Available online at www.sciencedirect.com
ScienceDirect
Information Processing and Management 43 (2007) 791–807
INFORMATION PROCESSING & MANAGEMENT
www.elsevier.com/locate/infoproman

Click data as implicit relevance feedback in web search

Seikyung Jung [a,*], Jonathan L. Herlocker [a], Janet Webster [b]

[a] School of Electrical Engineering and Computer Science, 1148 Kelly Engineering Center, Oregon State University, Corvallis, OR 97331-5501, USA

[b] Oregon State University Libraries, The Guin Library, Hatfield Marine Science Center, Newport, OR 97365, USA

Received 29 January 2006; received in revised form 11 July 2006; accepted 16 July 2006
Available online 17 October 2006

Abstract

Search sessions consist of a person presenting a query to a search engine, followed by that person examining the search results, selecting some of those search results for further review, possibly following some series of hyperlinks, and perhaps backtracking to previously viewed pages in the session. The series of pages selected for viewing in a search session, sometimes called the click data, is intuitively a source of relevance feedback information to the search engine. We are interested in how that relevance feedback can be used to improve the search results quality for all users, not just the current user. For example, the search engine could learn which documents are frequently visited when certain search queries are given.

In this article, we address three issues related to using click data as implicit relevance feedback: (1) How click data beyond the search results page might be more reliable than just the clicks from the search results page; (2) Whether we can further subselect from this click data to get even more reliable relevance feedback; and (3) How the reliability of click data for relevance feedback changes when the goal becomes finding one document for the user that completely meets their information needs (if possible). We refer to these documents as the ones that are *strictly relevant* to the query.

Our conclusions are based on empirical data from a live website with manual assessment of relevance. We found that considering all of the click data in a search session as relevance feedback has the potential to increase both precision and recall of the feedback data. We further found that, when the goal is identifying strictly relevant documents, that it could be useful to focus on *last visited* documents rather than all documents visited in a search session.

Keywords: Click data; Implicit feedback; Explicit feedback; Search engines; Information retrieval; Collaborative filtering; SERF

1. Introduction

Click data can be considered a form of relevance feedback. Classically, relevance feedback has referred to an information retrieval process whereby the user of the search engine indicates to the search engine that they would like ''more documents like this one''. The user is providing ''feedback'' to the system that ''relevant'' documents might look like the one indicated – thus *relevance feedback*. This is turn is used to improve the

* Corresponding author. Tel.: +1 541 231 5863.
E-mail addresses: jung@eecs.oregonstate.edu (S. Jung), herlock@eecs.oregonstate.edu (J.L. Herlocker), janet.webster@oregonstate.edu (J. Webster).

0306-4573/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ipm.2006.07.021

current search results for that user. Unlike the traditional work, we are interested in how relevance feedback from one user of a search engine can be used to improve the quality of search results for all users of the system.

Retrieval systems can collect relevance feedback from users in two different ways: explicitly or implicitly. Retrieval systems that collect *explicit feedback* ask users to mark documents in the search results that were relevant to their query. Systems that collect *implicit feedback* record and interpret users' behaviors as judgments of relevance without requiring additional actions from users.

Early retrieval systems that collected relevance feedback from users asked for explicit feedback (Rocchio, 1971). While explicit feedback from users clearly indicates what the user believes is relevant and useful, collecting it in sufficient quantity can be difficult. In order to get their own personal results, users often do not do the additional work to provide the feedback.

Inferring relevance from implicit feedback is based on the assumption that users continuously make tacit judgments of value while searching for information. Researchers have proposed or studied many forms of implicit feedback, including: clicks to select documents from a search results list (Smyth et al., 2005; Smyth, Freyne, Coyle, Briggs, & Balfe, 2003), scrolling down the text on a Web page (Claypool, Le, Wased, & Brown, 2001), book marking a page (Oard & Kim, 1998), printing a page (Oard & Kim, 1998), and the time spent on a page (Kelly & Belkin, 2001, 2004; Konstan et al., 1997; Morita & Shinoda, 1994; Oard & Kim, 1998; White, Jose, & Ruthven, 2003).

Inferences drawn from implicit feedback are often not as reliable as explicit relevance judgments. The potential for error in the additional inference step from the observed activity to the inferred relevance judgment increases the probability that there will be more documents that are erroneously marked as relevant. However, systems can often collect substantial quantities of implicit feedback without creating any additional burden on the user, and without changing the user experience. Thus by using implicit feedback we can achieve much greater coverage of relevance judgments over queries and documents. Furthermore, if we can collect sufficiently large quantities of data through implicit feedback, we should be able to separate the signal from the noise via aggregation.

Click data are particularly interesting implicit feedback data for several reasons. They are easy to collect in a non-laboratory environment (Joachims, 2002), and are more reliable than other forms of implicit feedback that are abundant (Joachims, Granka, Pan, & Gay, 2005). Most previous work has focused on clicks from search results list, but we believe that we can build even better search engines if we can incorporate implicit feedback based on each user's entire search session.

We investigated three major questions. (1) How using click data beyond the search results page might increase the precision and recall of a search engine over using just the clicks from the search results page; (2) Whether we can further subselect from this click data to get more reliable relevance feedback; and (3) How the reliability of click data for relevance feedback changes when the goal becomes finding one document for the user that completely meets their information needs (if possible). We refer to these documents as *strictly relevant*.

To answer our research questions, we analyzed three months of click data generated by the System for Electronic Recommendation Filtering (SERF), a university website search portal that tracks users' interactions with search results.

2. Related research

Three areas of related research are of particular interest: users' implicit behavior in retrieval as a relevance indicator; users' click data as evidence to judge search success; and collaborative filtering systems.

2.1. Users' implicit behavior in retrieval as a relevance indicator

One of our goals was to incorporate implicit relevance feedback that was abundant and reliable. Researchers have evaluated sources of implicit relevance feedback data. Though some have shown promise in experimental conditions, few have worked well in real world settings.

The authors of several early studies claimed that users' display time (duration) could be used as a document relevance indicator (Konstan et al., 1997; Morita & Shinoda, 1994; White et al., 2003; White, Ruthven, & Jose,

2002a, 2002b). Morita and Shinoda and others found that display time is indicative of interest when reading news stories (Konstan et al., 1997; Morita & Shinoda, 1994). White et al. (2003, 2002a, 2002b) used display time of a document's summary and claimed it was as reliable as explicit relevance feedback. However, other studies have shown that display time per document is not significantly related to the users' perception of the document relevance. Kelly and Belkin (2001, 2004) argued that display time was an unreliable indicator of relevance because factors unrelated to relevance – including tasks, the document collection, and the search environment – influenced display time.

