 User Modeling and User-Adapted Interaction (2004) 14: 383–423 
DOI 10.1007/s11257-004-5270-4

# Exploiting Query Repetition and Regularity in an Adaptive Community-Based Web Search Engine

BARRY SMYTH, EVELYN BALFE, JILL FREYNE, PETER BRIGGS, MAURICE COYLE and OISIN BOYDELL
*Adaptive Information Cluster, Smart Media Institute, University College Dublin, Dublin 4, Ireland. e-mail: barry.smyth@ucd.ie*

(Received: 8 June 2004; accepted in revised form: 1 December 2004)

**Abstract.** Search engines continue to struggle with the challenges presented by Web search: vague queries, impatient users and an enormous and rapidly expanding collection of unmoderated, heterogeneous documents all make for an extremely hostile search environment. In this paper we argue that conventional approaches to Web search – those that adopt a traditional, document-centric, information retrieval perspective – are limited by their refusal to consider the past search behaviour of users during future search sessions. In particular, we argue that in many circumstances the search behaviour of users is repetitive and regular; the same sort of queries tend to recur and the same type of results are often selected. We describe how this observation can lead to a novel approach to a more adaptive form of search, one that leverages past search behaviours as a means to re-rank future search results in a way that recognises the implicit preferences of communities of searchers. We describe and evaluate the I-SPY search engine, which implements this approach to collaborative, community-based search. We show that it offers potential improvements in search performance, especially in certain situations where communities of searchers share similar information needs and use similar queries to express these needs. We also show that I-SPY benefits from important advantages when it comes to user privacy. In short, we argue that I-SPY strikes a useful balance between search personalization and user privacy, by offering a unique form of anonymous personalization, and in doing so may very well provide privacy-conscious Web users with an acceptable approach to personalized search.

**Key words.** Meta search, personalization, social search, Web search

## 1. Introduction

Conservative estimates of the Web's current size speak about its 10 billion documents and a growth rate that tops 60 terabytes of new information per day (Roush, 2004). To put this into perspective, in 2000 the entire World-Wide Web consisted of just 21 terabytes of information. Now it grows by three times this figure every single day (Lyman and Varian, 2003). This makes for an extremely challenging information retrieval environment to begin with. In addition, however, Web searchers rarely produce high quality queries; typically they are vague or

ambiguous. For instance, the average query contains only about two query terms (Lawrence and Giles, 1998) and it is well known that people use a wide variety of terms to refer to the same types of information (Furnas et al., 1987). As a result, the terms used by searchers in their queries may not be found within the documents that they are searching for, and this lack of a strong correspondence between the query-space and the document-space introduces a major problem for the term-based matching approaches that form the kernel of modern search engines (Bollmann-Sdorra and Raghavan, 1993).

### 1.1. TERMS, QUERIES AND CONTEXT

In response, researchers have begun to look beyond the traditional IR perspective that has informed Web search in the main. In the recent past, for example, players such as Google (Brin and Page, 1998) argued for the need to consider factors such as link-connectivity information (in addition to the more traditional IR term-based factors) as a way to guide search towards more informative documents, and now many commercial search engines use similar ideas; see also Kleinberg (1998). More recently researchers have also focused on ways to exploit context in Web search as a way to disambiguate vague queries, either by explicitly establishing the search context up-front or by implicitly inferring the context as part of the search process (see Lawrence, 2000). Yet others have focused on leveraging knowledge about the query-space to directly address the correspondence problem, by mining query logs in order to identify useful past queries that may help the current searcher (e.g., Fitzpatrick and Dent, 1997; Glance, 2001; Raghavan and Sever, 1995; Wen and Zhang, 2002).

### 1.2. SOCIAL AND PERSONAL

Currently, most of the main Web search engines adopt a one size fits all approach to search – two users with the same query receive the same result-list, regardless of their preferences or context – and while there is broad agreement that this is far from optimal, real developments towards *practical* personalization techniques, that are both capable of coping with Internet-scale search tasks, and that are likely to be acceptable to today's privacy conscious users, have been slow to emerge.

However, one area of research that may have the potential to offer a practical form of personalization comes from recent work that has considered the intersection between social networking and Web search. Social networking applications such as *Friendster* (www.friendster.com) or *Orkut* (www.orkut.com) allow users to create, maintain and participate in online communities, and provide a range of applications and services to help these communities socialise more effectively on- and off-line. Inevitably, the members of a given community will all share certain characteristics and interests. For example, the members of a caving and potholing community will all have an interest in caving and potholing activities, obviously

enough, but they are likely to share a range of peripheral preferences too; a general interest in outdoor activities, they may be regular travelers, etc. The point is that these well-defined communities of like-minded individuals provide a backdrop for *social search* applications that are sensitive to the needs and preferences of a specific community of users operating within a well-defined topical domain (see also (Terveen et al., 1997) for related ideas in the recommender systems area).

### 1.3. COLLABORATIVE WEB SEARCH

In this paper we describe the evolution of a novel approach to Web search that relies on past search histories in order to re-rank search results for the needs of communities of users. It implements a powerful model of adaptive search and forms the basis of the I-SPY project, which aims to develop personalized search technologies that take advantage of the social aspects of community searching. In particular we describe the evolution of the I-SPY system over the past 4 years, by presenting two detailed case-studies: one describing I-SPY's original *collaborative search* technique and the results of a live-user trial, and the second describing how this technique has been recently extended and evaluated.

## 2. Background

Fundamentally our research is motivated by the need for better ways to cope with the vague queries that are commonplace in Web search. With this in mind, it is useful to consider our work from two different perspectives. On the one hand we take the view that the prevalence of vague queries leads to a basic mismatch between the query-space and the document-space, thus breaking a golden rule of information retrieval. This observation has led a number of researchers to look to the query-space as a source of retrieval knowledge and relevancy information. On the other hand, one can view the essence of the vague query problem to be one of absent context, a view that has led other researchers to develop context sensitive search techniques that seek to supplement vague queries with additional contextual retrieval cues. In this section we will consider both of these strands of research as a backdrop to our own work.

### 2.1. MINING THE QUERY-SPACE

A basic assumption of our work is that in Web search the gap between the query-space and the document-space can be especially wide. Generally, there is poor overlap between query terms and document terms and these query terms may change and evolve over time. We believe that, where in the past information retrieval research has focused largely on the document as the fundamental source of index terms, now, and in the context of Web search, we must also look

towards the queries themselves as a new and more dynamic source of indexing knowledge.

### 2.1.1. *The Correspondence Problem in Web Search*

The work of Furnas et al. (1987) highlights the so-called *vocabulary problem* and the surprising degree of variety with which people refer to the same thing, which implies that the terms that might be used in a query cannot be relied upon to match the terms used to index a given document based on its content. In related work Bollmann-Sdorra and Raghavan (1993) argue that the structure of the query-space can be very different from the structure of the document-space and that it makes sense to consider documents and queries to be elements from different term spaces. Indeed Cui et al. (2002) attempt to quantify the differences between the document-space and the query-space by measuring the typical similarities between document vectors and query vectors by using two months worth of query logs from the Microsoft Encarta search engine. They found very low average similarities between queries and document vectors, with the vast majority of similarities less than 0.4 (a cosine similarity metric was used) and a mean similarity value of only 0.28 across the collection as a whole.

Any lack of correspondence between the query-space and the document-space is likely to be amplified in Web search for a number of reasons. Firstly, it is well known that Web search queries tend to be short and vague, often with an average length of no more than two or three query terms (Croft et al., 1995; Lawrence and Giles, 1998). This leads directly to a low degree of overlap between query and document terms. At the same time, the Web is a very dynamic and largely unmoderated information environment without strong authorship controls, and this makes it very difficult for Web search engines to accurately and reliably index the ad-hoc content that makes up a significant portion of the Web. In turn, Web content is sensitive to changing trends and fashions and this can mean that documents that were originally produced for one purpose, and indexed accordingly, might later become relevant in an entirely different context, one that is not reflected by their index terms, in the future.

### 2.1.2. *Query-Log Analysis*

Recently, a number of researchers have begun to explore the possibility of leveraging knowledge about the query-space to improve search performance. For the most part this involves the mining of query logs in order to identify useful past queries that may help the current searcher. For instance, Raghavan and Sever (1995) propose the reuse of past optimal queries, stored in a query base, to either respond to new queries or to help formulate optimal queries. They highlight the importance of query-query similarity metrics and argue that existing query-document

metrics are inappropriate in this context. New user queries can be matched against persistent queries stored in the query base and the results of these prior successful queries can be recommended to the user if sufficient query similarity is detected.

A related idea has been subsequently exploited in Web search by Fitzpatrick and Dent (1997) and Glance (2001). The former describes how the use of past queries can improve automatic query expansion by using automatic feedback from the top documents returned in a result-list. The basic hypothesis explored was that the top documents retrieved by a query are themselves the top documents retrieved by past similar queries and are therefore a good source of data for automatic query expansion. In general the use of past queries to drive query expansion was found to deliver improved precision-recall performance. Glance (2001) describes a related community search assistant which enables communities of searchers to search in a collaborative fashion by using a form of query recommendation. Briefly, a query graph is constructed at runtime, connecting related queries by virtue of overlaps in their result-lists. Related queries are recommended to a searcher based on their degree of relatedness to the their original query. Thus, new queries are suggested to searchers rather than their existing queries being expanded.

Wen and Zhang (2002) describe how query logs can be mined to discover relevancy information that can be used during query clustering, which is especially useful in the identification of FAQs. Their basic principles include: recognising similar queries by virtue of the fact that these queries led to the same document selection behaviour; and recognising correspondences between document terms and query terms in the cases where a set of documents is frequently selected for the same queries. Their results indicate that this use of query logs can help to cluster queries more effectively than by using term-overlap approaches alone.

### 2.2. CONTEXT-BASED SEARCH

Context-based search techniques try to remedy the vague-query problem by adding new terms to queries in an effort to identify the user's context or their personal preferences. Context information can be explicitly provided by the user or search engine, or it can be implicitly inferred from the local search environment.

#### 2.2.1. *Capturing Explicit Context*

Perhaps the simplest way to capture explicit user context is to ask users to provide context terms as part of their search query. For example, Inquirus 2 (Glover et al., 2001) asks users to select from a set of categories such as 'research paper', 'homepage' etc. and uses the selected context categories to choose target search engines for the user's query; as such, Inquirus 2 is a meta-search engine. The category information can also be used for query modification (e.g., a query for research papers on '*web search*' might be modified to include terms such as '*references*').

At the time of writing, Google had just launched its prototype personalized search service (labs.google.com/personalized), designed to use information about a user's personal preferences to guide search. It works by asking users to explicitly indicate their preferences by selecting from a range of subject categories, and Google uses these terms to re-rank a limited set of search results (the first page of results only), promoting results that relate to these categories.

A second option for introducing explicit context into Web search is to use a specialised search engine whose index has been designed to cover a restricted information domain (eg., www.profusion.com, www.MP3.com, www.ireland.com etc.), essentially fixing the context prior to searching. Some specialised search engines automatically maintain their indexes by using information extraction techniques to locate and index relevant content (see Kushmerick, 1997). Good examples include CiteSeer (Lawrence and Giles, 1999), for searching scientific literature, and DEADLINER (Kruger et al., 2000), for conference and workshop information. For example, CiteSeer crawls the Web looking for scientific articles in a variety of document formats (HTML, PDF, Postscript etc.) and builds an index that is well suited to literature searching.

### 2.2.2. *Inferring Implicit Context*

Since many users are unwilling to provide explicit context information alternative approaches are needed. What if context could be automatically inferred? This question is being answered by a wide range of research focusing on different techniques for capturing different types of context. In fact two basic approaches have become popular depending on whether *external* or *local* context sources are exploited.

