Artificial Intelligence Review **21:** 229–252, 2004.

Further Experiments on Collaborative Ranking in Community-Based Web Search

JILL FREYNE, BARRY SMYTH, MAURICE COYLE, EVELYN BALFE & PETER BRIGGS
Smart Media Institute, Department of Computer Science, University College Dublin, Belfield, Dublin 4, Ireland (E-mail: barry.smyth@ucd.ie)

Abstract. As the search engine arms-race continues, search engines are constantly looking for ways to improve the manner in which they respond to user queries. Given the vagueness of Web search queries, recent research has focused on ways to introduce context into the search process as a means of clarifying vague, under-specified or ambiguous query terms. In this paper we describe a novel approach to using context in Web search that seeks to personalize the results of a generic search engine for the needs of a specialist community of users. In particular we describe two separate evaluations in detail that demonstrate how the collaborative search method has the potential to deliver significant search performance benefits to end-users while avoiding many of the privacy and security concerns that are commonly associated with related personalization research.

Keywords: context, personalization, relevance, web search

1. Introduction

The Web is rapidly becoming a victim of its own success. As it continues to grow, users are finding it more and more difficult to locate the right information at the right time. Even with the help of the most advanced search engines we regularly fail to locate relevant information in a timely manner. Many factors contribute to this access problem. Certainly, the sheer quantity of Web information, and its growth rate, tax even the most advanced search engines. For example, various estimates indicate that even the largest search engines cover only a fraction of the available information space (Lawrence and Giles, 1999a, b). They simply cannot keep up with the growth of the Web when it comes to the indexing of newly created documents or the re-indexing of recently updated documents. However, this search engine coverage issue is just part of the problem, and indeed can be relieved by using meta-search methods (Drillinger and Howe, 1997; Selberg and Etzioni, 1997).

Perhaps a more pressing problem stems from the fact that the average Web user is not an information retrieval expert. Even when users choose reasonable query terms to guide search, the resulting queries are rarely complete in the way that they reflect the search needs of a given user; their queries are often vague and imprecise. For example, a query might include terms that identify the primary information target, but might exclude terms that usefully describe the search *context*. For example, a simple query for 'cbr' does not indicate whether the user is interested in Case-Based Reasoning or the Central Bank of Russia, and queries for 'D. McSherry' do not distinguish between the University of Ulster lecturer and the well-known novelist, Frank D. McSherry. Thus, many researchers have recently focused on ways to exploit context in Web search as a means of resolving ambiguity (e.g. Rhodes and Starner, 1996; Budzik and Hammond, 2000; Glover et al., 2000, 2001; Lawrence, 2000; Haveliwala, 2002).

In this paper, we describe a novel, simple, yet powerful technique to exploit context during search (Section 3). This *collaborative search* method acts as a post-processing service for existing search engines and re-ranks results based on the learned preferences of a community of users; see also (Smyth et al., 2003a). We describe its implementation in the I-SPY system (http://ispy.ucd.ie) and show how I-SPY achieves this level of personalization in an anonymous fashion; it avoids storing individual user profiles, thus relieving many of the usual privacy issues associated with personalization techniques. In Section 4 we discuss the results of an extended evaluation of I-SPY. Two separate and complementary evaluations are presented. First we demonstrate how I-SPY's collaborative ranking engine can significantly improve result precision and recall, compared to benchmark search engines. We do this through a large-scale, multi-domain evaluation based on realistic artificial search models. This evaluation was first presented in (Smyth et al., 2003b) and is extended here by a new live-user evaluation of I-SPY which provides additional evidence in support of the performance advantages of collaborative search. We go on to argue that these advantages make collaborative search particularly well suited to device-limited information retrieval tasks, for example, search on mobile devices such as WAP phones and PDAs.

2. Background

For the most part, recent search engine advances have focused on improving existing indexing and ranking techniques (e.g. Brin and

Page,1998; Kleinberg, 1998). However, vague queries remain a significant problem and have led to a growing body of research looking at ways to supplement such queries with missing context terms (see also Lawrence, 2000). Context information can be generated according to two basic approaches: either it can be explicitly provided by the user or search engine or it can be implicitly inferred from the local search environment.

2.1. *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., 2000) 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').

The second option for introducing context into Web search is to use a specialised search engine whose index has been designed to cover a restricted information domain, essentially fixing the context prior to searching. For example, Tripadvisor (TripAdvisor, Inc.,) allows its users to locate information about destinations, CiteSeer (Lawrence and Giles, 1999a, b), focuses on searching scientific literature, and DEADLINER (Kruger et al., 2000) targets conference and workshop information. Some specialised search engines automatically maintain their indexes by using information extraction techniques to locate and index relevant content (Kushmerick, 1997).

2.2. *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 Hammond, 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 a document he/she views, and is guided by the text surrounding the marked query in that document, i.e. 'the 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. (Haveliwala, 2002) describes a method that uses categories from the Open Directory Project (ODP) (www.dmoz.org) as a source of context to guide a topic-sensitive version of PageRank (Brin and Page, 1998). 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. These techniques attempt to use 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 (Bharat, 2000) describes the SearchPad system which extracts context information, in the form of useful queries and promising result-lists, from multiple search sessions. Similarly, (Bradley et al., 2000) describes the CASPER search engine for job advertisements, which maintains client-side user profiles that include job cases that users have liked and disliked in previous searches. These profiles are used to classify and rerank the results of future searches. 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.

