SERF: Integrating Human Recommendations with Search

Seikyung Jung[1], Kevin Harris[1], Janet Webster[2] and Jonathan L. Herlocker[1]
[1]Northwest Alliance for Computational Science and Engineering
[2]Oregon State University Libraries

Oregon State University
218 CH2M Hill Alumni Center
Corvallis, OR 97331
541-737-6601

{jung, harriske, herlock}@nacse.org, Janet.Webster@oregonstate.edu

ABSTRACT

Today's university library has many digitally accessible resources, both indexes to content and considerable original content. Using off-the-shelf search technology provides a single point of access into library resources, but we have found that such full-text indexing technology is not entirely satisfactory for library searching.

In response to this, we report initial usage results from a prototype of an entirely new type of search engine – The System for Electronic Recommendation Filtering (SERF) – that we have designed and deployed for the Oregon State University (OSU) Libraries. SERF encourages users to enter longer and more informative queries, and collects ratings from users as to whether search results meet their information need or not. These ratings are used to make recommendations to later users with similar needs. Over time, SERF learns from the users what documents are valuable for what information needs.

In this paper, we focus on understanding whether such recommendations can increase other users' search efficiency and effectiveness in library website searching.

Based on examination of three months of usage as an alternative search interface available to all users of the Oregon State University Libraries website (http://osulibrary.oregonstate.edu/), we found strong evidence that the recommendations with human evaluation could increase the efficiency as well as effectiveness of the library website search process. Those users who received recommendations needed to examine fewer results, and recommended documents were rated much higher than documents returned by a traditional search engine.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; H3.7 [**Information Storage and Retrieval**]: Digital Libraries.

CIKM'04, November 8-13, 2004, Washington, DC, USA.
Copyright 2004 ACM 1-58113-874-1/04/0011...$5.00.

General Terms

Algorithms, Experimentation, Human Factors.

Keywords

Digital libraries, collaborative filtering, web search, information retrieval, user studies.

1. INTRODUCTION

University libraries are traditionally viewed as physical repositories of information such as books, maps, and journals, and more recently as aggregators of proprietary databases and journal indexes. However, the current generation of students (and many faculty members) has come to expect that information resources should be accessible from their own computer, visiting the library virtually, but not physically. To address this evolving need, we (the OSU Libraries and the School of Electrical Engineering & Computer Science) are researching new interfaces for integrating the existing library interfaces into a single, highly effective user interface.

Today's research university library has many digitally accessible resources, from web-based interfaces for the traditional journal indexes and card catalogs to new digital special collections that incorporate highly interactive map-based interfaces. The quantity of these resources is large – they cannot usefully be enumerated on a single web page. To provide access to these resources, classic web search technology has been applied, but we have found its utility is poor and its usage is low. The work reported in this paper represents our attempt to design a new type of search engine and accompanying user interface that could successfully serve the needs of the library community and potentially many other domains as well.

To understand the need for a new approach to search, we must consider the weaknesses of existing search technology. Existing search technology is *content-based* – it is based on matching keywords in a user query to keywords appearing in the full-text of a document (the content). Such an approach fails to retrieve the best results in many well-known cases. Examples include the inabilities to recognize word context or higher level concepts, recognize synonyms, and identify documents that have little or no text. Many of the most popular search results within our library have little or no text associated with them. Examples of such popular pages include journal indexes (database search interfaces) and the scanned pages of Linus Pauling's research notebooks[1]. However, most notable is the lack of ability of content-based search to differentiate quality of search results.

Researchers have proposed several possibilities to overcome some of these weaknesses of content-based search. Example approaches include explicit link analysis [17], implicit link analysis [21], popularity-ranking [7], and page re-ranking with web query categorization [9]. Example systems include FAQ finder [5] and Meta search engines [14]. The most commercially successful of these approaches, the link analysis techniques (i.e., Google), does not appear to provide significant performance increases compared to traditional keyword-only search, when applied to our Library web site. We believe that this is because our library domain does not have enough cross-links to effectively alter the search result rankings.

However, we can consider why link-based analysis works for highly-linked collections, and use that to motivate a new type of search that is not limited to highly-linked collections. The success of link-based analysis in global web search is based on the premise that links are implicit records of human relevance judgments. The intuition is that a web page author would not link to an external page, unless that page is both relevant and perceived to have some value. Thus, link-based analysis improves upon content-based analysis by including humans "in the loop" of identifying relevant and high-quality documents. How can we create a search engine that incorporates significant human analysis into the search results, without relying on hyperlinks in the content?

In this paper, we propose a System for Electronic Recommendation Filtering (SERF), which is a library website search portal that incorporates explicit human evaluation of content on a large, de-centralized scale and tracks users' interactions with search results in a sophisticated manner. We are attempting to establish a middle ground, taking advantage of the shared information needs among users as well as using traditional content analysis.

The operation of the system is as follows. First, the user issues a human-readable question or statement of information need (a query) in text. If previous users have issued similar queries, then SERF recommends documents, sites, or databases that SERF believes those previous users found relevant and useful. SERF determines that resources are relevant to a question by observing either a) an explicit user statement that the resource is valuable or b) some activity by the user that implies that a resource is useful. To find similar information needs, we use a keyword matching technique.

We have deployed an experimental prototype of SERF on the OSU Libraries web site[2] that is available to all users of the library. This paper presents some early results, based on analysis of the usage of SERF over a three month period. We examine how the users participated in the system and we focus on the following research question: *can a user find the right information more efficiently and effectively when given automatically detected recommendations from past users in a library website?*

2. RELATED WORK

Two areas of related research are of particular relevance: work on collaborative filtering systems and work on document search engines.

2.1 Collaborative Filtering Systems

Collaborative Filtering (CF) is the process whereby a community of users with overlapping interests work together to separate interesting information from non-interesting information. In CF, each member of the community shares their evaluation of each content item they experience. Then 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 SERF approach in essence is adapting CF for library resource searching.

Early studies of CF have focused on recommending items to individuals in entertainment related domains, such as music [19], movies [11], jokes [8], and books (http://www.amazon.com/). More recently, CF has been applied to the problem of recommending scientific literature in the context of the ResearchIndex system [6, 15]. However, the recommenders built for ResearchIndex only support query by example: users specify examples of scientific articles, and the citations of those articles are used to locate other related articles.