Several researchers have studied combining display time with other behaviors in order to overcome these limitations (Claypool et al., 2001; Fox, Kamawat, Mydland, Dumais, & White, 2005; Oard & Kim, 1998). These researchers claim that examining multiple behaviors simultaneously could be sufficient to predict users' interest. Oard and Kim (1998) explored users' behavior including display time, printing, saving, scrolling and bookmarking. They found that display time together with whether a page was printed was a useful indicator of user interest. Others found that the combination of display time with the amount of scrolling can predict relevance in Web page browsing (Claypool et al., 2001). Finally, in recent work, Fox et al. (2005) found in a non-laboratory environment that the overall time that users interact with a search engine as well as the number of clicks users make per query seemed to indicate users' satisfaction with a document.

In summary, previously published work does not consistently support the hypothesis that users' display time alone is an adequate implicit measure of relevance. Furthermore, while we believe that printing, saving, bookmarking, and emailing are likely reliable indicators of relevance, such events are not abundant. Thus, we did not attempt to address any of those measures in this work.

2.2. Applying users' click data as evidence to judge document relevance

Researchers have studied ways to improve document retrieval algorithms for search engines by leveraging users' click data. Their underlying assumption is that clicked documents are more relevant than the documents that users passed over, so users' click data could be used as relevance feedback.

Researchers have used click data to train retrieval algorithms to re-rank results based on users' clicks (Cui, Wen, Nie, & Ma, 2002; Smyth et al., 2005, 2003; Xue et al., 2003). Specifically, Cui et al. (2002) claimed that extracting candidate terms from clicked documents and using those terms later to expand the query could improve search precision. Alternatively, Xue et al.'s (2003) approach automatically infers a link between two documents if a user selected both documents from the same set of search results. The I-SPY project (Smyth et al., 2005, 2003) re-ranked results based on the selection history of previous searchers and claimed that this approach improves search performance.

Approaches that clustered similar queries based on users' click data have also been successful (Wen, Nie, & Zhang, 2001, 2002). The assumption underlying these studies is that if users click on the same document for two different queries, then the two queries are likely to be similar. Wen et al. applied this approach to find frequently asked questions (FAQ). They claimed that a combination of both keyword similarity and click data is better than using either method alone to identify similar queries. On the other hand, Balfe and Smyth (2005, 2004), who also clustered queries utilizing users' click data, attempted to improve search precision by expanding the query with other terms in a cluster. They found that expanding queries based on keyword matching techniques performed better than expanding queries based on users' click data. However, the majority of work published claims that click data can be valuable.

Most recently, Joachims et al. (2005) studied why and how users click documents from the search results list. They claimed clicked documents potentially contain better information than the documents users passed over. This and the majority of the other studies described in this section suggest that click data could be an abundant and valuable source of relevance information.

2.3. Collaborative filtering systems

Collaborative Filtering (CF) is the process whereby a community of users with overlapping interests work together to separate interesting information from the non-interesting information. Each user can tap into the collection of all past evaluations by all other members of the community, and use those evaluations to help

select new, unseen information. Our work is motivated by the early studies of CF that focused on recommending items to individuals in entertainment related domains, such as music (Shardanand & Maes, 1995), movies (Hill, Stead, Rosenstein, & Fumas, 1995), jokes (Goldberg, Roeder, Guptra, & Perkins, 2001), and books (Linden, Jacobi, & Benson, 2001). However, applying CF to document search is more challenging than applying it to entertainment. In entertainment, people's taste change slowly, making predicting users' taste based on their previous preferences relatively easy. In document search, every time the user issues a new query, they may have a different information need than the previous query.

When relevance feedback is used to benefit all users of the search engine, then it can be considered collaborative filtering. Relevance feedback from one user indicates that a document is considered relevant for their current need. If that user's information need can be matched to others' information needs, then the relevance feedback can help improve the others' search results.

CF has been applied to the problem of recommending scientific literature in the context of the ResearchIndex system (Cosley, Lawrence, & Pennock, 2002; McNee et al., 2002), but only in the context of search-by-example. The AntWorld system was a web search support tool that collected users' explicit ratings on pages they visited, but it did not incorporate the users' implicit feedback (Boros, Kantor, & Neu, 1999; Kantor, Boros, Melamed, & Menkov, 1999; Kantor et al., 2000; Menkov, Neu, & Shi, 2000).

In the next section, we describe exactly how we apply the concept of collaborative filtering to document search, incorporating both explicit and implicit relevance feedback.

3. Experimental system: SERF

The System for Electronic Recommendation Filtering (SERF) is a prototype of a document search system developed at Oregon State University (OSU) that applies the technique of collaborative filtering – where the system improves in capabilities just by observing users' search sessions, both the queries and the subsequent navigation.

3.1. Searching in SERF

Users log in with their university accounts or use SERF anonymously. For users that have logged in, the search interface page includes a list of links to previous queries asked by the user, resources that are frequently visited by the user, and bookmarks that the user has stored (Fig. 1).

Users enter their text queries indicating their information need in the search box on the initial search page. Belkin et al. (2003) indicated that users are more likely to issue more keywords when given a larger, multi-line query input box. With this feature, we hoped to get longer and more descriptive queries from users.

Fig. 1. The initial search screen of SERF.

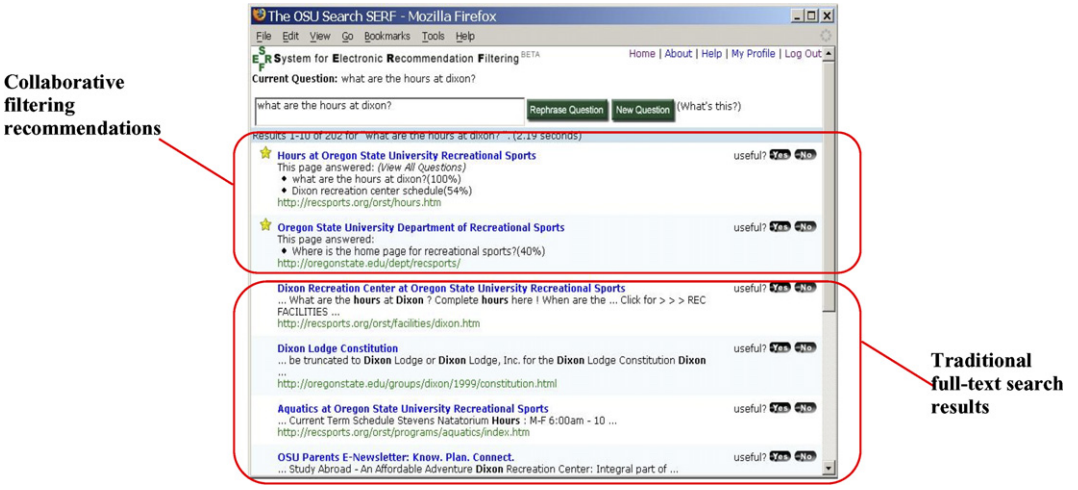

Fig. 2. The recommendations and search results screen of SERF. The stars indicate recommendations based on previously asked similar questions. The rest of the results are from the search engine NUTCH.