3. Collaborative Search and I-SPY

Collaborative search is motivated by two key ideas. First, specialised search engines 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 field 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. 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.

The 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. This approach has been fully implemented in the I-SPY search engine and will be detailed and evaluated in the following sections.

3.1. *The I-SPY system architecture*

The I-SPY collaborative search architecture is presented in Figure 1. 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. 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. I-SPY's key innovation involves the capture of search histories and their use in ranking metrics that reflect user behaviour.

The unique feature of I-SPY is its ability to personalize its search results for a particular community of users without relying on content-analysis techniques (e.g. Lawrence and Giles, 1998; Bradley et al., 2000). To achieve this, I-SPY borrows ideas from collaborative filtering re-

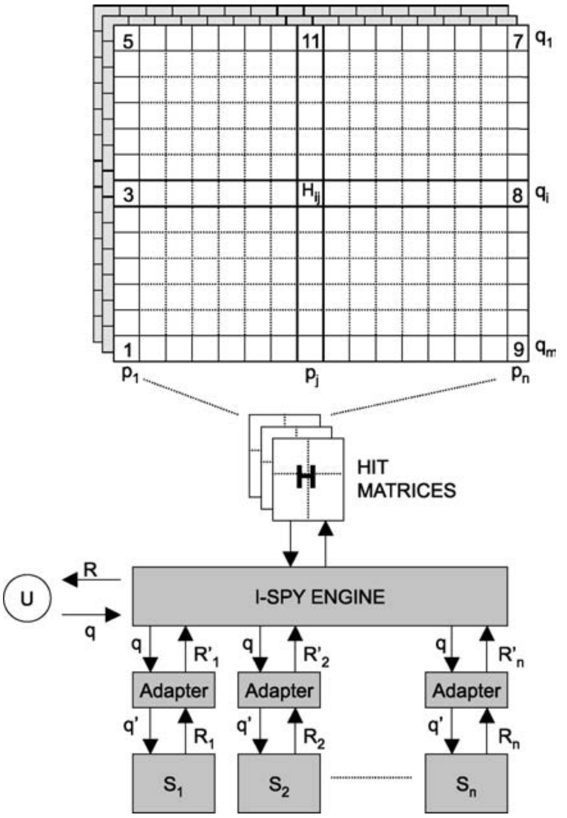


Figure 1. The I-SPY system architecture.

search 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 1). Each element of the hit matrix, H_{ij}, contains a value v_{ij} (that is, $H_{ij} = v_{ij}$) to indicate that v_{ij} users have found page p_j relevant 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. I-SPY maintains its hit matrix using a relational database and an efficient encoding for result URLs and query terms.

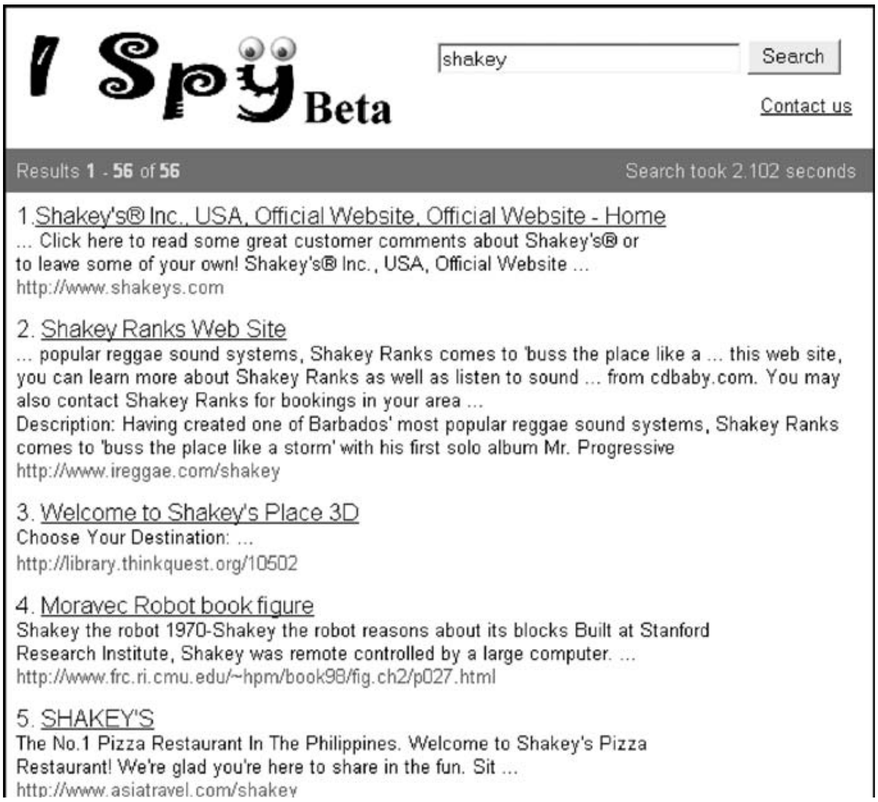
3.2. *Collaborative ranking*

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 concrete relevancy judgments by users with respect to query-page mappings. 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_{ij} 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 probability that p_j will be selected for query q_i (see Equation (1)).

$$\text{Relevance}(p_j, q_i) = \frac{H_{ij}}{\sum_{\forall j} H_{ij}} \tag{1}$$

Figures 2 and 3 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 single term query 'shakey' (refering to the robot developed at the Stanford Research Institute). Figure 2 shows the result-list returned before I-SPY has built-up its hit matrix data, and so the 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, AllTheWeb, WiseNut and HotBot. Clearly not all of the results presented are relevant and in fact only the 4th result is on-target. Since 'shakey' is a vague query it is no surprise that these results lack precision.