Perhaps most related to our SERF was a research system called AntWorld, which is designed to help users manage their web searching better and to share their findings with other people [4, 12, 13, 16]. AntWorld was a web search support tool, where users describe their "quests" before browsing or searching the web. When a user enters a new quest, that quest is compared to previously entered quests. At any point during a quest, a user may choose to "judge" the currently viewed web page, in essence rating its relevance to the quest. To our knowledge, AntWorld has never been evaluated in an empirical user study.

2.2 Document Search Engines

Most studies of information retrieval in document search have been based on keyword-based full text search [2]. Most recently, with a rapidly growing number of web documents, many researchers and products have been exploring other possibilities that could help separate useful information from less useful or useless information.

Examples of approaches that are appropriate for searching the global web include the PageRank algorithm of Google that takes advantage of the link structure of the web to produce a global importance ranking of every web page [16]. The DirectHit web search engine extracted useful information from users' access logs [7]. Gravano et al. [9] examined geographical locality associated with a search query to re-rank search results. Meta search engines, which filter search results returned by several web search engines, are also another effort to refine searching [14].

For narrower domains, Xue et al. [22] automatically inferred link analysis between two documents if a user selected both documents from the same set of search results. FAQ Finder matches a user's query with questions in the Frequently Asked Question files [5].

However, none of these techniques incorporate human evaluations as explicitly as we do in SERF.

[1] http://osulibrary.oregonstate.edu/specialcollections/rnb/

[2] OSU Library, http://osulibrary.oregonstate.edu/

3. SERF

The design and implementation of SERF was motivated by our desire to have a more effective web search for the library and our approach is strongly influenced by previous work in collaborative filtering.

The central intuition behind the design of SERF is that many users of the library will have very similar or even identical information needs. For example, we may have 300 students who all need to find the same materials for the COM101 class. For a more general example, we have a population of researchers in the biological sciences who all would be interested in accessing research resources related to biology. If we have multiple people with the same or similar information need, why should all of them have to dig through the library web site to locate the appropriate information? Rather, we should learn from the experience of the first person to encounter the information need and use that learning to decrease the time and effort needed by the remaining users who have the same need.

In traditional collaborative filtering systems, we assume that users have information needs that remain mostly consistent over time. We then match users with similar interests, and transfer recommendations between them. Essentially, traditional CF assumes that a user's query is always the same. This is clearly not appropriate for a resource search system that we needed for our library.

Our approach is to apply collaborative filtering in a novel way. Rather than matching users with similar interests, we match *information contexts*. An information context includes not only a user's profile of past interests, but also some representation of their immediate information need. In our approach, we use the user-specified text query as the indicator of their immediate need.

Once users log in and submit text queries to the system (thus establishing their information context), their activity is tracked. They can rate resources as valuable to their information context, or SERF can infer from their activity that a resource is valuable. After receiving a search query, SERF takes the current user's information context, locates past information contexts that are the most similar, and recommends those resources that were valuable to those past, similar information contexts. Associated with each recommendation is the original text question from the previous information context. The question is displayed alongside the recommendation. The user can then personally determine, by examining those questions, if the recommended past information contexts are truly related to their immediate information need.

In the next four sub-sections, we examine different aspects of SERF in detail.

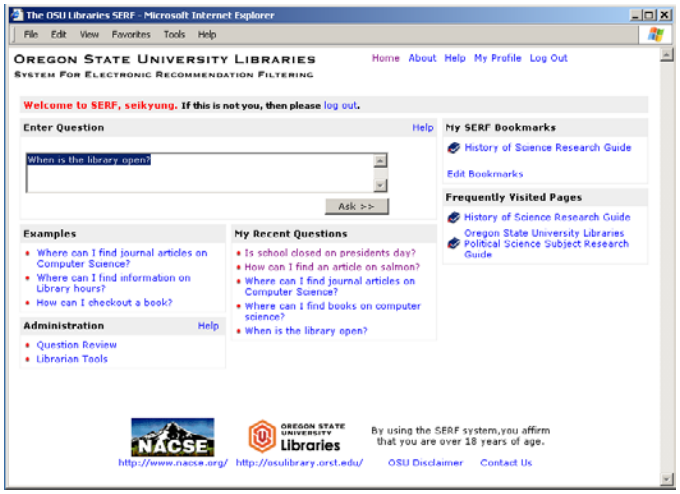
3.1 The Initial Search Page

The initial search interface is where the user enters the text query indicating their immediate information need. Initially, we gave no instructions to users regarding query formulation and used a traditional small text entry box. However, we quickly determined that most users entered short queries consisting of a few keywords out of context. When we displayed a previous user's query as part of a recommendation, the current user generally was not able to determine if the past users' information need had been similar.

We addressed this issue by encouraging users to utilize complete natural language sentences to specify their information need, rather than just a few keywords, as one might use with a popular search engine. We called these natural language queries *questions* (question format queries).

We investigated ways that we could use user interface elements to encourage the "proper" behavior [3, 21]. One report, by Belkin, et al. [3], indicated that users are more likely to issue more keywords when given a larger, multi-line query input box. We chose to create a query input box consisting of 3 rows of 74 columns and prompted the user to enter a question (Figure 1). Furthermore, we place a randomly selected question from a manually selected set of questions of the appropriate format in the text box to illustrate an example query. This question is also highlighted so that the user can type a new question and automatically erase the one existing. Finally, we list out several examples of good search questions below the search box.

Users can log in or utilize SERF anonymously. For users that have logged in, the search interface page also contains a list of links to previous questions asked by the user, resources that are frequently visited by the user, and explicit bookmarks that the user has created. In the future, we intend to allow students and faculty to authenticate with their university accounts, which will give us basic demographic information to use in matching their information contexts.


The OSU Libraries SERF - Microsoft Internet Explorer
File Edit View Favorites Tools Help
OREGON STATE UNIVERSITY LIBRARIES
SYSTEM FOR ELECTRONIC RECOMMENDATION FILTERING
Home About Help My Profile Log Out
Enter Question
Help
Ask >>
Examples
My Recent Questions
Administration
Help
My SERF Bookmarks
Frequently Visited Pages
NACSE
OREGON STATE UNIVERSITY Libraries

Welcome to SERF, seikyung. If this is not you, then please log out.

