(Web Search)$^{\text{shared}}$
Social Aspects of a Collaborative, Community-based Search Network

Maurice Coyle and Barry Smyth

Adaptive Information Cluster, School of Computer Science & Informatics,
University College Dublin, Belfield, Dublin 4, Ireland
{maurice.coyle,barry.smyth}@ucd.ie

Abstract. Collaborative Web search (CWS) is a community-based approach to Web search that supports the sharing of past result selections among a group of related searchers so as to personalize result-lists to reflect the preferences of the community as a whole. In this paper, we present the results of a recent live-user trial which demonstrates how CWS elicits high levels of participation and how the search activities of a community of related users form a type of social search network.

1 Introduction

The Web is evolving into a much more social place, with user-driven sites such as Wikipedia and Flickr and social networking sites such as FaceBook and Bebo[1] connecting users and facilitating communication in a community-oriented environment. At the same time, the field of Web search is also changing, with the traditional *one-size-fits-all* paradigm for result selection and ranking being abandoned in favour of a more personalized approach where, for example, *user profiles* store the preferences of each searcher and this profile information is reused in the future to formulate more specific queries [4] or re-rank the results returned by a search engine to reflect the profile contents [11]. Past user behaviour is also increasingly being used to inform Web search processes, with a *mass consensus*-style approach being used to determine which items within a repository are most relevant to a user's query [12]. Clickthrough data and other implicitly-collected user data have been shown to be useful for inferring global user preferences [1, 7] and for identifying useful query expansion terms [3].

Collaborative Web search (CWS) is a technique that combines both personalization and implicit feedback reuse with today's *social Web* ethos, operating at the level of a *search community* of users with overlapping search interests to generate focussed, relevant result rankings. CWS harnesses implicitly-collected *search knowledge* in the form of past queries and their associated result selections to enhance future search sessions by promoting and inserting previously-selected results. This ensures not only that users' natural searching behaviour is

[1] http://www.facebook.com, http://www.bebo.com

not interrupted but also that contributions towards the collective store of search knowledge are made by a large proportion of the community, which reduces the *participation inequality* often observed in user-driven online sites [8, 13].

In the past we have presented detailed accounts of the collaborative Web search approach [10], including a number of evaluations that have demonstrated the potential benefits of CWS at the level of the individual searcher. More recently we have begun to explore what might be termed the *social benefits* of CWS. For example, in [5] we focused on the hypothesis that much of the benefits of CWS were derived from the sharing of search histories among community members, in the sense that many users seemed to benefit from result promotions that come from the histories of *other* users. In fact we found that searchers more frequently selected promotions that came from the search histories of other community members, rather than their own. In this paper we present some evaluation results related to the search performance of CWS in addition to examining more of the social aspects of CWS. To this end we describe the *social search network* that evolves in a CWS setting as users forge and strengthen connections with other community members as a direct result of their search activities with emphasis on the participation levels across the community. We identify 2 different searching roles that emerge within this network: *search leaders* produce high quality search knowledge in the form of their result selections that are promoted and selected in future sessions, while *search followers* tend to select the promotions that have been derived from the search histories of other users.

The remainder of this paper is organised as follows. In Section 2 we will briefly describe the core CWS technique and mechanisms for result promotion. In Section 3, we present the results of a live-user trial, using the employees of a company as a search community. We will briefly highlight some performance benefits of CWS, along with an examination of participation levels within the community before presenting a visualisation of the CWS social search network that emerges, which illustrates graphically the social dimension of the technology.

2 Collaborative Web Search

Collaborative Web search (CWS) is a technique for personalizing the results returned by an existing Web search engine. Instead of maintaining individual preference profiles as some of the techniques mentioned in the introduction to this paper do, CWS maintains a community profile in which the contributions of individual users are unknown. In essence, CWS uses the implicitly-collected searching histories of a community of like-minded users to tailor the result selection and ranking processes. The computational details of CWS have been presented previously (see [5, 10]) and so in this section only a brief description of the core technique is provided.