Users rarely perform searches in isolation. It is much more likely that the search will be related to some other task that they are currently performing. Perhaps they are reading a Web page, replying to an email, or writing a document when they need to search for some associated piece of information. By taking advantage of a user's activity immediately prior to the search it may be possible to determine a suitable search context. This is the goal of systems such as Watson (Budzik and Mammond, 2000), the Remembrance Agent (Rhodes and Starner, 1996), IntelliZap (Finkelstein et al., 2001) and Letizia (Lieberman, 1995).

Watson and the Remembrance Agent provide just-in-time information access by deriving context from everyday application usage. For example, as a Watson user edits a document in Microsoft Word, or browses in Internet Explorer, Watson attempts to identify informative terms in the target document by using a heuristic term-weighting algorithm. If the user then searches with an explicit query, Watson modifies this query by adding these newly derived terms. IntelliZap's search is initialised by a text query marked by the user in the document that he/she is viewing, and the search is then guided by the text in the local region of the marked query terms; this local text serves as a definition of the user's implied context. Similarly, Letizia analyses the content of Web pages that the user is currently browsing,

extracting informative keywords using similar term-weighting heuristics, and proactively searches out from the current page for related pages. In this sense, Letizia is more of a browsing assistant than a search assistant but it does exploit context in a similar manner; incidentally, Watson can also operate in this mode by continually searching the Web for related documents based on query terms extracted from the current document that the user is working on.

Other researchers have proposed a method to use categories from the Open Directory Project (ODP) (www.dmoz.org) as a source of context to guide a topic-sensitive version of PageRank (Haveliwala, 2002). Briefly, the URLs below each of the 16 top-level ODP categories are used to generate 16 PageRank vectors that are biased with respect to each category. These biased vectors are used to generate query-specific importance scores for ranking pages at query-time that are more accurate than generic PageRank scores. Similarly, for searches performed in context (e.g., when a user performs a search by highlighting words in a Web page), context-sensitive PageRank scores can be computed based on the terms and topics in the region of the highlighted terms.

The above refer to the use of external sources of context. Techniques also exist for the exploitation of local sources of context by using the results of a search as the basis for context assessment, extracting useful context terms that can then be used to supplement the user's original query. Typically, these context terms are those terms that are highly correlated in the initial search results. For example, the technique proposed by Mitra et al. (1998) extracts correlated terms from the top-ranking search results to focus context on the most relevant search results as opposed to the entire set. This idea of using the local search context can be extended beyond a single search episode. Many users will perform a sequence of searches on a specific topic and their response to the results can provide valuable context information. Thus, by monitoring and tracking queries, results and user actions it may be possible to model search context over an extended search session or even across multiple search sessions. For example the SearchPad system extracts context information, in the form of useful queries and promising result-lists, from multiple search sessions (Bharat, 2000). Similarly, the CASPER search engine for job advertisements, maintains client-side user profiles that include job cases that users have liked and disliked in previous searches, and these profiles are used to classify and re-rank the results of future searches (Bradley et al., 2000). CASPER can learn that a given user is interested in Dublin software-engineering jobs that require more than 5 years experience because in the past they have liked job cases in the Dublin region and consistently avoided jobs with lower experience requirements.

Of course, although methods that attempt to infer context information automatically may help matters, in the sense that at least the searcher does not have to perform additional work at search time, they do little to alleviate the privacy issue that is often associated with context-based or personalized search. Asking users to provide anything close to personal information is liable to alienate

them because of legitimate privacy concerns. And systems such as Watson, that are seen to monitor the user as they conduct other tasks, such as writing an email or a letter, may alienate users even more (see Kobsa, 2002 and also Section 6.7).

## 3. Repetition and Regularity in Web Search

Two key ideas about Web search inform our research: *query repetition* and *selection regularity*. First, we assume that the world of Web search is a repetitive place: similar queries tend to recur. Second, we assume that the world of Web search is a regular place: searchers tend to select similar results for similar queries. If these assumptions hold, we believe that significant performance benefits can be realised by reusing past search histories, but under what conditions do they hold? In this section we will provide evidence to suggest that the first of these assumptions does hold to a lesser or greater extent across a variety of Web search scenarios, and later in Section 4.4 we will show that the second assumption also holds in many of these scenarios.

In order to assess the degree of repetition among Web search queries we will make use of five very different sets of query logs for different types of search task (Jansen et al., 1998; Ozmutlu et al., 2000; Spink and Bateman, 1998) – general Web search using the Excite search engine (General), image search (Image), a more specialised topical search task (Nutrition), a focused fact-finding search task as part of a live-trial discussed in Section 4.4 (Live-Trial), and a second example of general Web search but this time from a small software development company of 50 people over a short (6 week) period of time (CW) – see Table I.

As a first step we consider query duplication by computing the percentage of queries for which there is an exact duplicate in each of the search logs. The results are shown in Figure 1. They indicate that in general Web search duplicate queries are relatively rare, accounting for just over 15% of the Excite queries and a similar proportion of the CW queries. However, in the more specialised search scenarios, represented by Image, Nutrition, and the focused Live-Trial duplicate queries are far more common. In these logs duplicate queries account for approximately 55–60% of the queries.

*Table I.* Search log use during query repetition analysis

| Name | Number of queries | Search scenario |
|---|---|---|
| General | 65,535 | General search using Excite.com |
| Image | 33,478 | Image search |
| Nutrition | 16,008 | Specialised search in the Nutrition domain |
| Live-Trial | 1705 | Fact-finding search task (see Section 4.4) |
| CW | 7696 | Local software company search log |

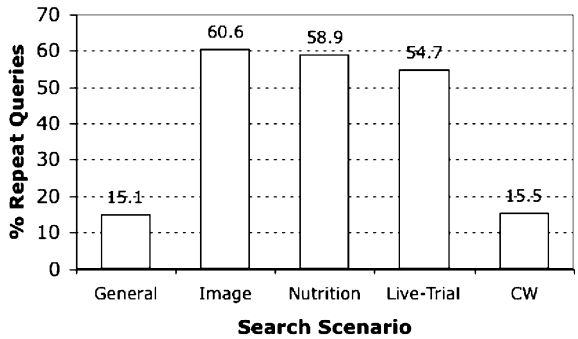


*Figure 1.* The percentage of queries that are exact duplicates in the General, Image, Nutrition, Live-Trial and CW query logs.

Of course a query may not be exactly duplicated, but it might be very similar to other queries, and we can consider this form of repetition by looking for query similarities. One way to measure query similarity is to compute the degree of overlap between query terms; see Equation 1. For example, the similarity between the query '*jaguar pictures*' and '*jaguar photos*' is 0.33.

$$\mathrm{Sim}(q, q') = \frac{|q \cap q'|}{|q \cup q'|} \tag{1}$$

Using this approach to query similarity we can consider a modified form of query duplication that considers two queries to be duplicates if they are within a given similarity threshold; for example '*jaguar pictures*' and '*jaguar photos*' are considered to be duplicates above the 0.25 similarity threshold but not above the 0.5 threshold. Of course this is a rather weak similarity metric that is liable to identify unrelated queries as similar, especially at low similarity values. For example, '*jaguar XK8*' and '*jaguar cats*' are also above the 0.25 threshold even though one obviously refers to cars and the other cats. We will return to this issue of query similarity in Section 5.1.

Figure 2 shows the percentage of queries that have duplicates above a range of similarity thresholds for each of the query logs. Note that the original values for exact duplicates are also shown and that a similarity threshold of 1 allows for duplication that involves permutations of the same query terms. The results indicate that there is a high level of overlap between queries even in the General search scenario. For example, 75% of general search queries share at least one term with other queries (that is similarity $> 0$) and approximately one third of these share 50% of their terms with other queries. In the more specialised search tasks the degree of overlap is even higher. In the Image, Nutrition and Live-Trial search logs, about 90% of queries share at least one search term, and more than 70% of queries share at least 50% of their terms. Interestingly the CW logs present with intermediate overlap statistics. Even though they refer to general searches, by virtue of the limited scope of the searches (50 employees of a small software

company) it is likely that their searches will turn out to be fairly clustered, which is found to be the case. That said, the CW logs cover a relatively short period of time (6 weeks) so we might legitimately expect overlaps to increase further over a longer time period.

The above results provide us with high-level information about the prevalence of repetition in the query-space but they do not tell us about the degree of repetition. For example, 70% of Nutrition queries may share 50% of their terms with other Nutrition queries, but how many other Nutrition queries? This is partly answered in Figure 3 where we report the average number of similar queries that are available at the different similarity thresholds. We see, for example, that in the General and Image search logs there are nearly 100 other similar queries, on average, for each duplicate query at the 0.5 similarity threshold. And even at the more stringent similarity thresholds the degree of repetition is surprisingly high. For instance, there are nearly 60 queries on average associated with every query that counts as a repeat at the 0.75 threshold in the Image search logs. In other words, Figure 2 tells us that nearly 70% of Image queries share at least 75% of their terms with other

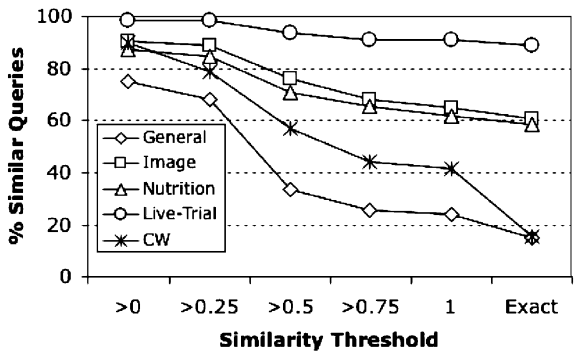


*Figure 2.* The percentage of repeat queries that exceed certain query similarity thresholds.

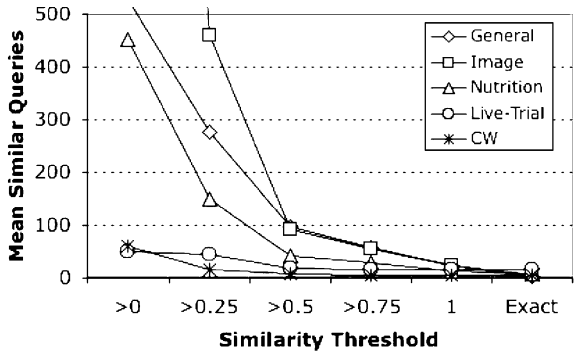


*Figure 3.* The average number of similar queries available at different similarity thresholds.

Image search queries, and Figure 3 tells us that, on average, for each of these 70% of the Image search queries there are nearly 60 such queries sharing more than 75% of their terms.

Although the mean figures above indicate that the degree of repetition is quite significant these figures are generally biased because of a small number of very common queries that have very large overlap sets. In this work we are interested in reusing previous search sessions that reflect searches that are similar to the current target query. Therefore it is important for us to know how many of these repeat queries can be associated with at some minimum number of similar queries. Figure 4 graphs the percentage of repeat queries that can be associated with more than 10 other similar queries at a given similarity threshold. For example, we see in the Live-Trial data, that of the queries that have overlaps of more than 0.5 with other queries, 70% of these can be associated with more than 10 other similar queries. In general the results are similarly impressive for the other focused search tasks, 70% of the Nutrition queries and about 90% of the Image queries can be associated with more than 10 similar queries at the 0.25 overlap threshold.

This analysis indicates two important things. First, similar queries do tend to recur frequently in Web search. Second, the degree of repetition varies from being especially high in very focused search tasks or at moderate overlap thresholds, although the degree of repetition does fall off for more general search tasks. This provides a firm foundation for any approach to Web search that attempts to exploit query-repetition – at least repetition occurs, and it is especially frequent in focused tasks – but, of course, whether this repetition reflects an underlying regularity of search behaviour that can be exploited remains to be seen.