When is the library open?

Examples
- Where can I find journal articles on Computer Science?
- Where can I find information on Library hours?
- How can I checkout a book?

My Recent Questions
- Is school closed on presidents day?
- How can I find an article on salmon?
- Where can I find journal articles on Computer Science?
- Where can I find books on computer science?
- When is the library open?

Administration
- Question Review
- Librarian Tools

My SERF Bookmarks
- History of Science Research Guide

Edit Bookmarks

Frequently Visited Pages
- History of Science Research Guide
- Oregon State University Libraries Political Science Subject Research Guide

http://www.nacse.org/ http://osulibrary.orst.edu/

By using the SERF system,you affirm that you are over 18 years of age.

OSU Disclaimer Contact Us

Figure 1. The Initial Search Screen of SERF

3.2 Recommendations

The search results page in SERF has two separate regions. The first region at the top contains recommendations based on previous information contexts and the region below that contains search results from a traditional search engine (Figure 2). Here we describe the first region; the second region is described in Section 3.3.

Recommendations consist of similar questions that have been asked before and associated documents rated high for that question. The similarity between the current question Q_c, and a previously asked question Q_p, is computed as follows:

$$Sim(Q_{p_j}, Q_c) = \frac{\vec{Q}_{p_j} \cdot \vec{Q}_c}{|\vec{Q}_{p_j}||\vec{Q}_c|} = \frac{\sum_{i=1}^{t} w_{i,p_j} \times w_{i,c}}{\sqrt{\sum_{i=1}^{t} w_{i,p_j}^2} \times \sqrt{\sum_{i=1}^{t} w_{i,c}^2}}$$

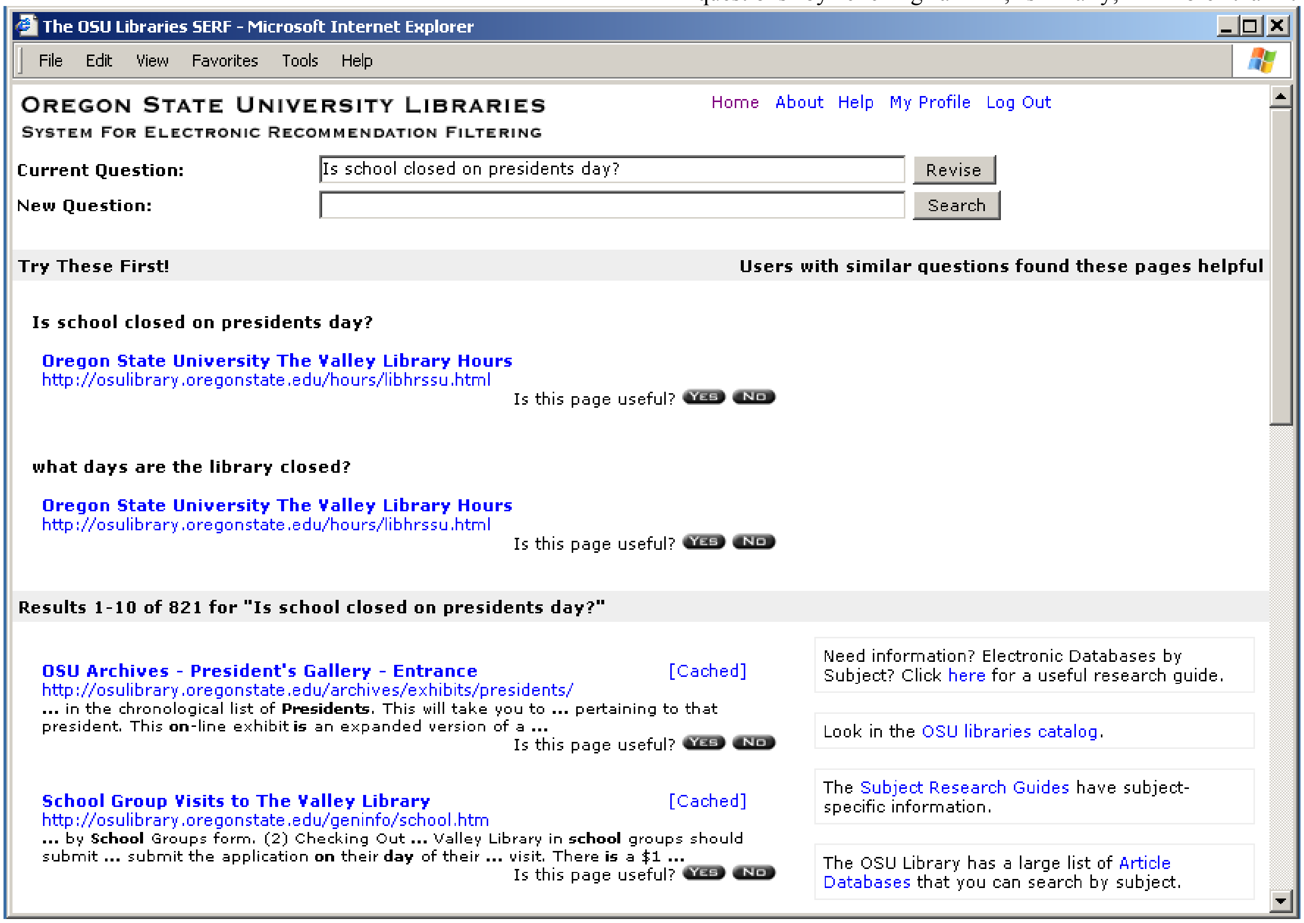
where Q_{pj} is the j-th previously asked question, $w_{i,pj}$ is the weight of keyword k_i in Q_{pj}, and $w_{i,c}$ is the weight of keyword k_i in Q_c. The $w_{i,pj}$ is computed as follows:

$$W_{i,p_j} = f_{i,p_j} \times idf_i$$

where $f_{i,pj}$ is the normalized frequency and idf_i is the inverse question frequency penalizes common words. The $f_{i,pj}$ and idf_i are computed as follows:

$$f_{i,p_j} = \frac{freq_{i,p_j}}{\max_{\forall v} freq_{v,p_j}}$$

$$idf_i = \log \frac{N+1}{n_i}$$



OREGON STATE UNIVERSITY LIBRARIES

SYSTEM FOR ELECTRONIC RECOMMENDATION FILTERING

Home About Help My Profile Log Out

Current Question: Is school closed on presidents day? Revise

New Question: Search

Try These First! Users with similar questions found these pages helpful

Is school closed on presidents day?

