# A Community-Based Approach to Personalizing Web Search

Barry Smyth
University College Dublin

**Researchers can leverage the latent knowledge created within search communities by recording users' search activities—the queries they submit and results they select—at the community level. They can use this data to build a *relevance model* that guides the promotion of community-relevant results during regular Web search.**

O ver the past few years, current Web search engines have become the dominant tool for accessing information online. However, even today's most successful search engines struggle to provide high-quality search results: Approximately 50 percent of Web search sessions fail to find any relevant results for the searcher.

The earliest Web search engines adopted an information-retrieval view of search, using sophisticated term-based matching techniques to identify relevant documents from repeated occurrences of salient query terms. Although such techniques proved useful for identifying a set of potentially relevant results, they offered little insight into how such results could be usefully ranked.

How then should documents be ranked and ordered? Some researchers[1,2] solved this problem when they realized that ranking could be greatly improved by evaluating the importance or authoritativeness of a particular document. By analyzing the links in and out of a document, it became possible to evaluate its relative importance within the wider Web. For example, Google's famous PageRank metric assigns a high page-rank score to a document if it is itself linked to by many other documents with a high page-rank score, and it iteratively evaluates the page-rank scores for every document in its index for use during results ranking.

Other researchers began exploring alternative ranking options. One notable alternative, implemented in the Direct Hit search engine, argued that search results should be ranked by their popularity among searchers. All other things being equal, results with higher click-through rates should be preferred during ranking. Unfortunately, Direct Hit's so-called *popularity engine* did not play a central role on the modern search stage (although the technology does live on as part of the Teoma search engine) largely because the technology proved inept at identifying new sites or less well-traveled ones, even though they may have had more contextual relevance to a given search.

Despite Direct Hit's fate, the notion that searchers themselves could influence the ranking of results by virtue of their search activities remained a powerful one. It resonates well with the ideas that underpin the *social web*, a phrase often used to highlight the importance of a suite of technologies and ideas that see users playing a more active role in Web content creation and management. From blogs to wikis, social networks to tagging, the social web emphasizes the importance of *community*, *participation*, and *sharing* when it comes to the creation, organization, and dissemination of Web content.

These ideas have influenced research into how the search behavior of communities of like-minded users can be harnessed and shared to adapt the results of a conventional search engine according to the needs and preferences of a particular community. Ideally, this leads to an improved personalized search experience that can deliver more relevant result pages that reflect the experiences of a community of users, effectively forming a collective search wisdom.

At its heart, this *collaborative Web search* (CWS) approach promotes the idea that community search

activities can provide a valuable form of search knowledge and that facilitating the sharing of this knowledge between individuals and communities makes it possible to adapt traditional search-engine results according to the community's needs. Several natural community-based search scenarios motivate this work, and recent evaluation results speak to the potential of the collaborative Web search approach.

## PERSONALIZING WEB SEARCH

The one-size-fits-all approach typified by conventional search-engine technologies shows room for improvement. The vague queries that are commonplace in Web search do little to distinguish the searcher's real information needs, while recent advances in areas such as user profiling and personalization suggest potential solution strategies capable of delivering more relevant, personalized search experiences. In this context there have been numerous recent developments and practical applications of personalized search using several different approaches.

> The collaborative Web search approach promotes the idea that community search activities can provide valuable search knowledge, and sharing this knowledge makes adapting traditional search-engine results possible.

For example, one common approach seeks to leverage the search histories of individual users to personalize future search sessions. One study[3] introduces a technique that constructs a client-side index from all of the documents that a user creates, copies, or employs on a client machine. This index is treated as a type of *user profile* intended to disambiguate the user's search query terms and to improve result relevance by reranking relevant documents within search results. As users manipulate certain documents, this technique assesses the material as being of greater or lesser relevance to the user's needs and assigns a relative importance to terms used in the search.

A different approach leverages an individual user's search history in combination with a general profile gleaned from the Open Directory Project (http://dmoz.org).[4] The user's search history is mapped to a set of content categories drawn from the ODP. These categories then serve as a source of preference or context terms for future queries.

Other researchers[5,6] use search-selection histories to choose a *topic-sensitive PageRank* value for each returned search result, which is then used to rank those results. Previously selected search results serve as biased indicators of user interest; each page in the engine's index has a number of PageRank values calculated for it, one for each top-level category in the ODP. At query time, the search engine accesses the target user's stored history to select an appropriate PageRank value for each result, depending on the user's preferences. The engine then uses these PageRank values to rank the returned results so that they reflect any potential topic bias in the user's interests.

