AN ABSTRACT OF THE DISSERTATION OF

Seikyung Jung for the degree of Doctor of Philosophy in Computer Science presented on May 25, 2007.

Title: Designing and Understanding Information Retrieval Systems using Collaborative Filtering in an Academic Library Environment

Abstract approved: _____

Jonathan L. Herlocker

Accessing information on the Web has become ingrained into our daily lives, and we seek information from many different sources, including conference and journal publications, personal web pages, and others. Increasingly, web-based information retrieval systems such as web-based search engines, library on-line catalog systems, and subscription-based federated search systems are made available to provide an interface to collections of information from these sources. Because the quantity of new information available every day exceeds how much information individuals can handle effectively, we spend significant effort in locating information, often unsuccessfully.

This dissertation consists of three scholarly articles presenting a broad set of results with the goal of helping people find interesting information in large web document collections. The results cover three specific challenges: designing and utilizing Web document recommendation systems based on human judgment, improving recommendations based on users' web usage as a source of implicit relevance feedback data, and understanding and designing metasearch systems for academic materials. To address these challenges, a combination of offline analysis and user studies is used.

We recommend documents by determining the similarity between users' information needs and the previously viewed documents by other users. We conducted experiments and observational studies to evaluate the system that we developed, and in both cases we found that recommendations from prior users with similar queries could increase the efficiency and effectiveness of document search.

To improve recommendation effectiveness, we studied users' click data from complete search sessions, and found that applying all of the click data in a search session as relevance feedback has the potential to increase both precision and recall of search results. In particular, our data provides evidence that the *last visited document* of each search session is a highly reliable source of implicit relevance feedback data.

In understanding and designing systems for academic materials, we designed a metasearch system for retrieving materials from OSU library's subscribed databases and catalogs. We conducted a think-aloud usability experiment and found that modeling the familiarity and ease of use of commercial web search engines is an important factor to attract undergraduates. However, when undergraduates faced the interface that felt familiar, they expected similar performance to a web search engine, such as its quality of ranked results, its speed and other qualities.

Designing and Understanding Information Retrieval Systems using Collaborative Filtering in an Academic Library Environment

by
Seikyung Jung

A DISSERTATION
submitted to
Oregon State University

in partial fulfillment of
the requirements for the
degree of

Doctor of Philosophy

Presented May 25, 2007
Commencement June 2008

Doctor of Philosophy dissertation of Seikyung Jung presented on May 25, 2007

APPROVED:

Major Professor, representing Computer Science

Director of the School of Electrical Engineering and Computer Science

---

Dean of the Graduate School

I understand that my dissertation will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my dissertation to any reader upon request.

Seikyung Jung, Author

CONTRIBUTION OF AUTHORS

Dr. Jon Herlocker was involved in the research design of all work described and assisted with the writing of all chapters of the dissertation,
Janet Webster was assisted in editing of Chapter 3 and Chapter 4, and involved in usability experiment and editing of Chapter 5,
Kevin Harris was involved with the system implementation and writing of Chapter 3,
Margaret Mellinger was involved with the usability experiment and writing of Chapter 5, and
Jeremy Frumkin was involved with the system design of Chapter 5.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

Page

TABLE OF CONTENTS (Continued)

Page

LIST OF FIGURES

LIST OF FIGURES (Continued)

LIST OF TABLES

Designing and Understanding Information Retrieval Systems using Collaborative Filtering in an Academic Library Environment

1. GENERAL INTRODUCTION

Finding and using information has always been ingrained into our daily lives, but more and more often we rely on the Web to access the information we need. Information comes from different sources, including conference and journal publications, technical articles, news articles, personal web pages, and others. Increasingly, web-based information retrieval systems such as library online catalog systems, subscription-based federated search systems, and web-based search engines are made available to provide an interface to collections of information from these sources. These web-based information retrieval systems differ in terms of interfaces, search models, and data collections, but the key goal of such systems is to allow users to retrieve useful or relevant information. However, to retrieve appropriate information, it is necessary to decide which source or system is appropriate, and then search or browse for information from these systems. Because the quantity of new information available every day exceeds how much information individuals can handle effectively, we spend significant effort in locating information, often unsuccessfully. Yet as we locate useful information, we learn – we discover new information that is valuable and we discover new methods or processes that can successfully lead us to useful information. Amazingly, while this learned knowledge, if shared, would benefit us all, it most often never leaves the mind of the original person. In fact, today's information searching tools make it difficult for individuals to share search strategies, search queries, or the analysis of results with others.

People also expect to find information quickly and easily, and they assume these web information retrieval systems retrieve relevant information. In this dissertation, we investigate technologies that can better help users find the best information faster and more than existing search technologies.

The remainder of this introductory chapter is organized as follows. First, we will briefly define and describe web information retrieval, information filtering, collaborative filtering, metasearch, and federated search – all technologies used in the course of this research. Second, we will present a list of the research challenges that we have addressed. Third, we describe the research method that we applied towards each challenge. At the end of this chapter, we present an outline for the remaining chapters.

1.1. BACKGROUND

Technologies and processes for "finding" information are studied in the research of the domain of **information retrieval (IR)**, primarily text-based document search engines. Many universities and public libraries use IR systems to provide access to books, journals, web pages, and other documents. Traditionally, a user describes his/her information need in the form of a query to the IR system and the system attempts to find documents that match the query within a document collection. While IR is focused on storage, indexing, and retrieval technology for textual documents, IR is not limited to web-based commercial search engines.

Information retrieval systems that employ the filtering approach, called **information filtering (IF)** systems, are " … typically designed to sort through large volumes of dynamically generated information and present the user with sources of information that are likely to satisfy his or her information requirement" (Oard, 1997, page 2). Email filtering software that categorizes email based on statistical inference of the importance of incoming email is an example of IF. Receiving news articles based on previously specified interests is another example of information filtering. While information retrieval traditionally studies users actively searching for and interacting with information, with information filtering, the source of information is often dynamic, with new documents being added or updated regularly. As new documents appear, only those documents that match the interests of the user are presented to the user. Information is asynchronously pushed to the user.

While information filtering relies on the content of documents to select and rank results, **collaborative filtering (CF)** relies on users' judgments and opinions, instead of documents' content, to select and rank results. Collaborative filtering is the process whereby a community of users (collaborating) with overlapping interests work together to separate interesting information from non-interesting information. In CF, each member of the community shares their preference (e.g. votes) of each document they experience. Then each user can tap into the collection of all past preferences by all other members of the community, and use those preferences to help select new, unseen information. Collaborative filtering is used in many practical applications in entertainment related domains, such as movies (http://movielens.umn.edu) and books (http://www.amazon.com/). Recommending scientific literature (ResearchIndex) is another example using CF (http://citeseer.ist.psu.edu)

Relevance feedback is a process for refining a representation of a user's information need. In relevance feedback, users identify relevant documents within search results. The system then

creates a new refined query based on those sample relevant documents and the user's original query. Classically, relevance feedback has referred to an information retrieval process whereby the user of the search engine indicates to the search engine that he or she would like "more documents like this one". The user is providing "feedback" to the system that "relevant" documents might look like the one indicated. This is used to improve the current search results for that user. 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. Explicit feedback of user preferences requires the evaluator to indicate a value for the content on a rating scale. This added work may burden users, but it can mean that the feedback received is more accurate. Systems that collect *implicit feedback* record and interpret users' behaviors as judgments of relevance without requiring additional actions from users (e.g. click data, display time). 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.

Metasearch engines provide unified access to multiple existing search engines. A metasearch engine does not maintain its own index of documents, but a sophisticated metasearch engine may maintain information about the contents of its underlying search engines to provide better service (Meng et al., 2001). Since it is hard to catalog the entire web, the idea is that by searching multiple search engines, it is possible to search more of the web in less time and do it with only one click. Metasearch engines are an emerging feature of web based library and information retrieval systems. When a library-based metasearch engine receives a query from a user, it transforms a query, broadcasts the query to a group of proprietary databases, merges the results, and presents them in a unified format to users. A9 (http://a9.com) is an example of commercial metasearch engine. ExLibris and MetaFind are examples of metasearch engines for academic materials.

1.2. OBJECTIVE AND RESEARCH QUESTIONS

The main objective of this research is to investigate and develop solutions that better support people in finding relevant information.

My first research objective is to develop a domain-independent framework that describes how to create a recommender system that provides relevant document recommendations. In order to reach this objective, I explored the following research question.

- Can a user find documents that provide relevant information more efficiently and effectively when given automatically detected recommendations from past users?

My second research objective is to identify how implicit relevance feedback can be used to improve the search results quality for all users, not just an individual user. For example, the search engine could learn which documents are frequently visited when certain search queries are given. In working towards this objective, I investigated the following research question.

- Can click data reliably indicate users' implicit preferences?

My third research objective is to develop a metasearch system that has the look and feel of a web-based search engine but with the relevance and proven quality of content that is traditionally made available only through university libraries. In order to reach this objective, I investigated the following research question.

- How effective are existing academic metasearch systems, and which specific characteristics of a metasearch engines are actually important for end user usability?

1.3. APPROACH

As this thesis focuses on three different aspects concerning document retrieval systems, three different research approaches are used; for each an approach that best suits that specific aspect. The approach for the first study to answer the first objective was to find a generic model of CF techniques in document search. Using this generic model, we developed a framework that describes how to create a CF based recommender system, and then evaluated this system through a controlled experiment.

The approach for the second study to answer the second objective was based on analysis of log data from a CF-based web page recommender system which was available on the Oregon State University Library website. We collected and archived users' explicit and implicit feedback, and analyzed the log data to measure the effectiveness and efficiency of the recommender interfaces. From our analysis, we also identified an accurate form of implicit feedback.

The approach for the third study to answer the third objective was to conduct a extensively monitored user study and think-aloud experiment with various search engines and a metasearch system to identify factors that are important in choosing an academic search system.

1.4. THESIS ORGANIZATION

An outline of the remainder of the dissertation can be found in Table 1-1.

Table 1-1. Overview of the dissertation.

Chapter	Chapter Description
1	**Introduction** (this chapter)
2	**SERF: Integrating human recommendations with search.** This chapter introduces the SERF (System for Electronic Recommendation Filtering) which is a collaborative filtering system that recommends context-sensitive, high-quality information sources for document search. This chapter also describes an observational (not experimental) study related to the usage of SERF as deployed at the Oregon State University library websites. Log data was analyzed to evaluate the efficiency and effectiveness of the library website search process through the SERF system. *Proceedings of the Thirteenth Conference on Information and Knowledge Management (CIKM),2004, New York, NY: ACM Press*
3	**Click data as implicit relevance feedback in web search.** This chapter presents work that shows how implicit relevance feedback can improve search performance. This includes log data analysis from users' click data to identify better indicator of users' implicit preferences. *Information Processing and Management, 43(3): 791-807. Elsevier*
4	**LibraryFind: system design and usability testing of academic metasearch system.** This chapter introduces the LibraryFind metasearch system that is designed to serve the needs of the library community by providing information from multiple propriety databases and other content sources. A think-aloud experiment with OSU college students was performed to evaluate the performance of the LibraryFind and to understand what factors are important in choosing academic search systems *Under review by Journal of the American Society for Information Science & Technology. John Wiley & Sons Inc, NJ*.
5	**Conclusions**

SERF: INTEGRATING HUMAN RECOMMENDATIONS WITH SEARCH

Seikyung Jung*, Kevin Harris*, Janet Webster**, and Jonathan L. Herlocker*

*School of Electrical Engineering and Computer Science,
Oregon State University, Corvallis OR, USA
** Oregon State University Libraries
{jung, harriske, herlock}@eecs.oregonstate.edu, janet.webster@oregonstate.edu

Proceedings of the Thirteenth Conference on Information and Knowledge Management (CIKM), 2004, New York, NY: ACM Press

2. SERF: INTEGRATING HUMAN RECOMMENDATIONS WITH SEARCH

2.1. 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.

2.2. 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 (Page et al., 1998), implicit link analysis (White et al., 2003), popularity-ranking (DirectHit, http://www.directhit.com), and page re-ranking with web query categorization (Gravano et al., 2003). Example systems include FAQ finder (Burke et al., 1997) and Meta search engines (MetaCrawler). 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

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

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?*

2.3. RELATED WORK

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

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

2.3.1. 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 (Shardanand & Maes, 1995), movies (Hill et al., 1995), jokes (Goldberg et al., 2001), 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 (Cosely et al., 2002; McNee et al., 2002). 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 (Boros et al., 1999; Kantor et al., 1999; Kantor et al., 2000; Menkov et al., 2000). 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.

2.3.2. Document Search Engines

Most studies of information retrieval in document search have been based on keyword-based full text search (Baeza-Yates & Ribeiro-Neto, 1999). 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 (Menkov et al., 2000). The DirectHit web search engine extracted useful information from users' access logs (http://www.directhit.com/). Gravano

et al. (2003) 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 (MetaCrawler).

For narrower domains, Xue et al. (2003) 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 (Burke et al., 1997).

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

2.4. 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.

2.4.1. 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).

Figure 2-1. The Initial Search Screen of SERF

We investigated ways that we could use user interface elements to encourage the "proper" behavior (Belkin et al., 2003; White et al., 2003). One report, by Belkin, et al. (2003), 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 2-1). 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.

2.4.2. 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 2-2). Here we describe the first region; the second region is described in Section 2.4.3.

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:

$$f_{i,p_j} = \frac{freq_{i,p_j}}{\max_{\forall v} freq_{v,p_j}} \qquad idf_i = \log \frac{N+1}{n_i}$$

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 (Baeza-Yates & Ribeiro-Neto, 1999).

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 (Porter, 1980).

Figure 2-2. 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

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.

2.4.3. 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 2-2). 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 2.4.1, 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.

2.4.4. Revising Queries

On the search results page (Figure 2-2), 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.

2.4.5. 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 2-3). 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.

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.

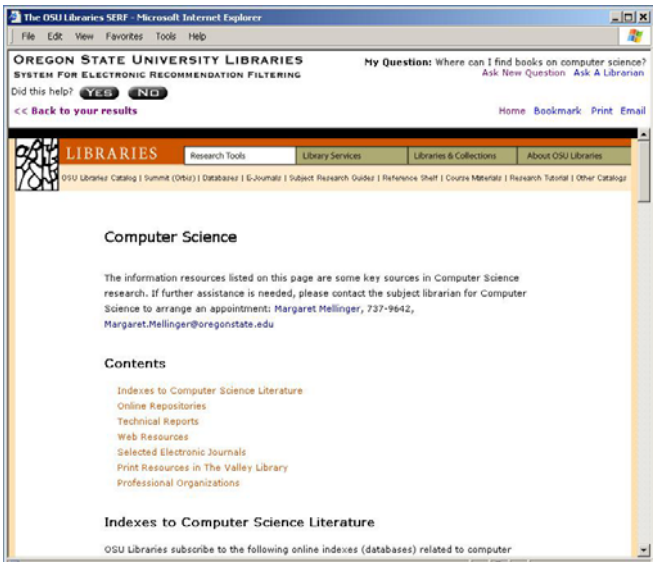

Figure 2-3. The interface for viewing web pages within the SERF. The upper frame is always present while browsing, regardless of what site the user is visiting and allows users to rate current document.

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 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.

2.5. 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 2.5.1) and results (Section 2.5.2).