Oregon State University The Valley Library Hours
http://osulibrary.oregonstate.edu/hours/libhrssu.html
Is this page useful? YES NO

what days are the library closed?

Oregon State University The Valley Library Hours
http://osulibrary.oregonstate.edu/hours/libhrssu.html
Is this page useful? YES NO

Results 1-10 of 821 for "Is school closed on presidents day?"

OSU Archives - President's Gallery - Entrance [Cached]
http://osulibrary.oregonstate.edu/archives/exhibits/presidents/
... in the chronological list of **Presidents**. This will take you to ... pertaining to that president. This **on**-line exhibit **is** an expanded version of a ...
Is this page useful? YES NO

School Group Visits to The Valley Library [Cached]
http://osulibrary.oregonstate.edu/geninfo/school.htm
... by **School** Groups form. (2) Checking Out ... Valley Library in **school** groups should submit ... submit the application **on** their **day** of their ... visit. There **is** a $1 ...
Is this page useful? YES NO

Need information? Electronic Databases by Subject? Click here for a useful research guide.

Look in the OSU libraries catalog.

The Subject Research Guides have subject-specific information.

The OSU Library has a large list of Article Databases that you can search by subject.

Figure 2. The recommendations and search results screen of the SERF. The top half of the screen shows recommendations based on previously asked similar questions. The bottom half shows results from a Google search

where $freq_{i,pj}$ is the raw word frequency of k_i in Q_{pj}, $\max_{\forall v} freq_{v,j}$ is the most frequent keyword appearing in previously asked questions Q_{pj}, N is the total number of questions in our database, and n_i is the number of questions in which keyword k_i appears. This computation is identical to the computation used to compare queries to documents in some search engines [2].

Before computing similarity between questions, we remove extremely common words (stop words) and we apply the Porter stemming algorithm to reduce morphologically similar words to a common stem [18].

Our goal with SERF is to provide very high precision recommendations. Thus we only want to recommend documents from a previous information context if that context (thus the query) is substantially similar to the current context. To achieve this, we have a configurable similarity threshold – we only make a recommendation when the similarity of a previous question to the current question is greater than that threshold. We currently consider a question to be similar if the similarity is greater than 0.5.

We display up to two of the most similar questions and up to three of their highest rated documents. If more than two similar questions exist, users may elect to view additional similar questions by clicking a link; similarly, if more than three documents are rated highly for a question, users can also choose to view the other documents by clicking a link next to the recommended question.

3.3 Document Retrieval

The SERF system learns as it interacts with users. Early in the lifetime of the system, SERF will not have a large database of information contexts from which to recommend documents. Thus we can expect that early users will frequently not receive recommendations, even if their question is relatively common. Even after the system has been running for some time, we expect to see questions asked that are unrelated to any previously asked questions.

To address this situation, we also present results from a full-text document search engine in the second (lower) region of the search results screen (Figure 2). We use the results from a local Google appliance that only indexes the library web site. We use the text string specifying the information need as the query to the search engine.

In Section 3.1, we described how we encourage users to enter complete natural language sentences to express their information need. In the process, we are encouraging users to enter longer queries. However, because the Google search appliance defaults to using AND to logically combine search terms, search queries with many keywords are likely to return no results. Thus we have to adapt the query before sending it to the Google engine. We transform the query by taking the original query and appending each term using an explicit Boolean OR. For example "Where is the map room?" becomes "Where is the map room OR where OR is OR the OR map OR room." With this approach we hope to maintain some of the adjacency context of the original query, yet ensure that results do not have to have all of the keyword terms. For example, this ensures that documents with the exact string "map room" appear above documents with the two words not adjacent, yet does not require the words to be adjacent. Of course in the previous example, the stop words "is" and "the" would be ignored by Google.

Users can rate each document, whether recommended or returned by Google, as being useful or not useful directly on the results screen. Although the user may not have viewed the document through SERF, this feature allows users who have existing knowledge of document contents to provide feedback without navigating to the document. Librarians requested this feature, because they could frequently tell immediately from the title and/or URL of a document (using their knowledge of the library resources) if the document was relevant to their question.

These ratings are used by the system to indicate if a document contributed to answering a specific information need, defined by the query.

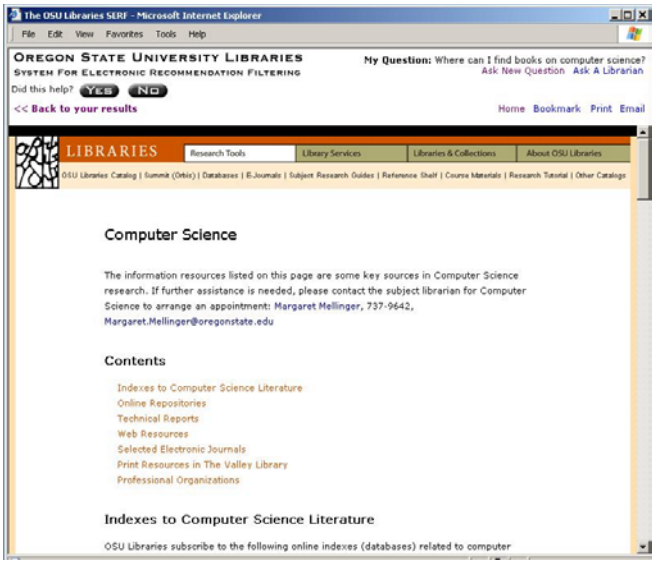
3.4 Revising Queries

On the search results page (Figure 2), there are two search boxes at the top of the page. In the first box, users have the option of revising their current question by entering new words or removing words. Any new keywords that are added are also added to the information context. At the same time, if the user removes words when revising a query, those words are retained in the information context. The intuition is that if somebody else has the same information need later, they may use the same set of keywords, even if those keywords in turn are not useful for matching with document text.