These exemplars show just a small subset of the ongoing research in the Web search personalization area. Indeed, leading commercial search engines Google (www.google.com/psearch) and Yahoo! (http://myweb2.search.yahoo.com) recently have undertaken related initiatives, which have offered their own particular demonstrations of personalized search.

## THE VAGUE QUERY PROBLEM

Web search represents a significant technological challenge. The Web's size and growth characteristics, and the sheer diversity of offered content types, represent formidable information-retrieval challenges in their own right. At the same time, as the demographics of the Web's user base continue to expand, search engines must be able to accommodate an increasingly diverse range of user types and skill levels. In particular, most users fail to live up to the expectations of the document-centric, term-based information-retrieval engines that lie at the heart of modern search technology. These engines, and the techniques they rely upon, largely assume well-formed, detailed search queries. But such queries are far from common in Web search today. Instead, most Web search queries are vague or ambiguous with respect to the searcher's true information needs, and queries often contain terms not reflected in the target documents.

For example, consider a search query for a common term such as "jaguar." Querying Google shows an emphasis on the importance of car-related meanings for this query, with nine of the top 10 results linking to car-related products and services. Only one of the top 10 results links to pages that relate to the wild cat. Other interpretations, from the NFL football team to Apple's OS X operating system, appear much further down the list.

A query such as "jaguar" is inherently vague, offering a search engine such as Google little insight into the searcher's intention. Nevertheless, such queries are commonplace, with many researchers noting that a typical Web query contains only two or three terms.[7,8] Certainly, if searchers need data on the NFL or the operating system, they will be disappointed with Google's first page of results and, at best, must continue their search to locate relevant results further down the listing.

Typically, developers respond to examples such as this by declaring that users must be taught to provide more meaningful and detailed queries. Although this makes perfect sense, it is not the perfect solution. Many users
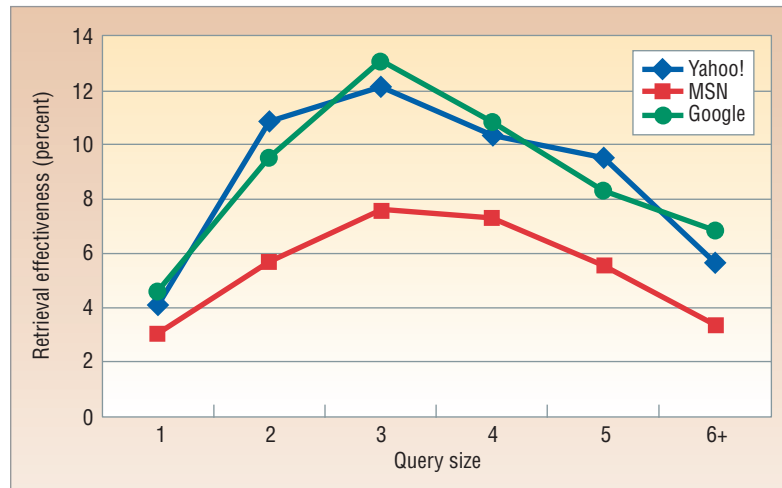
Figure 1. Retrieval effectiveness versus query size. The search engines performed poorly overall, at best retrieving the target results in their top 10 search results less than 14 percent of the time.
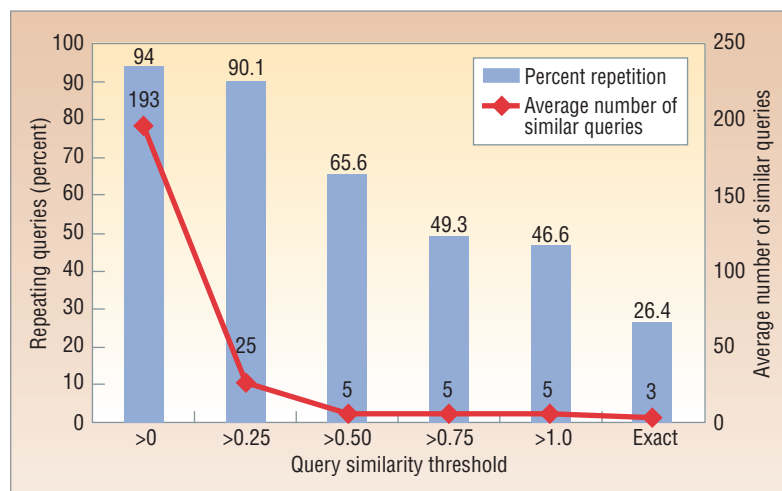


Figure 2. Search patterns. Results of a 17-week study of the search patterns for a set of about 70 employees at a local software company, showing the percentage of query repetition at various similarity thresholds and the mean number of similar queries.

will continue to provide vague queries. Recent evidence suggests that even when users do provide additional query terms, they might not select the types of terms that will help a search engine understand their needs. Essentially, a *vocabulary gap* exists because users will sometimes select terms that are not even present in their desired results. For example, researchers recently submitted just under 7,700 queries to the three leading search engines—Google, Yahoo!, and MSN—in an effort to locate a particular target page for each query. They estimated the effectiveness of each search engine in terms of the average percentage of times users retrieved the target page within the top 10 results returned.