For our purposes, a community may be defined as any ad hoc or structured group of searchers who share some set of interests; in this work we are not concerned with the precise origins of a community of searchers and only assume that such a community can be identified. Very briefly, given a target query

Fig. 1. For the query 'michael jordan', CWS promotes previously-selected results related to the Berkeley professor within a community of computer science researchers.

q_T^C submitted by some member of community C, CWS will identify a set of similar queries $\{q_1, ..., q_n\}$ previously submitted by the community; typically using a standard term-overlap query-similarity metric $Sim(q_T, q_i)$ [10]. Each similar query q_i is associated with a set of previously-selected results and the relevance score $Rel(p_j, q_T, q_1, ..., q_n)$ for each such result p_j can be calculated (see Equation 1) based on how often p_j has been selected for similar queries. Then the top ranking results can be promoted within a result-list that is returned by some underlying search engine; once again the details are purposefully light here and the interested reader is directed to [10] for a more detailed account of these relevance ranking and result promotion techniques.

$$Rel(p_j, q_T, q_1, ..., q_n) = \frac{\sum_{i=1...n}(\frac{H_{ij}}{\sum_{\forall j} H_{ij}}) \bullet Sim(q_T, q_i)}{\sum_{i=1...n} Exists(p_j, q_i) \bullet Sim(q_T, q_i)} \quad (1)$$

Figure 1 illustrates how the result-list returned by Google for the query *michael jordan* is re-ranked so that results that relate to the shared interests of a search community are promoted. Within a community of computer science researchers, results relating to the well-known Berkeley professor of artificial intelligence and machine learning have attracted selections in the past for similar queries and thus they are promoted ahead of results about the basketball star. Note that the promoted results are identified as such using graphical icons that summarise the result's interaction history (see [6] for more details and an evaluation of this explanation-oriented interface).

3 Evaluation

The objective of this trial was to evaluate the impact of CWS in a more realistic or natural search setting than previous evaluations (see [10]), involving live users over a significant period of time. To this end, the trial was conducted in conjunction with the 66 employees (ranging in age from their early 20s to early 50s) of a Dublin software company, who used the CWS system as their primary search front-end over a 17-week period. In addition to evaluating the baseline effectiveness of CWS we were particularly interested in exploring some of the social dynamics of the search community that evolves as a result of shared search behaviours.

3.1 Methodology

One of the key challenges in evaluating search technologies in a natural setting concerns the determination of relevance. Since we were unable to directly elicit feedback from the participants regarding the relevance of a particular result list, we used 2 indirect measures of search success which allow us to compare alternative search strategies in a systematic way. For our purposes, a *search session* is defined as a single query submission, including any result selections from the list returned by the underlying engine. We define a *successful session* as one in which at least one result was selected (i.e. the searcher found a search result which was at least *apparently* relevant to their search). In addition, we will consider the selection of the top result in a ranked list to be an indicator of success in so far as it tells us that a result which looked relevant enough for the user to select it was chosen by the ranking method as being the most relevant result; see [2] for an analysis of the importance of the top result. Thus, the percentage of sets of sessions which have the first result selected can be used as a means for comparing the success of different ranking functions.

It might be argued that the presence of explanation icons alongside promoted results (see Figure 1) might affect the selection probability of a result and thus skew the success rates of search sessions containing promotions. To control for this we disguise the promoted results in certain sessions by eliminating the explanation icons altogether. Thus we have 3 basic session types; a *standard session* (STD) is a search session for which CWS failed to identify any promotion candidates and so the default Google results were returned to the searcher, without modification; an *identified session* (ID) is one for which CWS made promotions and these promotions were annotated with appropriate explanation icons (see Figure 1); finally a *disguised session* (DISG) is identical to an identified session, except that promoted results were not annotated with explanation icons. Promotions were disguised for 20% of sessions containing promotions and this feature was not communicated to the trial participants.

