

The Oregon State University Digital Library Framework

Kevin Harris
Dr. Jon Herlocker
School of Electrical Engineering and Computer Science
Oregon State University
August 22, 2003

ABSTRACT

The Oregon State University Digital Library Framework (OSU DLF) is a system for performing experiments on recommendation algorithms in domain-specific web based searches. Since users' information needs vary with time, the system compares question similarity instead of user similarity. Additionally, the OSU DLF attempts to overcome concerns of users not voting by implementing criteria defining an implicit vote. Finally, it is being designed with the goals of portability, configurability, and ease of maintenance. This document provides a high-level overview of the system as well as future research goals using it.

INTRODUCTION

Search engine users have an information need that they attempt to satisfy using the engines. Information needs can be the answer to a simple question or to more complex questions that may require the synthesis of many pieces of information. Search engines attempt to filter the pages that they have information about through some algorithmic process. Traditionally, this is a form of content analysis, such as *keyword filtering*.

Keyword filtering occurs when a search engine takes a list of keywords from a user and uses them to determine the relevance of information to be displayed to the user. Information that is determined to closely match the keywords is given a higher priority than that which does not, and frequently this is indicated to the user by appearing closer to the top of a list of results. The major problem with this approach is that determining the quality of content returned by content-based analysis is difficult if high quality and low quality documents share the same keywords in similar frequencies.

Collaborative filtering, also known as *social filtering*, attempts to overcome this issue by having users rate content. This information can then be used to recommend information to other users based upon a recommendation algorithm. A successful example of collaborative filtering includes MovieLens [1]. In MovieLens, users rate movies that they have seen and then receive recommendations from the system. Recommendations are generated based upon similarities between user ratings on movies.

The Oregon State University Digital Library Framework (OSU DLF) has been in development for approximately two years and is primarily a test bed for new collaborative filtering algorithms. Anonymous and registered users enter *questions* and are recommended *question-document pairs* in addition to results from a keyword-filtered search engine. When viewing a page, users have the ability to place votes on a four-point scale. Figures 1-3 are screenshots of the initial page, the search results page, and of viewing a document.

In order to be an effective framework for collaborative filtering algorithms, the design focuses on *task-oriented searching*, *voting*, and *administrative ease*. Each of these goals will be discussed in the following sections followed by a brief plan of future research to improve the framework.

The OSU Libraries Digital Library System for Electronic Recommendation Filtering

Search [Help](#)

Entering a FULL question helps us find a better match for your question, so you'll get better results.

Log In

Login:

Password:

Save login & password. [Why save?](#)
[Forgot your password?](#)

Examples

- [Where can I find journal articles on Computer Science?](#)
- [Where can I find information on Library hours?](#)
- [How can I checkout a book?](#)

Register

Join the OSU Libraries Digital Library community!

[Why register?](#)

(a)

The OSU Libraries Digital Library System for Electronic Recommendation Filtering

Search [Help](#)

Entering a FULL question helps us find a better match for your question, so you'll get better results.

Recent Questions

- [Where am I?](#)
- [Why does zoology show up for popular culture?](#)
- [Where can I find journal articles on Popular Culture?](#)
- [Where can I find information on Library hours?](#)
- [journal of the american society for information science and technology](#)
- [Where are journal articles, databases, or other resources on Physical Science?](#)
- [Where are journal articles, databases, or other resources on Social Science?](#)
- [Where are journal articles, databases, or other resources on Popular Culture?](#)

Frequently Visited Documents

- ◆ [Zoology Research Guide](#)
- ◆ [The Valley Library Guide](#)
- ◆ [Oregon State University electronic journal subject list](#)
- ◆ [Oregon State University: OSU Libraries Database Trials & News Page](#)
- ◆ [History of Science Research Guide](#)
- ◆ [Popular Culture: Race, Rights, and Responsibility - Video | Valley Library - Oregon State University](#)

Examples

- [Where can I find journal articles on Popular Culture?](#)
- [Where can I find information on Library hours?](#)
- [How can I checkout a book?](#)

Administration

[Administrative Tools](#)

(b)

Figure 1: (a) The initial page of the OSU DLF. Users enter their questions in the text box on the left side and are given the option to log in using the interface on the right side. (b) The initial page after a user has logged in. Note that the administration box only appears for users that have been flagged as Administrators.