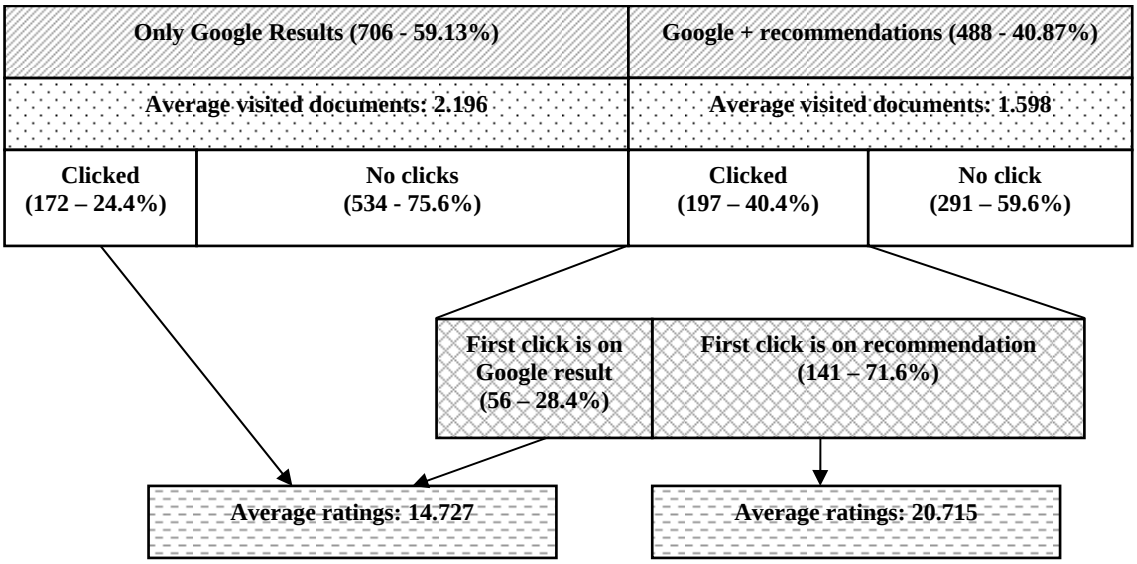

Figure 2-4. Summary of the data collected

2.5.1. 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.

2.5.2. Results

Figure 2-4 summarizes the data collected during the period in question of our trial. The rest of this section explains our findings with respect to Figure 2-4. Each subsection represents an important research question that we investigated.

2.5.2.1. 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 2-1). 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 2-1 does not indicate if the recommendations presented were actually useful or not. The next sections examine if the recommendations were actually useful.

Table 2-1. How often did users get recommendations?

Figure 2-4 pattern	Only Google results	Google + recommendations
Number of transactions	706 (59.13 %)	488 (40.87%)

2.5.2.2. 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 2-2 shows that users were more likely to first click a document from the recommendations than a document from the Google search results.

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 2-2 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 2-2, 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.

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

 Figure 2-4 pattern	First click is on Google result	First click is on recommendation
Number of transactions	56 (28. 4%)	141 (71.6%)

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 2-3 shows that users were more likely click at least one search result (Google result or recommendation) when there was a recommendation.

Table 2-3. How often did users click a document?

 Figure 2-4 pattern	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 2-3 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.

2.5.2.3. Do recommendations increase searcher effectiveness or efficiency?

In previous section, 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 2-4 shows the number of visited documents when there are recommendations and when there are only Google results.

Table 2-4. How many documents did users visit?

Figure 2-4 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 2-4 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 2-4 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 2-5 shows the rating values of the first visited documents between the recommended documents and Google results for a query.

Table 2-5. Rating values of the first visited document between recommended documents and Google results for a query (30: relevant, 0: not relevant)

 Figure 2-4 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 2-5 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 2-5, 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.

2.5.2.4. 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 2-6 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.

Table 2-6. Question format queries vs. keyword format queries

Question format queries	Keyword format queries
67.68%	32.32%

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.

2.5.2.5. 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 2-7. Percentage of transactions where users provided some form of feedback (a rating) as to the relevance of a search result.

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

Table 2-7 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.

The rating percentages shown in Table 2-7 are almost 30% of the time users provided us with at least one rating for a document, good or bad.

2.5.2.6. 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 (Jung et al., 2004).

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 2-8 shows the relationship between the "last viewed document" and the "non-last viewed documents", and the resulting rating on the data.

Table 2-8. 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.

2.6. 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 (2003) indicates that more keywords of a search query can increase search performance. Belkin et

al (2003) 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.

2.7. 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.

2.8. REFERENCES

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

Baeza-Yates, R. & Ribeiro-Neto, B. (1999) *Modern Information Retrieval*, ACM Press.

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

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

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

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

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

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

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

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

Kantor, P.B., Boros, E., Melamed, B. & Menkov, V. (1999) 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).*

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

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

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

MetaCrawler, http://www.metacrawler.com.

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

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

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

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

White, R.W., Jose, J.M. & Ruthven, R. (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

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. P*roceedings of the 26th international ACM SIGIR conference*, 56-63.

CLICK DATA AS IMPLICIT RELEVANCE FEEDBACK IN WEB SEARCH

Seikyung Jung*, Jonathan L. Herlocker*, and Janet Webster**

*School of Electrical Engineering and Computer Science,
Oregon State University, Corvallis OR, USA
** Oregon State University Libraries
{jung, herlock}@eecs.oregonstate.edu, janet.webster@oregonstate.edu

Information Processing and Management, 2007, 43 (3): 791-807. Elsevier.

3. CLICK DATA AS IMPLICIT RELEVANCE FEEDBACK IN WEB SEARCH

3.1. 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.

3.2. 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 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 it 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., 2003, 2005), scrolling down the text on a Web page (Claypool et al., 2001), book marking a page (Oard and Kim, 1998), printing a page (Oard and Kim, 1998), and the time spent on a page (Kelly and Belkin, 2004, 2001; White et al., 2003; Oard and Kim, 1998; Morita and Shinoda, 1994; Konstan et al., 1997).

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 et al., 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.

3.3. 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.

3.3.1. 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 (Morita and Shinoda, 1994; Konstan et al., 1997; White et al., 2002a, 2002b, 2003). Morita and Shinoda and others found that display time is indicative of interest when reading news stories (Morita and Shinoda, 1994; Konstan et al., 1997). 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 (Oard and Kim, 1998; Claypool et al., 2001; Fox et al., 2005). These researchers claim that examining multiple behaviors simultaneously could be sufficient to predict users' interest. Oard and Kim explored users' behavior including display time, printing, saving,

scrolling and bookmarking (1998). 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. 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 (2005).

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.

3.3.2. 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 (Xue et al., 2003; Cui et al., 2002; Smyth et al., 2003, 2005). Specifically, Cui et al. claimed that extracting candidate terms from clicked documents and using those terms later to expand the query could improve search precision (2002). Alternatively, Xue et al's approach automatically infers a link between two documents if a user selected both documents from the same set of search results (2003). The I-SPY project (Smyth et al., 2003, 2005) 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 et al., 2002, 2001). 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, who also clustered queries utilizing users' click data, attempted to improve search precision by expanding the query with other terms

in a cluster (2005, 2004). 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.

3.3.3. 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 and Maes, 1995), movies (Hill et al., 1995), jokes (Goldberg et al., 2001), and books (Linden et al., Amazon.com US Patent 6,266,649, 1998). 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 et al., 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 et al., 1999; Kantor et al., 1999; Kantor et al., 2000; Menkov et al., 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.

3.4. 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.

3.4.1. 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 (Figure 3-1).

Figure 3-1. The Initial Search Screen of SERF.

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.

3.4.2. Search Results

We presented results from Nutch[3], a full-text document search engine as a base-line in the lower region of the search results screen (Figure 3-2). 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

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

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 (Figure 3-2). 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.

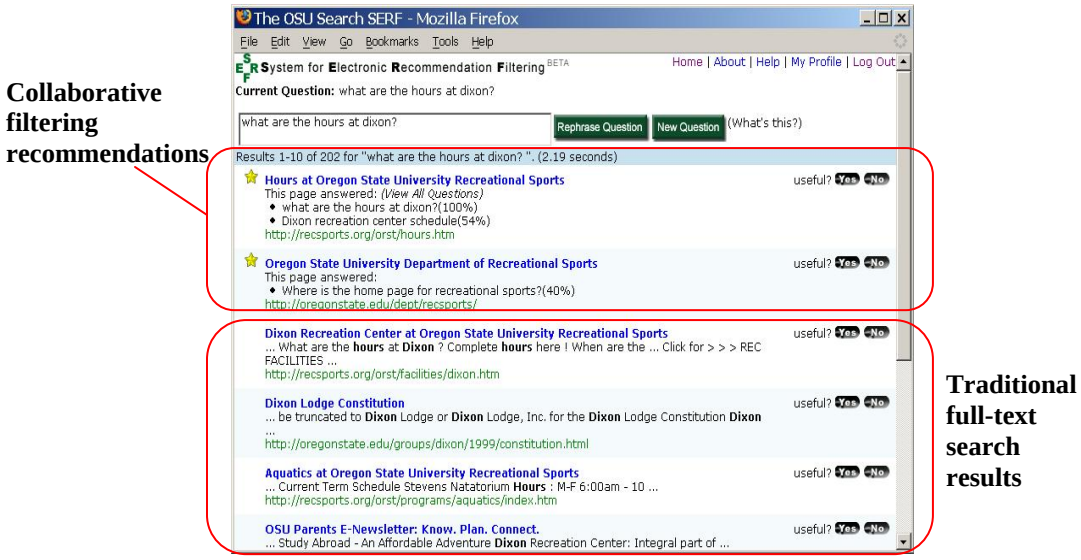

Figure 3-2. The recommendations and search results screen of SERF. The stars indicate recommendations based on previously asked similar questions. The rest of the results are from the search engine NUTCH.

3.4.3. 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 (Figure 3-3Figure 3-3). 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 (Figure 3-2). This allows them to provide feedback, if the relevance of a search result is clear from the displayed metadata and excerpt.

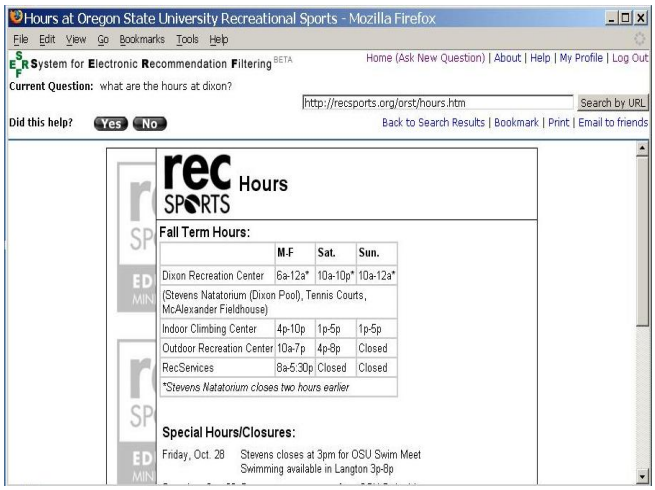

Figure 3-3. The interface for viewing web pages within the SERF. The upper frame is always present while browsing, regardless of what site the user is visiting and allows users to rate the current document.

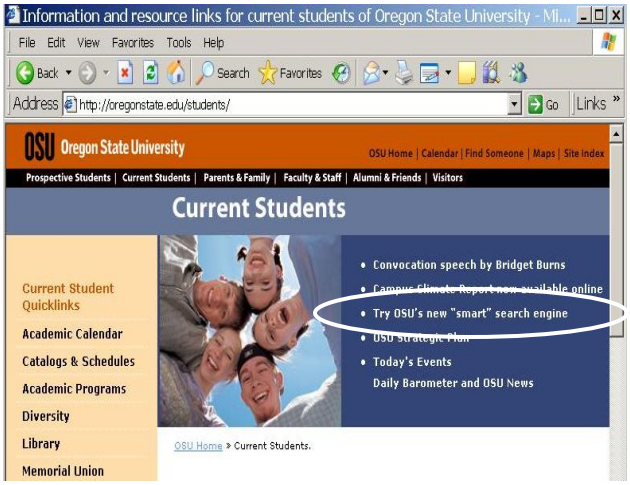

Figure 3-4. The OSU site for student with the ink to the experimental prototype of SERF that is available to all OSU users.

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.

3.5. METHODOLOGY & DATA SUMMARY

3.5.1. 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 (Figure 3-4Figure 3-4).

3.5.2. 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 (Figure 3-5).

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

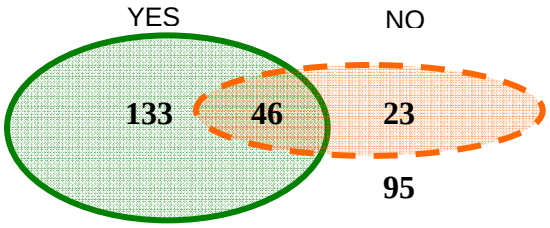

Figure 3-5. Search sessions from collected data based on users' ratings. Numbers count search sessions (Total: 297 search sessions, 133 search sessions with at least one [useful=]YES rating and no NO ratings, 46 search sessions with at least one YES rating and at least one NO rating, 23 search sessions 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).

3.5.3. Query Statistics

Figure 3-6 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 (Spink et al, 2002; Jansen et al, 2000), 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.

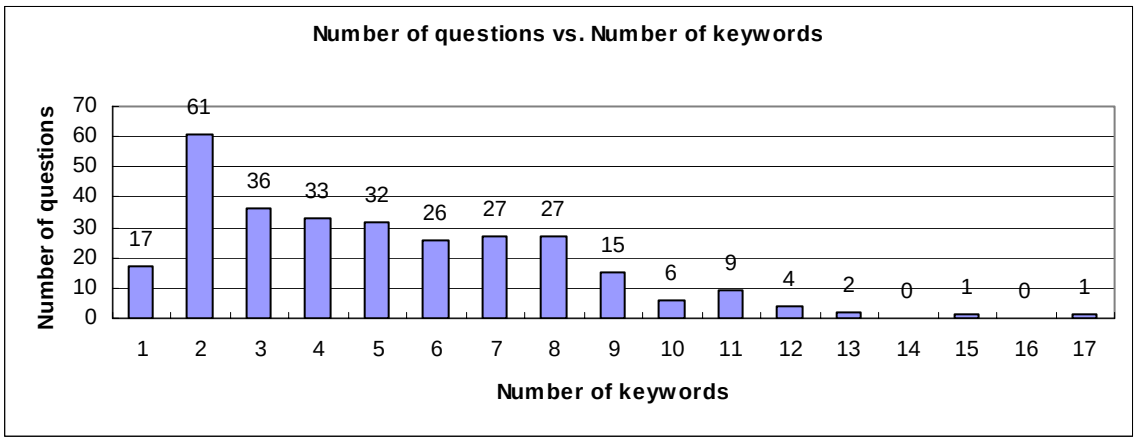

Figure 3-6. The frequency of questions by the number of keywords

3.5.4. 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 binary 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 either 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.

3.6. 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 Figure 3-7.

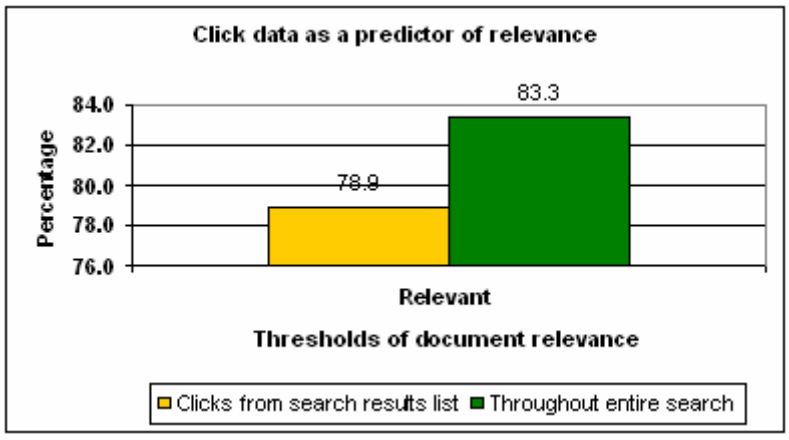

Figure 3-7. Comparison of percentage of relevant documents as a predictor of relevance between clicks from search results and clicks throughout entire search.