3.2. Search results

We presented results from Nutch,[1] a full-text document search engine as a base-line in the lower region of the search results screen (Fig. 2). Nutch is open source web-search software. It builds on Lucene Java, adding features specific for web search, such as a crawler, a link-graph database, and parsers for HTML and other document formats. Nutch returns links to web pages that contain the keywords in the user's query. We only indexed web sites affiliated with OSU.

In addition to the Nutch search results, SERF provides collaborative filtering recommendations. After receiving a search query, SERF takes the current user's query, locates past queries that are the most similar, and recommends those documents that were valuable to those past similar queries (Fig. 2). The associated past queries are displayed alongside the recommended document. Users can then personally determine, by examining those queries, if the recommended past queries are truly related to their current information need.

3.3. Capturing click data and explicit feedback

Once users submit queries to the system, their activity is tracked. When users click on a document from the search results list, that document is displayed within a frame controlled by SERF in the web browser (Fig. 3). The upper frame reminds the user about their entered query and provides links to rate, print or email the currently viewed document. Users may rate the currently viewed page's relevance for their current information need by clicking either the "YES" button to indicate the document was relevant, or "NO" button to indicate the document was unrelated. Navigation controls allow users to return directly to their search results or to submit new queries. Users also can rate each document useful or not useful directly from the results screen (Fig. 2). This allows them to provide feedback, if the relevance of a search result is clear from the displayed metadata and excerpt.

To collect click data and ratings beyond the search results list, all web pages are transferred through the SERF engine, which serves as a proxy. In the process, we rewrite all hyperlinks found in HTML so that when future hyperlinks are selected, the request is first routed to SERF. Thus, when users click on a link, SERF fetches the requested page from its original source, rewrites all hyperlinks found in the requested page, and displays the result within the rating frame. As long as users are searching with SERF, the rating frame never disappears, no matter which server is providing the original copy of the page.

[1] http://lucene.apache.org/nutch/

Fig. 3. The interface for viewing web pages within the SERF. The upper frame is always present while browsing, regardless of what site the user is visiting and allows users to rate the current document.

4. Methodology and data summary

4.1. Collecting and filtering data

We collected data logs from the SERF portal from March to May 2005. To promote usage, we linked to SERF from several prominent pages on the Oregon State University (OSU) website. Thus, the data reported are from "opt-in" users: users had to choose the SERF interface instead of using OSU's normal search engine (Fig. 4).

4.1.1. Search sessions

Over the span of three months, we gathered 297 search sessions from anonymous users and 104 sessions from logged-in users. Each *search session* represents a query from users (possibly the same query from different users), all the documents that were visited, and the ratings that were entered, until the user submitted a new query or left the system. Of the 297 search sessions, 202 search sessions include at least one visited document with at least one explicit rating. Of the 202 search sessions, 179 search sessions include at least one visited document rated by users as useful, while 69 search sessions have at least one document rated negatively (Fig. 5).

Approximately 60% of search sessions (179 out of 297) have at least one explicit positive rating compared to 30% in our previous study (Jung, Harris, Webster, & Herlocker, 2004).

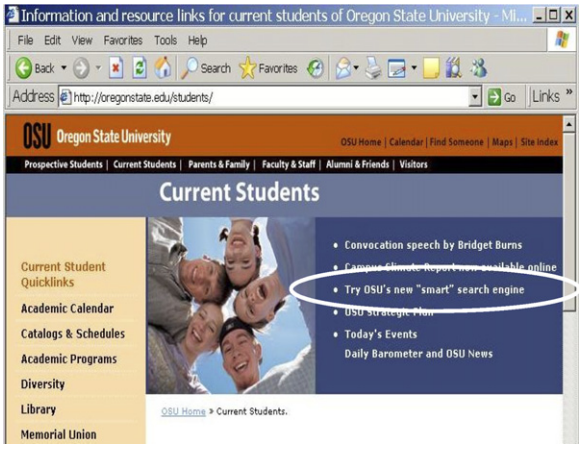

Fig. 4. The OSU site for student with the ink to the experimental prototype of SERF that is available to all OSU users.

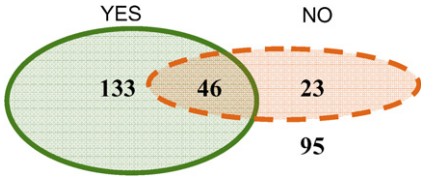

Fig. 5. Search sessions from collected data based on users' ratings. Numbers count search sessions (Total): 297 search sessions, 133 search sessions with at least one [useful=] YES rating and no NO ratings, 46 search sessions with at least one YES rating and at least one NO rating, 23 search sessions with at least one NO rating and no YES ratings, and 95 search sessions without ratings.

For analysis, we selected the 179 unique search sessions that included at least one visited/clicked document, and at least one document explicitly rating as relevant. We limited our analysis to this subset, because we wanted to compare how the relevance of documents clicked with that of documents explicitly rated as useful. Thus, we removed from consideration sessions that had only explicit negative ratings (23 search sessions) or did not have any ratings (95 search sessions).

4.1.2. Query statistics

Fig. 6 describes the queries in terms of the number of keywords in each query. Through the design of the system, we encouraged users to pose queries detailed enough so that others viewing those queries are able to identify the information need of the query. Thus, unlike most previous IR studies, in which users typically submit short queries of two to three words (Jansen, Spink, & Saracevic, 2000; Spink, Jansen, Wolfram, & Saracevic, 2002), our users' queries were more descriptive. The average length of queries issued from SERF was five words. Users of SERF frequently (70%) entered descriptive queries (queries in which we understood what users' information needs relatively well), and 47% of queries started with an interrogative word, such as "when", "where", "who", "what", "which", and "how". Query examples include:

- How can I join Dixon Recreation Center?
- Where is Kidder Hall?
- When does Weatherford dorm close for the summer?
- What is the Milne computing center's phone number?
- Student pay rate guidelines.

4.2. Measuring relevance

In order to understand how effective click data would be as a source of implicit relevance feedback, we measured the relevance of documents that were visited (clicked) during a search session. We manually reviewed each of the queries and the associated 691 visited documents. Within each search session, we assessed the bin-

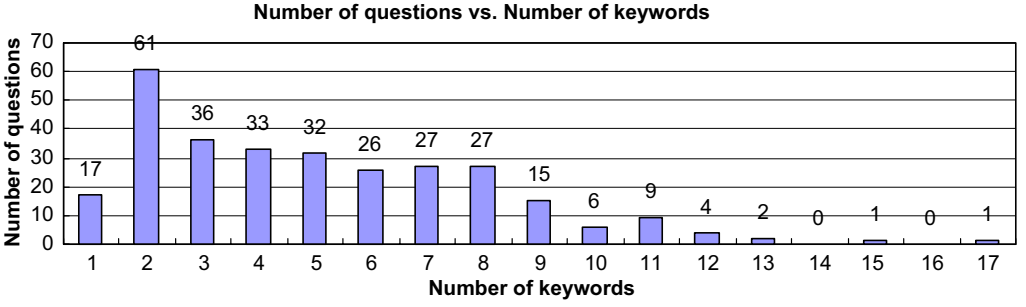

Fig. 6. The frequency of questions by the number of keywords.