It should also be noted here that although the core CWS technique requires no user identification to operate, for certain parts (see Sections 3.6 and 3.7) of this evaluation the identities of participants were extracted and logged.

3.2 Preliminary Observations

Over the 17 weeks a total of 20,448 search sessions were generated, covering a total of 15,977 result selections. The average query contained 2.73 terms, which is in line with the findings of Silverstein et al. [9]. One of the basic assumptions of this trial was that the participants would behave as a search community with broadly similar search interests, on the basis that the vast majority of their searches would be work-related and thus somewhat aligned to a shared set of business interests. In Figure 2, we graph the percentage of (stopword-stripped) queries whose terms overlap to various degrees with at least one other query and we see, for example, that more than 65% of queries share at least half of their terms with other queries (for this trial, a query similarity threshold of 0.5 was applied). We can also see, in Figure 2, how each query sharing half of its terms with at least one other query, actually shares half of its terms with 3 other queries, on average. These results indicate the trial participants search for similar information in similar ways on a regular basis.

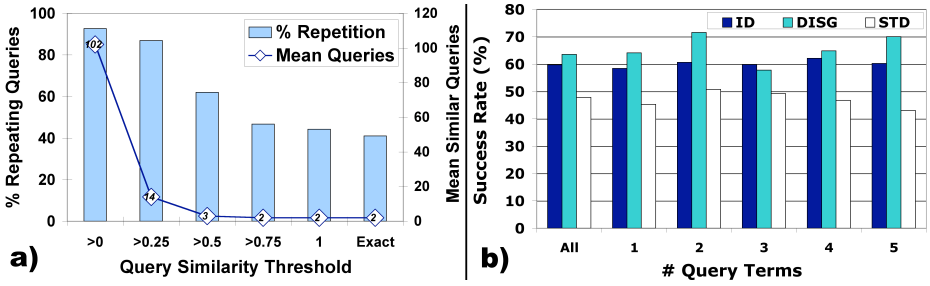

Fig. 2. a) Percentage of query repetition at various similarity thresholds and mean number of similar queries, b) The percentage of sessions with at least 1 result selected for identified (ID), disguised (DISG) and standard (STD) sessions.

3.3 Session Success Rates

To begin with, we compare the success rates of the three different types of session (standard, identified, and disguised) across all queries and also for queries of different lengths. The results are presented in Figure 2b) and clearly indicate a performance advantage for both types of promoted sessions (identified and disguised) compared to the standard (Google) sessions. For example, on average (across all query lengths — "All" in Figure 2b)) we found a 48% success rate for the standard sessions containing the default Google result-lists; this means that searchers failed to find any apparently relevant results in more than half of the standard Google sessions. In comparison, the success rates for promoted sessions are significantly higher (at the 99% confidence level), with success rates

of 60% and 64% reported for the disguised and identified sessions, respectively. Improvements of between 25% and 33% are found across different query lengths and once again all differences are statistically significant at the 99% confidence level.

It is worth noting that the success rates for disguised sessions are found to be higher than those for identified sessions for all but one of the query lengths considered. The increased success rates due to the disguised sessions may suggest a *Google bias* inherent in participant selection behaviour: the participants appear to be somewhat sceptical of the annotated CWS promotions in identified sessions and, all other things being equal, are more likely to select results from the more Google-esque disguised sessions; see also Section 3.4.

3.4 Ranking Success

Figure 3a) shows, for each session type, the percentage of successful sessions that had the top result selected. We can see that across all query lengths (i.e. the bars labelled *All* in Figure 3a)), successful identified and disguised sessions will have the top result selected 23% more frequently than successful standard sessions.