OREGON STATE UNIVERSITY LIBRARIES [Home](#) [About](#) [Edit Profile](#) [Log Out](#)
 SYSTEM FOR ELECTRONIC RECOMMENDATION FILTERING

Original Question: **WHERE CAN I FIND JOURNAL ARTICLES ON COMPUTER SCIENCE?** [Ask New Question](#)

Revise Question :

Users with similar questions found these pages helpful

[+] Where are **journal articles**, databases, or other resources on Mathematics & **Computer Science**?

[Database] [Applied Science & Technology Index](#) Index to journal articles from over 400 research journals covering aeronau... more ▶

[Database] [Compendex \(Engineering Index\)](#) Engineering Index (Compendex) covers all areas of engineering including: ac... more ▶

[Database] [Science Citation Index \(Web of Science\)](#) Index of cited articles from over 5000 scientific journals in all areas of ... more ▶

[+] Where are **journal articles**, databases, or other resources on Physical **Science**?

[Database] [Academic Search Elite](#) Academic Search Elite provides full text for over 1,250 academic, social sc... more ▶

[Database] [Applied Science & Technology Index](#) Index to journal articles from over 400 research journals covering aeronau... more ▶

[Database] [ArticleFirst](#) Indexes articles from over 12,500 journals in all fields. Use ArticleFirst ... more ▶

[+] Where are **journal articles**, databases, or other resources on Environmental **Science**?

[Database] [Academic Search Elite](#) Academic Search Elite provides full text for over 1,250 academic, social sc... more ▶

[Database] [AGRICOLA \(FirstSearch\)](#) Index to journal articles, government reports and extension publications. ... more ▶

[Database] [Applied Science & Technology Index](#) Index to journal articles from over 400 research journals covering aeronautics; chemistry; computer technology; engineering; environment; food science; geology; mathematics; metallurgy; minerology; oceanography; petroleum/gas; physics; plastics; textiles; transportation; etc. Abstracts and index: 1983-present

[More Similar Questions]

Results 1-10 of 57 for "Where can I find journal articles on computer science?"

Page 1 [2](#) [3](#) [4](#) [5](#) [6](#) [Next >>](#)

Oregon State University Finding Articles in Electronic ...
<http://osulibrary.oregonstate.edu/research/findart.htm>
 ... are using a **computer** on campus, you ... authorized user and **can** search any of ... for magazine or **journal articles** on a ... 1987-1997; **Science** Citation Index ...

Oregon State University Libraries Class Assignment Help Page ...
<http://osulibrary.oregonstate.edu/instruction/classign/ans430.htm>
 ... Terms for **Computer** Searching (Boolean ... and the **articles** are not ... fields of **science**. Only

The Oregon State University Library [Research Gateway](#) is the perfect place to start researching any topic, whether you're looking for books, journals, articles, or any other form of media.

Figure 2: The results screen. Recommendations appear in the top half of the display, and search results appear in the bottom half. The *[database]* label indicates that the document links to a topic database. If descriptions are available, we display an excerpt to the user and give them the option of mousing over the *more* label to read more, as shown by the last recommendation. Normal documents appear with a *[page]* label. To the right of the search results are the system links, which display links relevant to registered users regardless of question based upon their profile information.



Figure 3: Viewing a document from the results page. The OSU DLF header is in a frame at the top of the page. The requested page is displayed in the bottom pane after being wrapped.

TASK-ORIENTED SEARCHING

A *task* in the OSU DLF is defined as the process of entering an initial *question* followed by *question revision* and *document selection*. A *question* is a complete sentence, while *question revision* includes altering an initial question with the goal of retrieving the same information. *Document selection* occurs when a user clicks on a result from the *recommender* or *search result*. Documents currently consist of HTML or PDF files. A task ends when a user enters a new question from the initial page.

An idealized user interaction with the system is as follows: Pat is looking for a computer science article on collaborative filtering. At the initial screen, Pat enters “Where can I find computer science articles?” and is presented with a results screen with recommendations as well as traditional search results. Deciding that the shown documents are too general, Pat then changes the question to “Where can I find computer science articles on collaborative filtering?”. This time, Pat sees a few documents that sound promising. Clicking on the first recommended document for the similar question “Where are social filtering resources?”, Pat accesses a document with the desired information.