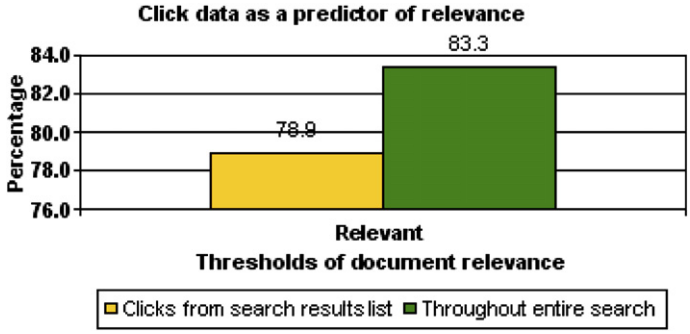

Fig. 7. Comparison of percentage of relevant documents as a predictor of relevance between clicks from search results and clicks throughout entire search.

ary relevance of each document visited to the query that initiated the session. Thus we separated visited documents into one of two categories: relevant or non-relevant.

We defined relevant documents to be documents that had cither a complete response to a user's query, had a partial, yet incomplete response to a user's query, or a link to a page with a partial or complete response to a user's query. For queries that were not self-describing, such as those with very small numbers of keywords, the relevance assessors were asked to imagine all the different types of responses that might be valuable to users who would ask the particular query. Any documents that fit that category were considered relevant. This was compatible with how SERF functions, where documents rated by past users with similar queries are presented to the current user as potentially relevant.

5. Click data beyond the search results page

Previous studies have examined how users' clicks originating from the search results list could be used as implicit relevance feedback for collaborative filtering-based search engines. Our hypothesis is that we can collect substantially more and better relevance feedback information by including the clicks that occurred after the user had followed a link from the search results pages. In other words, we believe that using the click data from the entire search is going to provide a better source of implicit relevance feedback. To understand this, we examined the differences in the number and percentage of relevant documents throughout the user's entire search compared to the subset of those clicks that originated from the search results page. The results from our SERF data are shown in Fig. 7.

The number of relevant documents found in a set of implicit relevance feedback data can be thought of as the *recall* of that relevance feedback. This terminology is particularly appropriate for SERF, where relevance feedback documents are candidates for recommendation to later users. The more relevant documents in the relevance feedback collection, the greater the "reach" of the recommendation system. In the SERF data, we see there are 261 relevant documents in the click data from the search results page, but we can double that number by including clicks beyond the search results (520 relevant documents).

So using all clicks as relevance feedback will increase the number of recall of our implicit relevance feedback data. As we increase the quantity of implicit feedback data, we would expect to also get an increased quantity of noise – irrelevant documents in the relevance feedback. Fig. 7 shows that while the total number of irrelevant documents in the relevance feedback increases, the percentage of relevant documents *actually increases!* 78.9% (261 out of 331) of documents reached by dicks from search results lists contained information relevant to the user's query. We can think of this as the *precision* of the implicit relevance feedback. When we include every document visited throughout the entire search session, we see that the precision increases to 83.3% (520 documents out of 624[2]).

[2] Among 691 total documents, 36 were not found (HTTP 404 error) and 31 were documents rated YES directly from search results list without clicking (Fig. 2, Section 3.1.2, so 67 documents are excluded in this result).

6. Last visited documents

While our data from SERF indicates that using all click data should have an advantage over just using clicks from search results, we are still interested in further refining the quality of implicit relevance feedback. With relevance feedback, search is very much like a classical machine learning problem – the goal is to take training examples and attempt to learn a function that maps from queries to documents. The relevance feedback provides the training examples, and most machine learning methods perform poorly when 16.7% of the training examples are erroneous. Can we further subselect from all click data to get more precise relevance feedback? In particular, we hypothesized that *last visited documents* in a search session might be a more precise subset of relevance feedback data. Our intuition was that the most relevant document may be the last place where users looked. Previous studies have shown that a majority of Web users tend to visit about eight web documents on average per query (Jansen & Spink, 2003; Jansen et al., 2000; Spink et al., 2002). Intuitively, if users were satisfied with the first document that they clicked on, then they would not need to see seven more documents. To explore this issue, we compared four different subsets of the click data, each described in Table 1.

Table 1
Subsets of users' click data that we compared

Clicks from the search results list	Documents reached directly from the search results
Last visited documents	The document last requested by users before initiating a new search or leaving the system
Explicitly rated useful	Document within click data explicitly rated as useful
Clicks beyond search results	Documents reached by following a link from a page other than the search results page

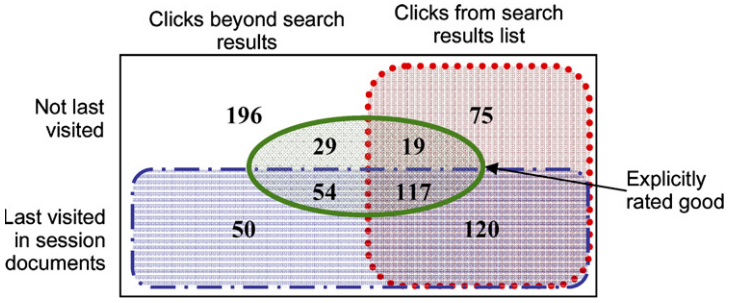

Fig. 8. Venn diagram of three sets (clicks from search results list, last visited documents, and explicitly rated good). Numbers count documents.

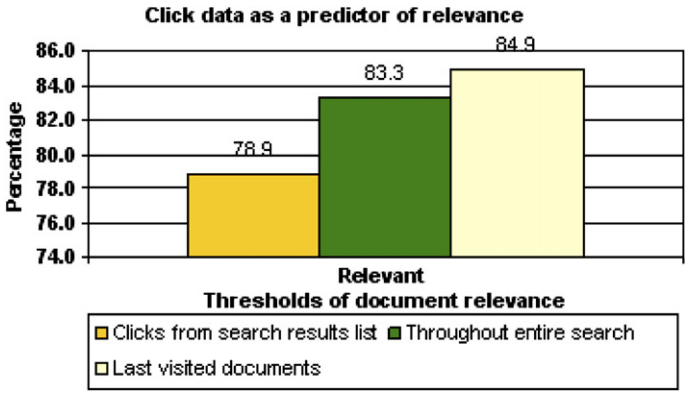

Fig. 9. Comparison of percentage of relevant documents as a predictor of relevance among clicks from search results, clicks throughout entire search, and last visited documents.