Figure 1 shows these results as a graph of retrieval effectiveness against query size. The search engines performed poorly overall, at best retrieving the target results in their top 10 search items less than 14 percent of the time. However, this is a particularly tough measure of relevance. It is also clear that both Google and Yahoo! perform consistently better than MSN across all query sizes. All three search engines perform best for queries with three terms, suggesting that modern search-engine technology has been optimized for typical query lengths.

Retrieval effectiveness increases rapidly as query size increases from one to three terms, which supports the idea that encouraging users to provide more detailed queries can improve search-engine performance. However, this is true only to a point. For queries longer than three terms, a gradual decline in retrieval effectiveness occurs that appears to relate to the additional terms users add to their queries. More often than not, these extra terms offer search engines little help when it comes to identifying their target document, and users frequently choose very specialized terms that do not even occur in the target document.

## REPETITION AND REGULARITY IN SEARCH COMMUNITIES

It is common practice to think about search as an isolated single-user activity, one that relies on the services of a generic search engine. In reality, there are many scenarios in which search can be viewed as a type of community activity. For example, consider a wildlife information portal designed to provide users with access to a host of wildlife-related resources. The portal pages also host several search boxes so that visitors can easily initiate standard Web searches as they browse; this is common practice with all the main search engines. Visitors to this portal constitute an ad hoc community with a shared interest in wildlife. All other things being equal, searches originating from this portal will more likely be wildlife-related—a fact that the search engines providing these search boxes typically ignore—but that this research seeks to exploit as a means of improving the quality of subsequent result lists.

Many other examples of naturally occurring *search communities* exist. For example, the employees of a small- or medium-size company, or a group in a larger multinational, or even a class of students, might each constitute a search community with individuals searching for similar information in similar ways. Indeed, with the advent of social networking services, thousands of
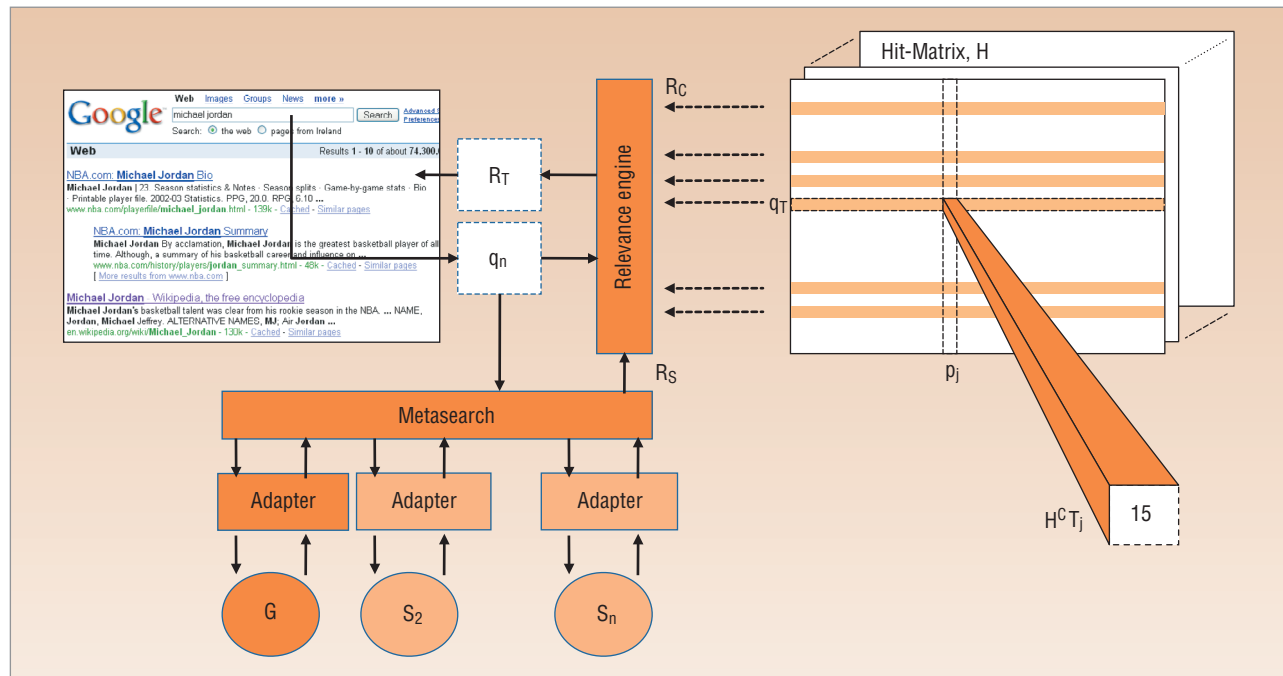
Figure 3. Metasearch architecture. The collaborative Web search architecture provides for a form of metasearch with the result of one or more underlying search engines augmented by community-based result promotions.