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. Figure 3-7 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 clicks 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[4]).

3.7. 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 and Spink, 2003; Spink et al., 2002; Jansen et al., 2000). Intuitively, if users were satisfied with the first document that they

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

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 3-1.

Table 3-1. 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	Documents within the 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

Figure 3-8 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 query was issued by multiple people.

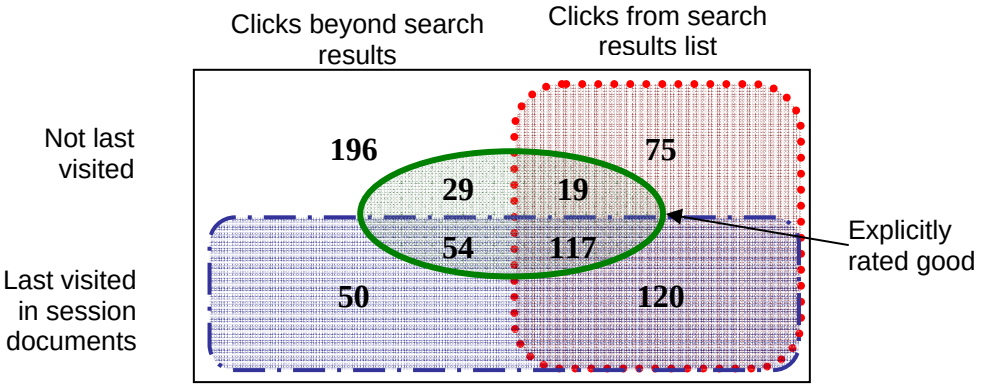

Figure 3-8. Ven diagram of three sets (clicks from search results list, last visited documents, and explicitly rated good). Numbers count documents.

We see there are 281 relevant documents from the last visited documents among 331 (84.9%, Figure 3-9). 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).

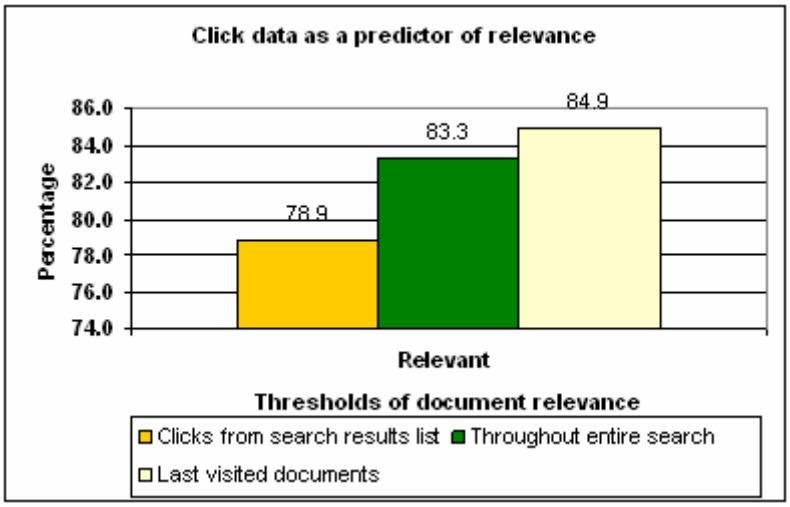

Figure 3-9. Comparison of percentage of relevant documents as a predictor of relevance among clicks from search results, clicks throughout entire search, and last visited documents

3.8. 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.

Table 3-2. Two thresholds of document relevance[5].

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

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 (Table 3-2). Strictly relevant documents are

[5] 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.

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.

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

Table 3-3. 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	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		

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 3-4). 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.

Table 3-4. Categorizing relevance of documents from a general searches

Query	Document	Relevance thresholds	
Parking	Home page of transit/parking services	Strictly relevant	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…"		

3.9. STRICT RELEVANCE BEYOND THE SEARCH RESULTS

Figure 3-10 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 3.6. 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[6]. 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 3.6, 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.

[6] 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%)

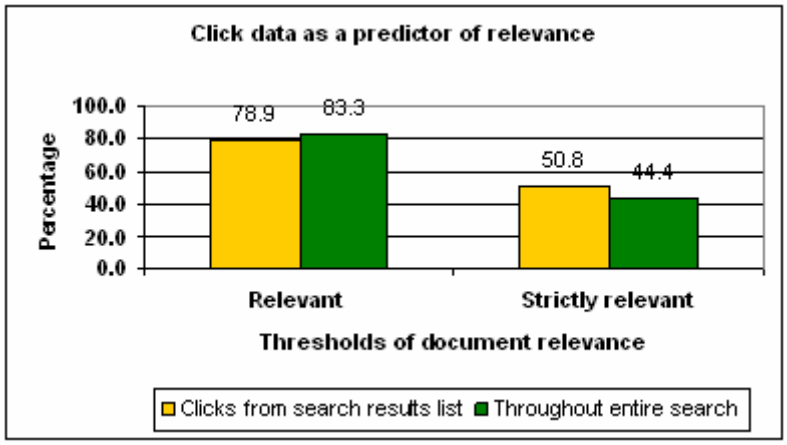

Figure 3-10. Comparison of percentage of relevant documents as a predictor of relevance between clicks from search results and clicks throughout entire search.

3.10. 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 3-5 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.

Table 3-5. The percentage of clicked documents among the users' data according to relevance judgments. Description of relevance categories can be found in Table 3-2. 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.

	Lowest Relevance → Highest Relevance			
	Implicit			Explicit
	Clicks beyond search results	From the search Results	Last visited	Ratings
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)

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 3-5 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. 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.

3.11. 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 3-6.

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.

Table 3-6. 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

3.12. 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, 1982), and may not choose query terms that precisely represent their information needs, resulting in a need to reformulate their query. Spink et al. (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.

3.13. REMAINING ISSUES

3.13.1. 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.

3.13.2. Feedback Aggregation and Query Clustering

As we introduced in the previous subsection (3.13.1), 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.

3.13.3. 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.

3.14. 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.

3.15. 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.

3.16. REFERENCES

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

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

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

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

Boros, E., Kantor, P.B. & Neu, D.J. (1999). Pheromonic representation of user quests by digital structures. In Hlava, M.K., Woods, L. (Eds.). *Proceedings of the 62nd Annual Meeting of American Society for Information Science,* 633-642. Medford, NJ: ASIS.

Claypool, M., Le, P., Wased, M., & Brown, D. (2001). Implicit interest indicators. In Sidner, C, Moore, J. (Eds.). *Proceedings of the 6th International Conference on Intelligent User Interfaces*, 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*, 35-46. 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,* 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 Nulti-Tasking Knowledge Workers. *In International Conference on Intelligent User Interfaces,* 75-82.

Fox, S., Karnawat, 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. & Furnas,G. (1995). Recommending and evaluating choices in a virtual community of use. In Katz, I., Mack, R., Marks, L., Rosson, M.B., Nielsen, J. (Eds.), *Proceedings of SIGCHI on Human Factors in Computing Systems*, 194-201. New York, NY: ACM Press.

Jansen, B.J., & Spink, A. (2003). An analysis of web documents retrieval and viewed. *The 4th International Conference of Internet Computing*, Las Vegas, Nevada, 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 Zaiane, O.R.(Ed.). *ACM International Conference on knowledge Discovery and Data Mining*, 133-142. New York, NY: ACM Press.

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

Jung, S., Harris, K., Webster, J. & Herlocker, J.L. (2004) SERF: Integrating Human Recommendations with Search. *In Preceedings of the Thirteenth ACM International Conference on Information and Knowledge Manangement (CIKM)*, 571-580.

Kantor, P.B., Boros, E., Melamed, B. & Menkov, V. (1999). The information quest: A dynamic model of user's information needs. In Hlava, M.K., Woods, L. (Eds.). *Proceedings of the 62nd Annual Meeting of American Society for Information Science,* 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. (2004). Display time as implicit feedback: Understanding task effects. In Sanderson, M., Jarvelin, K., Allan, J., Bruza, P. (eds.). *Proceedings of the 27th annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 377-384. New York, NY: ACM Press.

Kelly, D & Belkin, N.J. (2001). Reading time, scrolling and interaction: Exploring implicit sources of user preferences for relevance feedback. In Kraft, D.H., Croft, W.B., Harper, D.J., Zobel, J. (Eds.). *Proceedings of the 24th annual international ACM SIGIR Conference on Research and Development in Information Retrieval*, 408-409. New York, NY: ACM Press.

Konstan, J., Miller, B., Maltz, D., Herlocker, J., Gordon, L., & Riedl, J. (1997). GroupLens: Applying collaborative filtering to Usenet News. *Communications of the ACM*, 40(3): 77-87.

McNee, S.M., Albert, I, Cosley, D., Gopalkrishnan, P., Lam, S.K., Rashid, A.M., Konstan, J.A. & Riedl, J. (2002). On the recommending of citations for research papers. In E.F. Churchill, J. McCarthy, C. Neuwirth and T. Rodden (Eds.). *Proceedings of the 2002 ACM Conference on Computer Supported Cooperative Work*, 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 and H. Unger (Eds.). *Proceedings of the 2000 Workshop on Distributed Communications on the Web*, 13-22. Berlin: Springer-Verlag.

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

Oard, D., & Kim, J. (1998). Implicit feedback for recommender systems. In Kautz, H.A. (Ed.). *Recommender Systems: Papers from a 1998 Workshop,* 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,* 313–323. Englewood Cliffs, NJ: Prentice-Hall.

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

Smyth, B., Freyne, J., Coyle, M., Briggs, P., Balfe, E. (2003). I-SPY: Anonymous, community-based personalization by collaborative web search. In Bramer, M.A., Ellis, R. (Eds.). *Proceedings of the 23rd SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence*, 367-380. London: Springer-Verlag.

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.

Spink, A., Jansen, B. J., & 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. (2002). Query clustering using user logs. *ACM Transactions on Information Systems*, 20(1): 59-81.

Wen, J-R, Nie, J-Y., & Zhang, H-J. (2001). Clustering user queries of a search engine. In Shen, V.Y., Saito, N., Lyu, M.R., Zurko, M.E. (Eds.). *Proceedings of the 10th International Conference on World Wide Web*, 162-168. New York, NY: ACM Press.

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 Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 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 Clarke, C, Cormack, G. (Eds.). *Proceedings of the 26th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 56-63. New York, NY: ACM.

LIBRARYFIND: SYSTEM DESIGN AND USABILITY TESTING OF ACADEMIC METASEARCH SYSTEM

Seikyung Jung*, Janet Webster**, Margaret Mellinger*, Jeremy Frumkin*, and Jonathan L. Herlocker*

*School of Electrical Engineering and Computer Science,
Oregon State University, Corvallis OR, USA
{jung, herlock}@eecs.oregonstate.edu,

** Oregon State University Libraries
{Janet.Webster, Margaret.Mellinger, Jeremy.Frumkin}@oregonstate.edu

Under review by Journal of the American Society for Information Science and Technology (JASIST). John Wiley & Sons Inc, NJ.

4. LIBRARYFIND: SYSTEM DESIGN AND USABILITY TESTING OF ACADEMIC METASEARCH SYSTEM

4.1. ABSTRACT

Using off-the-shelf search technology provides a single point of access into library resources, but we found that such commercial systems are not entirely satisfactory for the academic library setting. In response to this, Oregon State University (OSU) Libraries designed and deployed LibraryFind, a metasearch system. We conducted a usability experiment comparing LibraryFind, the OSU Libraries website and Google Scholar. Each participant used all three search systems in a controlled setting and we recorded their behavior to determine the effectiveness and efficiency of each search system. In this paper, we focus on understanding what factors are important to undergraduates in choosing their primary academic search system for class assignments. Based on a qualitative and quantitative analysis of the results, we found that mimicking commercial web search engines is an important factor to attract undergraduates. However, when undergraduates use these kinds of search engines, they expect similar performance to web search engines, including factors such as relevance, speed, and the availability of a spell checker. They also expected to be able to find out what kinds of content and materials are available in a system. Participants' prior experience using academic search systems also affected their expectations of a new system.

4.2. INTRODUCTION

Today's academic library websites serve as gateways to digitally-accessible library resources, ranging from full-text newspaper and journal articles to on-line catalogs of physical collections. Yet, undergraduate students often use web-based search engines as a primary tool to find class related information, not only because they are familiar with web-based search engines (Notovny, 2004; Augustine & Greene, 2002), but also because they struggle to use library websites as a gateway for finding information (Stephan, et al., 2006; Notovny, 2004; Yakel, 2004; Chisman et al, 1999; Eliason et al., 1997).

Many researchers in library and information science focus on improving library websites so that navigating library services is easy. These include local resources (e.g., the library catalog), as well as databases offered by third-party vendors (e.g., Academic Search Premier). Yet, navigation

remains difficult for novice users unfamiliar with exploiting the multiplicity of library services. They do not know where to start in part because they do not know which databases are proper for their current information need (Stephan et al., 2006; Eliason et al., 1997). They also find it difficult to identify useful documents from the returned results (Stephan et al., 2006; Eliason et al., 1997). Undergraduates expect that disparate library resources be aggregated so that a single query will return satisfactory results (Augustine & Greene, 2002; Notovny, 2004).

One promising approach to these persistent issues is simplifying the search rather than constantly revising the library website to improve navigation. For that reason, researchers are experimenting with metasearch and federated search systems. In this paper, we distinguish metasearch systems based on whether the system is able to index locally. When a metasearch system (such as LibraryFind) receives a query from a user, it invokes multiple library service providers at the time of the search request to retrieve useful information. Another type of metasearch system (such as Google Scholar) requires those multiple content sources to provide their indices so that they can be searched locally and the results can be retrieved remotely. Researchers vary in how they define metasearch and federated search systems. Christenson & Tennant (2005) refer to the first kind of search engine as a metasearch and the second kind as a federated search. Interestingly enough, this usage of the term "federated search" is inconsistent with previous definitions of the term from database research, where federated search referred to the first kind of search (Lim et al., 1995; Hwang et al., 1993).

We are designing LibraryFind to successfully serve the needs of the library community and specifically college undergraduates. LibraryFind has the look and feel of a web-based search engine but provides access to the high quality content that is traditionally made available only through libraries. One of our goals is to make a system that OSU undergraduates will choose as their primary web-based service to find appropriate information for class assignments. In addition, we expect that OSU undergraduates will find the relevant databases for their information needs so that they can associate their information needs with databases in the future.

To investigate our progress towards these goals, we conducted formal usability testing in the fall of 2006. We assumed that current undergraduates seek academic materials from either the library website or from commercial web-based search engines. Thus, we decided to use three different search systems. The first system is the OSU Libraries website, which is a traditional online library website that serves as a gateway to library resources. The second system is

LibraryFind. For the third system, we chose Google Scholar as a scholarly web search engine because it returns academic materials yet has the familiar Google interface.

Using three different search systems in our usability testing, we investigated three research questions to determine how undergraduates search for academic materials depending on the search system they use.

- Which system do undergraduates choose to use?
- What factors are important in choosing the academic search system?
- How effectively and efficiently did each system work when used by undergraduates?

The following describes the LibraryFind system (4.3), related work (4.4), explains our usability testing (4.5) and then explores the results of that testing (4.6).