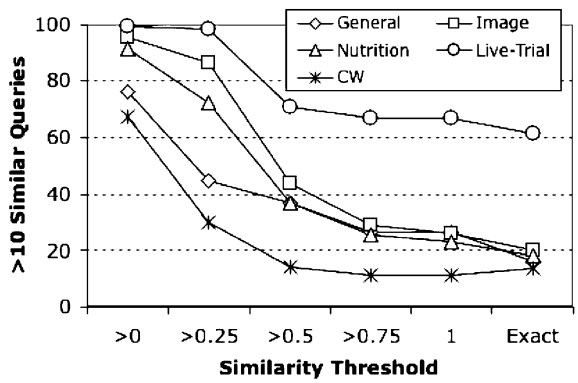


*Figure 4.* The percentage of repeat queries that can be associated with at least 10 other similar queries at a given similarity threshold.

## 4. Direct Query Reuse for Collaborative Search

Our research is motivated by two key ideas. First, specialised search facilities attract communities of users with similar information needs and so serve as a useful way to limit variations in search context. For example, a search box on an AI Web site is likely to attract queries with a computer-related theme, and queries such as '*cbr*' are more likely to relate to Case-Based Reasoning than to the Central Bank of Russia. Certainly the results in the previous section indicate that specialised or topical search scenarios are subject to a significant degree of query repetition. Second, by monitoring user selections for a query it is possible to build a model of query-page relevance based on the probability that a given page $p_j$ will be selected by a user when returned as a result for query $q_i$ (see Section 4.2). It is worth clarifying here that $q_i$ refers to a single query, which may be composed of multiple terms; in this work we do not decompose queries into their constituent terms.

Our collaborative search approach combines both of these ideas in the form of a meta-search engine that analyses the patterns of queries, results and user selections from a given search interface serving a well-defined search context or an ad-hoc community of like-minded individuals. The resulting selection patterns are used to capture the relationship between queries and documents in a given search context to bridge any gap that might exist between the query and document space. In this section we concentrate in particular on our initial work on collaborative search, which employed a simple model of query reuse by focusing on exact matches between the target query and past search sessions. Nevertheless, the results of constrained live-user trials are impressive, even with this simple reuse model, as we will demonstrate. In Section 5 we will go on to describe our latest work on extending this reuse model to provide for a more powerful instantiation of collaborative search.

### 4.1. THE I-SPY SYSTEM ARCHITECTURE

The I-SPY collaborative search architecture is presented in Figure 5. It presents a meta-search framework in which each user query, $q$, is submitted to base-level search engines ($S_1 - S_n$) after adapting $q$ for each $S_i$ using the appropriate adapter. To be clear, this query adaptation involves a simple reformatting of the original queries in line with each underlying search engine's query interface, but no query elaboration or expansion is carried out. Similarly, the result set, $R_i$, returned by a particular $S_i$ is adapted for use by I-SPY to produce $R_i'$, which can then be combined and re-ranked by I-SPY, just like a traditional meta-search engine. Similarly, this result adaptation involves converting the results from each search engine into a common I-SPY format.

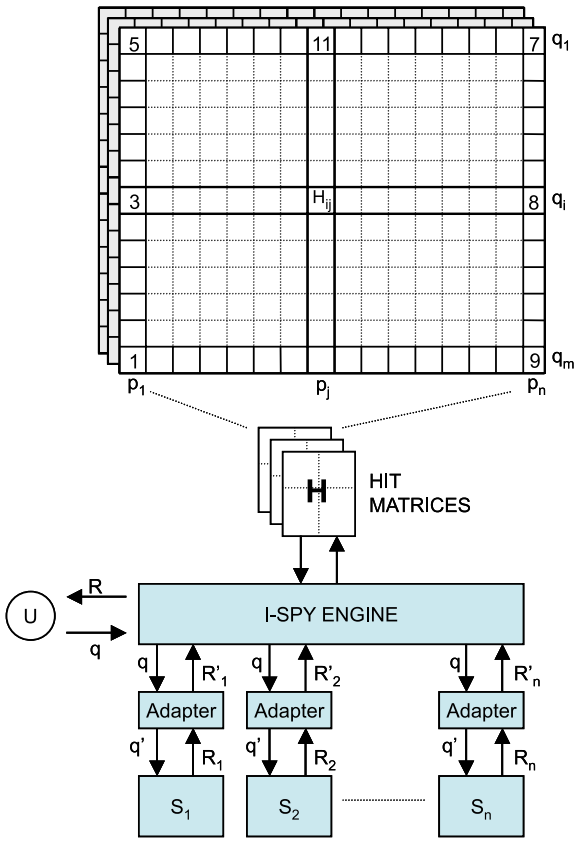


*Figure 5.* The I-SPY system architecture combines meta-search with a facility for storing the search histories of individual communities of searchers.

I-SPY's key innovation involves the capture of search histories and their use in ranking metrics that reflect user behaviour. This allows it to personalize its search results for a particular community of users without relying on content-analysis techniques (e.g., Bradley et al., 2000; Lawrence and Giles, 1998). To achieve this, I-SPY borrows ideas from collaborative filtering research to profile the search experiences of users. Collaborative filtering methods exploit a graded mapping between users and items and I-SPY exploits a similar relationship between queries and result pages (Web pages, images, audio files, video files etc.). This relationship is captured as a *hit-matrix* (see Figure 5). Each element of the hit-matrix, $H_{ij}$, contains a value $v$ (that is, $H_{ij} = v$) to indicate that $v$ users have found page $p_j$ relevant (at least in the sense that they have selected the page) for query $q_i$. In other words, each time a user selects a page $p_j$ for a query $q_i$, I-SPY updates the hit matrix accordingly; of course users may select multiple pages from a given result-list in which case each selection results in an update. I-SPY maintains its hit

table using a relational database and an efficient encoding for result URLs and query terms.

### 4.2. COLLABORATIVE RANKING

I-SPY's key innovation is its ability to exploit the hit matrix as a *direct* source of relevancy information – after all, its entries reflect relevance judgments by users – that helps to bridge the gap between the query-space and the document-space. Most search engines, on the other hand, rely on *indirect* relevancy judgments based on overlaps between query and page terms, but I-SPY has access to the fact that, historically, $v$ users have selected page $p_j$ when it is retrieved for query $q_i$. I-SPY uses this information in many ways, but in particular the relevancy of a page $p_j$ to query $q_i$ is estimated by the relative frequency with which this page has been selected in the past in response to this query (see Equation 2). Results that have been previously selected for a given query are *promoted* ahead of other search results returned by the base-level search engines and these results are ordered according to the relevance values.

$$\text{Relevance}(p_j, q_i) = \frac{H_{ij}}{\sum_{\forall j} H_{ij}} \tag{2}$$

Figures 6 and 7 show two screen-shots of the I-SPY system and serve as a simple example of the system's potential. Each presents part of the results page for a query by a computer science student for the (vague) single term query '*shakey*' (refering to the robot developed at the Stanford Research Institute). Figure 6 shows the result-list returned before I-SPY has built-up its hit matrix data, and so the

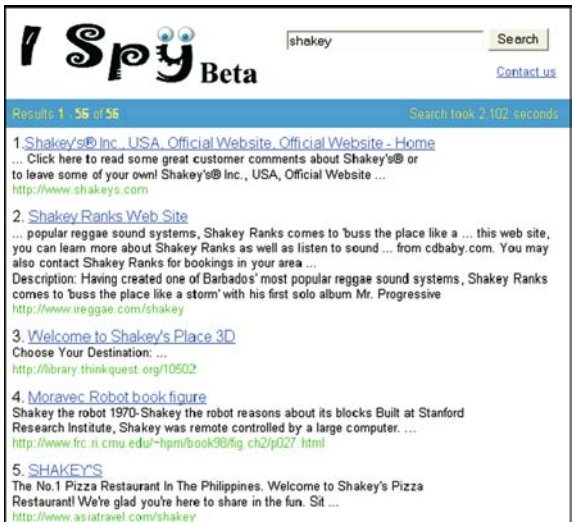


*Figure 6.* I-SPY search results before training.

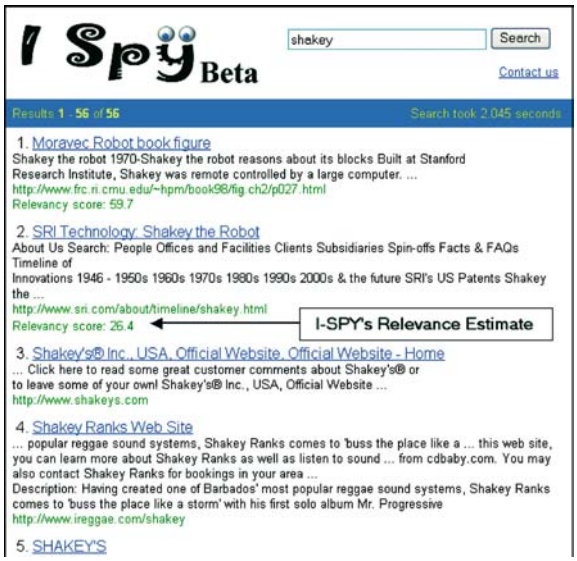


*Figure 7.* I-SPY search results after training. The annotation indicates the I-SPY relevancy score.

results are ordered using a standard meta-search ranking function, giving preference to results that are highly ranked by I-SPY's underlying search engines; in this case, Google, Yahoo, AllTheWeb, Teoma, WiseNut, HotBot and AltaVista. Clearly not all of the results presented are relevant and in fact only the 4th result is on-target. Of course since '*shakey*' is obviously a very vague query, it is no surprise that these results lack precision.

In contrast, Figure 7 shows the results for the same query, but after I-SPY has been *trained* by a community of computer science students; that is, the query-result patterns of a set of computer science students have been used to build the hit matrix. The results are now ranked by I-SPY's relevance metric, as discussed above, rather than by the standard meta-search ranking function. The point is that this time the results are more relevant; the top two results now refer to Shakey the robot rather than other interpretations of the '*shakey*' query. For example, the second ranking result, '*SRI Technology : Shakey the robot*' is the developer's page and has an I-SPY relevance value of 26.4. In other words, this page has been selected 26.4% of the times that *shakey* has been used as a query. This page previously would have been ranked in 36th position by the standard meta-search ranking function.

### 4.3. COMMUNITY-BASED FILTERING

A key point to understand about I-SPY is that its relevancy metric is tuned to the preferences of a particular set of users – the community of I-SPY users using a particular version of I-SPY in a well-defined context – and the queries and pages

that they tend to prefer. Deploy I-SPY on a wildlife Web site and its hit matrix will be populated with query terms and selected pages that are relevant to wildlife fans. Over time the value-space of the relevancy metric will adapt to fit the appropriate query-page mappings that serve this target community. For example, queries for '*jaguar*' will tend to result in the prioritisation of sites about the wild cat, as opposed to sites related to cars, because previously when users have submitted this query term they will have selected these wildlife sites. The other sites may still be returned but will be relegated to the bottom of the result-list.

In fact the I-SPY architecture has been designed to make it easy to deploy and manage multiple I-SPY search agents, each devoted to a specific search context. The central I-SPY server is configured to service many different search services and each search service can be provided to a different community of users and will naturally adapt to the preferences of these communities. I-SPY users can configure their own search service thereby creating a new hit matrix and service identifier. Users can use their newly configured service as a personal search agent that adapts to their own needs and preferences. Alternatively, they can deploy an I-SPY search box on a Web site, or part of a Web site, that services a well-defined community of users. For instance, in a large directory portal placing a search box on a '*programming languages*' directory page will naturally tend to attract queries from this domain. Consequently, the behaviour of the users providing these queries will gradually adjust I-SPY's relevancy metric and ranking function in favour of programming languages pages.