The second search box is used to specify a completely new question, thus creating a new information context. We hypothesize that without the two different boxes it would be very challenging to detect when a user was modifying the query, without changing their information need, and when they were asking something entirely different.

3.5 Viewing Documents

When a document is clicked from a search results list, that document is displayed within a frame controlled by SERF. SERF displays informational and rating information in a separate frame above the document (Figure 3). The upper frame reminds the user about the entered question and provides links to rate, print or email the currently viewed document. Navigation controls allow users to return directly to their search results or to the home search page to enter a new question. Logged in users may also add the document to their SERF bookmarks.


The OSU Libraries SERF - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Oregon State University Libraries
System For Electronic Recommendation Filtering
Did this help? YES NO
<< Back to your results
Ask New Question Ask A Librarian
Home Bookmark Print Email
LIBRARIES
Research Tools
Library Services
Libraries & Collections
About OSU Libraries
OSU Libraries Catalog | Summit (Orbis) | Databases | E-Journals | Subject Research Guides | Reference Shelf | Course Materials | Research Tutorial | Other Catalogs

My Question: Where can I find books on computer science?

Computer Science

The information resources listed on this page are some key sources in Computer Science research. If further assistance is needed, please contact the subject librarian for Computer Science to arrange an appointment: Margaret Mellinger, 737-9642, Margaret.Mellinger@oregonstate.edu

Contents

Indexes to Computer Science Literature
Online Repositories
Technical Reports
Web Resources
Selected Electronic Journals
Print Resources in The Valley Library
Professional Organizations

Indexes to Computer Science Literature

OSU Libraries subscribe to the following online indexes (databases) related to computer

Figure 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 current document.

Users may rate the currently viewed page's helpfulness for their current information need by clicking either a "yes" or "no" button. Additionally, logged in users may send a request for assistance to the library staff. This request consists of the user's current question, any revisions to the question, and the current page the user is viewing.

In order to be able to collect ratings on all pages, we must route all links in the viewed document to go through SERF. This allows the rating frame to always be displayed, no matter what server is providing the original copy of the page. To do so, we find the HTML link tags and rewrite them to point to our system. When a link is clicked, SERF fetches the requested page from its original

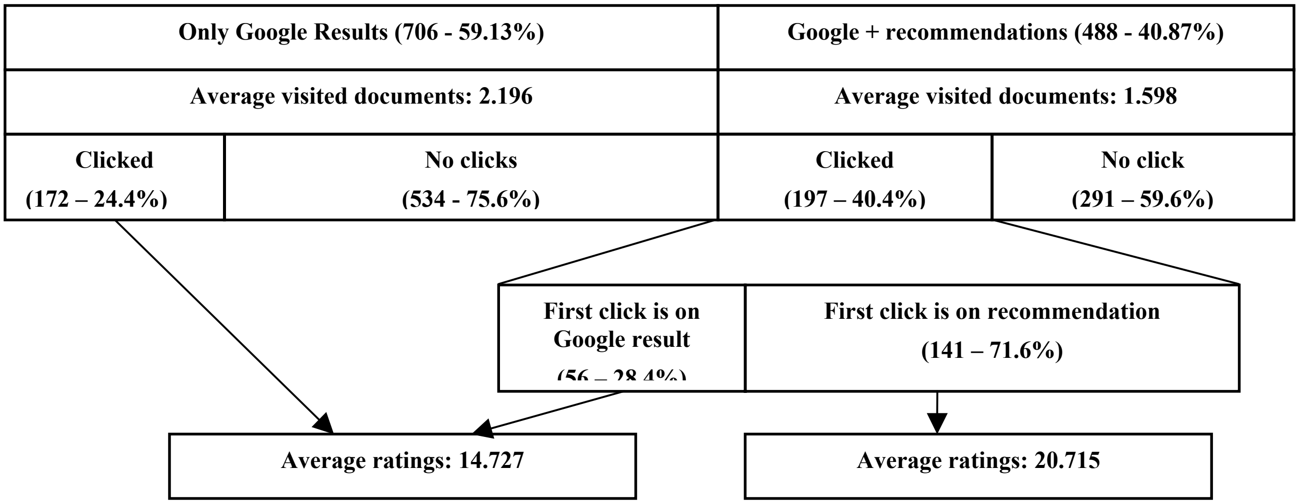

Only Google Results (706 - 59.13%)
Google + recommendations (488 - 40.87%)
Average visited documents: 2.196
Average visited documents: 1.598
Clicked (172 – 24.4%)
No clicks (534 - 75.6%)
Clicked (197 – 40.4%)
No click (291 – 59.6%)
First click is on Google result (56 – 28.4%)
First click is on recommendation (141 – 71.6%)
Average ratings: 14.727
Average ratings: 20.715

Figure 4. Summary of the data collected

source, rewrites the link tags, and displays the result within the rating frame. SERF does not currently handle Java Script or VB Script rewriting, and this information is purged from the page being processed. Additionally, since SERF utilizes frames to accept ratings, HTML tags that specify removing parent frames are also rewritten.

Simply discarding JavaScript or VBScript by default greatly reduces the complexity of wrapping HTML pages and works for most library pages. However, some of the highest value services that the library provides are commercial journal indexes, which frequently require the usage of Java and JavaScript. We maintain a list of these services and SERF returns documents from those services unprocessed; however, any ratings for a document within such a retrieval system are applied instead to its home page. By coincidence, this turns out to be desirable behavior. For the majority of the proprietary database interfaces, URLs are dynamically generated and session dependent. Attempting to return directly to those URLs can cause an error. As a result, recommending those URLs to later users is not desirable. By not wrapping the proprietary databases, any rating for a URL within those databases will be treated as a URL for the root search page of that database.

4. LOG DATA ANALYSIS

In this study, we analyzed the log data to attempt to understand if the SERF approach has potential to improve the efficiency and effectiveness of search.

This section reports the details of the system's deployment (Section 4.1) and results (Section 4.2).