4.3. OSU LIBRARYFIND SYSTEM – DESIGN AND IMPLEMENTATION

The design and implementation of LibraryFind was motivated by our desire to have a more effective and efficient library metasearch engine. In 2004, OSU Libraries implemented Innovative Interfaces' MetaFind search system, knowing that the search technology was not fully developed or standardized (Boock, Nichols & Kristick, 2006). Over the next two years, deficiencies in the commercial product led to low use statistics, even though MetaFind was introduced in the library segment of Writing 121, the introductory writing course taken by most first year students. Previous studies have shown that searches from metasearch engines were too slow, that results were difficult to interpret, and that users did not always understand what they were searching (Tallent, 2004; Lee, 2006; Cervone, 2005). OSU Libraries developed LibraryFind to address searching and display issues identified in previous usability studies.

4.3.1. The Initial Search Page

We designed the initial search page to look like any commercial web-based search engine so that potential users could initiate interaction with our system without difficulty. We assume that potential users are accustomed to commercial web-based search systems and have trouble using traditional library search systems because they do not understand what it means to search various library services, or because they do not know which services are appropriate for their current information needs. We also tried to avoid library jargon that users might not understand and instead tried to use terms which are commonly used in any web-based search system.

We minimized text in the initial search page (Figure 4-1) to make the search text box prominent so that users will quickly notice the box. We included three tabs; “General”, “Images and more”, and “Books and more”, and filtered based on document types.

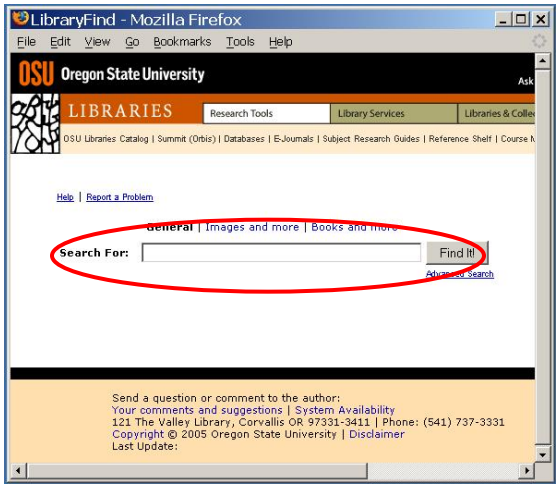

Figure 4-1. LibraryFind Initial Page

4.3.2. Document Retrieval

Once a user submits a query, the metasearch engine sends it to multiple library services. When the number of library services is small, this is a reasonable strategy. However, as the number and diversity of available library services increases, it becomes increasingly probable that results will be included from services that are not relevant to the user’s information need. This large number of irrelevant resources will not only degrade the performance of the system, but also they will increase the number of potentially irrelevant results displayed to the user. This problem of identifying potentially useful information providers to search for a given query is known as the information provider selection problem (Meng et al., 2002). The goal is to select as many potentially useful library resources to search as possible while minimizing the search of less useful library resources. At this time, we have not yet implemented any automated method of determining the best services. Instead, we included a subset of local resources and databases chosen by librarians as the best starting point for a majority of undergraduate research.

4.3.2.1. Query dispatcher

The query dispatcher (Figure 4-2) establishes a connection with the server of each selected library service and passes the query to it. Access to each library service is most commonly

accomplished through the Z39.50[7] or SOAP[8] protocols. Ideally, we would like to harvest metadata directly from OSU's vendors utilizing OAI-PMH[9], as this would allow quicker retrieval through local indexing. However, few vendors currently provide this type of information access to their databases. Each library service has its own query format, meaning that the original user query may need to be translated to a new query before being sent to each service.

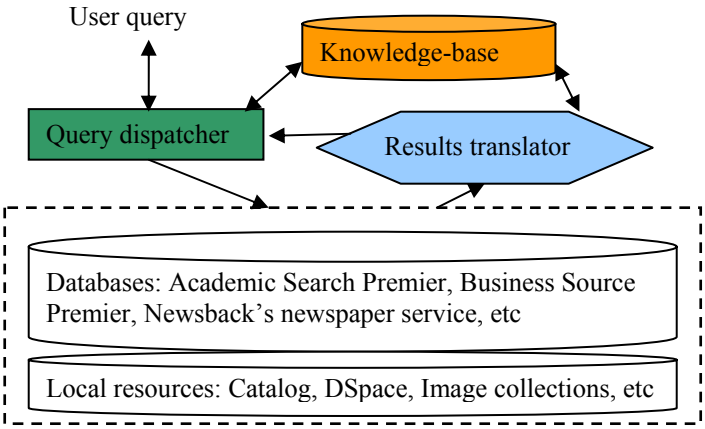

Figure 4-2. LibraryFind software Component architecture.

LibraryFind addresses these issues by utilizing a central knowledge-base, which stores relevant configuration details of each service, such as the connection information, query format, and the format of search results (Figure 4-2). For example, Academic Search Premier uses the Z39.50 protocol and returns results in MARCXML[10]. The knowledge-base uses this information to connect to the service and extract citation information from the returned record sets.

The query dispatcher utilizes information from the knowledge-base to send the transformed query to the target services directly. The search results from each database and local resource are then passed into the results translator where individual results are extracted and normalized for display and ranking.

---

[7] Z39.50 is a client server protocol for searching and retrieving information from remote computer databases.
[8] Simple Object Access Protocol (SOAP) is a protocol for exchanging XML based messages over computer network.
[9] Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) is a protocol to harvest the metadata descriptions of the records in an archive so that services can be built using metadata from many archives.
[10] In library and information science, MARCXML is an XML schema based on the bibliographic MARC standards.

4.3.2.2. Results translator

After retrieving results from each database, the results translator merges the results into a single ranked list and presents the merged results. Implementing the results translator presents two challenges. The first challenge is to filter out duplicate records. For example, the same article may be represented with slightly different metadata across different databases. To overcome this issue, we compare metadata to create a confidence value that two records refer to the same publication. The second challenge is to globally rank the results in the merged list. To do this, filtered records are passed through a simple ranking algorithm to sort records within the results set according to the query terms. Results where primary titles, authors or subject metadata exactly match query terms are weighted the highest. More emphasis is given to keyword matches that occur at the beginning of each field, or keyword matches that occur often in each field.

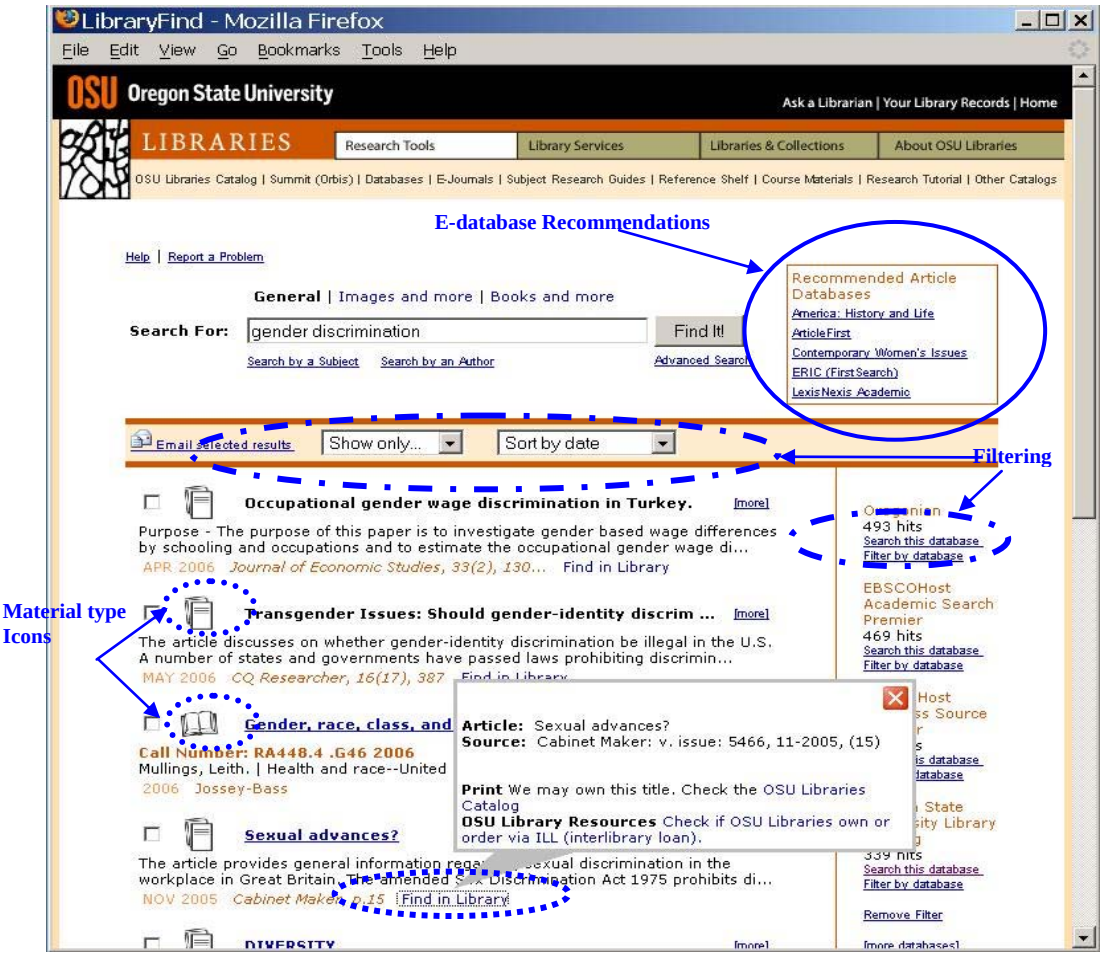

Figure 4-3. LibraryFind results page.

4.3.3. Viewing search results

We present the title of each result along with metadata appropriate for the type of material (Figure 4-3). Users can access materials directly by clicking the title and can see detailed metadata information by clicking a "find in library" link (bottom-center of Figure 4-3).

4.3.3.1. Accessing documents

For some records, we can provide a direct link to the full-text of the resource, often using OpenURL[11] to create links to the full-text when available. These links can be passed to a library's OpenURL Resolver[12], which determines if the organization has access to that particular title. If full text access can be located, then the title of the resource is linked to the full text. If the title is unavailable, an additional "find in library" link provides links to additional content location options returned by the library's OpenURL Resolver. If no access points exist, the tool provides a link to the parent organization's Inter Library Loan.

4.3.3.2. Additional features

One of our goals is to recommend appropriate databases (upper-right of Figure 4-3) so users will be able to find the right gateway to the information they need, even if they did not find that information directly in the LibraryFind search results. "Database recommendations" are determined by matching the user's query against collections of keywords descriptive of each database. The keywords collections for each database are constructed from 120,000 Library of Congress Subject Headings and abstract information about each database.

Users can filter returned results by their material type and database. They can also sort results by date and relevance (upper-center of Figure 4-3). These options were created in response to perceived deficiencies within current metasearch vendor products and librarians' feedback.

4.4. BACKGROUND AND RELATED WORK

Our goal in developing LibraryFind was to create an academic information search tool that students would use. The working prototype of LibraryFind was installed on the OSU Libraries website in the summer of 2006. Previous work on usability studies of library websites and work on evaluation of metasearch systems suggested issues to explore.

---

[11] OpenURL is a type of URL that contains resource metadata for use in libraries. National Information Standard Organization (NISO) has developed OpenURL and its data container (the ContextObject) as international ANSI standard Z39.88.

[12] OpenURL Resolver offers context sensitive services based on that metadata.

4.4.1. Usability studies in academic library websites

Recent studies show that users of library websites have difficulty understanding library and archival jargon and lack familiarity with the structure or contents of these interfaces (Augustine & Greene, 2002; Yakel, 2004; Hamburger, 2004; Cobus et al., 2004). Users also have trouble deciding where to start searches, perceiving where they are in their search, other areas of the website (Battleson et al., 2001; Stephan et al., 2006; Thomsett-Scott, 2005; Eliason et al., 1997, Chisman et al., 1999; Travis & Norlin, 2002). Additionally, college students find it difficult to choose an appropriate database from library websites (Battleson, et al., 2001). Novice and experienced users have different goals when using library websites, which complicates library website design (Cockrell & Jayne, 2002; Cunningham, 2005). Some researchers noted that current students' familiarity with web search engines leads them to expect similar functionality and performance in library search systems (Augustine & Greene, 2002; Notovny, 2004).

To remain competitive, library websites must integrate external resources from vendors and provide a user experience similar to web search engines. Despite extensive efforts in overcoming these challenges, library websites still struggle to attain the usability of their competition.

4.4.2. Usability of academic library metasearch systems

While metasearch systems have been around for many years, their importance and use within the library community is a relatively new phenomenon. Metasearch tools allow users to enter their search once and avoid making the user choose a suitable database from a list of hundreds (Crawford, 2004). These users are accustomed to the ease, convenience and speed of web searching (Christensen & Tennant, 2005). Bringing these resources together in an easily searched interface addresses the need users have for clear starting place. For these reasons, librarians have looked to metasearch systems as one solution for bringing together their heterogeneous resource collections (Luther, 2003). At present, libraries spend a large portion of their budgets on library resources that users do not discover, or learn how to access (Luther, 2005). Very few libraries have built their own metasearch systems (Reese, 2006). Most libraries have purchased and modified commercial products, such as ExLibris's MetaLib, WebFeat and Fretwell Downing (Boss, 2002; Breeding, 2005).

Few formal usability studies have been conducted on metasearch systems given their relatively recent implementation in libraries. Also, as most systems are proprietary, testing may occur but results are rarely publicly available. An exception to this is a study done by Endeavor

(commercial metasearch engine) that demonstrated a need for accurate sorting by relevance and then date, and the interest in personalization features (Randall, 2006). Other studies described positive findings including that college students liked having descriptive metadata to aid in evaluating the relevance of items (Reeb, 2006; Randell, 2006), and that they valued the ability to access full text (Lee, 2006). Some negative findings about the use of metasearch engines were also revealed in previous studies. Mandatory user log-ins and the necessity of selecting databases and other resources from a list were barriers that kept college students from using a metasearch system (Tallent, 2004). Users were also dissatisfied with the way search results were returned database-by-database (Lee, 2006; Tallent, 2004), with the slowness of the search (Lee, 2006), and with an interface that they found unintuitive (Tallent, 2004; Cervone, 2005).

In summary, previous user studies examined how students interacted with metasearch systems and identified which features students liked and did not like. None of the studies conducted tested whether a state-of-art metasearch system can serve college students as a primary academic search system or compared it with other search systems.

4.4.3. Usability of scholarly web search engines

A recent study shows that users in an academic setting want more relevant results with the ease of a web search engine (Marshall, et al., 2006). In response to these demands, web search engines with a scholarly focus such as Google Scholar, Citeseer, PubMed, and others emerged.

Much of the literature on these scholarly web search engines (as opposed to Section 4.4.2 library metasearch systems) describe strengths and weakness of Google Scholar (Neuhaus, et al., 2006; Jacso, 2005; Tennant, 2005; Notess, 2005; Burright, 2006; Bosman, et al, 2006). Positive findings include that Google Scholar leverages successful attributes of the Google interface: the interface is highly familiar, and the search experience is intuitive to anyone who has used Google previously. Full-text is often available; users can sometimes find free web versions of otherwise inaccessible full-text articles (Notess, 2005). OpenURL implementation allows users to link to their library's subscribed full-text content from within Google Scholar (Tennant, 2005). Highly cited articles are ranked to appear near the top of the results list, which is good for novice users but perhaps not as useful to the expert (Notess, 2005). Negative aspects of Google Scholar are its lack of transparency about which journals and publishers are included in the search and from what years, its English language bias, its absence of authority control, and a time lag in uploading new content (Meltzer, 2005; Burright, 2006; Neuhaus, et al., 2006; Jacso, 2005).