### 4.4. LIVE-USER EVALUATION

It is our contention that many search problems are a direct result of the prevalence of poor queries that articulate the user's search needs in only very vague terms. The basic hypothesis of I-SPY is that:

1. Such queries tend to recur;
2. It is possible to learn implicit search context information by monitoring the selection behaviour of users with respect to these recurring queries;
3. This search context information can be used to re-rank standard search results to better reflect the relevancy of a set of retrieved documents in a certain context.

While we demonstrated the validity of (1) through an analysis of user search logs (see Section 3), we evaluated (2) and (3) with a live-user evaluation that focused on the search behaviour of a group of computer science students.

This experiment involved 92 computer science students and took place in October 2003. It was designed to evaluate the benefits of I-SPY in the context of a fact-finding or question-answering exercise that would ensure a high frequency of recurring queries. To frame the search task we developed a set of 25 general knowledge AI and computer science questions, each requiring the student to find

*Table II.* Sample questions used in the live-user trial

| |
|---|
| What game did Arthur Samuels help computers to play? |
| Who was Shakey? |
| How many bits are in a nibble? |
| What was the name of the first microprocessor? |
| Who introduced minimax? |
| Who co-founded Apple with Steve Jobs? |
| Where does Michael Jordan teach? |
| Who wrote the first e-mail? |
| Who invented the concept of a 'universal machine'? |
| Who founded Firefly? |

out a particular fact (time, place, person's name, system name etc.); see Table II for a number of sample questions.

#### 4.4.1. *Setup*

Each student received all 25 questions in a random order and they were asked to use the I-SPY search engine to locate their answers; the students were actively monitored to ensure that they used I-SPY only. They were instructed to attempt as many questions as possible within the allotted time, and were asked to record their answers and the URL where their answers were found on their question sheets. They were allowed to answer the questions in any order and in fact each student received their questions in random order. They were also allowed to skip questions and return to questions. I-SPY was set up with an empty hit-matrix and configured to draw on results from four underlying search engines (Google, Hotbot, Wisenut and AllTheWeb).

#### 4.4.2. *Methodology*

The students were randomly divided into two groups. Group 1 contained 45 students and Group 2 contained the remaining 47. Group 1 served as the *training group* for I-SPY, in the sense that their search histories were used to populate the I-SPY hit-matrix but no re-ranking occurred for their search results. This group also served as a control against which to judge the search behaviour of the second group of users, who served as the *test group*. They benefited from I-SPY's re-ranking based on their own incremental usage and the hit-matrix produced by the first group; that is, at the start of the Group 2 test, the hit-matrix was initialised to that produced by the Group 1 users and the hit-matrix continued to evolve as a result of the Group 2 usage. The results presented here correspond to the performance of each group of users during a 45 min time-slot.

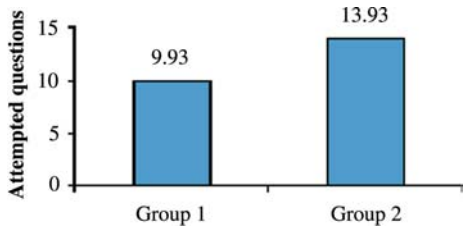


*Figure 8.* Mean number of questions attempted per user.

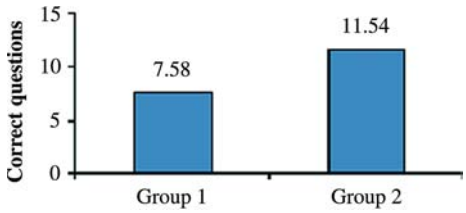


*Figure 9.* Mean number of questions answered correctly per user.

### 4.4.3. *Questions Attempted and Answered*

We are particularly interested in a number of key issues related to the effectiveness of I-SPY's collaborative search functionality. First and foremost, is there any evidence that collaborative search benefits users when it comes to helping them to more efficiently locate relevant information? In other words is there any difference between the groups in terms of the number of questions attempted or correctly answered?

Figure 8 presents the mean number of questions attempted for Group 1 and 2 users. There appears to be an obvious advantage for the Group 2 users, who answered 40% more questions on average when compared to Group 1 (9.9 vs. 13.9 for Group 1 and 2, respectively); this difference is statistically reliable at $p < 0.01$ with $n(89) = -5.39$. Perhaps more importantly, in comparison to Group 1, Group 2 users answered 52% more questions correctly (see Figure 9). Indeed, the average Group 2 user answered more questions correctly (11.5) than the average Group 1 user even attempted (9.9); again this difference is statistically reliable at $p < 0.01$ with $n(67) = -5.84$.

Thus, I-SPY's collaborative search technique appears to have helped students to answer questions more efficiently and more successfully. Below, we will attempt to better understand the reasons behind this advantage by looking for biases within the user groups and by comparing their granular search behaviors at a result selection level.

#### 4.4.4. *A Question of Bias?*

Of course one reason for a performance difference between the user groups might simply be the result of some bias in students to begin with. Although the test conditions were carefully controlled and the student selection process was random, we did attempt to shed light on such concerns by examining the type of queries used by the different user groups; perhaps the Group 2 users were capable of formulating more effective queries than the Group 1 users?

First, it is worth highlighting that more than 70% of the unique queries used by the Group 1 users were also used by the Group 2 users and these repeated queries were used by approximately 65% of the 97 users. This high repeat-rate would suggest a close correspondence between the vocabulary and query formation capabilities of each group. However, it is also worth pointing out that when we compared the average query length (terms per query) for each group of users we found Group 2 users to be using slightly shorter queries (see Figure 10); Group 1 users used 2.57 terms per query on average compared to 2.16 terms per Group 2 query, a difference that is statistically significant at $p < 0.01$. In other words, aside from repeating queries, the Group 2 users appeared to provide queries that have fewer terms. One possibility is that the Group 2 users discovered that shorter queries worked better, although this does not appear to be the case when we look for changes in average query length for each searcher as the experiment goes on. So, all other things being equal, the Group 2 queries are likely to be *more* vague than the queries provided by Group 1 users. If anything this suggests that the Group 1 users might have been slightly advantaged.

Finally, given that I-SPY has no ability to re-rank the results that are returned for non-repeating queries we would expect no real difference in the selection behaviour of the users for such queries if both groups are truly similar. When we examined the average position of the results selected for these non-repeating queries we found that the Group 1 users had an average result selection position of 2.98 compared to 3.12 for the Group 2 users. Once again, this minor difference appears to operate in the favour of the Group 1 users and is consistent with the hypothesis that these users do seem to be using more precise, queries; more precise queries should see relevant results with lower positional values.

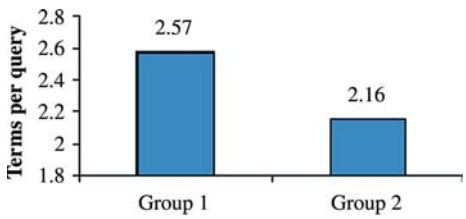


*Figure 10.* Average number of terms per query.

From this analysis it appears to be valid to conclude that there is no significant bias operating in favour of the Group 2 users that might go to explain the significant improvements in their search performance. If anything there is some evidence that perhaps the Group 1 users were at a slight advantage when it came to their query construction abilities, making the positive I-SPY results for the Group 2 users even more significant.

#### 4.4.5. *Selection Behaviours*

We contend that the success of the Group 2 users is not due to any inherent positive bias regarding their searching abilities, but rather that it is a direct result of I-SPY's re-ranking process. In short, I-SPY is frequently able to promote previous user result selections to the Group 2 users. These selections will appear higher up in the result lists and, assuming that they appear to be just as relevant to the Group 2 users as they did to the Group 1 users, we should find that the Group 2 users respond by selecting more of these results from higher positions within the result lists. In doing so, the Group 2 users are likely to follow fewer false-leads, because popular selections will be promoted higher than spurious selections that are more likely to be false-leads. This should help the Group 2 users to locate relevant results more quickly and more frequently and would explain their ability to attempt more questions, than the Group 1 users, as well as their ability to answer more of these questions correctly.

Figure 11 confirms this hypothesis: The Group 2 users selected results with an average position of 2.24 whereas the Group 1 users selected results with an average position of 4.26; a 47% reduction in the position of selected results for Group 2 users when compared to Group 1 users; this difference is statistically significant at $p < 0.01$. Moreover, we found that, for instance, Group 2 users select twice as many results in position 1 compared to the Group 1 users, presumably reflecting the fact that, for Group 2 users, twice as many position-one results are relevant, because many of these results have been promoted due to I-SPY's re-ranking. Indeed overall, 84% of the results selected by Group 2 users fall in the 1–3 position range. In contrast, to reach this figure for Group 1 we must consider a result range that goes as far as position 8.

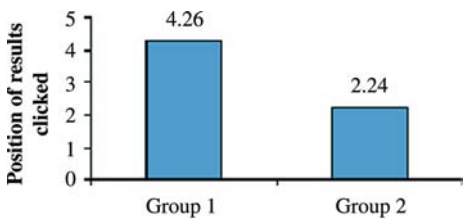


*Figure 11.* Average positions of selected results.

#### 4.4.6. *Summary*

The primary aim of this experiment has been to demonstrate the effectiveness of I-SPY with live-users rather than artificial-users; results of more limited live-user trials as well as comprehensive artificial-user evaluations have previously been reported and demonstrated similar positive benefits in a range of other search scenarios, including less constrained search tasks (Smyth et al., 2003; Smyth et al., 2003a,b). We believe that the results presented in this paper demonstrate a clear and significant advantage for I-SPY users when compared to the control group, indicating the benefits of I-SPY's collaborative search and re-ranking methods over more traditional meta-search approaches. The fact that the Group 2 users are capable of answering more questions correctly by examining fewer search results is a demonstration of I-SPY's enhanced search performance in the absence of any alternative explanation by way of any strong bias in the search skills of one of the user groups.

Of course it has to be noted that the current evaluation has its limitations. It refers to a closed search scenario in which a relatively homogeneous group of users take part in a constrained search task. Nevertheless we believe that it is a good demonstration of the potential of collaborative search. In reality, much larger communities will participate in more open-ended search tasks and this is likely to have an impact on the benefits due to collaborative search.

## 5. Beyond Exact-Match Query Reuse

The approach to collaborative search so far described is limited by its reliance of exact matches between the target query and the queries associated with previous search sessions. In the context of our experiment such matches were relatively commonplace ($> 70\%$ repeat rate), mostly because the experiment established a narrowly focused search domain. That said, the results from Section 3 do tell us that repeat queries are commonplace in specialised search tasks; for example we found 60% of queries in the Image and Nutrition query-logs to be exact repeats of earlier queries. This suggests that I-SPY may be able to enhance the performance of up to 60% of the searches in specialised domains, but at the same time it means that there are at least 40% of search sessions that cannot be enhanced.

Our aim in this section is to explain how a more flexible query-reuse policy can address this issue. The approach taken has its genesis in the observation that by allowing for partial query matches we can greatly enhance the applicability of collaborative search. In Section 3 we saw how similar queries are very common in both specialised and generic search. For example, more than 85% of queries share at least 25% of their terms with previous queries in the Nutrition and Image query-logs. And even in the generic search task (represented by the Excite logs) we find that 70% of queries share 25% of their terms with previous queries.