more structured communities of friends with shared interests emerge daily. These emergent search communities are interesting because of the high likelihood that similarities will exist among community members' search patterns. For example, Figure 2 shows the results of a 17-week study of the search patterns for a set of about 70 employees at a local software company. This study examined more than 20,000 individual search queries and almost 16,000 result selections.

Figure 2 looks at the average similarity between queries during the study. On average, just over 65 percent of submitted queries shared at least 50 percent (0.5 similarity threshold) of their query terms with at least five other queries; more than 90 percent of queries shared at least 25 percent of their terms with at least 25 other queries. Thus, searchers within this ad hoc corporate search community seemed to search in similar ways, much more so than in generic search scenarios, which typically show lower repetition rates of about 10 percent at the 0.5 similarity threshold.

This result, supported by similar studies of other search communities,[9] shows that, in the context of communities of like-minded searchers, Web search is a repetitive and regular activity. As individuals search, their queries and result selections constitute a type of community search knowledge. This in turn suggests that it might be possible to harness such search knowledge by facilitating the sharing of search experiences among community members.

As a simple example, when visitors to the wildlife portal search for "jaguar pictures," the collaborative search engine can recommend search results that other community members have previously selected for *similar* queries. These results will likely relate to the community's wildlife interests. So, without any expensive processing of result content, the search results can be personalized according to the community's learned preferences. This lets novice searchers benefit from the shared knowledge of more experienced searchers.

## COLLABORATIVE WEB SEARCH

The latent search knowledge created by search communities can be leveraged by recording the search activities of users—the queries they submit and results they select—at the community level. This data can then be used as the basis for a *relevance model* to guide the promotion of community-relevant results during regular Web search. A key objective here is to avoid replacing a conventional search engine, instead enhancing its default result lists by highlighting particular results that are especially relevant to the target community. For example, regarding the "jaguar" queries, it should be possible to promote some of the wildlife community's pages that relate to the wild cat ahead of those related to cars. Thus, the most relevant results from a community perspective can be promoted to the top of the default result list, while other results might simply be labeled as relevant to the community but left in place.

### How it works

To achieve a more granular relevance scale, the metasearch architecture shown in Figure 3 operates in cooperation with one or more underlying search engines. For

simplicity, assume Google is the single underlying conventional search engine.

For example, consider a user $u$ (as a member of some community $C$) submitting a query $q_T$. In the first instance $q_T$ is submitted to Google to obtain a standard set of ranked results, $R_S$. In parallel, $q_T$ is used to query the community's search knowledge base to produce another set of ranked results, $R_C$, judged to be especially relevant to members of $C$ based on their past search behavior.

Next, $R_S$ and $R_C$ are combined to produce a final results list, $R_T$, which is presented to the user. These result lists can be combined in many different ways. One strategy has worked well in practice: The top three results in $R_C$ are promoted to the top positions in $R_T$ with all other $R_C$ results retaining their default position in $R_S$ but being labeled as community-relevant.

### Capturing community search knowledge

Capturing a community's search behavior means recording the queries submitted and the results selected for these queries, as well as their selection frequency. This can be conceptualized as populating the community search matrix, $H^C$, called a *hit-matrix*, such that $H^C_{ij}$ refers to the number of times that a result page, $p_j$, has been selected for a query, $q_i$. Thus, each row of a community's hit-matrix corresponds to the result selections that have been made over multiple search sessions by members of $C$ for a specific query $q_i$. In turn, the column of the hit-matrix related to $p_j$ refers to the number of times that the community has selected $p_j$ for different queries.