In summary, while people have examined the positive and negative aspects of these search engines, we are aware of no studies exploring how undergraduates use scholarly web search engines compared with library metasearch systems and traditional online library search systems.

4.5. USABILITY EXPERIMENT

Our first goals with the study were to get a true picture of users' willingness to use LibraryFind as their primary academic search tool and to identify the factors affecting this willingness. Given this, we tested the usability of LibraryFind against two other systems that students would be likely to use in the academic setting, the OSU Libraries' website and Google Scholar. We wanted to get quantitative results in addition to the qualitative, so, we enlisted twenty-four users[13].

4.5.1. Participants

LibraryFind is targeted at OSU undergraduates who take lower level classes and need to use library services for writing papers. We recruited OSU undergraduates to participate for 90-minute time slots. Fourteen participants were upper level students (senior and junior) and ten were lower level students (sophomore and freshman). Nine students had database experience (prior database and catalog experience[14]), seven students had catalog experience (prior catalog experience, but not database experience), and eight students were novices (no prior experience with catalogs and databases). Surprisingly, only three students had prior Google Scholar (GS) experience, three students had prior Library Find (LF) experience, and eight students had not used the library website at all. Majors and genders were distributed relatively equally.

4.5.2. Procedures

We began each session by explaining the experiment's procedures, and then participants completed a background questionnaire. In this paper we define task as a work assigned to participants with one topic and one system. Participants completed each of the first three tasks using one of the pre-selected topics and systems. For the fourth task, participants were allowed to choose any combination of the three systems to search a final pre-selected topic. This choice allowed us to examine their preferences after experiencing all three systems.

To avoid any kind of learning effect from using systems in a certain order, and any advantages or disadvantages from the ordering of topics, each participant searched the three

[13] Nielson (http://www.useit.com/alertbox/quantitative_testing.html)
[14] All subjects who had prior database experience also had catalog experience

systems and four topics in a different order. The system and topic orders were matched according to six unique Latin squares (Kuehl, 1994). We randomly assigned participants to particular sequences. The study employed a within-participants (repeated measures) design.

The participants documented their findings on a session questionnaire after each task. Each participant wrote down three relevant information resources, identified the material type of each, and indicated their satisfaction level using a five point Likert scale. After this step, we asked questions about their experiences with the system, focusing on whether the system was easy to navigate and to retrieve relevant documents. After the fourth task (users' choice of systems), we asked why they chose their system or combination of systems among the three systems.

Two researchers observed each session and took notes. All four tasks were recorded using Morae[15] event recording software. Morae creates clips of experiments that include audio and video, a screen capture video, and a record of mouse and keyboard interactions with the computer. We also requested that participants say whatever they are looking at, doing and feeling as they are performing tasks. The goal of the think-aloud protocol was to allow the participants to report their thoughts while using each system giving us a more direct view of the mental processes searchers are engaged in while searching. After the last task, participants completed post-session questionnaires that solicited their comments on their experience about each of the three systems.

4.5.3. Description of the search systems

We opened the OSU libraries' website interface (Figure 4-4), and from there, participants could choose various types of library services except LibraryFind. For LibraryFind tasks, we opened the initial page of LibraryFind (Figure 4-1) and directed participants to use the search box. For Google Scholar tasks, we opened initial page of Google Scholar (http://scholar.google.com). We did not conduct any tutorial prior to the session because we wanted to investigate how well participants found information using each system without any instructions.

4.5.4. Search Topics

The four search topics were derived from class assignments from lower level OSU writing courses. Before selecting, we confirmed that each had adequate information resources in each

[15] Software by TechSmith corp. http://www.techsmith.com/morae.asp

system. We communicated each topic[16] verbally rather than in written form. We described each topic in depth as if it was an assignment for a writing course. One of the most challenging activities in information seeking is query formulation and most users struggle to formulate search queries (Belkin, 1982). In a controlled experiment, when search topics and guidelines are explicitly given, these tasks are often typed directly as queries. The initial formulation of a query strongly affects search results, and thus affects experimental results. Crestani & Du (2006) found that using speech to formulate one's information need provides a way to express it more naturally than written topics. We believe that using spoken topics prevents users from copying keywords directly from given topics to formulate their queries.

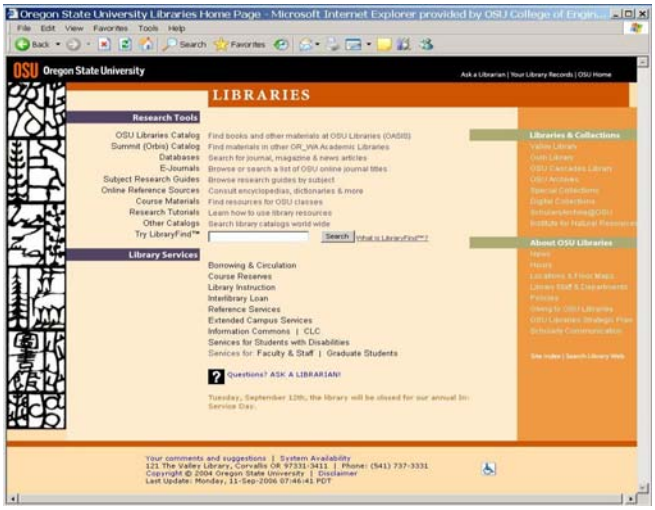

Figure 4-4. OSU Library Website

4.5.5. Data analysis

The Morae recordings and the observations from researchers were reviewed. Answers from open questions were transcribed and grouped into different themes by researchers, such as ease of use, familiarity, relevance, and descriptions of results. The number of answers was counted for each theme, thus yielding a measure of how many of the participants commented on each theme.

We analyzed several metrics per task from user actions and navigation events recorded via Morae. These metrics include time spent to complete each task, the number of queries issued, the number of documents clicked, the number of times results list were returned, the number of times "no results found" returned, and the number of times "next page" was clicked in a search results page.

[16] The topics consisted of 1) history of immigration in Oregon, 2) ethical concerns and political issues surrounding stem cell research, 3) fast food consumption and obesity, and 4) discrimination issues.

We also asked eight librarians to review the information resources that participants selected for their relevance to each topic. The librarians validated participants' material type (e.g. journal article, book) and rated documents in terms of appropriateness for OSU Writing 121 students, especially considering the authority/reliability of the information and its relevance to the topic. For each topic, two librarians reviewed participants' chosen information. We averaged the document ratings between librarians and we considered as "agreed" when at least one librarian agreed with a participant's written material type. Through this validation, we analyzed whether participants were able to recognize material types from each system and whether there is any satisfaction difference between participants and librarians, or between systems.

4.6. RESULTS

We first report on which systems of the three the participants used when given a choice (4.6.1). Then, we explore the factors involved in their decisions (4.6.2) and compared participants' preferences for each system with their success in finding relevant materials (4.6.3).

4.6.1. Which system do undergraduates choose to use?

LibraryFind overcomes some factors that previous studies identified as causing user dissatisfaction (Section 4.4.2). Consequently, we hypothesized that undergraduates would prefer LibraryFind to using the library website and would find LibraryFind as helpful as Google Scholar. To test our hypotheses, we examined which of the three systems participants chose to use in the fourth task. We then analyzed the experience level of the participants in relationship to their system choices, as shown in Figure 4-5.

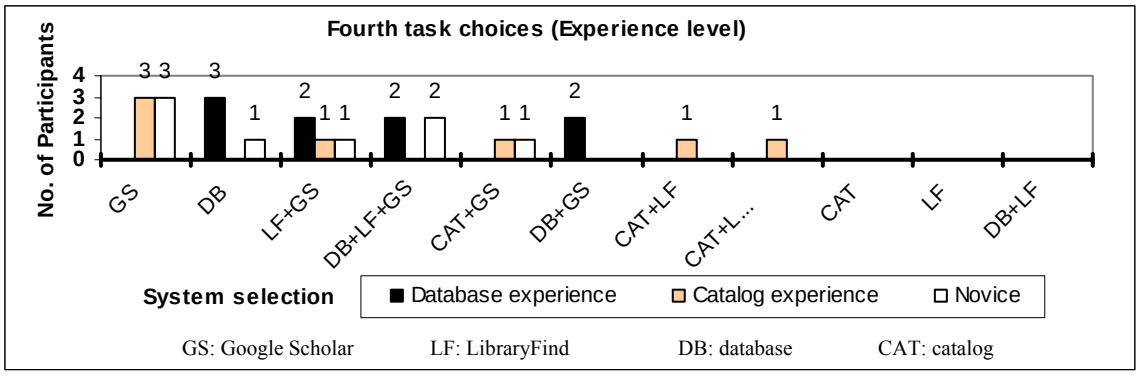

Figure 4-5. Fourth task choices based on participants' experience level as defined in Section 4.5.1.

We found that none of the participants chose LibraryFind or the library catalog from the Library website as the sole system to use in the fourth task. In contrast, four participants chose databases from the library website and six chose Google Scholar as the sole system. Many participants (58%, 14 out of 24) used more than one system when working on the fourth task. This use of multiple systems could suggest that participants were not completely satisfied with the results from one system in terms of relevance; as the task specified that they had to find relevant materials and not a range of material types.

The participants' class level was not related to their choice of system. However, their experience did appear to influence their choice of systems. Participants with database experience tended to select databases over Google Scholar or LibraryFind for the fourth task (78%, 7 of 9). Novice participants (no prior experience with catalogs or databases) favored using Google Scholar (88%, 7 of 8). Among seven catalog experienced participants, only three chose the catalog and all in conjunction with another system. These same participants preferred Google Scholar (6 of 7), while only three participants chose to use LibraryFind.

These results suggest that experience with a library catalog may not be as useful as experience with databases when undergraduates are exposed to new search systems. Only 10 of 24 participants chose LibraryFind for their fourth system, which was not as high as we hypothesized. It might be because they prefer to use familiar systems. However, our results also suggest that undergraduates will switch their primary system if a system can attract their attention through the interface and relevance of returned results. This is demonstrated by catalog experienced participants switching to Google Scholar when given the choice. This means that participants did not choose LibraryFind because they either did not find it compelling enough, they found issues with it, or they found the databases or Google Scholar more compelling.

4.6.2. What factors are important in choosing an academic search system?

A majority of our participants did not choose LibraryFind as their primary academic search system, yet 10 of the 24 did use it in conjunction with another system. We explored why participants chose certain systems for the fourth task by reviewing their stated reasons for choosing a system and their impressions of all three systems. Based on previous studies (Section 4.4.2), we hypothesized that participants would value speed, simple user interface, descriptive metadata, and the availability of full text. We also hypothesized that participants' prior experience would influence their choice of the system.

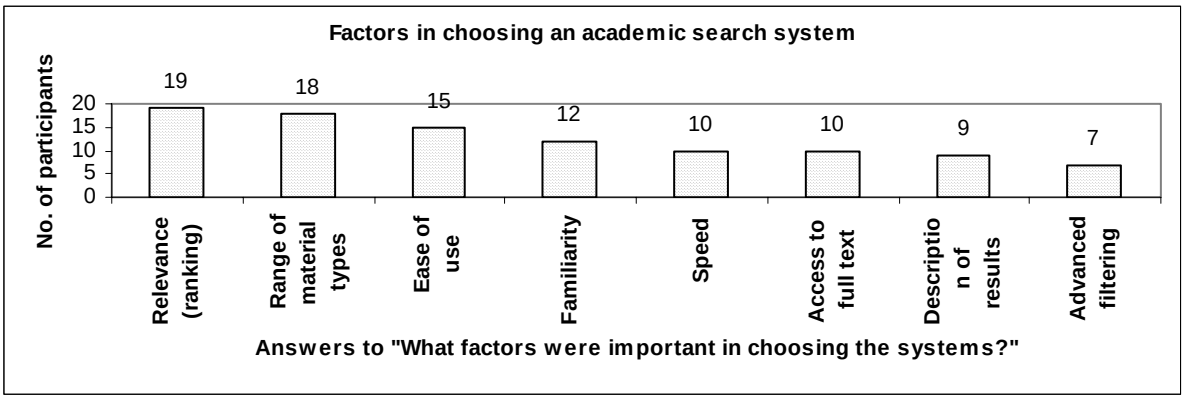

Figure 4-6. Grouped answers from open question

We grouped the comments into eight major factors (Figure 4-6). Getting relevant results with high precision[17] is the most important factor (79%, 19 out of 24). Finding the type of material needed was also important with 75% (18 out of 24) of participants voicing that they considered what each system searched or covered. For example, they chose a search system based on whether they needed a newspaper article, a book, or journal articles. We considered the factors in Figure 4-6 when reviewing the comments on each search system.

4.6.2.1. What factors affect usability of the library website?

Multiple factors emerged that appear to affect usability of the library website as shown in Figure 4-7. Participants value the library website as a means to actually get materials, whether a physical book or a digital article. However, they also perceive a limited range of library resources often due to confusion about various resources, e.g. the differences between the catalog and a database, or experience with one or the others. A majority of the participants still think of the library as a physical place to get actual articles rather than accessing online articles, corroborating Travis and Norlin's findings (2002).

[17] The proportion of retrieved and relevant documents to all the documents retrieved.

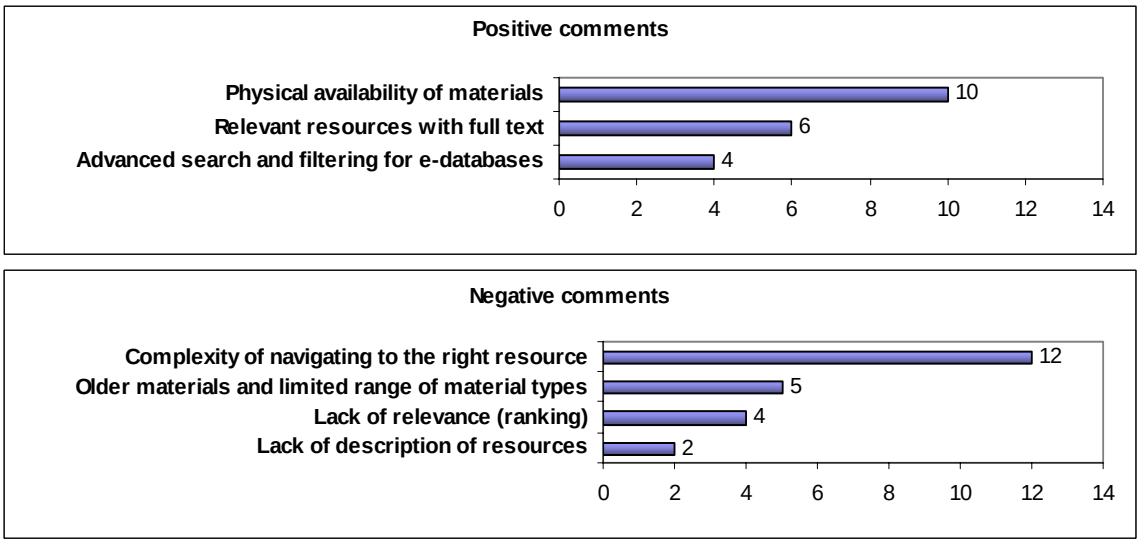

Figure 4-7. Grouped participants' comments about library website

Participants were not satisfied with the catalog as they had difficulty understanding what the library catalog system retrieved and were confused when the returned results were mostly books. For example, P14 searched the catalog to retrieve newspaper article and got confused when he could not find newspapers in the catalog results. This corroborates the findings of previous studies (Eliason, et al., 1997; Chisman, et al., 1999).