4.1 Deployment Details

For the results reported in this paper, we used real data from the SERF portal deployed as a link from the OSU Libraries web site from January 2004 to mid-April 2004 containing 1433 search transactions. Each search transaction represents a single information context and the associated user behavior (which was tracked). The data reported are from "opt-in" users. At the bottom of the main page of the OSU Libraries a link was placed titled "Try our new experimental search interface." Users had to click this link to reach the SERF interface.

Of the 1433 search transactions, we discarded 239 because they either represented usage by a member of our research group, or because the data in the transaction was corrupt or inconsistent. These filtering steps left us with 1194 search transactions total.

Prior to launching the system we conducted a training session for OSU librarians. The purpose of this training session was to encourage the librarians to support this system. The librarians' training data are not included in our analysis, but were used to generate recommendations for later users.

4.2 Results

Figure 4 summarizes the data collected during the period in question of our trial. The rest of this section explains our findings with respect to Figure 4. Each subsection represents an important research question that we investigated.

4.2.1 How frequently are recommendations given?

If the SERF system only gives recommendations infrequently, then it provides little value above a traditional search engine. To identify recommendation usage, we analyzed the percentage of transactions where recommendations were available on the search results page. We found that approximately 40% of the transactions had at least one recommendation while the remaining 60% of the transactions had only Google results (Table 1). This is a surprisingly high number (40%), given that these data are from the first three months of operation and no information contexts were available at the beginning (with the exception of the data from the librarians' training). This could suggest that many people are asking similar questions. However, the data shown in Table 1 does not indicate if the recommendations presented were actually useful or not. The next sections examine if the recommendations were actually useful.

Table 1. How often did users get recommendations?

▨ Figure 4 pattern	Only Google results	Google + recommendations
Number of transactions	706 (59.13 %)	488 (40.87%)

4.2.2 *Do users make use of recommendations?*

For recommendations to be valuable, users must perceive them to have potential value. To understand better whether recommendations were perceived to be potentially useful or not, we looked at what percentage of users first clicked on a recommendation rather than a Google search result. Table 2 shows that users were more likely to first click a document from the recommendations than a document from the Google search results.

Table 2. How often did users click recommendations first when there are recommendations?

▩ Figure4 pattern	First click is on Google result	First click is on recommendation
Number of transactions	56 (28. 4%)	141 (71.6%)

If we assume that users are more likely to click the most perceived relevant documents after examination of the returned results, then the data in Table 2 provides some evidence that our recommendations were perceived by the users to provide more relevant information to the current information need than the Google search results.

It is also possible that users selected a recommended document first because the document was recommended by the system regardless of its relevancy, or because the document was at the top of the results. However, based on the results of Table 2, we can conclude that users actually examined returned results and did not simply click the first results, in spite of them being recommended (at least 24.8% of first clicks). We also may assume that for some portion of the 71.6% who clicked on a recommendation, the users only did so after examination of returned results.

To try and understand better what portion of the 71.6% might have selected a recommendation for the right reasons (because it appeared to be relevant), we looked at how often users actually chose to click on any of the search results or recommendations. Table 3 shows that users were more likely click at least one search result (Google result or recommendation) when there was a recommendation.

Table 3. How often did users click a document?

□ Figure 4	Only Google results		Google + recommendations	
	Clicked	No clicks	Clicked	No clicks
Number of transactions	172 (24.4%)	534 (75.6%)	197 (40.4%)	291 (59.6%)

Table 3 shows that only in 24.4% of the search results presented without recommendations were one of the search results selected! In the remaining 75.6% cases, users left the system, reformulated their query, or issued a new query. In contrast, when a user was presented with page having some recommendations, there was a 40.4% chance that the user would select one result (recommendation or Google result).

One thing we can learn from this data is that users are not just clicking on recommendations because they are displayed at the top. If that was the case, then we should see the same frequency of clicks when only Google search results were displayed.

Another thing that we see is that in many cases, users do not select any results. This gives us some evidence that users are actually reading (or at least scanning) the summary details given with each search result and recommendation. This is necessary to make the decision that none of the results are relevant. Thus we have more confidence that when users click on a recommendation result, they do so because it appears to have potential relevance to their need.

4.2.3 *Do recommendations increase searcher effectiveness or efficiency?*

In Section 4.2.2, we report our evidence that users perceive recommendations to be more relevant than Google search results, based on examination of summary information regarding the document (author, title, URL, text snippet, etc). However, how good are those recommended results, once the user has had a chance to read or visit the recommended resource? Or, in general, are recommendations actually helping users to find more relevant information, and to find it faster?

First let us examine efficiency. Table 4 shows the number of visited documents when there are recommendations and when there are only Google results.

Table 4. How many documents did users visit?

⬚ Figure 4 pattern	Only Google results	Google + recommendations
Average documents visited	2.197	1.598

An analysis of variance (ANOVA, $p < 0.05$) on the data summarized in Table 4 indicates that the mean number of visited documents when there are recommendations is significantly smaller than the mean number of visited documents when there are only Google results.

Here we assume that if a user views more documents, it is more likely that a user is not satisfied with the first document(s) visited and had to view more. The data in Table 4 provide strong evidence that recommendations helped users to search more efficiently.

In terms of effectiveness of the recommendations, we would like to know just how relevant were the results that SERF recommended. Table 5 shows the rating values of the first visited documents between the recommended documents and Google results for a query.

Table 5. Rating values of the first visited document between recommended documents and Google results for a query (30: relevant, 0: not relevant)

⬚ Figure 4 pattern	Documents from recommendations	Documents from Google results
Average rating	20.72 (n = 40)	14.73 (n = 67)

For historical reasons, a "Yes" rating for relevance is recorded as having a value of 30 and a "No" rating is recorded as 0. An analysis of variance (ANOVA, $p < 0.05$) of the data in Table 5 indicates that the mean rating of the first visited document from the recommended documents is significantly higher than the mean rating of the first visited document from the Google results.

From Table 5, we can see that if the first selected document from the search results comes from recommendations, on average, it is more likely to be rated as relevant. Thus we have evidence that, by adding recommendations, the users will be more likely to initially encounter a document that is more relevant than if they just had the Google search results. This strongly suggests that recommendations from SERF could increase the effectiveness of user searching.