In contrast, Figure 3 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



Results **1** - **56** of **56** Search took 2.102 seconds

1.Shakey's® Inc., USA, Official Website, Official Website - Home
... Click here to read some great customer comments about Shakey's® or to leave some of your own! Shakey's® Inc., USA, Official Website ...
http://www.shakeys.com

2. Shakey Ranks Web Site
... popular reggae sound systems, Shakey Ranks comes to 'buss the place like a ... this web site, you can learn more about Shakey Ranks as well as listen to sound ... from cdbaby.com. You may also contact Shakey Ranks for bookings in your area ...
Description: Having created one of Barbados' most popular reggae sound systems, Shakey Ranks comes to 'buss the place like a storm' with his first solo album Mr. Progressive
http://www.ireggae.com/shakey

3. Welcome to Shakey's Place 3D
Choose Your Destination: ...
http://library.thinkquest.org/10502

4. Moravec Robot book figure
Shakey the robot 1970-Shakey the robot reasons about its blocks Built at Stanford Research Institute, Shakey was remote controlled by a large computer. ...
http://www.frc.ri.cmu.edu/~hpm/book98/fig.ch2/p027.html

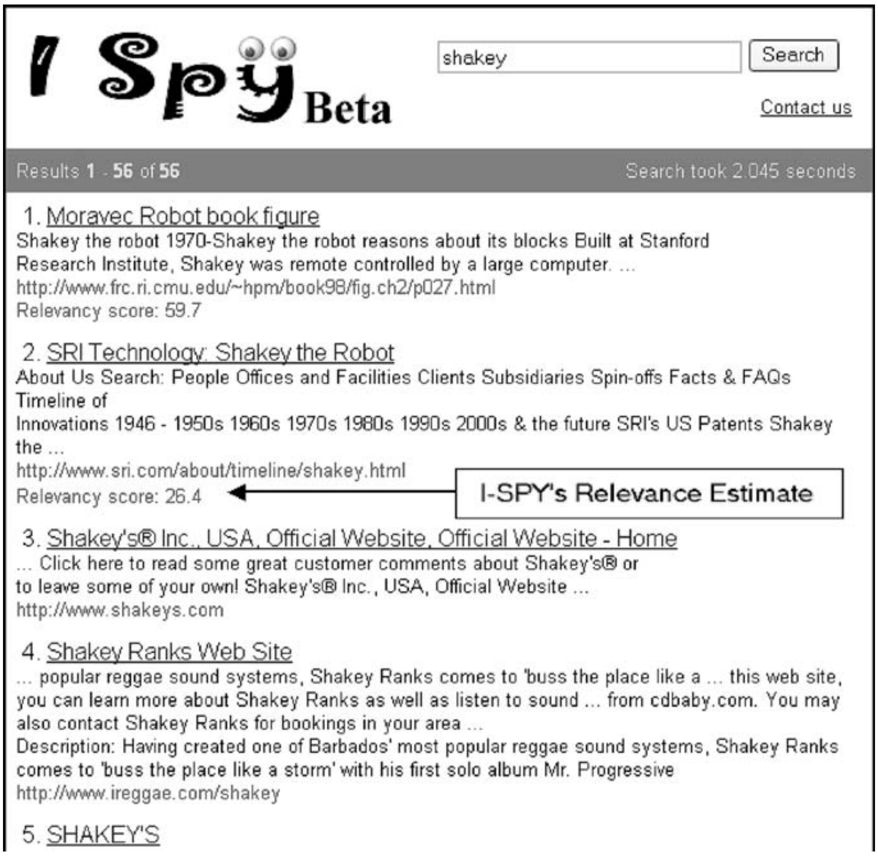
5. SHAKEY'S
The No.1 Pizza Restaurant In The Philippines. Welcome to Shakey's Pizza Restaurant! We're glad you're here to share in the fun. Sit ...
http://www.asiatravel.com/shakey

Figure 2. I-SPY search results before training.

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 2 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 developers 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.