They also were frustrated when they got frequent "no results found" errors, since they expected the catalog system to work like web-based search engines which return many results most of the time. When participants encountered the "no results found" error often, they said that catalogs are useful only when they know the call number, exact title, or authors of a book.

> (P06) "It was more specific for like finding a book or for like finding out what the book's about. The majority of all the links on it tended to lead to the call number and where to find it or like just bibliography on the book"

Participants who used the databases from the library website appreciated the relevance of results, full text access, and advanced features. Some gravitated to the databases because they were familiar resources. Once found, these databases could be readily and effectively used.

> (P2) "In middle school and high school, we'd used EBSCO a lot, we had specific library class. I used that because I felt most comfortable with it and because it came out with a lot of good results before"

Navigating to the databases proved frustrating to participants, but seemed to be worth the effort for the participants without database experience. Participants who used a database during the experiment were more satisfied with its performance than Google Scholar. Thus, these participants chose databases in fourth task.

> (P22) "If you find something you like, like here, this subject terms are actually listed there (*by transcriber: EBSCO – narrow results by subject*) and you can click on them too. I like that. I think if I learned how to use this more I'd be a lot better. I like how this set up after you pass all these crazy things (*referring to links necessary to get to EBSCO from library website)*"

Major factors in dissatisfaction and hence usability of the library website were its disorganization and complexity. Participants focused their searching on either the databases or the catalog, and reported differing experiences with each. Problems with the navigation and the interface precluded users from getting to resources so they could judge relevance and satisfaction.

> (P18) "OSU Library website was very disorganized and cluttered. It had bunch of different links to places that I didn't know what they were. Not having global search function really hinders the ability of myself to search. Because I like to be able to at least give general idea what I want"

Catalog-experienced participants struggled to find a starting link from the library website. Four out of seven catalog experienced participants tried other links, but failed to use them and came back to use the catalog. For example, even if they found the databases link, they did not know which database to choose. They also went to the *e-journal list*, which is a listing of holdings, but entered search keywords where the search box prompted for a journal title. Two catalog experienced participants succeeded in using databases and one used only the catalog during the task.

Most of the novice participants also faced obstacles in using the library website effectively. They usually did not find the databases (2 out of 8 novices used databases), and defaulted to using the catalog (4 out of 8 novices), which was not always the best tool depending on the task. Two novices gave up the task. Similar to catalog experienced participants, six (out of 8) novices struggled to use databases or e-journal list.

In summary, the complexity of the library website hindered usability; by attempting to make all resources visible, the library has perhaps made none very accessible. Familiarity through

experience seems to be the key to changing users' perceptions of the usability of the library website.

4.6.2.2. What factors affect the usability of LibraryFind?

Participants did not gravitate to LibraryFind as a first choice, but did appear to be intrigued by it. 42% of all participants (10 out of 24) used LibraryFind in conjunction with other systems when given the choice. The participants who favored LibraryFind noted the range of material types and its ease of use as shown in Figure 4-8. Most of these participants who made favorable comments had no experience with databases.

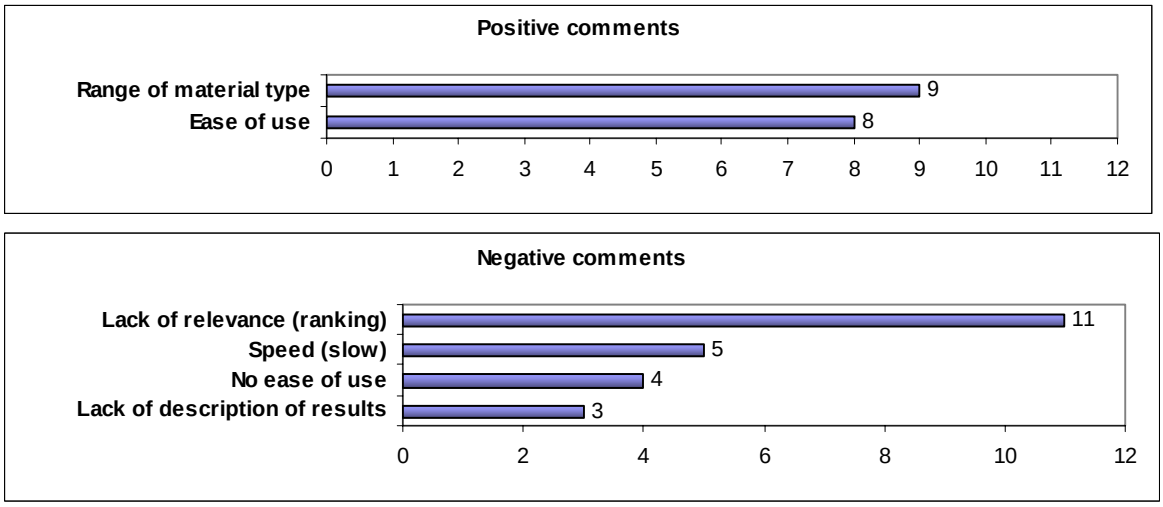

Figure 4-8. Comments about LibraryFind.

Many found it simple to use due to the interface. This reinforces the trend towards most users expecting a search box, rather than navigating links. Participants also liked the range of resources and the sense of discovery.

> (P16)"It's always easy. You type in what you're after, click the button, and it shows you what you can find, where you can find it, and so it's really a while lot less work. LibraryFind ends up finding not only newspapers but also a publication. That's why I like LibraryFind, it finds all sorts of stuff. And then I wanted to find a book. And LibraryFind just had a book, too"

There was disagreement over the relevance of results. From the participants' comments, the dissatisfaction derived from several issues. Several mentioned not understanding what was being searched. There also were comments about knowing enough about the returned results to judge relevance; this argues for longer descriptive summaries or abstracts. Finally, some thought

that LibraryFind results were biased too much towards certain sources such as the local newspaper.

> (P14) “It didn’t really seem to explain or let me have any options for deciding what I was searching in” (P03) “I just didn’t feel like the search results that I was able to identify what was pertinent or relevant to what I had wanted just from the titles or the short summaries” (P05) “It never, never really helps much. When I used it, I was just pulling up articles form the Oregonian”

Slow response time annoyed some participants. Often, speed is a key factor in system choice and satisfaction. Although the majority of participants did not mention slow response time as a dissatisfaction of LibraryFind, most metasearch systems have a speed issue that is not easy to solve given their current configuration.

> (P18) “Search time was a little bit long. I guess it searches through a lot of stuff, but that’s generally kind of annoying, if you have paper due tomorrow”

Even though participants did not choose LibraryFind as their primary system, most did not reject it categorically. They perceived some value and expressed interest in using it more.

> (P04) “If I was looking for things that I could pick up here in a short period of time or local articles, I would definitely be on LibraryFind”

4.6.2.3. What factors affect the usability of Google Scholar?

Participants who favored Google Scholar were satisfied with the relevance, the ease of use, description of results, and familiarity (Figure 4-9). Negative factors were the lack of links to full text and results that participants felt were not appropriate (generally relevant but too advanced) material. Interestingly, participants did not cite relevance as the most important reason they chose Google Scholar. Instead, participants explained that they liked Google Scholar because it returns only academic journal articles and books, unlike Google itself (14 participants). We believe that participants were heavy Google users and so we were surprised that they did not know about Google Scholar.

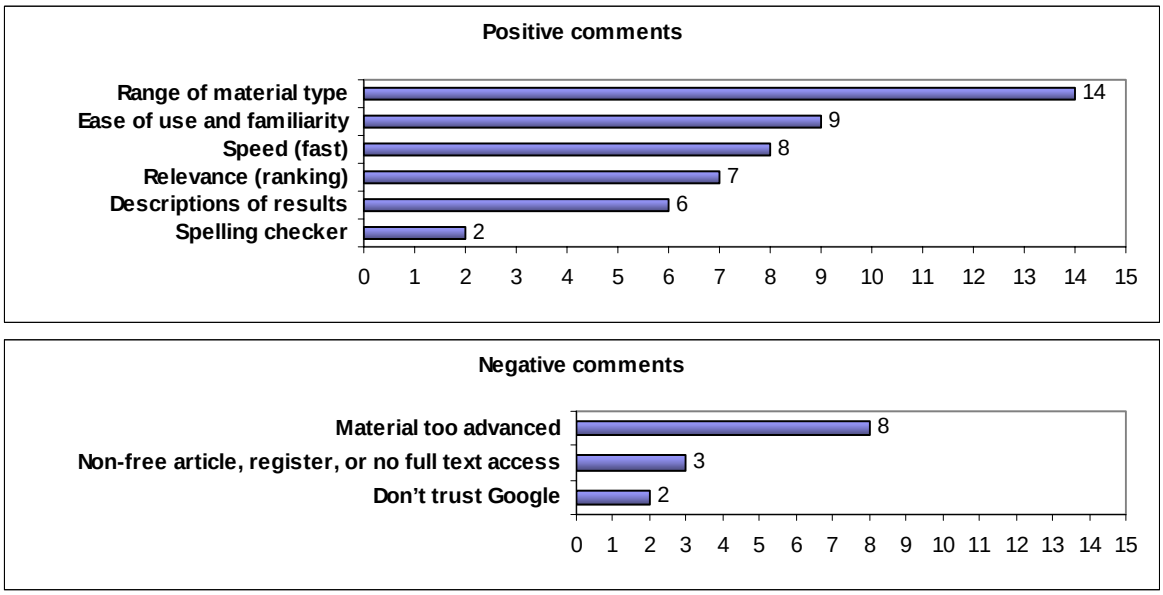

Figure 4-9. Comments in Google Scholar

Our findings suggest that most database experienced participants prefer using databases to Google Scholar, whereas participants without database experience chose Google Scholar because they were not satisfied with LibraryFind or the library website. The following transcript is indicative of participants' reasons for favoring Google Scholar over the other systems:

> (P09) "I was more comfortable with Google Scholar. LibraryFind tended to be a lot slower than Google Scholar. I'm impatient. Also, LibraryFind and Library website (*by transcriber: this participant used catalogs*) are a little bit more complicated and difficult to understand. I think Google Scholar just included the summary. You didn't have to spend your entire time looking through the entire document"

Although participants praised how Google Scholar limited the results to academic journal articles, they also complained that the returned results were targeted too much towards experts (8 participants). Since most journal articles are written for researchers or graduate students, we found that the content was too advanced for undergraduates.

> (P06) "Google Scholar is kind of specific for whatever your topic is. So if you don't have a fundamental grasp of your topic it's a little harder to navigate through it all"

There was also some confusion about access to materials. The links to full text often brought up a request for payment rather than direct access to the information.

> (P14) "The one thing I don't like about Google Scholar is that I don't have easy access to the articles like I do when I'm logged in through the library's website"

Once again, familiarity and past experience influenced choice and impression. Although only three of the participants were familiar with Google Scholar, it is difficult to assess how much their Google experience influenced their Google Scholar comments. Participants were obviously comfortable with Google Scholar's interface even though some had problems with results. Although relevance is important, many participants do not understand how that relevance is decided.

> (P24) "I do like how they show you how the other sources cited it, because that might help me find the relevance of that source. Is it actually sorted by that? So, I'm like wow if there are a lot of other papers mentioning this one, it must be important. So that was good"

4.6.3. How efficiently and effectively did each system work when used by undergraduates?

Our qualitative data identified factors that participants told us they used when choosing search systems. However, the participants' self-reported preferences among systems may or may not reflect their actual usage and success with searching. In this section, we investigate whether there are quantitative differences in participants' search performance among systems. We examined participants' search performance using several metrics described in Section 4.5.5 (4.6.3.1). We also analyzed how participants rated their satisfaction with the documents they selected for each task. We compared their ratings with those of eight librarians as described in Section 4.5.5 (4.6.3.2).

4.6.3.1. Efficiency of searching

We were unable to show any statistically significant difference in the average amount of time that participants spent on each system and on each task (ANOVA, $p > 0.05$). The topic did not affect the amount of time participants spent within each system (ANOVA, $p > 0.05$). This result is not surprising because participants were asked to work on each task for approximately 20 minutes, a time limit we did not strictly enforce. If they were unable to find three satisfactory documents as requested, they wrote down whatever documents they could find into the session questionnaire along with their satisfaction scale. So, while participants spent a similar amount of time with each search system and each topic, their success varied which may lead to perceived differences in the quality and functionality of these search systems.

For the first three tasks, there was no significant difference in the number of queries participants issued (ANOVA, $p > 0.05$). However, participants issued approximately five queries

(4.87) for the fourth task (participants' choice of system), versus seven queries issued on average for each of the first three tasks. Although there is no statistical difference in terms of time spent and number of queries issued, participants spent less time and issued less queries when allowed to choose their system or combination of systems during the last task. This may be related to a learning effect or a comfort level with the testing by the fourth task.

Table 4-1. Three metrics from search results page among systems.

	Search results page		
	Number of times users got first search results page	Number of times users got "no results found" error when queries are issued	Number of times users visited beyond first search results page
Users' choice (fourth task)	114	12 (9.5% of times)	28 (24.6% of times)
Google Scholar	162	14 (8.6%)	52 (32.1%)
LibraryFind	195	37 (19.0%)	84 (43.1%)
Libraries Website	175	51 (29.1%)	34 (19.4%)

Given that the time spent, the number of issued queries, and the number of viewed documents did not significantly differ among systems, we looked for other metrics that might explain participants' different perceptions of each system. We examined the number of search results pages that participants viewed (Table 1). We distinguished between the number of times that participants received at least one page of search results in response to a query (second column of the Table 1) and the number of times participants reached at least one subsequent results pages by clicking "next page" link or clicking on a results page number (fourth column of the Table 1). All of the systems displayed ten documents per search results page as did many of the databases. These numbers suggest how much effort a participant had to expend to find satisfactory results.

Participants viewed more than one results page (e.g. clicking "next page" link) significantly more often when they used LibraryFind than the other two systems (Pearson's Chi-Square test, $\chi 2(3) = 20.13$, $p < 0.01$). Participants clicked "next page" 43.1% of the time (84 out of 195) when they used LibraryFind. This means that users had to look at more pages of results to find satisfactory documents from LibraryFind

Google Scholar returned "no results found" errors significantly less often than the other systems (Pearson's Chi-Square test, $\chi 2(3) = 14.08$, $p < 0.01$), while participants got "no results found" mostly when they used the library website (29.1% (51/175) of the time). Because users

spent effort formulating search queries (Belkin, 1982), they were more likely to get frustrated when they received "no results found" errors.

Overall, the findings suggest that the three systems have similar efficiencies in terms of time spent, number of queries issued, and number of documents viewed. Google Scholar seems to have an edge as participants using LibraryFind viewed more pages, and participants using the library website were more likely to encounter "no results found" errors.

4.6.3.2. Effectiveness of searching

Overall, participants were quite satisfied with the documents they selected (Figure 4-10). Participants were significantly more satisfied with the documents selected during the fourth task when they had more control over their searching choices (ANOVA, fourth task vs. library website, $F[1,131]=16.14$, $p<0.01$; fourth task vs. LibraryFind, $F[1, 134]=15.95$, $p<0.01$; fourth task vs. Google Scholar, $F[1,131]=4.51$, $p=0.04$). Comparing the three systems, participants were the most satisfied with Google Scholar. They tended to be more satisfied with what they found than the librarians reviewing their selections. The difference in ratings between participants and librarians was the least for Google Scholar (0.35) and greatest for the library website (0.95).

