

Provided by the author(s) and University College Dublin Library in accordance with publisher policies. Please cite the published version when available.

Title	Remote asynchronous collaborative web search : a community-based approach
Author(s)	Smyth, Barry; Coyle, Maurice
Publication Date	2009-05
This item's record/more information	http://hdl.handle.net/10197/1890

Downloaded 2012-05-16T20:37:59Z

Some rights reserved. For more information, please see the item record link above.



Remote Asynchronous Collaborative Web Search

A Community-Based Approach

Barry Smyth

CLARITY: Centre for Sensor Web Technologies
School of Computer Science and Informatics
University College Dublin
barry.smyth@ucd.ie

Maurice Coyle

CLARITY: Centre for Sensor Web Technologies
School of Computer Science and Informatics
University College Dublin
maurice.coyle@ucd.ie

ABSTRACT

Recently researchers have argued that the prevailing view of web search, as a solitary activity, is flawed: that, in reality, web search is often an inherently collaborative task. In this paper we describe and evaluate an approach to collaborative web search that seeks to enhance mainstream search engines by harnessing the past search experiences of communities of like-minded searchers in order to adapt the result-lists of traditional search engines so that they reflect the niche interests of community members.

1. INTRODUCTION

The world of web search is usually viewed as a solitary place. Although millions of searchers use services like Google and Yahoo everyday, their individual searches take place in isolation, leaving each searcher to fend for themselves when it comes to finding the right information at the right time. Recently, researchers have begun to question the solitary nature of web search, proposing a more collaborative search model in which groups or users can cooperate to search more effectively. For example, studies in specialised information seeking tasks, such as military command and control tasks or medical tasks, have found clear evidence that search type tasks can be collaborative as information is shared between team members [14, 15, 17, 16].

Recent work by [9] highlights the inherently collaborative nature of more general purpose web search. For example, during a survey of just over 200 respondents, clear evidence for collaborative search behaviour emerged. More than 90% of respondents indicated that they frequently engaged in collaboration at the level of the *search process*. For example, 87% of respondents exhibited "back-seat searching" behaviours, where they watched over the shoulder of the searcher to suggest alternative queries. A further 30% of respondents engaged in search coordination activities, by using instant messaging to coordinate searches. Furthermore, 96% of users exhibited collaboration at the level of *search products*, that is, the results of searches. For example, 86% of respondents shared the results they had found during searches with others by email. Indeed almost 50% of respondents telephoned colleagues directly to share web search results, while others prepared summary docu-

ments and/or web pages in order to share results with others.

Thus, despite the absence of explicit collaboration features from mainstream search engines there is clear evidence that users implicitly engage in many different forms of collaboration as they search, although, as reported by [9], these collaboration "work-arounds" are often frustrating and inefficient. Naturally, this has motivated researchers to consider how different types of collaboration might be supported by future editions of search engines. The resulting approaches to *collaborative information retrieval* can be usefully distinguished in terms of two important dimensions, *time* and *place*. In terms of the former, collaborative search systems can be designed to support *synchronous* or *asynchronous* collaborative search. And in terms of the latter, systems can be designed to support either *co-located* or *remote* forms of collaborative search.

Co-located systems offer an collaborative search experience for multiple searchers at a single location, often a single PC (e.g. [1]) or, more recently, by taking advantage of computing devices that are more naturally collaborative, such as table-top computing environments (e.g. [19]). In contrast, remote approaches allow searchers to perform their searches at different locations across multiple devices; see e.g. [10, 11, 21]. While co-located systems enjoy the obvious benefit of an increased faculty for direct collaboration that is enabled by the face-to-face nature of co-located search, remote services offer a greater opportunity for collaborative search.

Synchronous approaches are often characterised by systems that broadcast a "call to search" in which specific participants are requested to engage in a well-defined search task for a well defined period of time; see e.g. [18]. In contrast, asynchronous approaches are characterised by less well-defined, ad-hoc search tasks and provide for a more open-ended approach to collaboration in which different searchers contribute to an evolving search session over an extended period of time; see e.g. [10, 20].

In this paper we will focus on a community-based approach to collaborative web search in which the *asynchronous* search experiences of communities of like-minded *remote* searchers are harnessed to provide an improved search experience that is more responsive to the learned preferences of a community of searchers. We will describe how this approach can be integrated with a mainstream search service and summarise the results of a recent live-user trial that serve to highlight the some of the end-user benefits of this approach to collaborative information retrieval.