Questions vs. Queries

Questions are of primary importance in task-oriented searching. Comparing user similarities like in MovieLens is not appropriate in a system where user information needs are constantly changing. New questions are thus compared with questions that have already been asked in the

system according to the current recommendation algorithm. Training the user to enter complete questions is thus a secondary research goal of the system.

The ambiguity of keyword based queries limits their usefulness in such a design. For example, the query “popular culture” could have a variety of meanings that only become apparent in complete questions: “Where can I find resources on popular culture?” or “How does popular culture influence students?” The answers to these explicit questions are likely answered by different documents, and thus comparisons based on keywords alone would potentially damage users' trust in the system when the recommended documents do not meet their information need.

Encouraging Questions

Belkin, et. al., found that prompting the user to enter a question, as opposed to key terms, positively correlated with users entering complete questions [2]. We are using these findings throughout the user interface by always referring to *questions* instead of queries or keywords. Additionally, initial questions are typed in a large box as opposed to a single line. Question revision does use a single row box for input, but this is used to differentiate it from asking an initial question.

Question comparison by itself does little to help the user. As such, when users vote for a document it is in terms of how well it answered their question, forming a *question-document pair*. Similarly, recommendations display questions similar to the user's and the documents that received the highest ratings for answering each recommended question. In order to increase the chances of question overlap, the OSU DLF only operates in a limited domain, such as tsunamis or the OSU Library web site.

Operating in a limited domain also allows us to optimize search results based upon user profile information. Currently, this is done using *system links*. System links appear to the right of the search results and contain links to pages that are frequently visited regardless of a user's question. For example, when the OSU DLF is used to manage a library web site, links to the card catalog and interlibrary loan appear (Figure 2).

VOTING

Voting is essential for recommendation systems to function. Ideally, users would always make explicit votes when viewing documents. This is unlikely to happen in actual usage, so in addition to an explicit voting system we have also implemented an implicit voting mechanism. Finally, we have taken steps to ensure the integrity of votes to guard against malicious users.

Explicit Voting

As voting on question-document pairs is an integral part of the system, voting must be made as easy and unobtrusive as possible. The current mechanism is a voting bar in the header (Figure 4). With this method, the user indicates the degree that a document answers the current question on a four-point scale displayed in an English sentence: Did this *Answer*, *Help Answer*, *Somewhat Answer*, or *Not Answer* your question?



Did this Answer, Help Answer, Somewhat Answer, or Not Answer your question?

Figure 4: The vote bar. Each of the underlined items is a link that, when the user clicks it, records the vote for the current question-document pair.

Votes are recorded for each registered user for each question-document pair. New votes by the same user on the same question-document pair overwrite old votes. This allows the system to record users changing their perception of the usefulness of the document. Anonymous user votes are, however, never replaced; instead, new votes are averaged with existing anonymous votes on each question-document pair. This is done as a compromise between vote storage requirements and tracking anonymous vote information.

Implicit Voting

Since collaborative filtering depends upon user votes, we have implemented a system that places implicit votes based on user browsing behavior. The time users spend browsing a page, the number of times they visit a document, and whether or not the document is the final document for a given task are the components that we have identified as browsing behavior. In order to record this information, document requests are altered to go through the OSU DLF.

This alteration, or *wrapping*, consists of rewriting HTML tags to either be absolute references, such as for images, or an absolute reference that uses the OSU DLF similar to a proxy server. This allows the user's browser to directly retrieve images, style sheets, and similar data referenced by a web page while ensuring that viewing a new page is done through our system. Our current implementation also strips scripting from HTML documents, which has the side effect of making some pages unviewable but helps guarantee that the user is still in our system.

The time required to wrap documents varies with document contents, so wrapping a document each time it is accessed requires between 1 and 90 seconds. This delay is unacceptable to users that expect to quickly view a page; therefore, each time a document is viewed the wrapped version is cached, or stored for future display. Reloading cached documents takes approximately zero to five seconds. To remove the initial wrapping delay from users, we have added a helper program that visits documents in the OSU DLF and caches them before users use the system. This program can be re-run during off-peak usage periods in order to update the cached versions of the pages.