Previous studies indicated that librarians were reluctant to use Google Scholar, preferring to use familiar licensed article databases where they were sure of the coverage (Meltzer, 2005; Burright, 2006). Given that our librarians did not know which search systems were used, it is interesting that the librarians were significantly more satisfied with the documents selected from Google Scholar than from any other system (ANOVA, Google Scholar vs. LibraryFind, $F[1,131]=10.66$, $p<0.01$; Google Scholar vs. library website, $F[1,123]=10.87$, $p<0.01$). Recently, the integration of Google Scholar in university library websites indicates that librarians are beginning to accept and promote Google Scholar as a search tool (Mullen & Hartman, 2006; Tenopir, 2005).

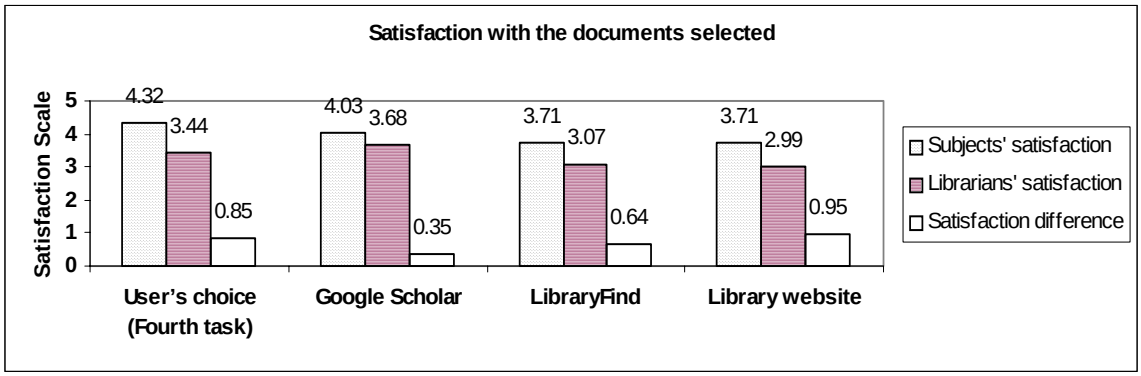

Figure 4-10. Ratings (1 to 5) of participants' satisfaction with the documents they selected for each task, and librarians' review of participants' selections

In terms of material type validation, the librarians classified the selected documents by material type, just as the participants had done during the search sessions. We found that there was a significant variance in librarian-participant rating consistency concerning material types among the systems (ANOVA, $F[3,263]=4.14$, $p<0.01$). The greatest variance was with materials found using the library website (54% agreement) where participants were not confident about assigning material types. For example, the library catalog does not report material types, which confused students. LibraryFind and Google Scholar had similar agreement (79%, and 72%, respectively), and users' choice (fourth task) had 75% agreement. Except for the library catalog, these systems indicate the material type with a word or an icon, which may help explain why LibraryFind and Google Scholar had similar agreement. This feature assisted participants and is valuable when assignments require a range of material types.

The findings in this subsection could suggest that participants are more satisfied with their choice of system while librarians did not agree with users in terms of both satisfaction ratings and material type.

4.7. DISCUSSION AND FUTURE WORK

Throughout this experiment we focused on exploring what makes undergraduates choose their primary academic search system. We also investigated if LibraryFind helps undergraduates find appropriate information efficiently and effectively. Examining the three systems revealed how familiarity, expectations and prior experience inform undergraduates' preferences when choosing a search system. We identified persistent issues that need to be addressed to increase users' satisfaction with LibraryFind and the library website as described in the following sub sections.

4.7.1. Familiarity and Ease of Use

Our findings confirm that familiarity is important, as it shapes the expectations of users. The current generation of undergraduates is familiar with commercial web search engines. Consequently, an interface similar to these web search engines helps users get started and feel confident. For example, users expect results with enough information to make a decision, obvious links to more information or full text, and spell checking. Metasearch engines, including LibraryFind, should develop interfaces that reflect the user's past experience.

4.7.2. Performance Expectations

When metasearch engines incorporate the familiar into their interface, then users expect those systems to perform in a similar fashion. The undergraduates in this study expected relevant results returned quickly. They were disappointed that the simple search box in LibraryFind did not consistently lead to fast, relevant results. These two performance issues are difficult for metasearch engines to overcome because they pass user queries through to multiple sources and return results only as soon as those sources respond.

Improving relevance suggests using more sophisticated ranking algorithms. With preprocessed document indices, we could implement the variants of TF/IDF algorithm (Baeza-Yates & Ribeiro-Neto, 1999). This entails developing the means to index databases sources locally as scholarly web search engines do. Co-citation information is also used by several scholarly web search engines to rank results including Google Scholar. Again, this approach requires local indexing.

Thus, the two obstacles of speed and relevance make metasearch systems less desirable to students as their primary search systems for academic materials. We find that users would accept slower performance only if the results are highly relevant to them.

4.7.3. Recognizing Resources

Users' understanding of what databases or resources they are searching affects their satisfaction with search systems. Participants voiced dissatisfaction when they got results that they were not expecting, when baffled by what they were searching, or when simply not getting results.

When searching the library catalog, some participants received no results when they did not distinguish between keyword searching and title searching, and when they issued misspelled queries. They lost confidence in the system when they got repeated "no results found" errors after

trying to correct spellings several times. These observations suggest that it may be effective to add automatic spell-checking to library search systems.

The library catalog also does not explicitly describe material types in terms that students always understand. The underlying assumption is that the catalog primarily searches for books in the OSU Libraries, whereas undergraduates do not share that assumption.

The lack of clarity regarding what is being searched was observed when users searched the library's e-journals and databases. Almost all the participants without database experience failed to retrieve acceptable results when using the OSU Libraries' e-journals search page. They input their keywords in the search box expecting articles as results, when the e-journals search page is designed to return titles of e-journals. They also avoided using the databases because they were overwhelmed with the number of choices and the lack of guidance on which to select. For these examples, clustering databases by subject may help. Appropriate e-journals or databases could be inferred from the user's search query. One possibility is using metadata and representative keywords to index the e-journals or databases. In that way, we can apply state-of-the-art ranking algorithms to recommend relevant e-journals and databases.

4.7.4. Users' Experience

Users' success with the search systems relates to their experience with those systems. Participants experienced with databases favored using them, even preferring them over Google Scholar. Preference for databases was driven by the availability of advanced search features and the ability to filter results sets. These preferences may help determine whether there is a set of minimum features to implement in academic metasearch engines.

Novices had a great deal of difficulty finding databases on the library's website and identifying which one to use. Given this lack of experience and knowledge of these resources, it may be useful to investigate whether novices would find LibraryFind more helpful if it only searched databases returning just journal and newspaper articles rather than including the catalog. Limiting the range of resources may lead to more satisfaction with the system.

In this study, we focused on participants' prior experiences and class level as predictors of their success. However, we could also examine whether the participant's searching discipline influence their satisfaction with the system.

4.7.5. Future work

This paper reports our preliminary results. One possibility for future work is to analyze query behavior: do users formulate their queries differently when searching academic metasearch tools than when searching the Web? Previous studies showed that users exhibit different query behavior in the home and work environment using the same web search engines (Rieh, 2004). We want to see if our results support the idea of users' query behavior remaining consistent across all types of systems, or whether they adapt their queries significantly based on the systems that they are interacting with. If the search context changes users' searching behavior, then we may need to incorporate a different search algorithm for an academic metasearch system.

It is also possible to design the library website to handle those who are experienced with the different specialized search engines offered (databases and catalogs), and those who are not. For experienced users, the library website could have direct links to databases and the catalog. Such a design is supported by evidence that users who had experience with databases continued to use them. For novices, the most prominent link could take them to a more interactive system, which would educate them as to what the library offers and why they might want to use them. As novices gain experience, they could use direct links (or short trails) from the library home page.

4.8. CONCLUSION

In this study, we explored what makes undergraduates choose their primary academic search system for class assignments and if LibraryFind can benefit them. We conducted a usability experiment comparing three systems. This approach gave us a more objective picture of user response and satisfaction with the academic search systems that college undergraduates actually use. This approach also revealed important general issues with search interfaces and LibraryFind.

Our study reinforced that college undergraduates use what is familiar. Consequently, a new academic metasearch system needs to meld familiarity while capitalizing on the varying experience levels of users. However, when undergraduates face the familiar interface, they expect similar performance to a web search engine, such as its relevance ranking, its speed and certain features. Our participants also expected to know what contents and materials they were searching in a system. Participants' prior experience using an academic search system affects their expectations for and satisfaction with using a new system.

We continue to believe that LibraryFind has the potential to entice undergraduates to use library resources. Until libraries can remove the technical barriers for retrieving highly relevant

materials from multiple resource providers, we will have to settle for offering a good academic search tool, but one that does not yet meet the majority of our users' expectations.

4.9. ACKNOWLEDGMENTS

Funding for this research has been provided by the Gray Family Chair for Innovative Library Services and the Oregon State Libraries. We thank all our LibraryFind development and usability members for their hard work in making the LibraryFind work.

4.10. REFERENCES

Augustine, S., & Greene, C. (2002). Discovering how students search a library web site: A usability case study. College & Research Libraries, 63(4), 354-364.

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). Modern Information Retrieval. ACM Press.

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

Boss, R. (2002). How to plan and implement a library portal. Library Technology Reports, 38(6).

Battleson, B., Booth, A., & Weintrop, J. (2001). Usability testing of an academic library web site: A case study. Journal of Academic Librarianship, 27(3), 188-198.

Boock, M., Nichols, J. & Kristick. L. (2006). Continuing the quest for the quick search holy grail. Internet Reference Services Quarterly, 11(4), 139-153.

Bosman, J., Mourik, I., Rasch, M., Sieverts, E., & Verhoeff, H. (2006). The coverage and functionality of the citation database Scopus including comparisons with Web of Science and Google Scholar. Utrecht University Library. Retrieved January 20, 2007, from http://www.info.scopus.com/news/coverage/utrecht.pdf

Breeding, M. (2005). Reshuffling the desk. Library Journal, 131(6), 40-54.

Burright, M. (2006). Google Scholar. Issues in Science & Technology Librarianship, 45.
Cervone, F. (2005). What we've learned from doing usability testing on OpenURL resolvers and federated search engines. Computers in Libraries, 25, 10-14.

Chisman, J., Diller, K., & Walbridge, S. (1999). Usability testing: A case study. College & Research Libraries, 60, 552-569.

Christenson, H. & Tennant, R. (2005). Integrating information resources: Principles, technologies and approaches. Oakland, CA: California Digital Library. Retrieved February 26, 2007, from http://www.cdlib.org/inside/projects/metasearch/nsdl/nsdl_report2.pdf

Cobus, L., Dent, V. F., & Ondrusek, A. (2004). How twenty-eight users helped redesign an academic library website? Reference & User Services Quarterly, 44(3), 232-246.

Cockrell, B. J., & Jayne, E. A. (2002). How do I find an article? Insights from a web usability study. Journal of Academic Librarianship, 28(3), 122-132.

Crestani, F., & Du, H. (2006). Written versus spoken Queries: A qualitative and quantitative comparative analysis. Journal of the American Society for Information Science and Technology, 57(7), 881-890.

Crawford, W. (2004). Meta, federated, distributed: Search solutions. American Libraries Online, Retrieved November 14, 2006, from http://www.ala.org/ala/alonline.thecrawfordfiles/crawford2004/crawfordAug04.htm

Cunningham, H. (2005). Designing a web site for one imaginary persona that reflects the needs of many. Computers in Libraries, 25(9), 15-19.

Eliasen, K., McKinstry, J., & Fraser, B. M. (1997). Navigating online menus: A quantitative experiment. College & Research Libraries, 58(6), 509-516.

Hamburger, S. (2004). How researchers search for manuscript and archival collections. Journal of Archival Organization, 2(1/2), 79-97.

Hwang, S-Y., Huang, J., & Srivastava, J. (1993). Concurrency Control in Federated Databases: A Dynamic Approach. Conference on Information and Knowledge Management, 694-703.

Jacso, P. (2005). As we may search - comparison of major features of the Web of Science, Scopus, and Google Scholar citation-based and citation-enhanced databases. Current Science, 89(9), 1537-1547.

Kuehl, R.O. (1994). Statistical principles of research design and analysis. Belmont, Calif.: Duxbury Press.

Lee, J. (2006). Earth Sciences metasearch portal usability testing. Retrieved November 14, 2006, from http://www.cdlib.org/inside/projects/metasearch/nsdl/UCLA_NSDL_062006.pdf

Lim, E-P., Hwang, S-Y., Srivastava, J., Chements, D., Ganesh, H. (1995). Myriad: Design and Implementation of a Federated Database Prototype. Software: Practice and Experience, 25(5), 533-562.

Luther, J. (2003). Trumping Google? Metasearching's promise. Library Journal, 128(16), 36-39.

Narshall, P., Herman, S., & Rajan, S. (2006). In search of more meaningful search. Serials Review, 32(3), 172-180.

Meltzer, E. (2005). UC Libraries use of Google Scholar. Retrieved November 14, 2006, from http://www.cdlib.org/inside/assess/evaluation_activities/docs/2005/googleScholar_summary_0805.pdf

Meng, B., Yu, C. & Lie, K.L. (2002). Building efficient and effective metasearch engines. ACM Computing Surveys, 34(1), 48-89.

Mullen, L. B. & Hartman, K. A. (2006). Google Scholar and the library web site: The early response by ARL Libraries. College & Research Libraries, 67(2), 106-122.

Neuhaus, C., Neuhaus, E., Asher A., & Wrede, C. (2006). The depth and breadth of Google Scholar: An empirical study. Portal: Libraries and the Academy, 6(2), 127-441.

Notovny, E. (2004). I don't think I click: A protocol analysis study of use of a library online catalog in the Internet age. College and Research Libraries, 65(6), 525-537.

Notess, G. R. (2005). Scholarly web searching: Google Scholar and Scirus. Online, 29(4), 39-41.

Randall, S. (2006). Federated searching and usability testing: building the perfect beast. Serials Review, 32(3), 181-182.

Reeb, B., D'lgnazio, J., Law, J., & Visser, M. (2006). Federated search observed in the context of student writing: Taking steps towards improving user experience. College & Research Libraries News, 67(6), 352–355.

Reese, T. (2006). Metasearch: Building a shared metadata-driven knowledge base system. Ariadne, (47)

Rieh, S-Y (2004). On the Web at home: Information seeking and Web searching in the home environment. Journal of the American Society for Information Science and Technology, 55(8), 743-753.

Stephan, E., Cheng, D. T., & Young, L. M. (2006). A usability survey at the University of Mississippi Libraries for the improvement of the library home page. Journal of Academic Librarianship, 32(1), 35-51.

Tallent, E. (2004). Metasearching in Boston College Libraries: A case study of user reactions. New Library World, 105(1/2), 69-75.

Tennant, R. (2005). Is metasearching dead? Library Journal, 28.

Tenopir, C. (2005). Google in the academic library. Library Journal, 32.

Thomsett-Scott, B. (2005). Providing a complete menu: Using competitive usability in a home page usability study. Technical Services Quarterly, 23(2), 33-47.

Travis, T.A. & Norlin, E. (2002). Testing the competition: Usability of commercial information sites compared with academic library web sites. College & Research Libraries, 63(5), 433-447.

Yakel, E. (2004). Encoded Archival Description: Are finding aids boundary spanners or barriers for users? Journal of Archival Organization, 2(1/2), 63-77.