Allowing for partial query matches has two important advantages. Obviously it allows collaborative search to influence a much greater number of search sessions; more than 85% of sessions if a 25% similarity threshold is allowed for partial matches, for example. In addition, however, it means that far more past search sessions are available as a source of relevance evidence; instead of relying on the search sessions for a single matching query, we can now draw on the past sessions for a whole range of related queries. Of course we must compensate for the fact that the search behaviour associated with related queries is likely to be less reliable than those behaviours associated with the exact matching query. Accordingly, in what follows, we describe how to extend I-SPY to allow it to reuse similar queries and how their influence can be appropriately weighted.

### 5.1. REUSING SIMILAR QUERIES

Normally, when a searcher submits a new *target* query, $q_T$, I-SPY locates an individual row in the hit-matrix that corresponds to this exact query. To take advantage of similar queries I-SPY must now locate each row of the hit-matrix that relates to a similar query; queries that share terms with the target query. These are the rows that contain search behaviours that *may* be useful to guide the ranking of the new result-list. To do this we compute the overlap between the terms used in $q_T$ and the terms used in every other query recorded in the hit-matrix, as shown in Equation (1). I-SPY then selects all queries that exceed a given similarity threshold to produce its list of related queries. For instance, setting a similarity threshold of $> 0$ will result in the selection of all queries that share at least one term with the current query. Of course such a lenient threshold is likely to result in the selection of unrelated queries which may interfere with result quality. Alternatively we could configure I-SPY to select the top $Q$ queries rather than all queries above a certain similarity threshold.

### 5.2. WEIGHTED RELEVANCE

In Section 4.2 we described how the relevance of a result-page, $p_j$, to a query, $q_i$, can be estimated from the number of selections received by $p_j$ for $q_i$, relative to the total number of page selections that have occurred for the page, $p_j$ (see Equation (2)). By extending collaborative search to take advantage of multiple similar queries, there are potentially multiple search histories to inform the relevance of a given page. For example, the page www.sun.com may have a high relevance value (let's say, 0.8) for a past query 'java language' but it may have a lower relevance for another past query 'java' (let's say 0.33). The question is: how can these relevance values be combined to produce a single relevance score for this page relative to a related target query, say 'enterprise java'?

We propose a normalised weighted relevance metric that combines the relevance scores for individual page-query combinations. This is achieved using the

weighted-sum of the individual relevance scores, such that each score is weighted by the similarity of its corresponding query to the target query. Thus, in our example above, the relevance of the page www.sun.com is 0.516: the sum of 0.264 (that is, 0.8 page relevance to query 'java language', multiplied by the 0.33 query similarity between this query and the target, 'enterprise java') and 0.165 (0.33*0.5 for the past query, 'java'), divided by 0.83, the sum of the query similarities. Equation (3) provides the details of this weighted relevance metric with respect to a page, $p_j$, a target query, $q_T$, and a set of retrieved similar queries $q_1, \ldots, q_n$. Exists $(p_j, q_i)$ is simply a flag that is set to 1 when $p_j$ is one of the result pages represented in the solution of query, $q_i$.

$$W\text{Rel}(p_j, q_T, q_1, \ldots, q_n) = \frac{\sum_{i=1\ldots n} \text{Relevance}(p_j, q_i) \bullet \text{Sim}(q_T, q_i))}{\sum_{i=1\ldots n} \text{Exists}(p_j, q_i) \bullet \text{Sim}(q_T, q_i)} \tag{3}$$

This weighted relevance metric is far from perfect. For example, it does not contain any term to devalue isolated results that come from queries with low aggregate similarity. Nevertheless we believe that this metric will prove adequate for the majority of query-result pairings and we return to the issue of possible future enhancements in Section 6.4.

### 5.3. EVALUATION

To test our extended collaborative search approach we draw on the search data collected during the live-user trial of I-SPY reported previously. In total, the Group 1 users produced 1049 individual queries and selected a combined total of 1046 pages, while the Group 2 users used 1705 queries and selected 1624 pages; these numbers do not refer to unique queries or pages however. The resulting log-data provides the following key information: the queries submitted by each user; the pages that they selected from the subsequent result-lists; the position of these pages within the result-list; the pages where they located a correct answer to a particular question, and the hit-matrix produced by the Group 1 users.

Accordingly, we "re-run" the live-user experiment by responding to Group 2 queries with the new result-lists that are recommended by the extended version of I-SPY, and we can evaluate the quality of these result-lists with reference to our known set of correct pages, comparing the outcome to the standard I-SPY and meta-search performance results. This ground-truth of pages that are known to be correct was independently produced by manually inspecting selected pages for correct question answers. We actually evaluate five different variations of the extended I-SPY – which we refer to as CB because our extended model of collaborative search draws on ideas from *case-based reasoning* research – each with a different minimum similarity threshold (0, 0.25, 0.5, 0.75, 1) during query selection to limit the range of selected queries. Also, in this evaluation we take a more traditional

IR approach to measuring search engine performance by examining retrieval precision and recall, for example.

### 5.3.1. *Minimum Accuracy*

Perhaps the most basic measure of a search engine's accuracy concerns its ability to return at least a single relevant result in its result-list; we refer to this as *minimum accuracy* and we will look at more refined accuracy measures that focus on the number of relevant results and their positions in due course. To measure the minimum accuracy for each search engine (CB, I-SPY and Meta), we compare the top 30 results returned by these search engines (including the five extended I-SPY variations), for the 1705 test queries, to the list of known correct results associated with these queries. We compute the percentage of result-lists that contain at least one correct result.

The results are presented in Figure 12 as a graph of minimum accuracy (that is, the percentage of searches for which a particular approach produced a result-list with at least one correct result) against the similarity threshold. Each plot corresponds to a single search engine. The plots tell us that CB produced a result-list with at least one correct result more frequently than I-SPY or Meta. At a similarity threshold $>0$, it returns a correct result page in 93% of sessions, and in 92% of sessions with the similarity threshold $>0.25$. This is a relative improvement of 43% over the standard version of I-SPY and Meta, which only produce a result-list with at least one correct result for 65% of the searches. The benefit here is derived from the fact that the CB version of I-SPY is able to include additional pages beyond those found by the underlying search engines in the result-lists returned for a given query. These additional results come from the result-lists contained within the search histories of the related queries. In contrast, Meta and I-SPY are

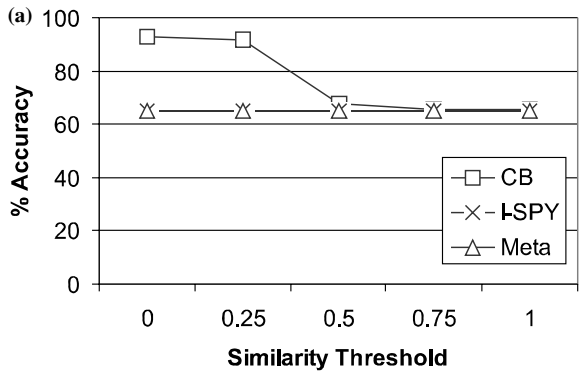


*Figure 12.* The minimum accuracy for different query-similarity thresholds. Each similarity threshold indicates the minimum acceptable similarity so that, for example, the 0 similarity threshold refers to queries that have a similarity $>0$.

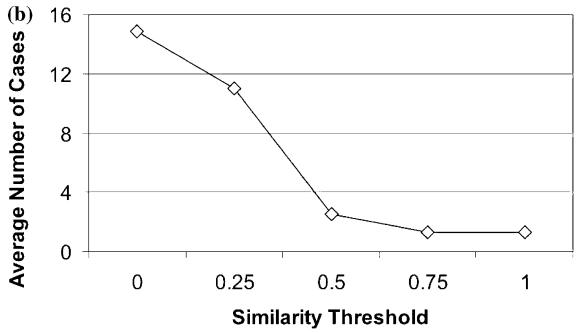


*Figure 13.* The average number of related query cases retrieved for different query-similarity thresholds. Each similarity threshold indicates the minimum acceptable similarity.

effectively limited to those results returned by the underlying search engines for the current target query; I-SPY simply reorders the Meta results using its relevance metric. The observed benefit proves that these additional results are frequently relevant to the target query.

It is interesting to note how the accuracy of the CB version of I-SPY drops off sharply with increasing similarity threshold. From a purely case-based reasoning perspective this appears strange at first glance. Normally we would expect that increasing the similarity threshold would improve the average similarity of the search-cases (the related queries and their selected results) being retrieved and we are conditioned to expect that this is likely to improve any 'solution' that is derived from these cases. Not so in our case-based view of search however, because the number of cases retrieved and the diversity of the results is likely to be important. When we plot the average number of similar search-cases retrieved, for a typical query, across the different similarity thresholds (see Figure 13) we can see that there is a sharp drop in available search-cases between the 0.25 and 0.5 thresholds. At the 0 and 0.25 thresholds, an average of 15 and 11 cases, respectively, are being retrieved for a target query, but this falls off to 2.5 for the 0.5 threshold and then 1.3 search-cases beyond this. At the higher thresholds there are simply not enough similar search-cases to make a meaningful additional contribution to the result-lists offered by the meta-search and so the benefits enjoyed by the CB version of I-SPY become marginal. So, even though low similarity thresholds may permit the reuse of unrelated search sessions (e.g., 'inventor java' would be considered similar to 'inventor ethernet' even though their associated selection histories are unlikely to be helpful), we find that the benefits of a greater number and variety of reusable cases easily outweighs any problems due to inappropriate retrievals. Moreover, our weighted relevance metric will tend to discount the selection histories associated with these less related cases.

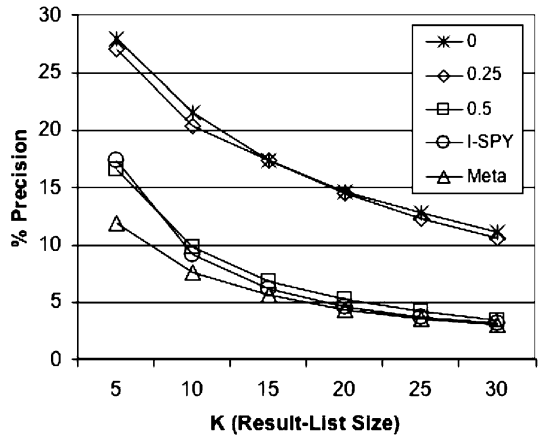


*Figure 14.* A comparison of precision values for different result-list sizes.

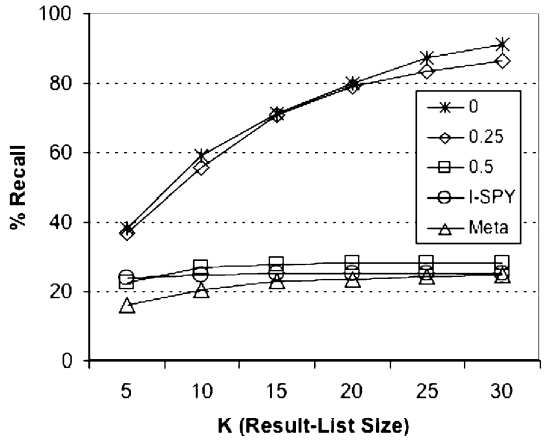


*Figure 15.* A comparison of recall values for different result-list sizes.

### 5.3.2. *Precision vs. Recall*

The standard objective test of search engine accuracy is the precision and recall test: the former computes the percentage of returned results that are relevant, while the latter computes the percentage of relevant results that are returned. We measure the percentage precision and recall values for each of the search techniques under review at different result-list sizes ($k = 5$–30).

