$\text{Relevance}(r_j, q_T, q_1, \ldots, q_n) = TFIDF(r_j, q_T) * (1 + \sum_{i=1}^{n} (\text{Rel}(r_j, q_i) \cdot \text{QuerySim}(q_T, q_i))) \tag{2}$$

In our current implementation, the final result list returned to the searcher, R_T, is the union of R_C and R_M with the R_C results returned before the R_M results. Thus the precision of the R_C results is critical and for this reason we use two techniques to filter the R_C results to enhance their precision; that is, in addition to standard stop-word removal and stemming during indexing and retrieval. First, we threshold the proportion of query terms that must be present in the document surrogate for that document to be retrieved as part of R_C. This effectively eliminates results that match on only a few of the query terms and can help to eliminate superficial results from being retrieved. By default we set this threshold to allow for the retrieval of results that match at least 50% of the query terms. Second, we also limit the total number of results returned in R_C to ensure that community promotions do not over-power the traditional meta-search results. Normally we set this limit at 5-10 promotions.

[1] http://lucene.apache.org

3. CONCLUSION

We have described an approach to personalizing Web search at the level of communities of like-minded searchers. The approach works by using the search behaviour of community members—their search queries, the results they select and their snippets and titles—to populate a local index. Each selected result document is represented by a surrogate that is made up of the various snippet texts that have been associated with each of its selections. These surrogates reflect a biased view of the document in terms of the community's implicit preferences. When responding to a new search query, previous community selections retrieved from the local index are used to complement the results returned by a standard meta-search. The former are promoted based on their overlap with the target search query and their relevance to the community estimated from their selection histories. Preliminary results from a live user trial show that using a community-based snippet index provides search results with a higher precision than both the original CWS system and standard Web search.

Finally, it is worth remarking on the privacy benefits of our approach to personalized Web search. Within any particular community, the search patterns of an individual cannot be identified and so their identity can remain anonymous. At the same time, personalized recommendations can still be made to the benefit of the individual searcher. Of course, whether users in general will perceive this as a reasonable privacy-personalization trade-off in practice remains to be seen.

4. REFERENCES

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

[2] J. Pitkow, H. Schutze, T. Cass, R. Cooley, D. Turnbull, A. Edmonds, E. Adar, and T. Breuel. Personalized search. *Communications of the ACM*, 45(9):50–55, 2002.

[3] 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.

[4] H. P. Luhn. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2:159–165, 1958.

[5] T. Sakai and K. Sparck-Jones. Generic summaries for indexing in information retrieval. In *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 190–198, New York, NY, USA, 2001. ACM Press.

[6] C. Kemp and K. Ramamohanarao. Long-term learning for web search engines. In *Proceedings of the 6th European Conference on Principles of Data Mining and Knowledge Discovery*, 2002.