### Making relevant promotions

How then can the current query, $q_T$, be used to identify results from a community's hit-matrix as potential promotion candidates? To begin with, any previous community history with respect to $q_T$ must be determined—have any pages been selected in the past for $q_T$? Assuming such pages exist, the hit-matrix will contain frequency selection information with respect to $q_T$, and this information can be used to estimate the relevance of each such page. For example, Equation 1 calculates the relevance of a result page $p_j$ with respect to the query $q_T$ as the relative proportion of selections that $p_j$ has received for this query:

$$Relevance^C(p_j, q_T) = \frac{H^C_{Tj}}{\sum_{\forall j} H^C_{Tj}} \quad (1)$$

As it stands, this exact query-relevance approach is limited because it restricts candidates considered for promotion to those pages previously selected for the specific target query ($q_T$). Certainly, the results shown in Figure 2 indicate that just over 25 percent of query submissions in the test community exactly match previous submissions. A more flexible approach would allow for the consideration of pages that have been selected for queries very similar to $q_T$. For example, Equation 2 provides a straightforward way to calculate query similarity by counting the proportion of terms shared by $q_T$ and some other query $q_i$.

$$Sim(q_T, q_i) = \frac{|q_T \cap q_i|}{|q_T \cup q_i|} \quad (2)$$

This query-similarity metric can then be used as the basis for a modified relevance metric, as Equation 3 shows:

$$W\,Rel^C(p_j, q_T, q_1, ... q_n) =$$
$$\frac{\sum_{i=1...n} Relevance^C(p_j, q_i) \bullet Sim(q_T, q_i)}{\sum_{i=1...n} Exists^C(p_j, q_i) \bullet Sim(q_T, q_i)} \quad (3)$$

The relevance of a page $p_j$, with respect to some target query $q_T$, is computed by independently calculating the exact query relevance of $p_j$ with respect to a set of queries $(q_1, ..., q_n)$ deemed to be sufficiently similar to $q_T$; in practice only queries that share 50 percent of their terms with the target query need be considered. The overall relevance of $p_j$ with respect to $q_T$ is then the weighted sum of the individual exact query relevance values, with the relevance of $p_j$ with respect to some $q_i$ discounted by the similarity of $q_i$ to $q_T$. In this way, pages frequently selected for queries very similar to $q_T$ are preferred over pages less frequently selected for less similar queries.

### Sample session

Figures 4 and 5 show an example of collaborative Web search in action. Figure 4 shows the results of a standard Google search for the vague query "O2," which refers to the European mobile operator. These results clearly target the average searcher by providing access to nearby stores, pricing plans, and various company information sites. In contrast, the results shown in Figure 5 correspond to the results returned by a collaborative Web search for a community made up of the employees of a local mobile software company. This time, the top three results have been promoted for this community. They target more specialized information that has proven to be of recent interest to community members for this and similar queries. These promoted results are annotated with several community icons to reflect their popularity, the number of related queries associated with the result, and the recency of the community history.

### PRACTICAL BENEFITS

The CWS technique for adapting a conventional search engine's results to conform with the preferences of a particular community of searchers reveals that these communities take many different forms. These range from ad hoc communities that arise from users visiting a

themed Web site to more structured communities such as those formed by a company's employees or a class of students. Some of the results emerging from a recent study of CWS in a corporate context show how it helped employees search more successfully as a result of sharing community search knowledge.

The trial participants included approximately 70 employees from a Dublin software company that deployed CWS for 10 weeks as the primary search engine covering more than 12,600 individual search sessions. During the trial all Google requests were directed to the CWS server and the standard Google interface was adapted to accommodate CWS promotions and annotations, as Figure 5 shows.

During this initial 10-week trial, approximately 25 percent of search sessions included CWS promotions, referred to as *promoted sessions*. The remaining 75 percent carried the standard Google result list, referred to as standard sessions.

While eliciting direct relevance feedback from trial participants proved infeasible, one useful indicator of search performance looked at the frequency of *successful sessions*. A search session is successful if the searcher selects at least one result—an admittedly crude measure of performance. Result selections can be good indicators of at least partial relevance, but not always. However, the lack of any result selections indicates that no relevant results have been noticed.