The results are presented as precision and recall graphs, for various similarity thresholds, in Figures 14 and 15. Each graph plots the precision (or recall) plot for the CB versions of I-SPY, along with I-SPY and Meta, against result-list size, $k$; for reasons of clarity we have omitted the plots for the $>0.75$ and 1 similarity thresholds as in each case they correspond closely to the plot for the $>0.5$ threshold. As expected we find that precision tends to fall-off with increasing result-list

sizes; typically the number of relevant results is much less than $k$, and the majority of these relevant results should be positioned near the top of result-lists. The critical point is that, once again the performance benefits due to the case-based approach are clear, especially at low similarity thresholds. For example, in Figure 14 we see that CB precision varies between nearly 28% (at $k = 5$ and for a similarity threshold >0) to 11% (at $k = 30$ for the same threshold). This is compared to precision values of between 17% and just over 3% for I-SPY, and values between 12% and 3% for Meta. These results indicate that the CB version of I-SPY enjoys a precision improvement of between 60% and 258%, relative to standard I-SPY at the similarity threshold $> 0$ level; similar benefits are indicated for a similarity threshold >0.25. These precision improvements are even greater (between 130% and 265%) when measured relative to Meta.

The recall results tell a similar story. The recall for the CB version of I-SPY (at similarity thresholds of $> 0$ and $> 0.25$) grows from approximately 37% ($k = 5$) to just over 91% ($k = 30$). At the same result-list sizes the standard I-SPY recall only grows from 23% to 25% and Meta's recall, from 16% to 25%. Obviously the CB version of I-SPY is locating a far greater portion of the relevant pages than the standard version of I-SPY or Meta, and it is gaining access to these additional relevant pages from the result-lists of its similar queries.

Once again we see that the benefits accruing to the CB version of I-SPY tend to fall away as the similarity threshold increases. For thresholds >0.5 and beyond, only minor precision and recall improvements are achieved (in the order of 7–12%). As discussed in the previous section, this can be readily explained by the sharp drop in similar cases for similarity thresholds of 0.5 and higher. Simply put, the limited number of search sessions produced during this trial meant that there was a tendency for very few similar cases at the $> 0.5$, $> 0.75$ and 1 thresholds and so there was very limited data available to I-SPY as a means to improve the result ranking, hence the minor variations in performance at these thresholds.

### 5.3.3. *Winners and Losers*

It is well known that result position is a major influencing factor on the behaviour of Web searchers, and various analyses of Web search behaviours have highlighted how users are reluctant to venture beyond the first few results (Jansen et al., 1998), regardless of how many are returned by a search engine; our own analysis of a variety of search logs indicates that up to 90% of result selections occur for the top five result positions (Freyne et al., 2004). Hence in this final experiment we investigate the position of relevant results within the result-lists returned by the various search engines.

This time we confine our analysis to I-SPY, Meta and the CB variation with a similarity threshold $> 0$. We also confine our result-lists to the top 20 results. For each result-list produced for a particular query, we note the position of its

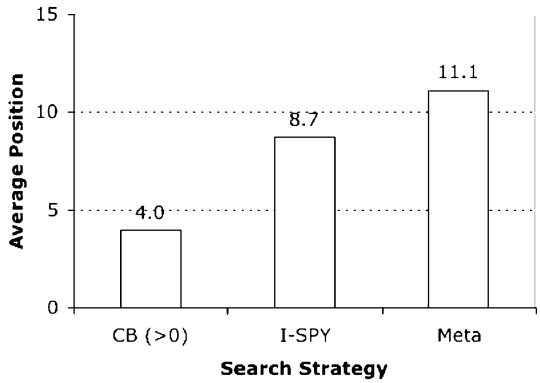


*Figure 16.* Average position of the first correct answer.

first correct result and average this across each search engine. If a search engine does not include a correct result in its top 20, then it is penalised with a positional score of 21. Obviously this is an underestimate that favours the poorly performing search engine because the first correct result might actually appear much further down its list, if at all. We also calculate the positional averages when we focus in on those sessions for which CB or I-SPY 'wins', 'loses' or 'draws': *CB Win* refers to the set of queries for which the CB version of I-SPY has its first correct result at a higher position than the standard version of I-SPY; *I-SPY Win* refers to those queries where the standard version of I-SPY has its first correct result at the higher position, and *Draw* refers to the queries for which both methods have their first correct result at the same position in the top 20.

The overall mean position results are presented in Figure 16 for the CB, I-SPY and Meta strategies. Once again it is clear that there is a significant benefit to the CB version of I-SPY. On average, its first correct results are found at position 4,

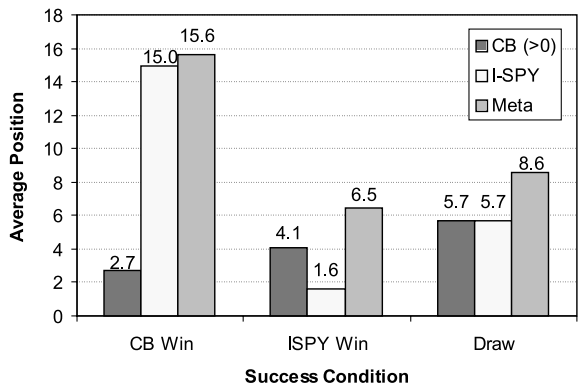


*Figure 17.* Average position for the first correct result under win, loose and draw success critiqeria.

compared to 8.7 and 11.1 for I-SPY and Meta, respectively. It is worth remembering here that we are using a penalty position of 21 in cases where a given search strategy does not retrieve the correct result within the top 20. So an average position of 4, in the case of CB for example, includes such penalties for the sessions where a correct result is not found (that is, about 7% of the sessions as discussed in Section 5.3.1). In the majority of sessions CB does retrieve a correct result in its top 20 and in these sessions the average position is approximately 2.7.

In Figure 17 we present similar results for the CB Win, I-SPY Win and Draw conditions introduced above. In CB Win, which occurs 43.4% of the time, the average CB position is only 2.7, compared to 15 and 15.6 for standard I-SPY and Meta respectively. In I-SPY Win – which happens only 23.4% of the time – its top correct result is at position 1.6, beating CB's position of 4.1. When both search techniques draw the average position is the same at 5.7, beating Meta's 8.6. So, from a positional viewpoint, the CB version of collaborative search does as good as, or better than, standard I-SPY 76.4% of the time, and when it does better it does a lot better with a positional difference of 12.3 (2.7 vs. 15). Only 23% of the time does the standard I-SPY produce a higher correct result than the CB version, and in these situations the positional differences are minor – 1.6 vs. 4.1 – with both techniques delivering the correct result within the top 5.

#### 5.3.4. *Summary*

These results are clearly very promising with the CB version of I-SPY producing accuracy, precision, recall and positional benefits that are statistically reliable at $p < 0.01$. This is clear and compelling evidence that by allowing I-SPY to reuse the search histories of similar queries, significant improvements in search performance can be achieved. Precision and recall are both significantly enhanced by the CB version of I-SPY. The CB version of I-SPY is able to influence a greater number of search sessions because it is no longer limited to exact query matches, and in doing so, it is able to provide improved result rankings because of its ability to draw on, and weigh up, a greater range of result-selection evidence.

However it is important to understand the limitations of our evaluation results. The evaluation corresponds to a very tightly controlled search task and a narrowly focused community of searchers with the same background. These conditions are unlikely to be duplicated in more realistic search tasks. That said, as an exploratory evaluation we believe that our trial results do add value by demonstrating the potential of the proposed techniques.

## 6. Discussion

During the development of the I-SPY technology a variety of researchers and commentators have highlighted a number of concerns and objections in relation to the

general applicability of the collaborative search approach. For example, a number of reviewers have drawn parallels between I-SPY and the failed DirectHit search engine – which also attempted to use result popularity to influence ranking – arguing that I-SPY may succumb to a similar fate for similar reasons. Others have highlighted key challenges that I-SPY must be seen to address in relation to its efficiency and privacy features, for example.

In this section we will consider these issues in detail. We will also look at I-SPY from a number of different perspectives, arguing that its collaborative search approach is likely to lead to a number of secondary benefits above and beyond its core performance advantages.

### 6.1. THE DIRECTHIT OBJECTION

A number of commentators have raised a series of objections that are related to the downfall of the ill-fated DirectHit search engine technology. In the late 1990's the first generation of Web search engines were based largely on traditional IR term-matching techniques with search engines like AltaVista leading the way. Around this time two new search engines emerged, each advocating the use of novel information in order to improve search engine performance. Google, advocated the use of link popularity, by mining connectivity information from the Web graph, whereas DirectHit proposed the use of result-popularity, by mining user selections.

Obviously Google won out, but the failure of DirectHit is often used to suggest that I-SPY is likely to face a similar fate given its reliance on user selection data. Fortunately the basis for this prediction does not hold up. DirectHit suffered from two important flaws:

1. In general search most queries are not repeated and those that are tend to be repeated very few times. So there is little data available to improve search in the case of the vast majority of queries.
2. Popularity-based ranking is open to abuse from users or agents if they make false selections in order to promote target results.

The important thing to remember about these issues is that they are exacerbated because of DirectHit's focus on generic search. At the same time they are ameliorated by I-SPY's community-oriented view of search. For instance, as we have seen in Section 3, while general-purpose search engines do not benefit from many repeat queries – only about 15% of Excite queries are repeats – query repetition is far more commonplace in more specialised search engines; 60% of queries were found to be repeats in the Image and Nutrition query logs. I-SPY's community-oriented view of search means that separate hit-matrices are created for different specialised search tasks and as such each hit-matrix benefits from far higher query repeat rates. Moreover, I-SPY's ability to work with related queries, instead of just exact-match queries, greatly increases the data that is available during re-ranking.

DirectHit was found to be extremely susceptible to fraudulent selections. For example, it was a relatively simple matter to design an automated agent to promote a target result by repeatedly submitting a query and simulating an arbitrary number of clicks for the target result. I-SPY is less susceptible to this type of fraud in the sense that it would be difficult for a malicious user to interfere with result rankings across all communities. While it may be relatively easy to influence result promotion for a single community of users it is far more laborious to achieve this across all communities. Of course this does not mean that I-SPY is impervious to this type of fraud. It is still an important issue and we are currently looking at a number of ways of detecting and discounting such activity. A potentially useful example of this is a simple coping strategy that discounts sequences of selections. Related work concerned with protecting collaborative filtering recommenders from rogue users provides further possibilities (see O'Mahony et al., 2003).

### 6.2. TRUST AND AUTHORITY

Issues of fraudulent selections aside, the relevance score computation used by I-SPY assumes that the contribution made by each member of a search community is equally important. However, it is more likely that certain members are more knowledgeable than others. The query terms and page selections of these people are likely to be more discriminating and informative than the selections made by a novice in the community. For instance, David Aha's selections for the query 'CBR' are likely to be more informative than the selections of a first year computer science student for the same query. This issue has been addressed in the area of knowledge management within a specialised user community by the work of Ferrario and Smyth (2001). They describe how community members can be explicitly recognised by their expertise and how this information can be leveraged when it comes to evaluating the reliability of submitted information items. The idea that certain users may be more authoritative than others is an issue in the context of I-SPY. However, in order to exploit this idea it is necessary to be able to distinguish between the selections of different searchers within the community. But, at the same time, it is worth noting that it relies on the identification of individual users, which I-SPY purposefully avoids; see Section 6.7. Nevertheless, this issue is worthy of further research and is likely to be investigated in the future.

### 6.3. EFFICIENCY

It may seem that I-SPY adds yet another layer of computational complexity to a world where instant search responses are demanded. After all I-SPY works to rerank search results that have already been ranked by some primary search engines. While this is certainly true it is worth highlighting that I-SPY's relevance computation is very efficient as it is carried out over only a small subset of the results

returned by the primary search engines, namely those results that have been previously selected according to their hit-matrix data.