Fig. 8 depicts the relationship among the different subsets of the click data. During the 179 search sessions that we analyzed, users clicked on a total of 691 documents. Among these 691 documents, 331 (75 + 19 + 117 + 120) documents were reached by *clicks from search results list,* 329 (196 + 29 + 54 + 50) documents were reached by *clicks beyond search results*, 341 (50 + 54 + 117 + 120) documents were the last visited documents and 250 (29 + 19 + 54 + 117) documents were *explicitly rated* useful. While there were 179 unique queries, there are 341 last visited documents because in some cases the exact same quay was issued by multiple people.

We see there are 281 relevant documents from the last visited documents among 331 (84.9%, Fig. 9). This means that the last visited documents get more precise relevance feedback (84.9%) than other subsets of users' click data. However, using the last visited documents as relevance feedback will decrease the number of recall because we see there are 281 relevant documents in the click data from the last visited document, but we can double that number by including clicks beyond the search results (520 relevant documents).

7. Considering strict relevance

Up to now, we have considered relevance as it might be computed for a traditional general purpose search engine. We have shown evidence that including click data beyond the search results page may lead to implicit relevance feedback that not only has more recall but is also more precise. *Last visited documents*, a subset of click data, shows an increased precision over all other categories of implicit data.

One of our objectives in using the implicit relevance feedback data with the SERF system was to provide improved performance for questions that were frequently asked by members of the community. Our goal was to make the improvement dramatic. As much as possible, we wanted users to find what they needed in the first result they examined. Towards this objective, we introduce a new class of relevance -strictly relevant. Strictly relevant documents are a subset of relevant documents that contain information sufficient to satisfy the complete information need of a query. In other words, we remove from the list of documents that were previously considered relevant those documents those "partially relevant" documents as well as documents that have links to documents with relevant information (Table 2).

To illustrate the difference between how relevant documents and strictly relevant documents were assessed, consider the example in Table 3. In this example, we have a descriptive user query. Documents that contain all

Table 2
Two thresholds of document relevance[a]

Manual category	Two relevance thresholds	Description
Relevant	Strictly relevant	Document had the right answer
	Relevant, but not strictly relevant	Document had a partial answer, or document did not contain an answer, but had a link to answered or partially relevant documents
Unrelated		Document did not have any answers and was not related at all

[a] We excluded documents that could not found at assessment time (those returning a HTTP 404 error) from our analysis, since we could not judge them.

Table 3
Categorizing relevance of documents from descriptive searches

Query	Document	Relevance thresholds	
How do I get to the student parking lot?	Document having the sentence "…when you pass 11th street, the next right will take you into the parking lot…"	Strictly relevant	
	Home page of transit/parking services. Web pages of parking maps, tickets, or permit. Home page of facilities services having link to transit/parking services		Relevant

Table 4
Categorizing relevance of documents from a general searches

Query	Document	Relevance thresholds
Parking	Home page of transit/parking services	Strictly relevant
	Web pages of parking maps, tickets, or permit. Home page of facilities services having link to transit/parking services. Document having the sentence "... when you pass 11th street, the next right will take you into the parking lot..."	Relevant

the information necessary to satisfy the information need are those documents that provide the answer to the question posed in the query. In this case, document containing directions to the student parking lot will be considered strictly relevant. When a user comes across a strictly relevant document, they should no longer have to continue their search. Web pages that have related and relevant information about parking but do not answer the specific question may be considered relevant but are not considered strictly relevant.

It is more challenging to determine how to assess strict relevance for queries that had few keywords or were very non-specific as to the information need. We called these *general queries*. General queries do not provide enough context for external relevance assessors to identify what users expected to find. An example might be the single word query "parking". Without additional information, the system cannot differentiate between people looking for parking lots from those looking for parking permits or jobs in the parking enforcement office. To assess the strict relevance of these documents, we used the design of the SERF system to guide us. In SERF, the most prominent results are documents that were rated by other users with similar queries. Thus, we tried to find the documents that, given all the possible reasons that users from the community would ask the given query, minimize the expected click distance to the final answer. In the parking example, the most strictly relevant result would be the home page of transit/parking services (Table 4). We categorized the documents for parking maps, parking tickets, and parking permit pages as relevant because they would be too detailed for many users who issue the query "parking". The home page of facilities services is also relevant because the page has a link to go the home page of transit/parking services.

8. Strict relevance beyond the search results

Fig. 10 shows the precision of click data as relevance feedback considering the two different categories of relevance, strict and regular. The first group of two bars shows the data that we reported in Section 5. The second group of two bars shows the data when considering strict relevance. In each group, the first bar is the precision of just the clicks from the search results page as relevance feedback, and the second bar is the precision when considering all click data. We see a significant drop in precision when we make our definition of relevance stricter. Only 50.8% (168 out of 331) of documents clicked from the search results were strictly relevant to the query.[3] In other words, only half of the clicked documents from the search completely met the information need of the user's query. A drop in precision is not unexpected when moving from relevant to strictly relevant, given that we have dramatically increased the stringency of what it means to be relevant. However, most machine learning algorithms will operate poorly with a 50% error rate in training examples. We need to find a way to increase that precision.

Considering recall, in the strictly relevant case, including all clicks instead of just clicks from the search results increased the recall substantially to 277 documents (164% of 168 documents). Once again, this is not surprising. However, unlike in Section 5, the precision of the relevance feedback data dropped when including the extra click data from 50.8% to 44.4%. We are faced with a notable tradeoff between precision and recall.

[3] The remaining 70 documents did not contain relevant information for the queries (unrelated 17.2%) or were out of date already (those returning a HTTP 404 error, 3.9%).

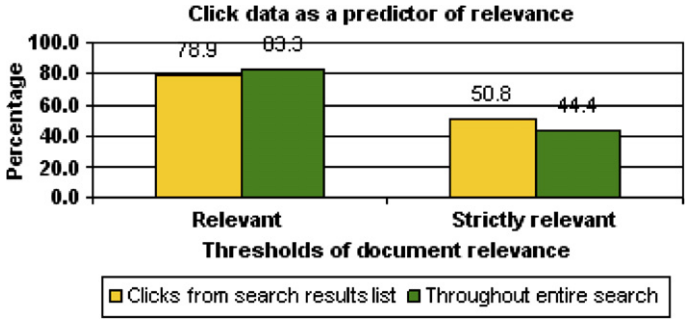

Fig. 10. Comparison of percentage of relevant documents as a predictor of relevance between clicks from search results and clicks throughout entire search.

9. Last visited documents and strict relevance

We have seen a significant change in precision and recall behavior when changing from measuring relevance to measuring strict relevance. It seems possible that we would see different results with respect to our hypothesis about last visited documents. Table 5 summarizes the precision of relevance feedback for the different subgroups of click data with respect to relevance and strict relevance. The results show that, among three types of implicit feedback, documents selected from *clicks beyond search results* were relevant the least often, and the *last visited documents* were relevant the most often.