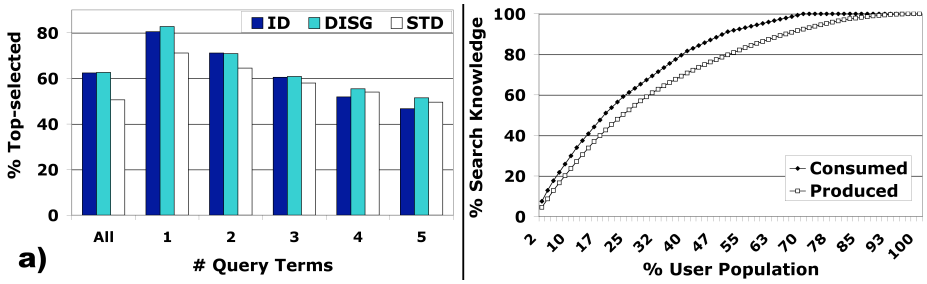

Fig. 3. a) Percentage of successful search sessions with top result selected for each session type, b) The percentage of the total number of selected promotions produced and consumed by different percentages of the user population.

When we examine the results across different query lengths, we see that the difference between promoted (i.e. identified and disguised) and standard sessions steadily decreases for longer queries; indeed for queries with 4 or more terms, standard sessions will have the top result selected more often than identified sessions[2]. This is an indication that within a CWS setting, performance is optimal for shorter queries and it is increasingly difficult to ensure that the *most* relevant result appears at the top as more terms are added to a query. We argue that this is an acceptable tradeoff in the context of Web search, since it has been shown that most user queries are of the order of 2-3 terms [9] and indeed for the trial described here, we found that over 75% of queries have 1, 2 or 3 terms.

[2] This could also be affected by the Google bias mentioned in Section 3.3

3.5 Discussion

One very important point to note before conclusions may be drawn regarding the performance metrics (i.e. success rates and top result selection) used here concerns the *nature* and ease of different search tasks. The presence of promotions in a search session correlates with a higher likelihood of at least one result selection. In addition, the top result in a promoted session (which, it should be noted, will always be a promotion by definition) is more likely to be selected that that in a standard session for shorter, more ambiguous queries. However, before firm conclusions can be drawn, further analysis into the characteristics of promoted and standard sessions is required to investigate whether search tasks that lend themselves to promotion (i.e. in which user queries overlap, enabling promotions to be made) are inherently *easier* for a search engine to satisfy for some reason.

In the context of the analysis of the social rather than the purely performance-based aspects of CWS as presented in this paper, the interested reader is directed to [5] for an investigation into the extent to which users select promotions that come from the search histories of other users above their own and the likelihood of promotions (when present) being selected over standard results.

3.6 Participation Levels in a CWS Search Community

One of the limitations of more traditional forms of user-generated content is that a relatively small number of users (i.e. in the order of 1-2% [8]) actively engage in content production. Such *participation inequality* [13] may be undesirable, since the few who contribute are unlikely to be representative of the views and opinions of the entire user base [8].

Before continuing with the discussions in this and the subsequent section, some basic definitions are required. A *search leader* is a user who is the first user to select a result which is not only promoted in a future search session, but also *selected*; that is, search leaders *produce* valuable search knowledge by selecting results that future searchers find useful. A *search follower* is a user who is presented with a result promotion which they then select; that is, they *consume* the search knowledge produced by previous users. Note that a search leader is the user who executes the *first click* on a result which is later promoted; all future selections of that result where it is promoted are examples of search knowledge consumption. Finally, since we are interested in the social nature of CWS, when examining a user's search history, we consider only search knowledge that is produced or consumed by community members other than the searcher themselves (the searcher's *peers*).

Figure 3b) shows the cumulative percentage of *first clicks* on promotions ascribed to different proportions of the test community along with the percentage of promotions consumed by increasing percentages of the community. We can see that 80% of valuable search knowledge is produced by 50% of the community. Also, the maximum contribution of any individual community member is just under 5%, and thus there are no users that dominate the production of search

knowledge. Similarly, we see a gradual increase in the percentage of promotions consumed by increasing percentages of users, with 80% of the search knowledge consumed by just under 38% of the community population. An examination of the most active producers and consumers finds that they share only 48% of their members, which is important because it highlights that the store of valuable search knowledge is not useful to only a small proportion of the community.