3.3. *Community-based filtering*

A key point to understand about this relevancy metric is that it is tuned to the preferences of a particular set of users – a community of I-SPY



Results 1 - 56 of 56 Search took 2.045 seconds

1. Moravec Robot book figure
Shakey the robot 1970-Shakey the robot reasons about its blocks Built at Stanford Research Institute, Shakey was remote controlled by a large computer. ...
http://www.frc.ri.cmu.edu/~hpm/book98/fig.ch2/p027.html
Relevancy score: 59.7

2. SRI Technology: Shakey the Robot
About Us Search: People Offices and Facilities Clients Subsidiaries Spin-offs Facts & FAQs Timeline of
Innovations 1946 - 1950s 1960s 1970s 1980s 1990s 2000s & the future SRI's US Patents Shakey the ...
http://www.sri.com/about/timeline/shakey.html
Relevancy score: 26.4

3. Shakey's® Inc., USA, Official Website, Official Website - Home
... Click here to read some great customer comments about Shakey's® or
to leave some of your own! Shakey's® Inc., USA, Official Website ...
http://www.shakeys.com

4. Shakey Ranks Web Site
... popular reggae sound systems, Shakey Ranks comes to 'buss the place like a ... this web site, you can learn more about Shakey Ranks as well as listen to sound ... from cdbaby.com. You may also contact Shakey Ranks for bookings in your area ...
Description: Having created one of Barbados' most popular reggae sound systems, Shakey Ranks comes to 'buss the place like a storm' with his first solo album Mr. Progressive
http://www.ireggae.com/shakey

5. SHAKEY'S

Figure 3. I-SPY search results after training.

users – 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 I-SPY can deploy multiple I-SPY search agents, each with its own separate hit matrix. Thus the central I-SPY engine can be used to service many different search services across a range of portals, for example, each one adapted for the needs of a particular user group through its associated hit matrix. Alternatively, different hit matrices could be associated with different regions of the same site to bias search with respect to different topics. Placing a search box on a 'programming languages' directory page will naturally tend to capture 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. Evaluation

In this section we describe two separate but complementary evaluations of I-SPY and its collaborative search method. The basic hypothesis of I-SPY is that it is possible to learn implicit search context information by monitoring the selection behaviour of users, and that it is possible to leverage this context information to re-rank standard search results in a useful way. We evaluate this hypothesis in two ways. First we describe the results of an artificial user evaluation to determine the precision and recall characteristics of I-SPY when compared to a benchmark search engine (in this case the HotBot search engine). Second, we describe a recent larger-scale live-user evaluation that focuses on the search behaviour of 92 computer science students, split into control and test groups, as they attempt to use their search expertise to answer a series of test questions.

4.1. *Experiment 1 – artificial users*

In this experiment we use HotBot (www.hotbot.com) as the basic underlying search engine and we demonstrate how HotBot's raw results can be re-ranked by I-SPY as implicit context information is learned from the selection behaviour of search users. It is important to highlight that the evaluation is conducted by using an artificial model of user search behaviour. The artificial user model is informed by the real search behaviour of live users and since the results of this study are in broad agreement with recent live-user trials, we argue that this artificial user study is useful and informative.

4.1.1. *Setup*

The evaluation is conducted over four different search domains, each corresponding to a different subject area (*topic*) with a set of selected query terms and known context terms. In the following sections we describe these different domains plus the generation of query and context terms, the establishment of result relevance, and the role of an artificial user model to simulate user search behaviour.

4.1.1.1. *Topic domains and query generation.* We focus on four distinct topic domains, each of which roughly corresponds to a community of Internet users that are interested in a particular subject or topic area. For each domain a set of sample queries are generated; these are the *raw* or uncontextualised query terms. In addition, for each domain we agree on a set of *context terms* which, when combined with the raw query terms, provide a set of *contextualised* queries. For example, in the 'programming languages' topic domain we generate 74 raw query terms (e.g. 'java', 'pascal', 'perl', etc.) from which we derive 74 contextualised queries (e.g. 'programming language java', etc.).

– *Mammals*
  No. of Queries: 211
  Type of Queries: Names of mammals
  Source: Mammals subdirectory in Yahoo
  Context: 'mammal'
  URL: http://dir.yahoo.com/Science/Biology/Zoology/Animals_Insects_and_Pets/Mammals

– *Travel*
  No. of Queries: 202
  Type of Queries: Country names
  Source: Family Education Network's Countries of the World page
  Context: 'travel'
  URL: http://www.infoplease.com/countries.html

– *CBR and ML Researchers*
  No. of Queries: 69
  Type of Queries: People involved in CBR and ML research
  Source: David W. Aha's CBR and ML Researchers page
  Context: Affiliation (e.g. 'University College Dublin')
  URL: http://www.aic.nrl.navy.mil/aha/people.html

– *Programming Languages*
  No. of Queries: 74
  Type of Queries: Names of programming languages
  Source: Programming Languages subdirectory in Yahoo
  Context: 'programming language'

URL: http://dir.yahoo.com/Computers_and_Internet
/Programming_and_Development/Languages

4.1.1.2. *Establishing relevance.* Each query term is used to generate two lists of search results from HotBot. The first list, called the *raw results*, corresponds to the results returned by HotBot for each of the raw queries; HotBot returns up to 1000 results per query. The second list, called the *context results*, corresponds to the results returned by HotBot for the contextualised queries. The essential point is that, for the purpose of our evaluation, the context results are assumed to be those results that are actually *relevant* to the user. For example, consider the query 'jaguar' in the 'mammals' domain. The raw results from HotBot (arising from the 'jaguar' query) contain a diverse set of results including pages that are related to cats, cars, and operating systems. The results returned for the query 'mammal jaguar' are assumed to be relevant for this topic domain and make it possible to identify a subset of the raw HotBot results as relevant to the user. Thus, for each list of raw results we have a way of identifying which of these results are likely to be relevant to a user searching in a given topic domain.