Voting Integrity

While receiving votes is essential to the success of the OSU DLF, a problem arises when users place *malicious votes*. These are votes that are designed to mislead the recommendation algorithm and generate incorrect recommendations for other users. For example, if a user constantly rated documents highly, low, or randomly, then the quality of documents cannot accurately be compared.

To address this issue, we have implemented a coarsely grained notion of *trustability*. Each user has a level of trust assigned by the system based upon voting habits. These levels are *trusted*, *neutral*, or *malicious*. Trusted users are those users whose votes tend to agree with *official votes*, or votes placed by administrators when reviewing questions, and malicious users are those that tend to disagree with the official votes (Figure 5). Neutral votes are those that do not clearly agree or disagree with official votes. We hypothesize that constant high votes or low votes will disagree with the official votes, while random votes will at worst be considered neutral votes.

OREGON STATE UNIVERSITY LIBRARIES [Home](#) [Edit Profile](#) [About](#) [Log Out](#)
 SYSTEM FOR ELECTRONIC RECOMMENDATION FILTERING [Admin Main](#) [Return to the LDL](#)

User Administration

Question Administration
[Review Questions](#) | [Seed Database](#) | [Answer Questions](#)

Review Questions

This page allows you to review recently asked questions and the documents that users selected as containing the answer to these questions. To examine these items in detail,

- Click on a question to ask that question through the system.
- Click on a document to go directly to that document. This allows you to place an "official" vote.
- Click on a URL to open the document in a new window *outside* of the system.
- Question/document pairs may be sorted by date or question by clicking on either **Date** or **Question** in the table header.

A total of 112 unique user-visited documents were found.

<u>Date</u>	<u>Question</u>	Page 1 2 3 4 5 6 7 8 9 10 11 12 Next >>	<u>Document</u>	<u>URL</u>	<u>Your Vote</u>	<u>Delete Question</u>
2003-08-21	What is the answer to Life, The Universe, and Everything?		Writing the Biography of a Living Sci...	http://osulibrary.oregonstate.edu/spe...	-	Delete
Vote Avg: 0		History: None				
2003-08-20	where can I find cultural information?		Oregon State University Libraries Hum...	http://osulibrary.oregonstate.edu/res...	-	Delete
Vote Avg: 0		History: None				
2003-08-20	where can I find cultural information?		AIHM 366: Cross Cultural Aspects of t...	http://osulibrary.oregonstate.edu/ins...	-	Delete

Figure 5: The question review screen. Clicking a question allows an administrator to run the question in the system. If an administrator clicks on a document, then the administrator can place an official vote for that question-document pair. This provides a baseline to which other votes can be compared to determine if a user is trustable or not. Inappropriate questions may be deleted by clicking on the delete link to the right.

When recommending documents, the recommender places more weight on votes that are from trusted users and ignores malicious user votes. Anonymous user votes are always neutral; as such, registered users with a history of good votes will have their question-document pairs recommended more than anonymous users.

ADMINISTRATION

The current implementation of the OSU DLF requires some administration in order to improve recommendation results. Additionally, it supports the viewing of statistics and is configurable without modifying the source code.

System Maintenance

As mentioned in Voting Integrity, official votes are used to establish a baseline to which all other votes can be compared in order to determine a user's intentions. Placing these official votes is thus the most maintenance-intensive requirement of using the OSU DLF.

Administrators or other users with special permission use the interface shown in Figure 5 to view question-document pairs. These questions may be filtered according to user information (school affiliation, department, and subject of interest). By clicking on a displayed document, administrators can review the page and vote based on how well they believe the question is answered. Administrators also have the option of receiving a daily email with lists of new questions that have been asked by registered users sharing the same subject of interest.

Information about the system is also provided through statistics (Figure 6). The statistics view shows frequently asked questions and viewed documents in addition to highest and lowest voted documents and user demographic information. The most common keywords are also available.

OREGON STATE UNIVERSITY LIBRARIES
SYSTEM FOR ELECTRONIC RECOMMENDATION FILTERING

[Home](#) [Edit Profile](#) [About](#) [Log Out](#)
[Admin Main](#) [Return to the LDL](#)

User Administration

[Site Statistics](#) | [System Statistics](#) | [Experiment Statistics](#)

System Statistics

Contents:

- [Demographics](#)
- [Most Visited Document\(s\)](#)
- [Highest and Lowest Voted Documents](#)
- [Most Frequently Asked Questions](#)
- [Highest and Lowest Voted Questions](#)
- [Most Frequent Words](#)
- [View All Documents](#)

Demographics

There are 25 registered users for the system.