It is interesting to note that, on average, documents selected *from the search results list* are strictly relevant significantly more than documents selected from *beyond the search results*, ($\chi^2(1) = 18.0$, $p < 0.001$). However, *last visited documents*, which represents a subset of click data taken from both the search results and beyond, shows an increased precision over all other categories of implicit data. The difference in precision between last visited documents and not last visited documents is significant for the strictly relevant case ($\chi^2(1) = 156.8$, $p < 0.001$). It is not significant for the non-strict case, but this could be due to lack of sufficient data points.

The data in Table 5 provides evidence that separating out the last visited documents will result in a set of relevance feedback data that is more precise. This in turn could be used to augment search engines and achieve search results that are more precise. In the relevant case, we see that the precision of the last visited documents is starting to approach that of the explicit ratings. This leads us to believe that the last visited documents do provide improved precision in both case of non-strict relevance and strictly relevance. However, the precision (68%) is considerably lower than the ideal precision that we achieve through explicit ratings (84.6%). The gap between those numbers indicates that there are more opportunities to refine the selection of implicit data, such that precision is maximized.

In the rightmost column, it is interesting to examine just how much disagreement there is. Roughly 7% (16/234) of documents explicitly rated by users as useful were determined by us to not be relevant to the query.

Table 5
The percentage of clicked documents among the users' data according to relevance judgments

	Lowest relevance → Highest relevance			
	Clicks beyond search results	Implicit		Explicit ratings
		From the search results	Last visited	
Strictly relevant	35.6% (109)	52.8% (168)	68.0% (225)	84.6% (198)
Relevant	84.6% (259)	82.1% (261)	84.9% (281)	93.1% (218)
Unrelated	15.4% (47)	17.9% (57)	15.1% (50)	6.8% (16)
Total	(306)	(318)	(331)	(234)

Description of relevance categories can be found in Table 2. Among 691 total visits to documents, 36 were not found (HTTP 404 error) and 31 were rated as useful directly from search results list without viewing the associated document, so 67 documents were excluded in this table.

Table 6
How does users' click data correspond to when users rate documents as useful?

	Clicks from search results list	Clicks beyond search results	Last visited	Not last visited
Rated YES	136 (41%)	83 (25%)	171 (50%)	48(15%)
Not rated or rated NO	195 (59%)	246 (75%)	170 (50%)	271 (85%)
Total	331	329	341	319

We do not know why users rated documents useful that, in our impression, were not relevant, but we would expect some number of random errors – users might have clicked the rating button by mistake, or perhaps they just wanted to see what happened. Or perhaps the errors were in our relevance assessments. We also would expect some disagreements between the users and our relevance assessments as to which documents are relevant. In any case, we could consider this as evidence that 93% (100 − 7) is the best possible result that we could hope to achieve with our methodology in the face of such noise in the data.

10. Using explicit ratings as relevance assessments

Relevance assessment is a process inherently full of variance, with many potential ways that bias could be unintentionally introduced into the process. To cross-validate our results, we applied the same methodology using a form of relevance assessment that could be considered authoritative: explicit ratings from the users. Our experimental platform allowed users to explicitly indicate, by clicking a button, if they found pages useful for their information need. If we assume that the user knows their own information needs best, then it follows that these explicit ratings for pages are authoritative relevance assessments. Results from this analysis are shown in Table 6.

The data parallels what we saw when we manually assessed relevance. The last visited documents category of click data has the highest percentage of explicit positive ratings, followed by the clicks from the search results list, then clicks beyond the search results list. This agrees with our previous assessment that the collection of last visited documents is a better collection of implicit relevance feedback than then all clicks or clicks from the search results page.

11. Discussion of results

One of the primary results of this work is evidence that the last visited documents in a search session will be better implicit relevance feedback data than other subsets of click data that have been explored. Furthermore, it is possible that our data actually underestimates the quality of last visited documents as relevance feedback. This is due to the fact that the last visited document of the current query, as we have computed it, might not be the last visited document of the session. We assumed that each query that a user submitted initiated a new search session, even if subsequent queries were actually reformulations of previous queries based on the same information need. Most users struggle to formulate search queries (Belkin, Oddy, & Brooks, 1982), and may not choose query terms that precisely represent their information needs, resulting in a need to reformulate their query. Spink, Jansen, and Ozmultu (2001) found that 33% of users reformulated their first query based on the retrieved search results. As a result, we may have attributed last-visited-document status to clicks that were followed by query reformulation that should have been considered part of the same search session because the information need did not change. This could have only increased the number of non-useful documents in the last-visited-documents category, and thus decreased the precision of relevance feedback we measured.

12. Remaining issues

12.1. Variance in quality of relevance feedback data

In traditional relevance feedback systems, the quality of a user's relevance feedback directly affects their search performance. Users have strong motivation to do their best to provide high quality relevance feedback.

Furthermore, if they provide bad relevance feedback, they immediately see the negative effects, and can correct their errors. The usages that we are proposing for relevance feedback do not have such self-correcting dynamics. One person's rankings will be changed based on relevance feedback from other users. There is a separation in space and time between the user who provides the relevance feedback and the user who views a ranked list that has been influenced by that feedback. The user providing the feedback may innocently provide incorrect data, and just never become aware of its effect. Or malicious users could intentionally attempt to influence the ranking algorithm through their misleading feedback. To make a robust system, further research is needed to identify the best methods for handling incorrect data in relevance feedback. Traditionally in collaborative filtering approaches this is done by requiring multiple corroborating sources of evidence towards the value of a particular piece of information. Through aggregation of data from multiple sources, we expect erroneous data to average out. Such approaches must be designed such that it is sufficiently hard for malicious users to automatically generate large volumes of relevance feedback. Other approaches include developing some sort of hierarchy or web of trust, where there are explicitly recognized trust relationships that can be leveraged to identify whose ratings are likely to be trustworthy.

If we follow the theme of collaborative filtering even further, we could enable the users to provide us with *metafeedback*, where users can view feedback that has been previously been given and inform the system when they feel that certain feedback is incorrect and should not be influencing the system. This could work particularly well, if users are shown examples of feedback that influenced their currently viewed search results. If they are unhappy with their results, and those results were strongly influenced by past feedback, then they can view the feedback and explicitly state their disagreement with the appropriateness or value of that feedback. We have begun to implement this approach with SERF. The strength of this approach is that humans are often much better than the computer at determining the relevance of a single contribution. The weakness is that the metafeedback itself must be monitored for potential misuse.

The challenge of ensuring the quality of relevance feedback data is must acute in search usage scenarios where significant financial gains can be had through manipulating search engine rankings. Many designers of commercial general search engines fear collaborative filtering in search as opening a door to a whole new world of "search engine spamming". The greatest opportunity for incorporating such collaborative filtering techniques into search engines is with smaller, more domain specific search engines, such as corporate intranet search engines. The economics of such environments will not encourage users to actively seek to manipulate the rankings. Furthermore, there may be more opportunities to explicit provide value to users who contribute strongly by providing feedback and metafeedback.