Indeed it is also worth highlighting that I-SPY can be configured to work in parallel with the underlying search engines. Accordingly it may be possible to deliver faster results than these search engines by compiling its two sets of results in parallel – one set from the underlying search engines (the *remote results*) and one set from the hit-matrix (the *local results*). The second set is computed locally and, because it is taken from a hit-matrix that records only previously selected results, this computation can be performed far faster than a search of a full document-index by a traditional search engine. Thus, the local results can be returned to the user almost immediately. These local results are likely to be better candidates than the remote results because they have been considered relevant in the past. We are currently exploring this mode of operation as an alternative to I-SPY's default post-processing style approach.

### 6.4. SIMILARITY AND RELEVANCE METRICS

This work has proposed a number of important metrics when it comes to evaluating query similarity (see Equation (1) and Section 5.1) and page relevance (in particular, see Equation (3) in Section 5.2). The first measures the similarity between a target and candidate query with a view to determining whether the latter may provide useful results in relation to the former. The second metric then uses this similarity information as a means to weight the influence of the results obtained from similar queries when computing an overall estimate of result relevance. Neither of these metrics are perfect – although both appear to perform well in practice – and there are many opportunities for improvement, which we shall discuss briefly here.

In relation to the query similarity metric, which implements a simple model of term overlap, it is clear that no consideration is given to term order or to phrasal structures. Moreover, no attempt is made to weight the query terms according to their relative importance. For example, in many search engines term order is considered to be very important, with terms at the start of a query gaining higher weightings than later terms. In the future we plan to investigate improvements to our standard overlap metric so that it can better account for ordering effects and phrasal structures. In addition, it is interesting to note that there are other sources of information available to I-SPY that could be used to evaluate query similarity. One option is to use I-SPY's selection information to discover query similarities even when there is no direct overlap between query terms. The correlation between the hit values of any pages that have been jointly selected for two queries can be used as the basis for a different type of similarity measure, for example. In other words, two queries are similar if the same pages are selected a similar number of times for each query. Our future work will also consider these issues in more detail.

As mentioned in Section 5.2, the weighted relevance metric does not contain any term to devalue isolated results that come from queries with low aggregate similarity and this may lead to biased relevance values in certain situations. For example, a page will have a maximum relevance value of 1 if it is the only page that has been selected for a given query. And if this query is reused as a candidate for some target query then its lone result is likely to benefit from a disproportionately high relevance value even if the query has a low similarity to the target query. In the future we plan to explore a number of possible solutions to this problem, such as devaluing the relevance of isolated results; see also Herlocker et al. (1999) for related ideas.

### 6.5. RELEVANCE BIAS AND PARADIGM CHANGE

Over time, the relevancy of particular pages is likely to change with respect to certain queries. A page that is considered to be very relevant to one query today might no longer be especially relevant in a few months or weeks time. However, within I-SPY there is an inherent bias toward older pages in the sense that these pages are more likely to have been retrieved in the past, and as such, have had a greater opportunity to accumulate user hits than new pages. In theory this means that these older pages may continue to be promoted ahead of more relevant, newer pages. To compound matters, if the older pages continue to be promoted then they are also more likely to attract further hits, by virtue of their improved position relative to the newer pages.

One of the ways that we plan to cope with this in I-SPY is to introduce an aging model for past selections so that the relevance scores of pages can be normalised with respect to their age. For example, the hits associated with a page might be gradually decayed over time so that hits from the past have less influence than more recent hits. Thus, pages that acquired all of their hits in the distant past will be discounted relative to newer pages that have received fewer, but more recent, hits.

This approach should also be of benefit to communities where novelty and new pages are likely to be far more important than older or even relatively recent pages. For example, in many news domains or perhaps entertainment and sports communities it is easy to imagine that users will be interested in obtaining the most up-to-date content and so the discounting that is applied to older pages can be increased accordingly.

### 6.6. QUERY-BASED INDEXING AND ALTERNATIVE RELEVANCE MODELS

The I-SPY approach to Web search represents a significant departure from more traditional IR-based approaches, which rely on a document index that relates document terms to documents. This index is static for a fixed set of documents. We can view the I-SPY hit-matrix as a different type of index, one that relates query

terms to documents, but one that is constantly evolving, even for a fixed set of documents, as it tracks the search behaviour of users. This supplementary index allows I-SPY to constantly adapt to its users' behaviours.

However, concerns are often raised about the scalability of I-SPY's hit-matrix data structure. While developing Internet-scale technologies is always a challenge there is no reason to expect I-SPY to suffer more than existing search engines. After all its hit-matrix data structure is equivalent to the standard term-based indexing structures used by traditional search engines, except, of course, that the hit-matrix is likely to be far smaller than a standard term-based index. The hit-matrix only stores information on result pages that have been selected (by a community of searchers), which is a tiny fraction of the pages that will have been retrieved by their searches. Moreover, the typical query contains far fewer terms than the typical document.

It is worthwhile considering an alternative view of I-SPY's hit-matrix as a general-purpose facility for imposing alternative relevance models on pre-existing search engines. In this paper we have focused on the use of a relevance model that is based on the selection histories of communities of users. However, it is also possible to train I-SPY with reference to very different relevance models by populating the hit-matrix with data from an alternative feedback strategy. In this way we can adapt the behaviour of I-SPY to suit a broad range of purposes. For example, one recent trend in the search engine arena focuses on the development of so-called local search engines that prioritise results from the locale of the searcher. Ordinarily, to add this functionality to a pre-existing search engine means updating its basic retrieval engine. However I-SPY presents an alternative: simply train an I-SPY community by populating its hit-matrix with hits for local pages for a range of common queries. In this way I-SPY will come to promote local results for popular queries, thereby mimicking the behaviour of a local search engine but without the need to alter its core retrieval functionality. Indeed, I-SPY can be used in this way to deliver a wide variety of retrieval behaviours from a single basic retrieval engine.

### 6.7. PROFILING, PERSONALIZATION AND THE PRIVACY SWEETSPOT

One of the key objectives behind I-SPY is to offer users a more personalized search experience. This is generally accepted as a key objective for today's *one-size-fits-all* search engines. The traditional approach to personalization is dominated by what might be termed the 'one-user-one-profile' philosophy. Accordingly, most personalized information services and recommender systems track the behaviour of individual users, or request these users to provide personal profile information. The resulting profiles are then used to filter, select and recommend items of information that are likely to be preferred by individuals according to their profiled needs and preferences; for example, at the time of writing Google's recent experimental foray

into personalized search involves the searcher declaring their personal preferences up-front, as discussed earlier in Section 2.2.1.

Of course the benefits of personalization come at a cost. More accurate recommendations may help us to locate relevant information more efficiently but, as individuals, we are faced with the prospect of revealing personal information to third-parties. And the issue of privacy is all too often swept aside in the quest for improved personalization; see Kobsa, (2002). If we are to bring personalized or context-sensitive search techniques to the Web on a large-scale then it is likely that privacy will play a vital role. Indeed we believe that any technique that requires users to explicitly provide information about their personal preferences, or any technique that appears to track user behaviour on a user-by-user basis, is likely to be acceptable to only an extremely restricted audience; the vast majority of privacy-conscious Web users choosing their anonymity over improved search accuracy. For instance, up to 87% of users indicate that they are very concerned about privacy threats on the Internet and 41% of users claim to have left sites that require registration information; see Kobsa (2002) for a recent review of privacy and personalization. It is worth noting that many governments are introducing (or have already introduced) data protection laws that may ultimately outlaw many approaches to personalization that require the maintenance of rich user profiles, or at best require users to provide their explicit consent for such information to be recorded or used; see for example www.dataprivacy.ie.

The real goal then must be to offer users search improvements that provide the right balance of personalization and privacy, and it is in this respect that we believe our collaborative search technique may offer a unique advantage. By providing a form of personalization that operates at the level of the community, rather than the individual, collaborative search offers improved search quality without compromising an individual's privacy. In short, we argue that the I-SPY approach strikes a novel and effective balance between personalization and privacy. From a personalization viewpoint, I-SPY offers communities of like-minded individuals the opportunity to benefit from search results that are adapted to their usage patterns and collective community preferences. Crucially, this is achieved without the need for profiling at the level of individual users. While the search histories of a community of users are recorded, I-SPY does not tag the search patterns of individual community members. Nor does I-SPY need to maintain a record of which members form part of a community; ad-hoc communities operate without the need for user logins or cookies or any other form of individual user tracking. In this sense we believe that I-SPY hits a unique *privacy sweetspot* by delivering effective personalization in an anonymous fashion.

## 6.8. BEYOND RESULT RANKING

In this work we have described how I-SPY reuses search histories to re-rank search results. It is worth pointing out that this data can also be pressed into service for

other search related tasks. For instance, in Balfe and Smyth (2004) we describe how I-SPY can leverage this information to make query recommendations alongside each of its promoted search results. In other words, in addition to promoting certain results within the result-list, I-SPY annotates these results with additional queries that have also led to the result in question being selected. The novelty in this query recommendation strategy stems from the way in which queries are scored and ranked by using relevance and coverage factors in order to prioritise those queries that are most likely to be successful in the current search context. Preliminary evaluation results, based on live-user studies, indicates that this particular approach to query recommendation has the potential to result in high-quality recommendations.

We are investigating how I-SPY's community-based partitioning of user selection data might be used as a form of result clustering. For example, during the normal course of operation, when a query is submitted by a user in a particular community, the search is conducted with reference to an individual hit-matrix and the results are ranked accordingly. However, the same query could also be simultaneously submitted to other hit-matrices in order to benefit from the behaviours of related communities of users. The end results can then be presented as different clusters of results, each labeled by their parent community's label. We are exploring techniques for automatically selecting related communities by comparing their selection patterns across shared queries and selected results. For example, a user might submit the query 'Val Thorens' through the I-SPY search-box on their favourite hill-walking site, and they would receive a list of results that had been preferred by other searchers using this site; presumably this would produce a ranking that prioritises information such as places to stay, trail map, nearby Alpine resorts, etc. In the case that there is a related community of searchers, let us say a community of skiers that use a 'European Skiing Information' version of I-SPY, then I-SPY can also return any hits from this related community, on the grounds that it also contains lots of queries for 'Val Thorens', labeling them as skiing related results. The point is that our searcher may benefit from this related alternative viewpoint.

## 7. Conclusions

In general, Web search engines are challenged by the vague queries that are commonly submitted by searchers. When a user enters the search query 'jaguar' are they looking for the car, the cat or the operating system? Does a searcher looking for 'Michael Jordan' want information on the basketball star, the Berkeley professor or the EDS chairman? Our response is motivated by two important observations:

1. Although query repetition is rare in general Web search, it is far more common-place in specialised search scenarios; our analysis indicates that in focused

search tasks duplicate queries are not at all uncommon and similar queries are downright prevalent.
2. Many searches can be considered to be specialised, even though they may be dealt with by a general-purpose search engine. For example, many searches orig-inate from search-boxes that are placed on niche content sites and as such they can be considered as serving the specialised needs of an ad hoc community of focused searchers.

Our approach to Web search relies on query repetition and, further, hypothesises that such repetition is accompanied by selection regularity; in other words, not only do searchers often user similar queries, they tend to select similar results in response to these queries. The collaborative search idea takes advantage of this repetition and regularity to improve the relevance of search results. In short, we have described a collaborative approach to Web search that captures the search histories of communities of related users and reuses this information to re-rank search results to better reflect community preferences. In particular, we have explained how the technology has evolved from one that relies on exact-match query reuse to one that allows for a more flexible approach to query reuse. The results of live-user trials have been reported and demonstrate that performance benefits are available when compared to more traditional approaches to Web search, at least in communities of searchers that do share a common information domain and where they are likely to represent their needs with similar queries. Our trial results also indicate that even relatively small communities of tens or hundreds of users can benefit from significant personalization, especially if the community is focused. However, in reality it is likely that many communities will be far less focused than our trial community and so we anticipate longer training times and the need for larger numbers of participating users.