3.7 Search Relationships

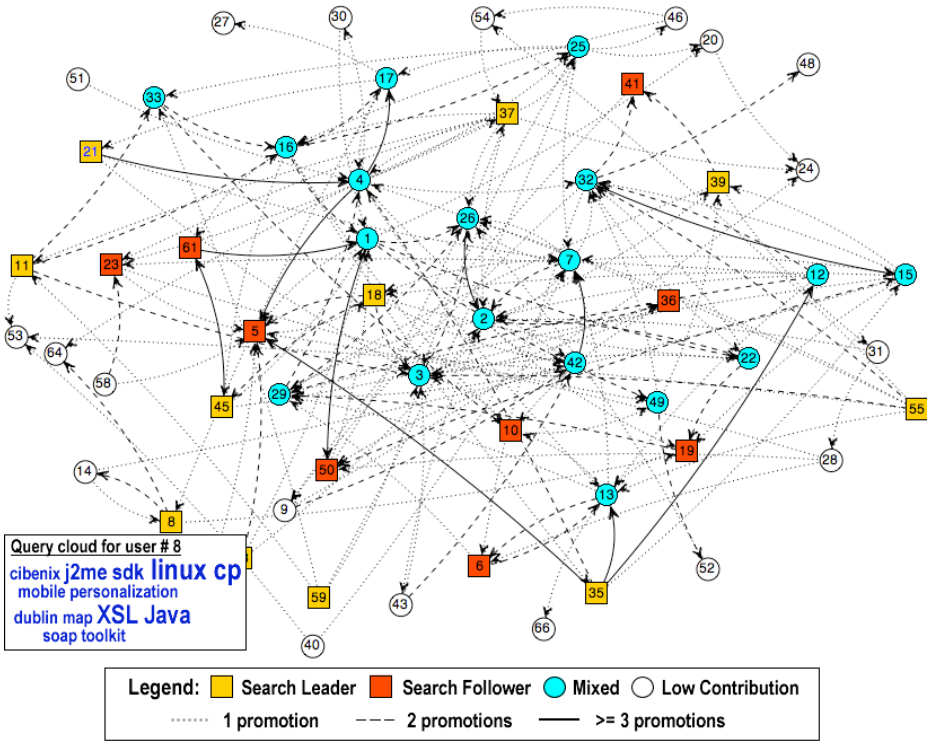

Fig. 4. The social search network that evolved from a CWS deployment within an organisational setting. The colour-coded nodes and weighted edges allow at-a-glance determination of search leaders and search followers within the community. This information can be used to improve knowledge management, expert identification and internal communications.

In this final section, we will attempt to graphically depict the relationships between search knowledge producers and consumers in more detail. Figure 4 contains a visualisation of the CWS social search network (generated by the JUNG

suite of network analysis tools[3]) which shows the extent to which community members interact and provides at-a-glance recognition of the most active search knowledge producers and consumers.

The vertices correspond to community members and search relationships are shown as directed arcs between the producer and consumer of search knowledge. The strength of each such relationship is encoded by the weight of the arcs, as indicated. Community members are coded by shape and colour according to whether they are *search leaders* or *search followers*, or a mixture of the 2. For example, the link from user 11 (a search leader) to user 33 (a search leader *and* a search follower) in Figure 4 indicates that user 11 has provided promotions that user 33 has subsequently selected twice (i.e. the arc is a heavy dashed line). This visualisation supports the quantitative analyses of previous sections, by showing that CWS fosters a social searching environment, in which people are actively creating and using search knowledge in a manner that is not prone to *cliques* or solely local sharing.

Figure 4 also demonstrates the ability of CWS to produce *query clouds* for community members which could be used in an application setting to identify search *experts* on various topics, thereby enhancing communication and knowledge sharing opportunities within the community.