4.1.1.3. *A user selection model.* Whether a user is likely to select a search result in a given search session depends on whether the result is relevant, but also on the position of the result in the result-list (earlier/higher results are far more likely to be selected than later/lower results). Our user selection model is informed by the search behaviour of 179 real users, observed over a period of 8 weeks and approximately 1500 search sessions. From this data we are able to model the probability that a given user is likely to select a relevant search result *r* given that it is ranked in position *k* in the result-list; we assume that the majority of user selections are for relevant results. For example, Figure 4 illustrates this probability distribution and indicates that users are very likely to select relevant results that occur in the top 3–5 positions, but that this probability quickly degrades with increasing result position. Our user selection model also includes a small random component to allow for the selection of irrelevant results by users, and during each search session the artificial user is limited to the selection of a predefined number of results, in this case *3* ± *3* results, as informed by our live-user data.

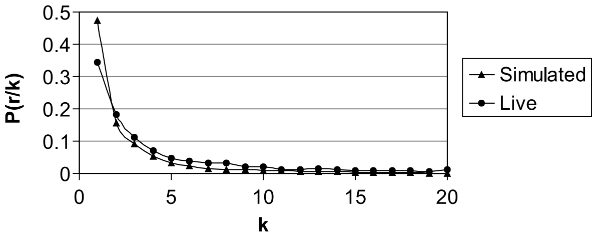


Figure 4. Observed user search behaviour shows a sharp decline in selection probability with increasing result position (k). The artificial selection model is tuned to closely match this selection behaviour.

4.1.2. *Methodology*

Our evaluation is carried out in the following way. For each topic domain, each query is submitted to HotBot between 100 and 200 times. Each raw result-list is processed, using our user model, to simulate user selections based on those results that are known to be relevant for the target query (according to the context results). These simulated user selections are used to populate an I-SPY hit matrix for the appropriate topic domain. The outcome is a hit matrix for each topic domain based on the selection behaviour of users in a given search context. Next, the queries are re-run (again between 100 and 200 times) but this time the result-lists are re-ranked by I-SPY, using the appropriate hit matrix to drive I-SPY's collaborative ranking engine. We calculate the precision and recall characteristics for these re-ranked I-SPY results, for different levels of k (result-list size), based on the known relevant-results data. Comparable precision and recall values are also computed for the raw results from HotBot as a benchmark.

4.1.3. *Results*

Figure 5 presents the results for each of the four domains, for both I-SPY and HotBot, as a graph of precision versus recall for each result-list size (k = 5–150). For clarity, the graphs have been partially annotated to indicate the k-values for individual data-points.

Overall the results demonstrate that there is a significant benefit to be derived from I-SPY's collaborative search technique – I-SPY's contextualised, re-ranked results have significantly higher precision and recall values (for a given k) than HotBot's original results. For example, for $k = 5$, the 'mammals' results (see Figure 5(a)) indicate that I-SPY delivers a precision of more than 0.8 and a recall of 0.12, as compared to a precision of less than 0.2 and recall of approximately 0.01 for HotBot.

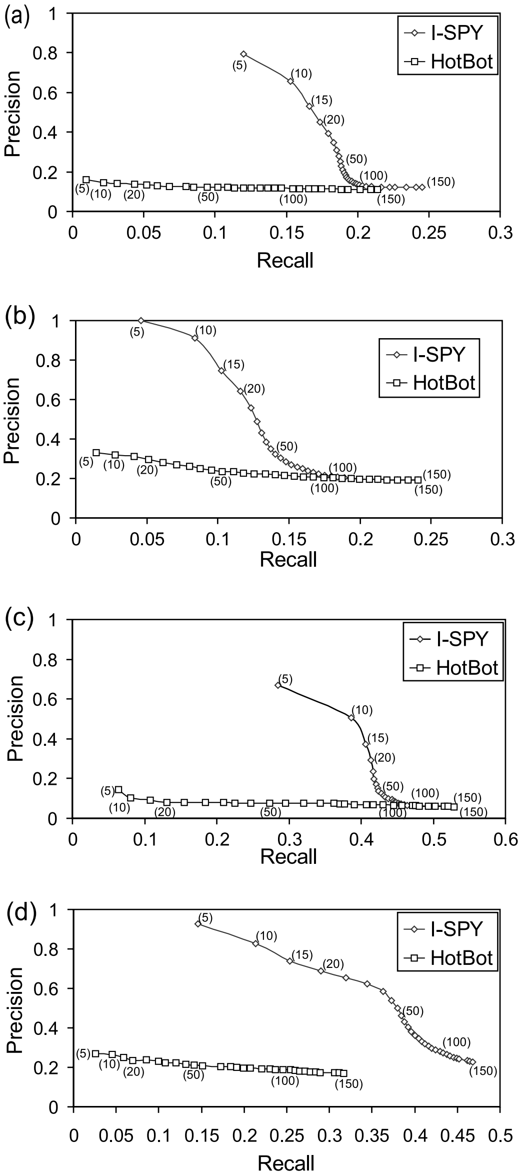


Figure 5. Precision vs. Recall for domain: (a) Mammals, (b) travel, (c) CBR and ML researchers, (d) programming languages.