When researchers analyzed the success rates of trial search sessions, they found marked differences between the promoted and standard sessions. For example, this analysis shows an average success rate of just under 50 percent for standard Google searches, compared to a success rate of just over 60 percent for promoted sessions—a relative advantage of approximately 25 percent directly attributable to CWS promotions. Thus, community promotions made by collaborative Web search helped users to search more successfully.
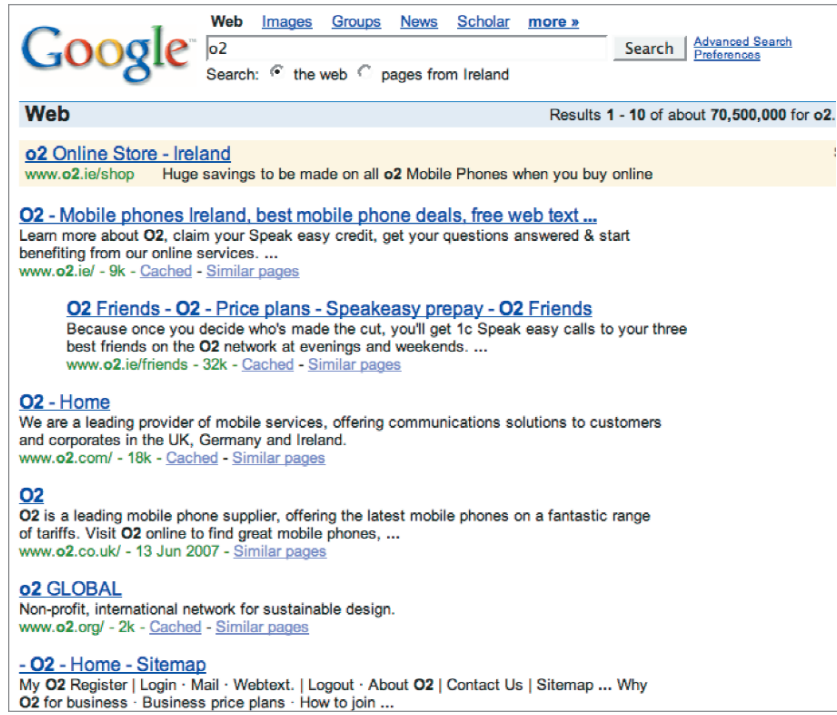


Figure 4. Standard vague query. Example of a one-size-fits-all search session for a vague query, "O2," which refers to the European mobile operator. The results returned clearly target the average searcher by providing access to nearby stores, pricing plans, and various company information sites.
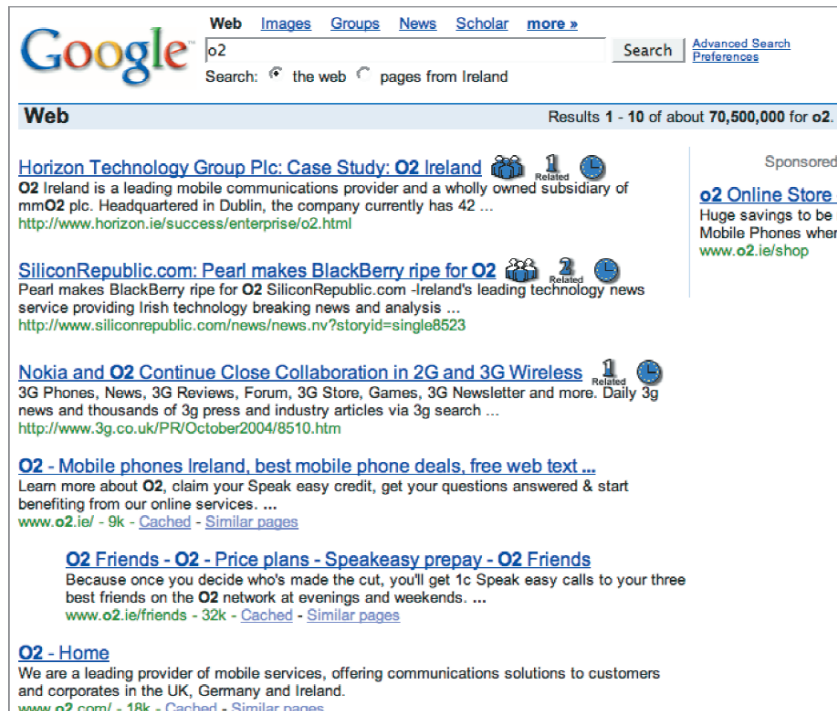