12.2. Feedback aggregation and query clustering

As we introduced in the previous Section 12.1, we can achieve more accurate and robust systems by aggregating multiple data points of feedback regarding a document's relevance to an information need. The analogy to book recommendations is that we have multiple people rating the same book. If we have enough people rating the book in an appropriate way, then we can average out a smaller number of cases where erroneous or misleading ratings have been given. With document search, the inherent overall value of a document is much less important. More important is the documents' value to the user's current information need. Thus, in order to aggregate feedback, we need to find multiple ratings for the exact same information need. This is extremely challenging if the only initial representation of the user's information need is the query. Two users might present the exact same query, yet have different information needs. This will be particularly true for very short queries. Two users with the same information need may word their queries differently.

There are several research threads that could be followed to improve the ability to aggregate feedback. If we assume that we can get users to issue a reasonable number of queries in longer natural language format, we could explore implementing some natural language analysis techniques, leveraging thesauri and knowledge of grammar. The danger of such approaches is that the queries issued by users are likely to violate many rules of grammar, even if the user thinks they are posing a well-formed query.

If we relax our desire to find strictly relevant pages, then we can consider query clustering to group together similar queries. Here we assume that our goal is to identify pages that are valuable for collections of very similar information needs. If we can identify means to effectively cluster queries, such that all queries in a cluster

have very similar information needs, then we can aggregate the feedback from all sessions initiated by a query from the cluster. Query clustering can be done based on the keywords in the query, but can also leverage the relevance feedback. If we find two different sessions with positive relevance feedback for the same documents, then we can infer that their information needs may be similar. One of the challenges of this approach is that we are adding another source of uncertainty – sometimes we will group together queries that may not really be that similar, leading to aggregation of data that may not be that similar.

12.3. Relevance can be time dependent

Another challenge of storing relevance feedback from one user and applying it to improve later searches is that some pages have a limited window of time in which they are valuable. For example, the relevant documents of the query "Who is the teacher of class CS411" might be different every year or term. It is crucial to update existing feedback periodically because some documents from users' feedback may be obsolete or may not exist any more. The system could easily detect pages that no longer exist, but there are many documents and pages on Intranets that contain outdated information. There are many techniques that could be used to handle these situations, and heuristics will probably work very well. For example, relevance feedback could be aged in the server, such that the influence of older data is slowly removed. Temporal analysis of the feedback data could identify when a particular document is starting to accumulate more negative ratings than positive votes, or when very similar but different documents are starting to attract more positive ratings. Or metafeedback could be provided to allow users to indicate when a recommendation page is out of date. We are implementing this as part of our overall metafeedback interface.

13. Conclusion

In this article, we have explored the reliability of click data as a source of implicit relevance feedback data and described a prototype system that uses that relevance feedback data to generate recommendations alongside traditional search results. Results from our SERF prototype suggest that using click data from the entire search session could be valuable, either because it increases the coverage of relevant documents (the recall), or because it increases the precision (for the non-strict relevance case). To achieve the maximal precision of feedback data, our data provides evidence that the "last visited document" of each search session is the more reliable source of implicit relevance feedback data. If we had been able to track query reformulations, we believe our results would have been even stronger. If you combine information about the last visited documents with further implicit feedback data, the reliability could be further increased. For example, did the user print, email, or copy and paste from the last visited document? Did they take notes about that document in another window? It is becoming increasingly possible to monitor those user actions (Dragunov et al., 2005). We conclude by stating that we continue to believe that integrating collaborative filtering ideas, such as we have described here, has the potential to create dramatically more effective search engines. However, there are many issues that still need to be resolved, most of them regarding the reliability of implicit feedback data in a complex human community.

Acknowledgments

Funding for this research has been provided by the National Science Foundation (NSF) under CAREER grant IIS-0133994, the Gray Family Chair for Innovative Library Services, the Oregon State Libraries and the NSF Research Experiences for Undergraduates program. We thank all our research members for their hard work in making the SERF happen.

References

Balfe, E., & Smyth, B. (2004). Improving web search through collaborative query recommendation. In R. Lopez de Mantaras & L. Saitta (Eds.), *Proceedings of the 16th European conference on artificial intelligence* (pp. 268–272). Amsterdam: IOS Press.

Balfe, E., & Smyth, B. (2005). An analysis of query similarity in collaborative web search. *Advances in Information Retrieval Lecture Notes, 3408*, 330–344.

Belkin, N. J., Cool, C., Kelly, D., Kim, G., Kim, J.-Y., & Lee, H.-J. (2003). Query length in interactive information retrieval. In *Proceedings of the 26th annual international ACM SIGIR* (pp. 205–212).

Belkin, N. J., Oddy, R. N., & Brooks, H. M. (1982). ASK for information retrieval: Part I. *Journal of Documentation, 38*, 61–71.

Boros, E., Kantor, P. B., & Neu, D. J. (1999). Pheromonic representation of user quests by digital structures. In M. K. Hlava & L. Woods (Eds.), *Proceedings of the 62nd annual meeting of american society for information science* (pp. 633–642). Medford, NJ: ASIS.

Claypool, M., Le, P., Wased, M., & Brown, D. (2001). Implicit interest indicators. In C. Sidner & J. Moore (Eds.), *Proceedings of the 6th international conference on intelligent user interfaces* (pp. 33–40). New York, NY: ACM Press.

Cosley, D., Lawrence, S., & Pennock, D. M. (2002). REFEREE: an open framework for practical testing of recommender systems using ResearchIndex. In *Proceedings of the 28th international conference on very large databases* (pp. 3546). San Francisco, CA: Morgan Kaufman.

Cui, H., Wen, J. R., Nie, J. Y., & Ma, W. Y. (2002). Probabilistic query expansion using query logs. In *Proceedings of the 11th international conference on world wide web* (pp. 325–332). New York, NY: ACM Press.

Dragunov, A., Dietterich, T. G., Johnstude, K., Mclaughin, M., Li, L., & Herlocker, J. L. (2005). Tasktracer: a desktop environment to support multi-tasking knowledge workers. *International Conference on Intelligent User Interfaces*, 75–82.

Fox, S., Kamawat, K., Mydland, M., Dumais, S., & White, T. (2005). Evaluating implicit measures to improve the search experiences. *ACM Transactions on Information Systems, 23*(2), 147–168.

Goldberg, K., Roeder, T., Guptra, D., & Perkins, C. (2001). Eigentaste: a constant-time collaborative filtering algorithm. *Information Retrieval, 4*(2), 133–151.

Hill, W., Stead, L., Rosenstein, M., & Fumas, G. (1995). Recommending and evaluating choices in a virtual community of use. In I. Katz, R. Mack, L. Marks, M. B. Rosson, & J. Nielsen (Eds.), *Proceedings of SIGCHI on human factors in computing systems* (pp. 194–201). New York, NY: ACM Press.

Jansen, B. J., & Spink, A. (2003). An analysis of web documents retrieval and viewed. In *The 4th international conference of internet computing, Las Vegas, NV* (pp. 65–69).