In other words, in this domain, for I-SPY, approximately 4 out of the top 5 results are relevant. For HotBot, on average, only about 1 out of the top 5 results are likely to be relevant.

We find that precision tends to fall for I-SPY for increasing values of k. This is to be expected because many relevant results will occur low down in the original HotBot results lists and so have a low probability of being selected by users. Such results, although relevant, are unlikely to make it into the hit matrix. The precision values for I-SPY and HotBot tend to converge at about $k = 100$ indicating that most of I-SPY's promoted results were originally ranked within the top 100 HotBot results. Perhaps the most important feature of these precision results is that I-SPY's maximum benefit tends to occur at low values of k. This is particularly important, and useful, in the context of Web search, and other consumer search applications (e.g. mobile-phone search applications) where only limited-size result-lists can be presented.

On average the recall results are low, as expected, especially for low values of k. For instance, in the 'mammals' domain there are an average of about 30 relevant results per query, so the maximum recall at $k = 5$ is 0.166 (i.e. $\frac{5}{30}$). I-SPY achieves more than 75% of this maximum recall value at $k = 5$, whereas HotBot achieves only 6% of this maximum. Similarly, at $k = 10$, there is a maximum recall of 0.33 and I-SPY achieves nearly 47% of this (0.155 recall) compared to HotBot, which again reaches only 6% of this (0.02 recall).

In the 'mammals' and 'researchers' domains, I-SPY's recall characteristics begin significantly ahead of HotBot's. For example, I-SPY achieves a recall of 0.27 in the 'researchers' domain for $k = 5$. HotBot only achieves a recall of 0.06 for this k value, and in fact requires the retrieval of about 50 results to match I-SPY's recall. Why do these two domains offer I-SPY improved recall from the outset? Both domains are characterised by a high level of ambiguity in their raw queries, leading to lower numbers of relevant results from the outset. We can estimate query ambiguity in terms of the average number of relevant results per query in the raw result-lists; if all of the raw results are relevant then the raw queries are not ambiguous, but if very few raw results are relevant then the raw queries must have high ambiguity. For example, in the 'mammals' domain, on average only 3% of results per query are relevant and for 'researchers' only 1.2% of results per query are relevant. This is in contrast to 10% and 18% of results per query being relevant for the 'travel' and 'programming languages' domains, respectively. Thus there is a strong negative correlation between I-SPY's recall and the number of relevant results per query and thus I-SPY's benefits are likely to increase with the level of ambiguity in a typical user query.

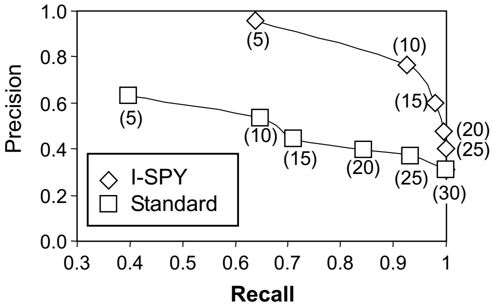


Figure 6. Precision vs. recall for live-users in the 'programming languages' domain.

4.1.4. *Summary*

In summary, we have shown that in theory I-SPY benefits from improved precision and recall characteristics when compared to HotBot. In particular, I-SPY enjoys vastly superior precision characteristics (>×2) for result lists up to $k = 20$. In addition, I-SPY can also achieve significantly superior recall, especially for ambiguous queries.

Of course, the above performance benefits are available 'in theory' only. The evaluation has been conducted using an artificial model of user search behaviour, and although this model has been designed with reference to the search behaviour of real users, doubts will naturally remain. In response it is worth drawing attention to an earlier evaluation of I-SPY (Smyth et al., 2003a) that does take advantage of live-user behaviour. The results of this earlier study are in broad agreement with the results presented here. For example, Figure 6 presents the precision-recall graph produced from this earlier study and the level of agreement with the current results should be clear. If such agreement was not present then there would be grounds to question the evaluation approach taken here, or at the very least, the user model used to simulate search behaviour. The fact that the results are so similar suggests that the user model is reasonable and the results valid.

However, this earlier live-user study also has its limitations. It was based on 20 computer science students and a very limited search domain, programming languages. In order to provide further evidence in support of collaborative search we have recently completed a larger-scale live-user evaluation, which we will describe in the following section.

4.2. *Experiment 2 – live users*

Our second experiment took place on Monday the 20th of October, 2003 and involved 92 computer science students from the Department of Computer Science at University College Dublin. It was designed to evaluate the benefits of I-SPY in the context of a realistic search task, a fact-finding exercise in this instance.

4.2.1. *Setup*

To frame the search task we developed a set of 25 general knowledge AI questions, each requiring the student to find out a particular fact (time, place, person's name, system name etc.); see Table 1 for a number of sample questions. Each student received a randomised list of these 25 questions – that is, 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 they were asked to record their answers and the URL where their answer was found on their question sheets. I-SPY was set-up with an empty hit matrix to draw on the results of 4 underlying search engines (Google, Hotbot, Wisenut and AllTheWeb).

4.2.2. *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. The second group served as the *test group*. They benefited from

Table 1. 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?

I-SPY's re-ranking based on their own incremental usage and the hit matrix produced by the first group.