5. GENERAL CONCLUSION

Our research objective is to explore search technologies that could provide substantial improvements in efficiency and effectiveness of document search. Consequently, the work presented in this dissertation focuses on technical solutions that support people in finding relevant documents in large collections. We addressed three different challenges in this dissertation: two focused on designing and developing document search engines, and one focused on improving recommendation accuracy.

The first challenge was: "**Can users find documents that provide relevant information more efficiently and effectively when given automatically detected recommendations from past users**?"

To address this challenge, we adapted the well-understood model of classic collaborative filtering to incorporate the concept of an immediate information need. To predict ratings for documents, we used the similarity between information needs (between queries) to locate records of users with past information needs and the associated documents that the past users rated. We evaluated the system that we developed experimentally and observationally, and in both cases we found that recommendations from prior users with similar queries could increase the efficiency and effectiveness of document search (see Chapter 2).

The second challenge was "**Can click data reliably indicate users' implicit preferences**?"

We have explored the reliability of click data as a source of implicit relevance feedback data. The results 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 a highly reliable source of implicit relevance feedback data. Combining information about the last-visited-documents with other implicit feedback data could increase the reliability of the feedback. This work was presented in Chapter 3.

The third challenge was "**How effective are existing academic metasearch systems, and which specific characteristics of a metasearch engines are actually important for end user usability**?"

Based on our previous experience with the SERF system, we found that it was impractical to offer recommendations for library databases offered by third-party vendors, because OSU library

does not have rights to access these proprietary databases locally and these sources change their URLs dynamically, making external linking challenging. This technical limitation precluded our use of SERF in a library setting, because the main objective of the library website is to recommend resources to patrons regardless of whether the resource is available locally or remotely through a third-party database. To solve this problem, we designed a metasearch system for academic materials. To achieve our research objectives, we conducted a formal usability study, which recorded audio and video with the concurrent desktop navigation of the participants. We also requested that participants think-aloud while they were searching for information. We found that modeling the familiarity and ease of use of commercial web search engines is an important factor to attract undergraduates. However, when undergraduates face a familiar interface, they expect similar performance to a web search engine, such as its relevance ranking, its speed and certain features. We also found that users expect to know what kinds of materials are available in a collection in which they were searching. Users' prior experience using an academic search systems also affected their expectations for and satisfaction with using a new system. This work was presented in chapter 4.

The systems discussed in this dissertation (SERF and LibraryFiind) have been designed specifically for the library environment, and thus the findings we report may not generalize to other environments. Many issues regarding the reliability of explicit and implicit feedback data still need to be resolved before we can apply our approach to general web search. Nevertheless, we believe that our approaches could be applied successfully in many other environments, and that integrating collaborative filtering ideas, such as those we have described in this dissertation, have the potential to create more effective information filtering systems.

6. BIBLIOGRAPHY

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

Augustine, S., & Greene, C. (2002). Discovering how students search a library web site: A usability case study. *College & Research Libraries*, 63(4), 354-364.

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. New York: ACM Press.

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

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

Battleson, B., Booth, A., & Weintrop, J. (2001). Usability testing of an academic library web site: A case study. *Journal of Academic Librarianship*, 27(3), 188-198.

Belkin, N.J., Cool, C., Kelly,D., Kim, G., Kim, J-Y., Lee, H-J., Muresan, G., Tang, M-C. & Yuan, X-J. (2003) Query Length in Interactive Information Retrieval. *Proceedings of the 26th annual international ACM SIGIR*, 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 Hlava, M.K., Woods, L. (Eds.). *Proceedings of the 62nd Annual Meeting of American Society for Information Science*, 633-642. Medford, NJ: ASIS..

Boss, R. (2002). How to plan and implement a library portal. *Library Technology Reports*, 38(6).

Boock, M., Nichols, J. & Kristick. L. (2006). Continuing the quest for the quick search holy grail. *Internet Reference Services Quarterly*, 11(4), 139-153.

Bosman, J., I van Mourik, M. Rasch, E. Sieverts, & H. Verhoeff. (2006). Scopus reviewed and compared: The coverage and functionality of the citation database Scopus including comparisons with Web of Science and Google Scholar. *Utrecht University Library*. Retrieved January 20, 2007, from http://www.info.scopus.com/news/coverage/utrecht.pdf

Breeding, Marshall, (2005). Reshuffling the desk. *Library Journal*, 131(6), 40-54.

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

Burright, M. (2006). Google Scholar. *Issues in Science & Technology Librarianship*, 45. Retrieved November 14, 2006, from http://www.istl.org/06-winter/databases2.html/

Cervone, F. (2005). What we've learned from doing usability testing on OpenURL resolvers and federated search engines. *Computers in Libraries*, 25, 10-14.

Chisman, J., Diller, K., & Walbridge, S. (1999). Usability testing: *A case study. College & Research Libraries,* 60, 552-569.

Christenson, H. & Tennant, R. (2005, August). Integrating information resources: Principles, technologies and approaches. Oakland, CA: *California Digital Library*. Retrieved February 26, 2007, from http://www.cdlib.org/inside/projects/metasearch/nsdl/nsdl_report2.pdf

Claypool, M., Le, P., Wased, M., & Brown, D. (2001). Implicit interest indicators. In Sidner, C, Moore, J. (Eds.). *Proceedings of the 6th International Conference on Intelligent User Interfaces*, 33-40. New York, NY: ACM Press.

Cobus, L., Dent, V. F., & Ondrusek, A. (2004). How twenty-eight users helped redesign an academic library web site. *Reference & User Services Quarterly*, 44(3), 232-246.

Cockrell, B. J., & Jayne, E. A. (2002). How do I find an article? Insights from a web usability study. *Journal of Academic Librarianship,* 28(3), 122-132.

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*, 35-46. San Francisco, CA: Morgan Kaufman.

Crestani, F., & Du, H. (2006). Written versus spoken Queries: A qualitative and quantitative comparative analysis. *Journal of the American Society for Information Science and Technology*, 57(7), 881-890.

Crawford, W. (2004). Meta, federated, distributed: Search solutions. *American Libraries Online,* Retrieved November 14, 2006, from http://www.ala.org/ala/alonline.thecrawfordfiles/crawford2004/crawfordAug04.htm

Cunningham, H. (2005). Designing a web site for one imaginary persona that reflects the needs of many. *Computers in Libraries*, 25(9), 15-19.

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,* 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 Nulti-Tasking Knowledge Workers. *In International Conference on Intelligent User Interfaces,* 75-82.

Eliasen, K., McKinstry, J., & Fraser, B. M. (1997). Navigating online menus: A quantitative experiment. *College & Research Libraries*, 58(6), 509-516.

Fox, S., Karnawat, 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.

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

Hamburger, S. (2004). How researchers search for manuscript and archival collections. *Journal of Archival Organization*, 2(1/2), 79-97.

Highsmith, A.L. &. Ponsford, B.C. (2006). Notes on Metalib® implementation at Texas A&M University. *Serials Review*, 32(3), 190-1194.

Hill, W., Stead, L., Rosenstein, M. & Furnas,G. (1995). Recommending and evaluating choices in a virtual community of use. In Katz, I., Mack, R., Marks, L., Rosson, M.B., Nielsen, J. (Eds.), *Proceedings of SIGCHI on Human Factors in Computing Systems*, 194-201. New York, NY: ACM Press.

Hwang, S-Y., Huang, J., & Srivastava, J. (1993). Concurrency Control in Federated Databases: A Dynamic Approach. *Proceedings of the second international conference on Information and Knowledge Management (CIKM)*, 694-703.

Jacso, P. (2005, November). As we may search - comparison of major features of the Web of Science, Scopus, and Google Scholar citation-based and citation-enhanced databases. *Current Science*, 89(9), 1537-1547.

Jansen, B.J., & Spink, A. (2003). An analysis of web documents retrieval and viewed. *The 4th International Conference of Internet Computing*, Las Vegas, Nevada, 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 Zaiane, O.R.(Ed.). *ACM International Conference on knowledge Discovery and Data Mining*, 133-142. New York, NY: ACM Press.

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

Jung, S., Harris, K., Webster, J. & Herlocker, J.L. (2004) SERF: Integrating Human Recommendations with Search. *In Preceedings of the Thirteenth ACM International Conference on Information and Knowledge Manangement (CIKM)*, 571-580.

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

Kantor, P.B., Boros, E., Melamed, B. & Menkov, V. (1999). The information quest: A dynamic model of user's information needs. In Hlava, M.K., Woods, L. (Eds.). *Proceedings of the 62nd Annual Meeting of American Society for Information Science,* 536-545. Medford, NJ: ASIS.

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

Kelly, D., & Belkin, N.J. (2004). Display time as implicit feedback: Understanding task effects. In Sanderson, M., Jarvelin, K., Allan, J., Bruza, P. (eds.). *Proceedings of the 27th annual*

International ACM SIGIR Conference on Research and Development in Information Retrieval, 377-384. New York, NY: ACM Press.

Kelly, D & Belkin, N.J. (2001). Reading time, scrolling and interaction: Exploring implicit sources of user preferences for relevance feedback. In Kraft, D.H., Croft, W.B., Harper, D.J., Zobel, J. (Eds.). *Proceedings of the 24th annual international ACM SIGIR Conference on Research and Development in Information Retrieval*, 408-409. New York, NY: ACM Press.

Konstan, J., Miller, B., Maltz, D., Herlocker, J., Gordon, L., & Riedl, J. (1997). GroupLens: Applying collaborative filtering to Usenet News. *Communications of the ACM*, 40(3): 77-87.

Kuehl, R.O. (1994). Statistical principles of research design and analysis. *Belmont, Calif.: Duxbury Press*.

Lee, J. (2006, May). Earth Sciences metasearch portal usability testing. Retrieved November 14, 2006, from http://www.cdlib.org/inside/projects/metasearch/nsdl/UCLA_NSDL_062006.pdf

Letnikova, G. (2003). Usability testing of academic library websites: A selective annotated bibliography. *Internet Reference Services Quarterly,* 8(4), 53-68.

Lim, E-P., Hwang, S-Y., Srivastava, J., Chements, D., Ganesh, H. (1995). Myriad: Design and Implementation of a Federated Database Prototype. *Software: Practice and Experience*, 25(5), 533-562.

Luther, J. (2003). Trumping Google? Metasearching's promise. *Library Journal*, 128(16), 36-39.

McNee, S.M., Albert, I, Cosley, D., Gopalkrishnan, P., Lam, S.K., Rashid, A.M., Konstan, J.A. & Riedl, J. (2002). On the recommending of citations for research papers. In E.F. Churchill, J. McCarthy, C. Neuwirth and T. Rodden (Eds.). *Proceedings of the 2002 ACM Conference on Computer Supported Cooperative Work*, 116-125. New York, NY: ACM Press.

Meltzer, E. (2005). UC Libraries use of Google Scholar. Retrieved November 14, 2006, from http://www.cdlib.org/inside/assess/evaluation_activities/docs/2005/googleScholar_summary_0805.pdf

Meng, W., Yu, C. & Liu, K. (2002). Building Efficient and Effective Metasearch Engines. *ACM Computing Surveys*, 34 (1): 48-89.

Menkov, V., Neu, D.J. & Shi, Q. (2000). AntWorld: A collaborative web search tool. In P. Kropf, G. Babin, J. Plaice and H. Unger (Eds.). *Proceedings of the 2000 Workshop on Distributed Communications on the Web*, 13-22. Berlin: Springer-Verlag.

MetaCrawler, http://www.metacrawler.com.

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

Mullen, L. B. & Hartman, K. A. (2006). Google Scholar and the library web site: The early response by ARL Libraries*. College & Research Libraries*, 67(2), 106-122.

Narshall, P., S. Herman, & S. Rajan. (2006). In search of more meaningful search*. Serials Review*, 32(3), 172-180.

Neuhaus, C., Neuhaus, E., Asher A., & Wrede, C. (2006). The depth and breadth of Google Scholar: An empirical study. *Portal: Libraries and the Academy*, 6(2), 127-441.

Nielsen, J. (2006). Quantitative studies: how many users to test? Jakob Nielsen's Alertbox. Retrieved November 14, 2006, from http://www.useit.com/alertbox/quantitative_testing.html

Notovny, E. (2004). I don't think I click: A protocol analysis study of use of a library online catalog in the Internet age. *College and Research Libraries*, 65(6), 525-537.

Notess, G. R. (2005, July/August). Scholarly web searching: Google Scholar and Scirus. Online, 29(4), 39-41.

Oard, D.W. (1997). The state of the art in text filtering. *User Modeling and User-Adapted Interaction*, 7: 141-178.

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

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

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

Resnick, P., & Varian H. (1997) Recommender Systems. *Communication of the ACM*, 40(3):56-58.

Randall, S. (2006). Federated searching and usability testing: building the perfect beast. *Serials Review*, 32(3), 181-182.

Reeb, B., D'lgnazio, J., Law, J., & Visser, M. (2006). Federated search observed in the context of student writing: Taking steps towards improving user experience, *College & Research Libraries News*, 67(6), 352–355.

Reese, T. (2006, April). Metasearch: Building a shared metadata-driven knowledge base system. Ariadne, (47), Retrieved November 14, 2006, from http://www.ariadne.ac.uk/issue47/reese/

Rieh, S-Y (2004). On the Web at home: Information seeking and Web searching in the home environment. *Journal of the American Society for Information Science and Technology*, 55(8), 743-753.

Rocchio, J. (1971). Relevance feedback in information retrieval. In G. Salton (Ed.). *The SMART Retrieval System: Experiments in Automatic Document Processing,* 313–323. Englewood Cliffs, NJ: Prentice-Hall.

Schafer, J.B., Konstan, J. & Riedl, J. (1999) Recommender Systems in E-Commerce. *Proceedings of the ACM 1999 Conference on Electronic Commerce*.

Schafer, J.B., Konstan, J. A. & Riedl, J. (2001) E-Commerce Recommendation Applications. *Journal of Data Mining and Knowledge Discovery.*

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

Smyth, B., Freyne, J., Coyle, M., Briggs, P., Balfe, E. (2003). I-SPY: Anonymous, community-based personalization by collaborative web search. In Bramer, M.A., Ellis, R. (Eds.). *Proceedings of the 23rd SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence*, 367-380. London: Springer-Verlag.

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.

Spink, A., Jansen, B. J., & 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.

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

Stephan, E., Cheng, D. T., and Young, L. M. (2006). A usability survey at the University of Mississippi Libraries for the improvement of the library home page. *Journal of Academic Librarianship*, 32(1), 35-51.

Tallent, E. (2004). Metasearching in Boston College Libraries: A case study of user reactions. *New Library World*, 105(1/2), 69-75.

Tennant, R. (2005, July 15). Is metasearching dead? *Library Journal*, 28.

Tenopir, C. (2005, February 1). Google in the academic library. *Library Journal*, 32.

Thomsett-Scott, B. (2005). Providing a complete menu: Using competitive usability in a home page usability study. *Technical Services Quarterly*, 23(2), 33-47.

Travis, T.A. & Norlin, E. (2002). Testing the competition: Usability of commercial information sites compared with academic library web sites. *College & Research Libraries*, 63(5), 433-447.

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

Wen, J-R, Nie, J-Y., & Zhang, H-J. (2001). Clustering user queries of a search engine. In Shen, V.Y., Saito, N., Lyu, M.R., Zurko, M.E. (Eds.). *Proceedings of the 10th International Conference on World Wide Web*, 162-168. New York, NY: ACM Press.

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 Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 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 Clarke, C, Cormack, G. (Eds.). *Proceedings of the 26th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 56-63. New York, NY: ACM.

Yakel, E. (2004). Encoded Archival Description: Are finding aids boundary spanners or barriers for users? Journal of Archival Organization, 2(1/2), 63-77.