Jansen, B. J., Spink, A., & Saracevic, T. (2000). Real life, real users and real needs: a study and analysis of users' queries on the web. *Information Processing and Management, 36*(2), 207–227.

Joachims, T. (2002). Optimizing search engines using clickthrough data. In O. R. Zaiane (Ed.), *ACM international conference on knowledge discovery and data mining* (pp. 133–142). New York, NY: ACM Press.

Joachims, T., Granka, L., Pan, B., & Gay, G. (2005). Accurately interpreting clickthrough data as implicit feedback. In R. Baeza-White & N. Ziviani (Eds.), *Proceedings of the 28th annual international ACM SIGIR conference on research and development in information* (pp. 154–161). New York, NY: ACM Press.

Jung, S., Harris, K., Webster, J., & Herlocker, J. L. (2004) SERF: integrating human recommendations with search. In *Proceedings of the thirteenth ACM international conference on information and knowledge management (CIKM)* (pp. 571–580).

Kantor, P. B., Boros, E., Melamed, B., & Menkov, V. (1999). The information quest: a dynamic model of user's information needs. In M. K. Hlava & L. Woods (Eds.), *Proceedings of the 62nd annual meeting of american society for information science* (pp. 536–545). Medford, NJ: ASIS.

Kantor, P. B., Boros, E., Melamed, B., Menkov, V., Shapira, B., & Neu, D. L. (2000). Capturing human intelligence in the net. *Communications of the ACM, 43*(8), 112–115.

Kelly, D., & Belkin, N. J. (2001). Reading time, scrolling and interaction: exploring implicit sources of user preferences for relevance feedback. In D. H. Kraft, W. B. Croft, D. J. Harper, & J. Zobel (Eds.), *Proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 408–409). New York, NY: ACM Press.

Kelly, D., & Belkin, N. J. (2004). Display time as implicit feedback: understanding task effects. In M. Sanderson, K. Jarvelin, J. Allan, & P. Bruza (Eds.), *Proceedings of the 21 annual international ACM SIGIR conference on research and development in information retrieval* (pp. 377–384). New York, NY: ACM Press.

Konstan, X., Miller, B., Maltz, D., Herlocker, J., Gordon, L., & Riedl, J. (1997). GroupLens: applying collaborative filtering to UsenetNews. *Communications of the ACM, 40*(3), 77–87.

Linden, G. D., Jacobi, J. A., & Benson, E. A. (2001). *Collaborative recommendations using item-to-item similarity mappings*. U.S. Patent No. 6, 266, 649. Washington, DC: Patent and Trademark Office.

McNee, S. M., Albert, I., Cosley, D., Gopalkrishnan, P., Lam, S. K., Rashid, A. M., et al. (2002). On the recommending of citations for research papers. In E. F. Churchill, J. McCarthy, C. Neuwirth, & T. Rodden (Eds.), *Proceedings of the 2002 ACM conference on computer supported cooperative work* (pp. 116–125). New York, NY: ACM Press.

Menkov, V., Neu, D. J., & Shi, Q. (2000). AntWorld: a collaborative web search tool. In P. Kropf, G. Babin, J. Plaice, & H. linger (Eds.), *Proceedings of the 2000 workshop on distributed communications on the web* (pp. 13–22). Berlin: Springer-Verlag.

Morita, M., & Shinoda, Y. (1994). Information filtering based on user behavior analysis and best match text retrieval. In W. B. Croft & C. J. van Rijsbergen (Eds.), *Proceedings of the 7th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 272–281). New York, NY: Springer-Verlag.

Oard, D., & Kim, J. (1998). Implicit feedback for recommender systems. In H. A. Kautz (Ed.), *Recommender systems: Papers from a 1998 workshop* (pp. 81–83). Menlo Park, CA: AAAI Press.

Rocchio, J. (1971). Relevance feedback in information retrieval. In G. Salton (Ed.), *The SMART retrieval system: Experiments in automatic document processing* (pp. 313–323). Englewood Cliffs, NJ: Prentice-Hall.

Shardanand, U., & Maes, P. (1995). Social information filtering: algorithms for automating "word of mouth". In I. R. Katz & R. Mack (Eds.), *Proceedings on Human Factors in Computing Systems* (pp. 210–217). New York, NY: ACM Press.

Smyth, B., Balfe, E., Freyne, E., Briggs, P., Coyle, M., & Boydell, O. (2005). Exploiting query repetition & regularity in an adaptive community-based web search engine. *User Modeling and User-Adapted Interaction: The Journal of Personalization Research, 14*(5), 383–423.

Smyth, B., Freyne, J., Coyle, M., Briggs, P., & Balfe, E. (2003). I-SPY: anonymous, community-based personalization by collaborative web search. In M. A. Bramer & R. Ellis (Eds.), *Proceedings of the 23rd SGAI international conference on innovative techniques and applications of artificial intelligence* (pp. 367–380). London: Springer-Verlag.

Spink, A., Jansen, B. X., & Ozmultu, C. (2001). Use of query reformulation and relevance feedback by excite users. *Internet Research, 10*(4), 317–328.

Spink, A., Jansen, B. J., Wolfram, D., & Saracevic, T. (2002). From e-sex to e-commerce: web search changes. *IEEE Computer, 35*(3), 107–109.

Wen, J.-R., Nie, J.-Y., & Zhang, H.-J. (2001). Clustering user queries of a search engine. In V. Y. Shen, N. Saito, M. R. Lyu, & M. E. Zurko (Eds.), *Proceedings of the 10th international conference on world wide web* (pp. 162–168). New York, NY: ACM Press.

Wen, J.-R., Nie, J.-Y., & Zhang, H.-J. (2002). Query clustering using user logs. *ACM Transactions on Information Systems, 20*(1), 59–81.

White, R. W., Jose, J. M., & Ruthven, I. (2003). A task-oriented study on the influencing effects of query-biased summarization in the web searching. *Information Processing and Management, 39*(5), 669–807.

White, R. W., Ruthven, I., & Jose, J. M. (2002a). The use of implicit evidence for relevance feedback in web retrieval. *Advances in Information Retrieval Lecture Notes, 2291*, 93–109.

White, R. W., Ruthven, I., & Jose, J. M. (2002b). Finding relevant documents using top ranking sentences: an evaluation of two alternative schemes. In K. Jarvelin, M. Beaulieu, R. Baeza-Yates, & S. H. Myaeng (Eds.), *Proceedings of the 25th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 57–64). New York, NY: ACM Press.

Xue, G.-R., Zeng, H.-J., Chen, Z., Ma, W.-Y., Zhang, H.-J., & Lu, C.-J. (2003). Implicit link analysis for small web search. In C. Clarke & G. Cormack (Eds.), *Proceedings of the 26th international ACM SIGIR conference on research and development in information retrieval* (pp. 56–63). New York, NY: ACM.