Each group was allotted a 55-min slot in a supervised laboratory to perform their searches. Due to some minor technical difficulties group 1 were actually only able to search for 45 min rather than the allocated 55 min and, as a result, all of the statistics reported here relate to a truncated 45-min session for the group 2 users.

4.2.3. *Results*

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 answered?

Table 2 presents the mean questions answered for the 45 group 1 users and the 47 group 2 users. The mean total questions attempted per user is shown alongside the mean number of questions each user answered correctly and the corresponding mean score per user; the mean score is simply the percentage of the 25 questions that the average user answered correctly. There appears to be a clear advantage for the group 2 users, who answered more questions on average (9.9 vs. 13.9 for group 1 and group 2, respectively). Perhaps more to the point, group 2 users also answered more questions correctly. Indeed while the average group 2 user attempted 40% more questions than the average group 1 user, the average group 2 user answered 53% more questions correctly than the average group 1 user. Indeed, the average group 2 user answered more questions correctly (11.5) than the average group 1 user even attempted (9.9). I-SPY's collaborative search technique not only allows users to answer more questions, it appears to help users to answer more of these questions correctly.

Table 3 presents the average number of results selected per attempted question and the average position of these selected results for the group

Table 2. Mean questions answered per user

	Group 1	Group 2
Answered	9.9	13.9
Correct	7.5	11.5
Score (%)	30	46

Table 3. URL selection

	Group 1	Group 2
Selections/question	3.07	2.57
Result position	4.26	2.24

1 and group 2 users. Once again the group 2 users appear to benefit from I-SPY's re-ranking of results. The group 1 users selected in excess of 15% more search results than the group 2 users for a typical query and we believe that is because group 2 users benefited from the promotion of relevant results and absence of false-leads among I-SPY's reranked results. We argue that the extra selections made by the group 1 users were sub-optimal in the sense that they did not lead them directly to an answer. This is further supported by the average position of these selected results. 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.

4.2.4. *Summary*

The primary aim of this experiment was to demonstrate the effectiveness of I-SPY under more realistic search conditions with live-users. The results presented demonstrate a clear and significant advantage for the group 2 users when compared to the control group, indicating the benefits of I-SPY's collaborative search and re-ranking methods. The fact that the group 2 users were capable of answering more questions correctly by examining fewer search results is a strong demonstration of I-SPY's enhanced search performance.

Evaluation is obviously a major issue in this work, and as always it a real challenge to gain access to live-users as a means of fully testing our ideas. In this paper we have attempted to supplement our artificial-user studies with a reasonable live-user study. The latter is certainly not perfect. The 92 test users hardly represents a large sample in the context of Internet-scale search engine usage. In addition, the search task that we have set is not an open-ended one as might normally be the case. Nevertheless, the clear consistency between the results of this live-user evaluation and our artificial-user evaluation is noteworthy and does go a long way to proving the case for collaborative search. Obviously in the future we will continue to evaluate I-SPY over longer periods of time, with more users, and by considering more open-ended search tasks.

5. Discussion

During the development of this work a number of issues have been raised in relation to the collaborative search idea and the manner in which it is implemented in I-SPY. In this section we consider these issues in more detail and outline possible options for dealing with the potential problems that they introduce.

5.1. *User reliability, trust and authority*

Relevance re-ranking allows the filtering work of past users to promote the most relevant pages for current users. One of the benefits (and drawbacks) of this approach is that relevance is a priori assumed to be established by the previous users of the system. The obvious benefit of this technique is that content analysis does not have to be carried out to establish relevance. However, the quality of the reordering is obviously now dependent on the quality of the users' selections. This of course leaves a system like I-SPY open to abuse by users intent on falsely promoting certain results by repeatedly selecting them for a range of queries. We are currently considering a range of strategies that may help to protect I-SPY from this type of fraudulent activity. For example, simple strategies that discount subsequent selections may prove to be reliable for less sophisticated result tampering. Related work in the area of collaborative filtering, which looks at how rogue users can influence collaborative filtering recommendations and how collaborative filtering recommenders may be protected from such users, is also likely to prove useful (see O'Mahony et al., 2003).

On a related topic, the relevance score computation used by I-SPY assumes that the contribution by each member of the 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 McSherry'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 (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 but, at the same time, it is worth noting that it relies on the identification of individual users, which I-SPY purposefully avoids. Nevertheless, this issue is worthy of further research and is likely to be investigated in the future.

5.2. *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 attract user hits than new pages. In theory this means that these older pages may continue to be promoted ahead of more relevant, but newer, pages. What is worse, 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 relevancy 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.

5.3. *Composite vs. atomic queries*

Currently I-SPY treats each query as an atomic object so that even if it contains multiple query terms it is considered to be a single query and is indexed in I-SPY's hit matrix as such. This obviously leads to a number of limitations, not the least of which being that it limits I-SPY's ability to recognise and reuse related past queries. We are currently looking at how I-SPY can be adapted so that individual query terms are separately rated and indexed. We believe that this will allow I-SPY to influence the re-ordering of many more search sessions to the benefit of end-users.