2. MOTIVATIONS

The *one-size-fits-all* approach adopted by conventional search engines is one area where there is significant room for improvement. The vague queries that are commonplace in web search do little to distinguish the real information needs of the searcher, for example, and recent advances in personalization technology speak

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

WOODSTOCK '97 El Paso, Texas USA

Copyright 200X ACM X-XXXXX-XX-X/XX/XX ...\$5.00.

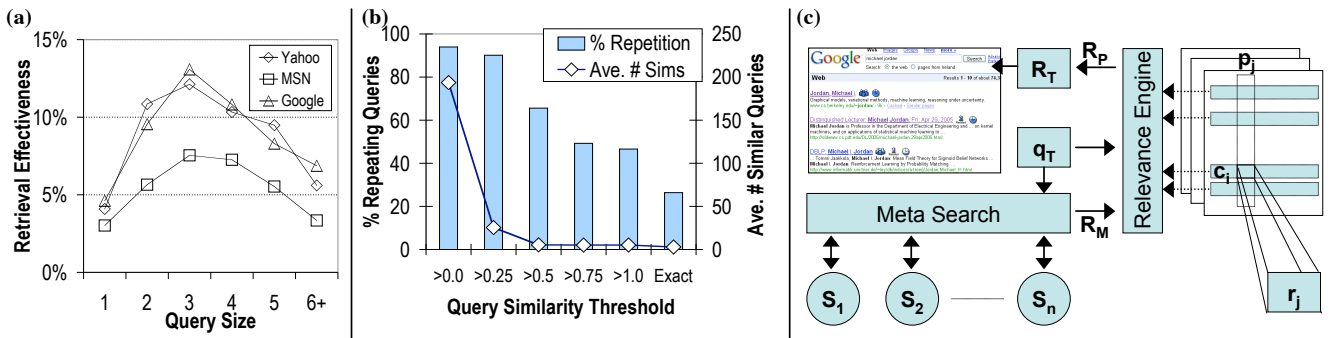


Figure 1: (a) Query size vs search engine effectiveness for the leading commercial search engines; (b) Repetition and regularity rates within a corporate search community; (c) The collaborative Web search architecture.

to the possibility of delivering a more personalized search experience in the future; see for example, [13, 8, 22]. Here we will focus on one particular approach to *community-based* personalization, and in this section we review our so-called collaborative Web search (CWS) technique, which attempts to exploit the natural repetition and regularity that exists in communities of searchers.

2.1 The Vocabulary Gap in Web Search

Early Web search engines adopted a term-based document-centric view of search that reflected their information retrieval origins. Then, in the late 1990's, researchers looked at the relationships between inter-linking documents, with Google's PageRank famously winning out as the 'best practice' in document ranking [6]. There is no doubt that these advances have served Web searchers well, but there remains considerable room for improvement. Today's failed searches are largely due to the mismatch between the query-space of the searcher and the document-space of the search engine index: users are prone to submit vague queries that often contain terms that are different from those used to index documents. The result is a significant *vocabulary gap* that leads to the poor search performance we find today.

To better understand the scale of this vocabulary gap, we recently submitted just under 7,700 queries to the three leading search engines (Google, Yahoo, and MSN) to locate a particular target page for each query; admittedly a particularly tough measure of relevance, but one that was straightforward to objectively measure. We evaluated the effectiveness of each search engine in terms of the percentage of times that the target page was retrieved in the top-ten results returned. The results, (Figure 1(a)) highlight how all 3 search engines struggle to perform, at best retrieving the target results in their top-ten less than 14% of the time. Indeed it is interesting to note how all three search engines perform best for queries with 3 terms, suggesting that modern search engine technology has been optimized for typical query lengths [20]. Importantly, we can also see how retrieval effectiveness increases, as query size grows from 1 to 3 terms, supporting the view that search engine performance can be improved if users provide more detailed queries. However, this is true only to a point. For queries with more than 3 terms we see a decline in retrieval effectiveness because oftentimes these extra terms are less than helpful when it comes to identifying a target document; for example, users frequently chose very specialised terms that do not even occur in the target document.

2.2 Repetition & Regularity in Web Search

There are many scenarios in which search can be viewed as

a community-oriented activity. For example, the employees of a company will act as a type of search community with overlapping information needs. Similarly, students in a class may serve as a search community as they search for information related to their class-work. Visitors to a themed web site (e.g., a wildlife portal or a motoring portal) will tend to share certain niche interests and will often use their site's search facilities to look for related information. And of course, groups of friends on a social networking site may act as a community with shared interests.