4.2.4 How frequently do users issue complete natural language sentences rather than just a few keywords?

We described earlier that it was important that users enter natural language questions or statements of their information need. However, in discussions with other researchers, we have encountered many who doubt that we will see a substantial use of longer questions, given that entering a long statement of need takes more time and cognitive effort from the user.

Table 6. Question format queries vs. keyword format queries

Question format queries	Keyword format queries
67.68%	32.32%

Table 6 shows that approximately 68% of the queries submitted were natural language queries, or questions (question format queries) and approximately 32% of queries submitted were just one or more simple keywords (keyword format queries). Although users have been trained by current search engines to only enter keyword queries (particularly search engines that combine keywords by default using the Boolean AND operator), this result shows that we effectively encouraged users to use complete natural language sentences. 68% of users is surprisingly high. Analysis on data from the AskJeves! Web by Spink and Ozmultu [19] indicated that about 50% of queries were full questions. The observed high rate in SERF could be attributed to the use of the previously described user interface cues in the SERF interface. Part of the high rate could also be attributed to the fact that the system was experimental and most users were new to the system. We expect the likelihood that consistent users of SERF will continue to issue full questions to drop, as they learn that SERF is still functional (from an immediate gratification perspective) even if they do not use a complete natural language sentence for a query. However, even if the rate drops to 50%, we believe there will be sufficient information contexts in the SERF database with natural language descriptions.

4.2.5 How often do users rate documents?

One of the common concerns with collaborative filtering systems is that they require explicit human participation in order to be successful. In particular, we must be able to collect from the user some indications of what resources are relevant and/or valuable to a particular information context. In previous domains that we have worked with (books, music, usenet news, movies), this concern has been proven unfounded; in each case there was a subset of the population that was willing and excited about providing ratings on an ongoing basis. However, in our current domain (web search for library resources), we expect the data to be exceptionally sparse for some topics – there will be many information needs that are only shared by a small number of people. If users with this information needs do not provide ratings, we are unable to capture what they learn through their searching and browsing. For more common information needs, there are more users involved, and as we increase the number of users, the likelihood that we will see ratings increases.

Table 7 shows how often users rate documents. The data only includes transactions where the user selected at least one result. Ratings are intended to be evaluations of documents; if users do not view a document, we do not expect them to evaluate it.

Table 7. Percentage of transactions where users provided some form of feedback (a rating) as to the relevance of a search result.

	First click was a recommendatio n	First click was a Google result
Percentage of transactions where a rating occurred	28.4%	29.4%

The rating percentages shown in Table 7 are almost 30% of the time users provided us with at least one rating for a document, good or bad.

4.2.6 What commonly observed activities can be used to infer that a user found a resource valuable?

In spite of the exceptionally high rating rate reported in the previous subsection, we are still concerned about the potential sparsity of the ratings in our domain. Thus we are trying to identify patterns of activity that is commonly observed from which we can infer that user found a resource valuable. In particular, some preliminary controlled studies suggested that the "last viewed document" might be a good indicator of a document that satisfied an information need [10].

The *last viewed document* is the web page that was last requested by a user before either initiating a new search (i.e. a new information context) or leaving the system. The intuition is that when a user finds sufficient information to meet their need, they are likely to leave or move on to a different topic. Of course, it may be misleading in many cases – for example, users might leave the system because they got frustrated since they could not find they wanted. However, if SERF observes that people with similar information contexts are "ending" on the same document, then there is strong evidence that the document is valuable and should be recommended.

Table 8 shows the relationship between the "last viewed document" and the "non-last viewed documents", and the resulting rating on the data.

Table 8. Rating values between 'last viewed documents' and 'not last viewed documents' (30: relevant, 0: not relevant)

	Last viewed documents	Not last viewed documents
Average rating	18.91	5.45

An analysis of variance (ANOVA, $p < 0.001$) indicates that the mean rating of the last viewed documents is significantly higher than the mean rating of the non-last viewed documents. Again, for historical reasons, a rating of "yes" was recorded as 30 and a rating of "no" was recorded as 0.

The data provide strong evidence that the "last-viewed document" is a good indicator of a highly relevant document. However, current SERF doesn't do this yet. In future versions of SERF, we will be investigating how to incorporate this information as a proxy for explicit ratings.

Other opportunities for inferring ratings come from "action" links that are available whenever the user is viewing a document. These action links include "print," "email," and "bookmark." If the user clicks one of these links, we can also use that as strong evidence that the document is valuable.

5. CONCLUSION

This paper reports the data analysis of our initial prototype. We focused on the two key issues: would people participate and would recommendations live up to their promise of improved efficiency and effectiveness in searching?

The SERF system as we have designed it requires a reasonable amount of participation from at least a small fraction of the user population. In particular, we were concerned that a) users would rate documents, and b) users would provide meaningful and understandable statements of information need for their queries.

In terms of rating, we were pleasantly surprised. In almost 30% of transactions where at least one document was viewed, we got at least one rating. Apparently, many users are participating in the rating of documents. Future research may examine exactly what factors compel users to provide ratings. For example, we chose to use a binary rating scheme (Yes/No) because we believe we would be more likely to get ratings with such a simple scheme. More complex rating scales might cause more cognitive load, resulting in less ratings provided. However, at some point, having more expressive ratings might outweigh the loss of some raters.

Analysis of the data showed that almost 70% of the users issued queries in natural language using at least one complete sentence. This is overwhelmingly positive. Complete sentences are important so we can explain the recommendation by displaying the question to which the recommended document holds the answer. The current user may not be able to determine the context of previous queries if only a few keywords are available. Further research is needed to investigate whether or not the longer question format queries actually improve search performance for the user issuing the query although we strongly believe that with SERF, users can increase the effectiveness of later users who receive recommendations. A study by Anick [1] indicates that more keywords of a search query can increase search performance. Belkin et al [3] have examined the relation between query length and search effectiveness, and found that query length is correlated with user satisfaction with the search. Given that we use the query to search against past queries, and that those past queries have very few keywords compared to a traditional document, we expect that more words in the query will substantially improve the quality of the search results.