Registered User Affiliations

User Affiliation	Number of Users
Faculty	1

Figure 6: The main statistics page. Available system-wide statistics are demographics, most visited documents, highest and lowest voted documents, most frequently asked questions, highest and lowest voted questions, most frequent words. Additionally, one may view all documents in the system.

Taken together, the statistics and question review allow administrators to see what questions are being asked frequently and to determine if a suitable page answers the questions. If not, a new document can be created and added to the system.

Look and Feel

The OSU DLF relies upon cascading style sheets [3] to handle the user interface look and feel. This allows administrators to change the colors, borders, and fonts without modifying the underlying program. Additionally, registered users are allowed to customize their look and feel by selecting a style sheet to use (Figure 7). Administrators wishing to make more comprehensive changes to the user interface can write new pages in PHP and interface with the existing core components.

Edit Your Profile

You may change your password here. If you do not want to change your password, leave the password fields as they are.

Required fields are marked with a (*).

Email address (this will be your LDL user name):

*

First name:

Last name:

Password:

*

Confirm Password:

*

Affiliation:

Main Subject Area:

Department:

Stylesheet:

Figure 7: The edit profile page. After registering, users can change the colors that are used by selecting a different style sheet. As the system uses style sheets more and more, this could lead to users being able to select what displays they want to see on the main page, the location of the voting bar, and more.

FUTURE RESEARCH

The OSU DLF is still in active development and is slated to be used in a variety of experiments in the future. The general design plans focus on improving the user experience and the statistical reporting. We also hope to perform experiments to test these changes in addition to testing recommendation algorithms.

Design plans

We plan to focus on three main improvements to the design: ease of voting, document accessibility, and code separation. Our current voting mechanism requires users to read a complete sentence and then answer the question. This string of text is located in close proximity to other lines of text possibly discouraging the user from reading it (Figure 8). To make the voting mechanism more visible, we are planning on implementing a graphical star-based rating system, similar to the system used in movies.



Figure 8: A close-up view of the header. With so much text, locating the voting menu at the bottom could be difficult for first-time users.

Our current document wrapper strips JavaScript from pages in addition to altering links on them. This causes some pages to behave incorrectly. Additionally, we do not wrap forms properly, so users may have difficulties submitting information to databases and similar sites. We plan to research alternative means of wrapping and tracking user information.

Finally, the OSU DLF is currently monolithic, meaning that administrators are restricted to using PHP to modify the user interface portions of the software. We plan to separate the software into a library with an application program interface and a user interface portion. This will allow administrators to further customize the software for their unique needs while still using the core functionality of the system.

Experimental plans

Our immediate experimental goal is to test the system in an uncontrolled environment. We have had encouraging results in unpublished controlled environments, but we have yet to test the system in an environment where users receive no extra training or questions to answer. We hope that this experiment will further correlate implicit voting with the parameters we established, that our recommendation algorithm is reasonably successful, and that the system is stable enough for “real world” use.

We also intend to run experiments to confirm that our user interface assumptions are valid. Namely, this includes that users can be trained to revise questions in the single row box and that a graphical voting mechanism will correlate with more explicit votes. Validating the research done by Belkin, et. al., and showing that users do enter complete questions is also an important experiment that we intend to perform.

REFERENCES

- 1.MovieLens. <http://www.movielens.org>
- 2.Belkin, N. J., et. al. (2003), “Query Length in Interactive Information Retrieval”. *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 205-212.
- 3.Cascading Style Sheets Home Page. <http://www.w3.org/Style/CSS/>

ACKNOWLEDGEMENTS

Kevin Harris was supported by the NSF Research Experience for Undergraduates Program and funded through National Science Foundation grant NSF-EEC-02-44205. The author would like to thank Seikyung Jung for providing direction and valuable testing for the project, Brandon Corry, Reyn Nakamoto, Kristina Pence, and Michael Tichenor for their assistance in developing the framework, and Leanne Lai for administering the computer network with a smile in spite of our constant demands.

BIOGRAPHY

Kevin Harris is a senior majoring in Computer Science at Oregon State University in Corvallis, Oregon. He plans on graduating in June of 2004.