Figure 5. Collaborative Web search. A search session personalized for the preferences of a particular community of searchers who work for a software company involved in developing mobile services and applications. The top three results have been promoted for this community and target more specialized information proven to be of interest to community members for this and similar queries.
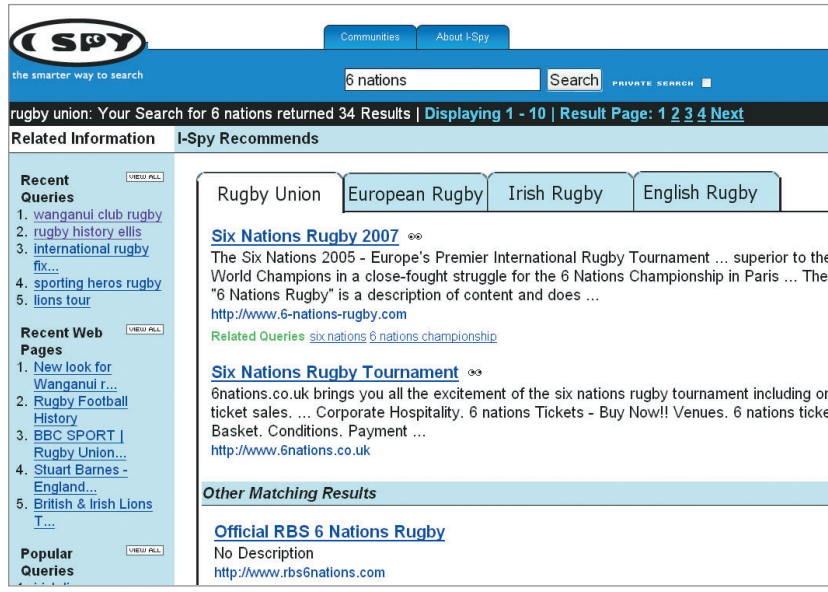
**Figure 6. Intercommunity collaboration facilitates the promotion of search results from multiple communities. The example shown presents rugby-related results for the host Rugby Union community.**



**Figure 7. Example of the results promoted to the Rugby Union community from the related Irish rugby community.**

Sharing is an important theme in collaborative Web search: Community members share past search experiences through result promotions. These promotions can come from two different sources:

- *The current searcher's past history.* One search might, for example, use a query similar to queries used in the past, which will gather promotions based on the user's previous selection history. These *self-promotions* are useful when helping searchers recover previously encountered results.

- *A different community member's past history.* When users receive such *peer promotions*, they share the search experiences of other community members. These promotions are especially useful for helping users discover new results, and they potentially help draw on the experiences of more informed searchers within the community.

While CWS does not store information about the individual searcher by default, during the trial, reconstructed information about the origins of promotions was used to investigate differences between users' behavior when it came to sessions made up of self- and peer promotions. This analysis generated revealing results.

For example, promoted sessions made up only of self-promotions have an average success rate of just under 60 percent. By comparison, sessions made up of peer promotions have a success rate of about 66 percent, while mixed sessions, made up of both self- and peer promotions, have an average success rate of more than 70 percent. This demonstrates that searchers do benefit from the search experiences of others within their community. Further analysis looked at how frequently sessions containing promotions from a given source led to those promotions being selected. Sessions containing peer promotions have higher click-through rates than sessions containing only self-promotions: a 60 to 70 percent click-through rate compared to only 30 percent for self-promotions.

At this trial's start it was not obvious how well participants would serve as a coordinated search community. For example, would their search activities break into small clusters of related activity, or would many individuals search in ways markedly different from their peers and not participate in creating or consuming search knowledge?

The trial's results showed that more than 85 percent of the participants became involved in the creation and consumption of search knowledge. About 20 percent of searchers behaved primarily as *search leaders* in the sense that many of their searches corresponded to discovery tasks in which there was little or no community

search knowledge to draw from. A similar percentage of searchers played the contrary role of *search followers*: These users generally searched on topics already well-known and thus benefited disproportionately from peer promotions. The remaining 60 percent of users displayed a mixture of roles, often producing new search knowledge by selecting fresh results and consuming existing search knowledge by selecting promotions.

## BEYOND THE COMMUNITY

Given the benefits of creating and sharing search knowledge within the community, the implications of cooperation between related communities of searchers must be considered. For example, a search community servicing the needs of skiers in Europe might benefit from promotions derived from the community search knowledge generated by a separate community of US skiers. A query for "late ski deals" by a member of the European community would likely be answered by promotions for the latest deals offered by European ski resorts. At the same time, the searcher might benefit from hearing about the latest snow conditions and special deals in the US, knowledge that would be better represented by the US ski community.

This idea has been explored in the context of the I-SPY search system,[10] a separate implementation of collaborative Web search that lets users easily create and deploy their own search communities. Figure 6 shows a screen shot of a result-list that has been generated for a member of one of several rugby-related communities, the Rugby Union. The query submitted is for "6 nations," a popular international rugby tournament, and the promoted results for the community appear ahead of other matching results provided by the underlying search engine. In addition, the screen shot also includes a set of search tabs, each containing the promotions from a community related to Rugby Union. Figure 7 presents the promotions from the Irish rugby community, which provide a different set of results for the "6 nations" query, results more appropriate for Irish rugby fans.