We became interested in these emergent *search communities* because of the potential for patterns to exist between the search behaviours of community members. For example, Figure 1(b) shows the results of a 17-week study of the search patterns for 70 employees of a local software company. During the study we examined more than 20,000 individual search queries and almost 16,000 result selections. On average, just over 65% of queries shared at least 50% (> 0.5 similarity threshold) of their query terms with at least 5 other queries; and more than 90% of queries shared at least 25% of their terms with about 10 other queries. In other words, searchers within this ad hoc corporate search community do search for similar things in similar ways, much more so than in generic search scenarios, where we typically find much lower repetition rates of about 10% at the 0.5 similarity threshold [20].

This is an important result, which is supported by similar studies on other communities of searchers [20]. It tells us that, in the context of communities of like-minded searchers, the world of web search is a repetitive and regular one. A type of community search knowledge is generated from the search experiences of individuals as they search. This suggests that it may be possible to harness this search knowledge by facilitating the sharing of search experiences among community members. So, as a simple example, when a visitor to the previously mentioned wildlife portal searches for "jaguar pictures" they can be recommended search results that have been previously selected by other community members for *similar* queries. These results will likely relate to the wildlife interests of the community and so, without any expensive processing of result content, we can personalize search results according to the learned preferences of the community. In this way, novice searchers can benefit from the shared knowledge of more experienced searchers.

3. COLLABORATIVE WEB SEARCH

The basic CWS architecture is presented in Figure 1(c). Briefly, when a new target query, q_T , is submitted, in the context of some community, result-list, R_T , is produced from the combined results of the underlying search engines (S_1, \dots, S_n), R_M , plus promoted

results (R_P) chosen because they have been previously selected by community members for queries that are similar to the target.

CWS adopts a case-based reasoning perspective [2]. The search history of a given community is stored as a case-base of search cases with each search case made up of a specification part and a solution part; see Equation 1. The *specification* part (see Equation 2) corresponds to a given query. The *solution* part (see Equation 3) corresponds to a set of selection-pairs; that is, the set of page selections that have been accumulated as a result of past uses of the corresponding query. Each selection-pair is made up of a result-page id and a hit-count representing the number of times that the given page has been selected by community members in response to the given query.

$$c_i = (q_i, (p_1, r_1), \dots, (p_k, r_k)) \quad (1)$$

$$Spec(c_i) = q_i \quad (2)$$

$$Sol(c_i) = ((p_1, r_1), \dots, (p_k, r_k)) \quad (3)$$

Given a new target query, q_T , CWS must identify a set of *similar* search cases from the community's search case-base. A standard term-overlap metric (Equation 4) is used to measure query-case similarity, to rank-order past search cases according to their target similarity, so that all, or a subset of, the similar cases might be reused during result ranking.

$$Sim(q_T, c_i) = \frac{|q_T \cap Spec(c_i)|}{|q_T \cup Spec(c_i)|} \quad (4)$$

Consider a page, p_j , that is associated with query, q_i , in some search case, c_i . The relevance of p_j to c_i can be estimated by the relative number of times that p_j has been selected for q_i ; see Equation 5.

$$Rel(p_j, c_i) = \frac{r_j}{\sum_{\forall r_m \in Sol(c_i)} r_m} \quad (5)$$

Then, the relevance of p_j to some new target query q_T can be estimated as the combination of $Rel(p_j, c_i)$ values for all cases c_1, \dots, c_n that are deemed to be similar to q_T , as shown in Equation 6. Each $Rel(p_j, c_i)$ is weighted by $Sim(q_T, c_i)$ to discount the relevance of results from less similar queries; $Exists(p_j, c_i) = 1$ if $p_j \in Sol(c_i)$ and 0 otherwise.

$$WRel(p_j, q_T, c_1, \dots, c_n) = \frac{\sum_{i=1 \dots n} Rel(p_j, c_i) \bullet Sim(q_T, c_i)}{\sum_{i=1 \dots n} Exists(p_j, c_i) \bullet Sim(q_T, c_i)} \quad (6)$$

This weighted relevance metric, $WRel$, is used to rank-order search results from the community case-base that are promotion candidates for the new target query. The top ranked candidates are then listed ahead of the standard meta-search results to give R_T ; see [7] for further details on the search interface and result presentation.

4. EVALUATION

So far we have described the CWS technique for adapting the results of an existing search engine(s) to conform to the preferences of a community of searchers. In this section we will describe the results emerging from a recent CWS trial in a corporate context, pointing out how CWS helped employees to search more successfully as a result of the sharing of community search knowledge.