Analysis of our data also provides evidence that recommendations from prior users with similar queries could increase the efficiency and potentially effectiveness of the OSU Libraries website search. In addition to decreasing the number of documents that users viewed, users were more likely to select a recommendation on the search result page. Moreover, documents recommended by SERF were rated higher than search results from Google.

One important caveat regarding this study is that the users of the study were self-selected; users had to choose to click the link for the "experimental" search which led them to the SERF interface. As a result, we expect that the participation of users is somewhat higher than what we would find in a true random population of library users. Nonetheless the results reported in this paper, both absolute and relative, are substantial enough that we believe we will see positive results when we sample a non-self-selecting population.

The data from this study is very promising, but is only the beginning. There is still substantial research and development to be done. One area that we are particularly concerned about is robustness of our search in the face of erroneous and malicious ratings. Malicious ratings are designed to mislead the recommendation algorithm and generate incorrect recommendations for other users. Currently, we do little automatically to detect or handle malicious or erroneous users. To make a robust system, further research is needed to judge users' ratings in terms of whether the rating is trustworthy or not. Furthermore, we are continuing to evolve SERF to meet the specific needs of the library environment. For example, our data shows that approximately 30% of all questions relate to library services that can be answered with a very short answer – questions such as "What are the library hours?" We are investigating ways to provide these answers directly with fewer clicks.

The SERF system has been designed specifically for the library environment, but we believe that much of our approach could be applied in many other environments. In particular, we believe that the SERF approach could be applied successfully to domains where extremely similar information needs reoccur frequently with different users of the domain. Because SERF incorporates human judgment into the search process, SERF results should become more and more accurate as time passes and users utilize the system, entering queries and ratings.

6. 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, and the NSF Research Experiences for Undergraduates (REU) program.

We thank all our research members for their hard work in making the SERF happen. Special thanks go the Margaret Mellinger and Mary Caughey for their valuable comments and answering our questions on the library system. We also would like to thank all OSU librarians who participated in the training session for their time and cooperation.

7. REFERENCES

[1] Anick, P. Using Terminological Feedback for Web Search Refinement: A Log-based Study, *Proceedings of the 26th annual international ACM SIGIR conference,* (2003), 88-95.

[2] Baeza-Yates, R. and Ribeiro-Neto, B. *Modern Information Retrieval*, (1999), ACM Press.

[3] Belkin, N.J., Cool, C., Kelly,D., Kim, G., Kim, J-Y., Lee, H-J., Muresan, G., Tang, M-C. and Yuan, X-J. Query Length in Interactive Information Retrieval. *Proceedings of the 26th annual international ACM SIGIR*, (2003), 205-212.

[4] Boros, E., Kantor, P.B. and Neu, D.J. Pheromonic Representation of User Quests by Digital Structures. *Proceedings of the 62nd American Society for Information Science (ASIS)*, (1999).

[5] Burke, R., Hammond, K., Kulyukin, V., Lytinen, S., Tomuro, N. and Schoenberg, S. Natural Language Processing in the FAQ Finder System: Results and Prospects, in Working Notes from *AAAI Spring Symposium on NLP on the WWW*, (1997), 17-26

[6] Cosley, D. Lawrence, S. and Pennock, D.M. REFEREE: An open framework for practical testing of recommender systems using ResearchIndex. *Proceedings of the 28th VLDB Conference* (2002)

[7] DirectHit, http://www.directhit.com/

[8] Goldberg, K., Roeder, T., Guptra, D. and Perkins, C. Eigentaste: A Constant-Time Collaborative Filtering Algorithm. *Information Retrieval*, 4, 2 (2001), 133-151

[9] Gravano, L., Hatzivassiloglou, V., and Lichtenstein, R. Categorizing Web Queries According to Geographical Locality. *Conference on Information Knowledge and Management (CIKM)* (2003), 325-333.

[10] Jung, S., Kim, J., and Herlocker, J. Applying Collaborative Filtering for Efficient Document Search. *The 2004 IEEE/WIC/ACM Joint Conference on Web Intelligence* (WI 2004)

[11] Hill, W., Stead, L., Rosenstein, M. and Furnas,G. Recommending and evaluating choices in a virtual community of use. *Proceedings of SIGCHI* (1995), 194-201

[12] Kantor, P.B., Boros, E., Melamed, B. and Menkov, V. The information Quest: A Dynamic Model of User's Information Needs. *Proceedings of the 62nd Annual Meeting of the American Society for Information Science (ASIS)*, (1999).

[13] Kantor, P.B., Boros, E., Melamed, B., Menkov, V., Shapira, B. and Neu, D.L. Antworld: Capturing Human Intelligence in the Net. *Communications of the ACM*, 43, 8, (2000).

[14] MetaCrawler, http://www.metacrawler.com.

[15] McNee, S.M., Albert, I, Cosley, D., Gopalkrishnan, P., Lam, S.K., Rashid, A.M., Konstan, J.A. and Riedl, J. On the Recommending of Citations for Research Papers. *Proceedings of ACM CSCW* (2002).

[16] Menkov, V., Neu, D.J. and Shi, Q. AntWorld: A Collaborative Web Search Tool. *Proceedings of the 2000 workshop on Distributed Communications on the Web*, (2000),13-22

[17] Page L., Brin S., Montwani R. and Winograd T. The PageRank Citation Ranking: Bringing Order to the Web. Technical report, Standford University Database Group, (1998)

[18] Porter, M.F. An Algorithm for Suffix Stripping. *Program*, 14, 3 (July, 1980), 130-137

[19] Shardanand, U. and Maes, P. Social Information Filtering: Algorithms for Automating "Word of Mouth", *In Conference Proceedings on Human Factors in Computing Systems* (1995), 210-217.

[20] Spink, A., and Ozmultu, H.C. Characteristics of question format web queries: an exploratory study. *Information Processing and Management*, 38, 4 (2002), 453-471.

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

[22] Xue, G-R., Zeng, H-J., Chen, Z., Ma, W-Y., Zhang, H-J. and Lu, C-J. Implicit Link Analysis for Small Web Search. *Proceedings of the 26th international ACM SIGIR conference*. (2003), 56-63.