Related communities can be identified and their promotions ranked during search by, for example, ranking communities according to their similarity to the *host community*—the community where a particular search originated.[10] Intercommunity similarity can be calculated based on the overlap between the results that have been selected between two different communities. For example, Rugby Union more closely resembles Irish rugby than a Manchester United community because the Irish rugby community will share many similar results with Rugby Union, which is unlikely in the case of Manchester United. In this way, a ranked set of similar communities can be produced, and those generating the most relevant results can be recommended to the host community as shown. The relevance of a result from a related community can be scored in the usual way, but further discounted by the related community's similarity to the host.

This technique offers two potentially important benefits. First, the related communities can provide an alternative source of interesting results, thereby improving the relevance and coverage of the results offered to the user. Second, partitioning the results according to their community provides a novel form of results clustering that does not rely on a detailed and computationally expensive analysis of a larger results set. Instead, each related community forms a coherent cluster from a results presentation perspective.

The collaborative approach to Web search offers a further advantage that many traditional approaches fail to provide: The vast majority of approaches to personalized search focus on the individual's needs and as such maintain individual user profiles. This represents a significant privacy issue because users' search activities can be revealing, especially if a third party maintains the profiles.[11,12]

In contrast, CWS avoids the need to maintain individual user profiles. The engine stores preferences at the community level, thereby providing individual users with access to an anonymous form of personalized search. In an increasingly privacy-conscious world, CWS can provide an effective balance between the user's privacy on the one hand and the benefits of personalization on the other. ∎

## References

1. S. Brin and L. Page, "The Anatomy of a Large-Scale Web Search Engine," *Proc. 7th Int'l World Wide Web Conf.* (WWW 01), ACM Press, 2001, pp. 101-117.
2. J.M. Kleinberg, "Authoritative Sources in a Hyperlinked Environment," *J. ACM*, vol. 46, no. 5, 1999, pp. 604-632.
3. J. Teevan, S.T. Dumais, and E. Horvitz, "Personalizing Search Via Automated Analysis of Interests and Activities," *Proc. 28th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval* (SIGIR 05), ACM Press, 2005, pp. 449-456.
4. F. Liu, C. Yu, and W. Meng, "Personalized Web Search for Improving Retrieval Effectiveness," *IEEE Trans. Knowledge and Data Engineering*, vol. 16, no. 1, 2004, pp. 28-40.
5. F. Qiu and J. Cho, "Automatic Identification of User Interest for Personalized Search, *Proc. 15th Int'l World Wide Web Conf.* (WWW 06), ACM Press, 2006, pp. 727-736.

6. T.H. Haveliwala, "Topic-Sensitive PageRank: A Context Sensitive Ranking Algorithm for Web Search," *IEEE Trans. Knowledge and Data Engineering,* vol. 15, no. 4, 2003, pp. 784-796.

7. S. Lawrence and C.L. Giles, "Accessibility of Information on the Web," *Nature*, vol. 400, no. 6740, 1999, pp. 107-109.

8. A. Spink et al., "Searching the Web: The Public and Their Queries," *J. Am. Soc. Information Science*, vol. 52, no. 3, 2001, pp. 226-234.

9. B. Smyth et al., "Exploiting Query Repetition and Regularity in an Adaptive Community-Based Web Search Engine," *User Modeling and User-Adapted Interaction: The Journal of Personalization Research*, vol. 14, no. 5, 2004, pp. 383-423.

10. J. Freyne and B. Smyth, "Cooperating Search Communities," *Proc. 4th Int'l Conf. Adaptive Hypermedia and Adaptive Web-Based Systems*, Springer-Verlag, 2006, pp. 101-110.

11. M. Teltzrow and A. Kobsa, "Impacts of User Privacy Preferences on Personalized Systems: A Comparative Study," *Designing Personalized User Experiences in eCommerce*, Kluwer Academic, 2004, pp. 315-332.

12. Y. Wang and A. Kobsa, "Impacts of Privacy Laws and Regulations on Personalized Systems," *Proc. CHI 2006 Workshop on Privacy-Enhanced Personalization*, 2006, pp. 44-45.

*Barry Smyth is the Digital Chair of Computer Science, School of Computer Science and Informatics, University College Dublin. His research interests include case-based reasoning, user modeling, and recommender systems with particular focus on personalization techniques. Smyth received a PhD in computer science from Trinity College Dublin. Contact him at barry.smyth@ucd.ie.*