4.1 Trial Setup

The trial included 70 employees from a local Dublin software company where CWS was deployed for an initial period of 10

weeks. During the trial all requests for Google were redirected to the CWS server. From a user perspective, the standard Google interface was adapted to accommodate CWS promotions; users saw their familiar Google results page with results promoted and re-ranked as appropriate by CWS. During the trial approximately 25% of search sessions included CWS promotions. We refer to these as *promoted* sessions. The remaining 75% of search sessions carried the standard Google result-list. We refer to these as *standard* sessions. While it was not permitted to capture direct relevance feedback from trial participants, one useful indicator of search performance is to look at the frequency of so-called *successful sessions*.

4.2 Promoted vs Standard Sessions

A search session is successful if at least one result is selected by the searcher. This is a very crude measure of performance — often result selections are good indicators of at least partial relevance but sometimes they are not — but the lack of any result selections is a good indication that no relevant results have been noticed. We found marked differences between the promoted and standard sessions. For example, Figure 2(a) shows an average success rate of $\sim 48\%$ for standard Google searches, compared to a success rate of $\sim 62\%$ for promoted sessions; a relative advantage due to CWS promotions of approximately 29%. In other words, when community promotions were available they were found to help users search more successfully.

4.3 Sharing Search Experiences

Sharing is a key to CWS: past search experiences are shared by community members through result promotions. A promotion may come from the past history of the current searcher: today I might search using a query that is similar to queries that I have used in the past, and I will receive promotions based on my own previous selection history. We call these *self promotions* and they are useful when it comes to helping searchers to *recover* results that they have previously encountered. On the other hand, a promotion may come from a different community member altogether. We call this a *peer promotion*, and when I receive peer promotions I am sharing the search experiences of other community members. Peer promotions are especially useful when it comes to helping me *discover* new results, and they potentially help me to draw on the experiences of more informed searchers within the community.

Figure 2(b) presents a summary analysis of the origin of promotions and their associated success rates. For example, we see that promoted sessions containing only self promotions have an average success rate of just under 60%. By comparison, sessions that contain only peer promotions have a success rate of about 66% while *mixed* sessions, containing both self and peer promotions, have an average success rate of more than 70%. Clearly searchers do benefit from the search experiences of others in their community. In fact, when we look at the how frequently sessions with promotions from a given source lead to at least 1 of these promotions being selected (Figure 2(c)), we find that sessions containing peer promotions have significantly higher 'click-thru' rates than sessions containing only self promotions: click-thru rates of 60-70% for peer promotion compared to only 30% for self promotions.

5. CONCLUSIONS

There are a number of contributions associated with this work. Firstly, we have clarified how naturally occurring communities of users tend to search in similar ways for similar things, motivating the reuse of their search experiences as a way to influence future result-lists as part of a collaborative approach to web search. Secondly, we have developed an approach to reusing these experiences

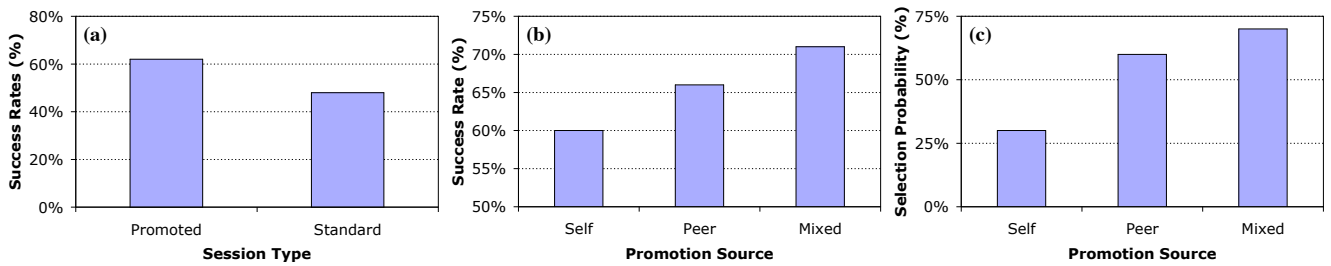


Figure 2: Key results: (a) success rates by session; (b) success rates by promotion type; (c) click-thru rates by promotion type.

that borrows heavily from case-based reasoning, thereby accommodating a very flexible approach to experience reuse. Thirdly, we have demonstrated the value of this approach over an extended period of time in a live-search scenario.