6. Conclusions

The collaborative search idea attempts to discover patterns in the activity of a community of searchers in order to determine general search context and prioritise search results accordingly. It makes no strong assumptions about the form of the underlying search engines, and is generally applicable across a range of content types. The proposed ranking metric is computationally efficient and requires no additional parsing of the result pages. Finally, the ability to personalize search results for the needs of a community is achieved without the need to store individualised search histories; no individual user profiles are stored and no user identification is necessary. This has significant security and privacy advantages compared to many more traditional approaches to personalization.

In this paper we have described and evaluated the I-SPY implementation of collaborative search. Two major evaluations are described including an artificial-user study based on a realistic real-user search model and a comprehensive live-user study. The artificial-user results indicate a clear potential for significant precision and recall improvements, when compared to traditional Web search engines, highlighting the potential for collaborative search to add value to existing Web search engines. These artificial-user results are corroborated by the results of a recent live-user study, which clearly demonstrates similar search benefits due to I-SPY's collaborative search and re-ranking methods. The results from the experiments described in this paper complement both each other, and the more limited live-user study described in (Smyth et al., 2003b). All three studies strongly support the hypothesis that I-SPY benefits from superior precision and recall when compared to traditional search engines or meta-search techniques. Moreover, I-SPY's benefits appear to be particularly significant for small result-list sizes, which in turn suggests that the collaborative search approach is especially well suited for search on devices such as mobile phones and PDAs, with their limited display and input capabilities.

Future research will focus on a number of areas, including further evaluation work. We plan to assess the likely impact of rogue users on I-SPY's ranking metric and to develop techniques for protecting I-SPY against such attacks by users. In addition, I-SPY's hit matrix is a valuable source of relevancy information, and we are also investigating how it can be used to develop new models of page and query similarity that can be used during query expansion and page recommendation.

Acknowledgements

The support of the Informatics Research Initiative of Enterprise Ireland is gratefully acknowledged.

References

Bharat, K. (2000). SearchPad: Explicit Capture of Search Context to Support Web Search. *Proceedings of the Ninth International World-Wide Web Conference* **33**(1–6): 493–501.

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

Brin, S. & Page, L. (1998). The Anatomy of A Large-Scale Web Search Engine. *Proceedings of the Seventh International World-Wide Web Conference* **30**(1–7): 107–117.

Budzik, J. & Hammond, K. (2000). User Interactions with Everyday Applications as Context for Just-In-Time Information Access. *Proceedings International Conference on Intelligent User Interfaces*, ACM Press 44–51.

Dreilinger, D. & Howe, A. (1997). Experiences with Selecting Search Engines Using Meta Search. *ACM Transactions on Information Systems* **15**(3): 195–222.

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

Finkelstein L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G. & Ruppin, E. (2001). Placing Search in Context: The Concept Revisited. *Proceedings of the Tenth International World Wide Web Conference.*

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

Glover, E. J., Flake, G. W., Lawrence, S., Kruger, A., Pennock, D. P., Birmingham W. P. & Giles, C. L. (2001). Improving Category Specific Web Search by Learning Query Modifications. *2001 Symposium on Applications and the Internet* (*SAINT 2001*) *January 08–12, 2001 San Diego, CA.*

Haveliwala, T. H. (2002). Topic-Sensitive PageRank. *Proceedings of the World-Wide Web Conference*, ACM Press 784–796.

Kleinberg, J. M. (1998). Authoritative Sources in a Hyperlinked Environment. *Proceedings of the Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, 668–677.

Kruger, A., Giles, C. L., Coetzee, F., Glover, E., Flake, G., Lawrence S. & Omlin, C. (2000). Building a New Niche Search Engine. *Proceedings of the Ninth International Conference on Information and Knowledge Management* 272–281.

Kushmerick, N. (1997). Wrapper Induction for Information Extraction. *Proceedings of the International Joint Conference on Artificial Intelligence*, 729–735, Morgan-Kaufmann.

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

Lawrence, S. & Giles, C. L. (1998). Context and Page Analysis for Improved Web Search. *IEEE Internet Computing* July–August: 38–46.

Lawrence, S. & Giles, C. L. (1999a). Accessibility of Information on the Web. *Nature* **400**(6740): 107–109.

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

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

Mitra, M., Singhal, A. & Buckley, C. (1998). Improving Automatic Query Expansion. *Proceedings of ACM SIGIR*, ACM Press.

O'Mahony, M. P., Hurley, N. J. & Silvestre, G. C. M. (2003). An Evaluation of the Performance of Collaborative filtering. *14th Irish Artificial Intelligence and Cognitive Science* (*AICS 2003*) *Conference*.

Rhodes, B. J. & Starner, T. (1996). Remembrance Agent: A Continuously Running Automated Information Retrieval System. *Proceedings of the First International Conference on the Practical Applications of Intelligent Agents and Multi-Agent Technologies*, 487–495.

Selberg, E. & Etzioni, O. (1997). The Meta-Crawler Architecture for Resource Aggregation on the Web. *IEEE Expert* Jan–Feb: 11–14.

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

Smyth, B., Freyne, J., Coyle, M., Briggs P. & Balfe, E. (2003b). Collaborative Ranking in Community-Based Web Search. *14th Irish Artificial Intelligence and Cognitive Science* (*AICS 2003*) *Conference* 199–204.

TripAdvisor, Inc. www.tripadvisor.com.