Finally, we believe that our collaborative search approach is an important advance in the area of personalized Web search. It provides users with a significant and appropriate degree of personalization. It also facilitates the modeling of different user moods or preference themes, by allowing a user to easily participate in different communities. But it does this without compromising their privacy and without the need to profile them as individuals. The result is an anonymous approach to personalized search that is likely to satisfy the concerns of even the most privacy conscious searchers.

## Acknowledgements

The support of the Informatics Research Initiative of Enterprise Ireland and Science Foundation Ireland is gratefully acknowledged. We would also like to thank our reviewers for their valuable comments and suggestions.

## References

Balfe, E. and Smyth, B.: 2004, Collaborative Query Recommendation for Web search. In: *Proceedings of 16th European Conference on Artificial Intelligence.* Valencia, Spain IOS Press, pp. 268–272.

Bharat, K.: 2000, SearchPad: Explicit Capture of Search Context to Support Web search. In: *Proceedings of the 9th International World Wide Web Conference.*

Bollmann-Sdorra, P. and Raghavan, V. V.:1993, On the Elusiveness of Adopting a Common Space for Modeling IR Objects: Are Queries Documents? *Journal of Americal Society for Information Science* **44**(10), 579–587.

Bradley, K., Rafter, R. and Smyth, B.:2000, Case-based User Profiling for Content Personalization. In: O. Stock, P. Brusilovsky and C. Strapparava (eds.): *Proceedings of the International Conference on Adaptive Hypermedia and Adaptive Web-based Systems.* Trento, Italy: Springer-Verlag, pp. 62–72.

Brin, S. and Page, L.:1998 The Anatomy of a Large-scale Web Search Engine. In: *Proceedings of the 9th International World Wide Web Conference*.

Budzik, J. and Hammond, K.:2000, User Interactions with Everyday Applications as Context for Just-in-time Information Access. In: *Proceedings of the 5th International Conference on Intelligent User Interfaces.* Louisiana, USA: ACM Press, pp. 44–51.

Croft, W. B., Cook, R. and Wilder, D.: 1995, Providing Government information on the internet: Experiences with THOMAS. In: *Proceedings of the Second Annual Conference on the Theory and Practice of Digital Libraries.* Austin, Texas, USA, pp. 19–24.

Cui, H., Wen, Ji- R., Nie, J. -Y. and Ma, W. -Y.: 2002, Probabilistic Query Expansion Using Query logs. In: *Proceedings of the 11th International World Wide Web Conference.* Honolulu, Hawaii, USA: ACM Press pp. 325–332.

Ferrario, M. A. and Smyth, B.: 2001, Distributing Case-base Maintenance, the Collaborative Maintenance approach. *Journal of Computational Intelligence: Special Issue on Maintaining Case-Based Reasoning Systems* **17**(2), 315–330.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G. and Ruppin, E.: 2001, Placing search in context: The concept revisited. In: *Proceedings of the 10th International World Wide Web Conference.* Hong kong, pp. 116–131.

Fitzpatrick, L. and Dent, M.: 1997, Automatic Feedback using Past Queries: Social Searching? In: *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* Philadelphia, ennsylvania, USA: ACM Press, pp. 306–313.

Freyne, J., Smyth, B., Coyle, M., Balfe, E. and Briggs, P.: 2004, Further Eexperiments on Collaborative Ranking in Community-based Web Search. *AI Review: An International Science and Engineering Journal* **21**(3–4), 229–252.

Furnas, G. W., Landauer, T. K., Gomez, L. M. and Dumais, S. T.: 1987, The Vocabulary Problem in Human–system Communication. *Communications of the ACM* **30**(11), 964–971.

Glance, N. S.:2001, Community Search Assistant. In: *Proceedings of the 6th International Conference on Intelligent User Interfaces.* Santa Fe, New Mexico, USA: ACM Press, pp. 91–96.

Glover, E., Lawrence, S., Gordon, M. D., Birmingham, W. P. and Lee Giles, C.: 2000, Web Search – Your Way. *Communications of the ACM* **44**(12), 97–102.

Haveliwala, T. H.:2002, Topic-sensitive page rank. In: *Proceedings of the 11th International World-Wide Web Conference.* Hanolulu, Hawaii, USA: ACM Press, pp. 517–526.

Herlocker, J. L., Konstan, J. A., Borchers, A. and Riedl, J.: 1999, An Algorithmic Framework for Performing Collaborative Filtering. In: *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* Bekkeley, California, USA: ACM Press, pp. 230–237.

Jansen, B. J., Spink, A., Bateman, J. and Saracevic, T.: 1998, Real Life Information Retrieval: A Study of User Queries on the Web. *SIGIR Forum* **32**(1), 5–17.

Kleinberg, J. M.:1998, Authoritative Sources in a Hyperlinked Environment. In: *Proceedings of the Ninth Annual ACM-SIAM Symposium on Discrete Algorithms.* San Francisco, California, USA, pp. 668–677.

Kobsa, A.: 2002, Personalized Hypermedia and International Privacy. *Communications of ACM* **45**(5), 64–67.

Kruger, A., Giles, C. L., Coetzee, F., Glover, E., Flake, G., Lawrence, S. and Omlin, C.: 2000, DEADLINER: Building a New Niche Search Engine. In: *Proceedings of the Ninth International Conference on Information and Knowledge Management.* McLean, Virgina, USA, pp. 272–281.

Kushmerick, N.: 1997, Wrapper Induction for Information Extraction. In: *Proceedings of the International Joint Conference on Artificial Intelligence*. Nagoya, Japan: Morgan-Kaufmann, pp. 729–735.

Lawrence, S.: 2000, Context in Web search. *IEEE Data Engineering Bulletin* **23**(3), 25–32.

Lawrence, S. and Giles, C. L.: 1988, Context and Page Analysis for Improved Web Search. *IEEE Internet Computing.* July–August, 38–46.

Lawrence, S. and Giles, C. L.: 1999, Searching the Web: General and Scientific Information Access. *IEEE Communications* **37**(1), 116–122.

Lieberman, H.: 1995, Letizia: An Agent that Assists Web Browsing. In: C. Mellish (ed.): *Proceedings of the International Joint Conference on Artificial Intelligence, IJCAI'95.* Montreal, Canada: Morgan Kaufman Publishers, pp. 924–929.

Lyman, P. and Varian, H. R.: 2003, How Much Information? Retrieved from http://www.sims.berkeley.edu/how-much-info-2003/.

Mitra, M., Singhal, A. and Buckley, C.: 1998, Improving Automatic Query Expansion. In: *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* Melbourne, Australia: ACM Press, pp. 206–214.

O'Mahony, M. P., Hurley, N. J. and Silvestre, G. C. M.: 2003, An Evaluation of the Performance of Collaborative Filtering. In: *14th Irish Artificial Intelligence and Cognitive Science Conference.* Dublin, Ireland, pp. 164–168.

Ozmutlu, S., Spink, A. and Ozmutlu, H. C.: 2000, Multimedia Web Searching Trends: 1997-2001. *Information Processing and Management* **39**(4), 611–621.

Raghavan, V. V. and Sever, H.: 1995, On the Reuse of Past Optimal Queries. In: *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* Seattle, Washington, USA: ACM Press, pp. 344–350.

Rhodes, B. J. and Starner, T.: 1966, Remembrance Agent: A Continuously Running Automated Information Retrieval System. In: *Proceedings of the 1st International Conference on the Practical Applications of Intelligent Agents and Multi-Agent Technologies.* London, UK, pp. 487–495.

Roush, W.: 2004, Search Beyond Google. *MIT Technology Review* **107**(2): 34–45.

Smyth, B., Balfe, E., Briggs, P., Coyle, M. and Freyne, J.: 2003, Collaborative Web Search. In: *Proceedings of the 18th International Joint Conference on Artificial Intelligence, IJCAI-03.* Acapulco, Mexico: Morgan Kaufmann, pp. 1417–1419.

Smyth, B., Freyne, J., Coyle, M., Briggs, P. and Balfe, E.: 2003, Collaborative ranking in community-based Web search. In: *14th Irish Artificial Intelligence and Cognitive Science Conference.* Dublin, Ireland, pp. 199–204.

Smyth, B., Freyne, J., Coyle, M., Briggs, P. and Balfe, E.: 2003b, I-SPY Anonymous, Community-based Personalization by Collaborative Meta-search. In: *Proceedings of the 23rd SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence.* Cambridge, UK: Springer, pp. 367–380.

Spink, A, Bateman, J and Jansen, B.J.: 1998, Searching Heterogeneous Collections of the Web: Behaviour of Excite Users. *Information Research* **4**(2). Retrieved from http://information.net/ir/4-2/paper53.html.

Terveen, L., Hill, W., Amento, B. and McDonald, D.: 1997, Phoaks: A System for Sharing Recommendations. *Communications of the ACM* **40**(3), 59–62.

Wen, J.-Y. J.-R. and Zhang, H.-J.: 2002, Query Clustering Using User Logs. *ACM Transactions on Information Systems* **20**(1), 59–81.

## Authors's vitae

**Prof. B. Smyth**

*Department of Computer Science, University College Dublin, Belfield, Dublin 4, Ireland*

Barry Smyth holds the Digital Chair of Computer Science. He is an ECCAI Fellow and is currently the head of the Department of Computer Science at University College Dublin. He is also a founder of Changing Worlds Ltd. and continues to serve as their Chief Technical Officer. He received a BSc in Computer Science from University College Dublin and a Ph.D. from Trinity College, Dublin. Prof. Smyth works in several areas of artificial intelligence including personalization, case-based reasoning and information retrieval. He has authored almost 200 technical papers and received best paper awards from conferences such as the International Joint Conference on Artificial Intelligence and the European Conference on Artificial Intelligence.

**E. Balfe**

*Department of Computer Science, University College Dublin, Belfield, Dublin 4, Ireland*

Evelyn Balfe is currently a Ph.D. candidate in the Department of Computer Science at University College Dublin where she received her B.Sc. in 2002. She is a member of the I-SPY research group, which is exploring the use of artificial intelligence techniques in collaborative Web search. Her particular research focuses on applying case-based reasoning methods to the area of personalized Web search, and specifically to query reuse and result ranking.

**J. Freyne**

*Department of Computer Science, University College Dublin, Belfield, Dublin 4, Ireland*

Jill Freyne is a Ph.D. candidate in the Department of Computer Science at University College Dublin. She received her B.Sc from University College Dublin in 2002. Jill's primary interests lie in the use of clustering techniques in Web search and personalization. Jill is a member if the I-SPY research group.

**P. Briggs**
*Department of Computer Science, University College Dublin, Belfield, Dublin 4, Ireland*
Peter Briggs is a Ph.D. candidate based in the Department of Computer Science at University College Dublin, where he has been working as part of the I-SPY group since graduating with a B.Sc in 2002. His research interests include the use of collaborative filtering and implicit profiling techniques in collaborative Web search.

**M. Coyle**
*Department of Computer Science, University College Dublin, Belfield, Dublin 4, Ireland*
Maurice Coyle is a Ph.D. candidate in the Department of Computer Science at University College Dublin. He received his B.Sc. from University College Dublin in 2002 and currently works as part of the I-SPY group. His primary research investigates the use of diversity enhancing techniques and user profiling in personalized Web search.

**O. Boydell**
*Department of Computer Science, University College Dublin, Belfield, Dublin 4, Ireland*
Oisin Boydell is a Ph.D. candidate in the Department of Computer Science at University College Dublin. He received his B.Sc. from University College Dublin in 2002 and after working in industry he has recently returned for take up a research position in the I-SPY group. Oisin's research focuses on the use of relevance feedback techniques in Web search.