As our research has taken shape, new opportunities have emerged to extend the basic CWS concept. For example, the work of [7] looks at how useful search communities can be automatically identified and how their promotions can be combined to deliver further improvements in search quality. Another issue concerns the susceptibility of CWS to spamming by malicious users. In this regard, recent research [12] quantifies how CWS offers some level of protection from such malicious users. Moreover, the work of [5] explains how a model of user-trust can be incorporated into CWS to offer further protection from malicious users by weighting community promotions according to the trustworthiness of the community members who originally selected them. Finally, it is worth highlighting the work of [3], which looks at profiling result selections according to their snippet terms, with significant improvements in promotion quality accruing to this richer representation format. In addition, this extended approach to CWS has led to a novel approach to generating community-based result summaries which have been shown to offer better precision/recall characteristics than more conventional summarisation techniques; see [4].

6. REFERENCES

- [1] S. Amershi and M. R. Morris. Cosearch: a system for co-located collaborative web search. In *CHI*, pages 1647–1656, 2008.
- [2] E. Balfe and B. Smyth. Case-based collaborative web search. In *ECCBR*, pages 489–503, 2004.
- [3] O. Boydell and B. Smyth. Enhancing case-based, collaborative web search. In *Proceedings of the International Conference on Case-Based Reasoning*, pages 329–343. Springer, 2007.
- [4] O. Boydell and B. Smyth. From social bookmarking to social summarization: an experiment in community-based summary generation. In *Intelligent User Interfaces*, pages 42–51. ACM, 2007.
- [5] P. Briggs and B. Smyth. On the role of trust in collaborative web search. *AI Review*, 25(1-2):97–117, 2006.
- [6] S. Brin and L. Page. The Anatomy of A Large-Scale Web Search Engine. In *Proceedings of the Seventh International World-Wide Web Conference*, 2001.
- [7] J. Freyne and B. Smyth. Cooperating Search Communities. In *Proceedings of the 4th International on Adaptive Hypermedia and Adaptive Web-based Systems*, Dublin, Ireland, 2006.
- [8] F. Liu, C. Yu, and W. Meng. Personalized Web Search for Improving Retrieval Effectiveness. *IEEE Transactions on Knowledge and Data Engineering*, 16(1):28–40, 2004.
- [9] M. R. Morris. A survey of collaborative web search practices. In *CHI*, pages 1657–1660, 2008.
- [10] M. R. Morris and E. Horvitz. S³: Storable, shareable search. In *INTERACT (1)*, pages 120–123, 2007.
- [11] M. R. Morris and E. Horvitz. Searchtogether: an interface for collaborative web search. In *UIST*, pages 3–12, 2007.
- [12] M. P. O’Mahony and B. Smyth. Collaborative web search: A robustness analysis. *AI Review*, Forthcoming, 2008.
- [13] F. Qiu and J. Cho. Automatic identification of user interest for personalized search. In *WWW ’06: Proceedings of the 15th international conference on the World Wide Web*, pages 727–736, New York, NY, USA, 2006. ACM Press.
- [14] M. C. Reddy and P. Dourish. A finger on the pulse: temporal rhythms and information seeking in medical work. In *CSCW*, pages 344–353, 2002.
- [15] M. C. Reddy, P. Dourish, and W. Pratt. Coordinating heterogeneous work: Information and representation in medical care. In *ECSCW*, pages 239–258, 2001.
- [16] M. C. Reddy and B. J. Jansen. A model for understanding collaborative information behavior in context: A study of two healthcare teams. *Inf. Process. Manage.*, 44(1):256–273, 2008.
- [17] M. C. Reddy and P. R. Spence. Collaborative information seeking: A field study of a multidisciplinary patient care team. *Inf. Process. Manage.*, 44(1):242–255, 2008.
- [18] A. F. Smeaton, C. Foley, D. Byrne, and G. J. F. Jones. ibingo mobile collaborative search. In *CIVR*, pages 547–548, 2008.
- [19] A. F. Smeaton, H. Lee, C. Foley, and S. McGivney. Collaborative video searching on a tabletop. *Multimedia Syst.*, 12(4-5):375–391, 2007.
- [20] 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.
- [21] B. Smyth, P. Briggs, M. Coyle, and M. P. O’Mahony. Google? shared! a case-study in social search. In *User Modeling, Adaptation and Personalization*. Springer-Verlag, June 2009.
- [22] 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, New York, NY, USA, 2005. ACM Press.