4 Conclusions

Collaborative Web search (CWS) harnesses the search behaviour of a community of users in order to adapt the result lists of a conventional Web search engine so that they reflect collective community interests. In this paper we have presented the results of a comprehensive evaluation of CWS within a corporate search setting. We have highlighted how CWS has the potential to deliver significant performance increases, in terms of session success rates and top result selection, when compared with standard Google rankings.

A key contribution of this paper has been to highlight how the production and consumption of valuable search knowledge (i.e. promotions that are reused rather than simply presented) is shared right across the community so that no small subset of users dominates either activity. We believe this highlights the utility of CWS as a means for effective implicit relevance feedback collection without the drawbacks that some other more explicit user-driven online services suffer from.

We have also defined distinct search roles within communities in the form of search leaders (users who create valuable search knowledge) and search followers (users who reuse the search knowledge produced by others). A visualization of the social search network for the particular community used in our evaluation enables the identification of the most active search leaders and followers, while highlighting the social interactions that occur throughout the community.

Combined with the results presented in [5] in which the value of leveraging the search histories of *other* users was highlighted, we feel the work presented

[3] http://jung.sourceforge.net

in this paper supports the view of CWS as a social medium for sharing valuable search knowledge within a search community.

References

1. E. Agichtein, E. Brill, S. Dumais, and R. Ragno. Learning user interaction models for predicting web search result preferences. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '06)*, pages 3–10, 2006.
2. E. Agichtein and Z. Zheng. Identifying "best bet" web search results by mining past user behavior. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '06)*, pages 902–908, 2006.
3. B. Billerbeck, F. Scholer, H. E. Williams, and J. Zobel. Query expansion using associated queries. In *Proceedings of the 12th International Conference on Information and Knowledge Management (CIKM '03)*, pages 2–9, 2003.
4. P.-A. Chirita, C. S. Firan, and W. Nejdl. Summarizing local context to personalize global web search. In *Proceedings of the 15th ACM international conference on Information and knowledge management (CIKM '06)*, pages 287–296, Arlington, Virginia, USA, 2006.
5. M. Coyle and B. Smyth. Information Recovery & Discovery in Collaborative Web Search. In *Proceedings of the 29th European Conference on IR Research (ECIR '07)*, pages 356–367, 2007.
6. M. Coyle and B. Smyth. On the Community-Based Explanation of Search Results. In *Proceedings of the 10th International Conference on Intelligent User Interfaces (IUI '07)*, pages 282–285, 2007.
7. T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay. Accurately Interpreting Clickthrough Data as Implicit Feedback. In *Proceedings of the 28th International Conference on Research and Development in Information Retrieval (SIGIR '05)*, pages 154–161, 2005.
8. Jakob Nielsen. Participation inequality: Lurkers vs. contributors in internet communities, 2006. http://www.useit.com/alertbox/participation_inequality.html.
9. C. Silverstein, M. Henzinger, H. Marais, and M. Moricz. Analysis of a Very Large AltaVista Query Log. Technical Report 1998-014, Digital SRC, 1998. http://gatekeeper.dec.com/pub/DEC/SRC/technical-notes/abstracts/src-tn-1998-014.html.
10. 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: The Journal of Personalization Research*, 14(5):383–423, 2004.
11. J. Teevan, S. T. Dumais, and E. Horvitz. Personalizing search via automated analysis of interests and activities. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information retrieval (SIGIR '05)*, pages 449–456, 2005.
12. M. Truran, J. Goulding, and H. Ashman. Co-active intelligence for image retrieval. In *Proceedings of the 13th ACM International Conference on Multimedia*, pages 547–550, 2005.
13. S. Whittaker, L. Terveen, W. Hill, and L. Cherny. The dynamics of mass interaction. In *Proceedings of the 1998 ACM conference on Computer-supported Cooperative Work (CSCW '98)*, pages 257–264, 1998.