# Collaborative Filtering for Digital Libraries

Jon Herlocker[*], Seikyung Jung[†], Janet Webster[§]

[*]Department of Computer Science

[†]Northwest Alliance for Computational Science and Engineering

[§]Oregon State University Libraries

102 Dearborn Hall

Oregon State University

Corvallis, Oregon 97370

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

## ABSTRACT

Can collaborative filtering be successfully applied to digital libraries in a manner that improves the effectiveness of the library? Collaborative filtering systems remove the limitation of traditional content-based search interfaces by using individuals to evaluate and recommend information. We introduce an approach where a digital library user specifies their need in the form of a question, and is provided with recommendations of documents based on ratings by other users with similar questions. Using a testbed of the Tsunami Digital Library, we found evidence that suggests that collaborative filtering may decrease the number of search queries while improving users' overall perception of the system. We discuss the challenges of designing a collaborative filtering system for digital libraries and then present our preliminary experimental results.

## Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; H.3.7 [**Information Storage and Retrieval**]: Digital Libraries.

## General Terms

Algorithms, Experimentation, Human Factors.

## Keywords

Collaborative filtering, digital libraries, user studies, tsunamis, natural hazards.

## 1. INTRODUCTION

Any digital library of substantial size relies on software-based search engines to help the users of the library locate documents related to their information need. These traditional search engines use content analysis to evaluate the relevance of documents with respect to the users' needs. Text search engines compare the keywords in a query to the keywords of documents in a database. Image search engines may compare the color histogram of a query image to the color histograms of images in a database.

These *content-based* search and retrieval techniques are inherently limited by the capability of computers to evaluate the relevance of content with respect to a user's stated need. For example, while humans can evaluate documents based on abstract concepts, we do not expect computers to reasonably recognize the aesthetics of artwork any time soon.

*Collaborative filtering* (CF) systems remove this limitation of content-based systems by placing the task of relevance evaluation in the hands of humans. Collaborative filtering can be described clearly using an analogy of a daily newspaper. A newspaper contains too many articles for an average individual to read completely every day, so individuals must decide which articles they will read. Each individual has specific tastes and information needs, and wants to be sure to read the articles that match his or her needs without wasting time reading articles that do not. Consider that by the time we (an average reader) pick up the newspaper, hundreds or thousands of people have already read articles from that paper. Imagine if we could instantaneously survey all of those previous readers, determine which of them have similar information needs to ours, and then ask them what articles were good and what articles were bad. We would then know exactly which articles we needed to read, and could avoid the bad ones. This is the essence of what CF does; by tracking and recording information usage by individuals, others can benefit from their experience and recommendations.

The computation in a CF system revolves around analysis of the ratings provided by human evaluators. Each participant in a CF community provides ratings for content items – each rating specifies how much that item meets the individual's information need. These ratings are used for two purposes: as examples of the individual's information need, and as recommendations for other users with similar information needs. No analysis is performed on the content itself, so CF techniques are universally applicable, regardless of content domain. The evaluation of content is performed by humans resulting in more complex analysis, including measuring the quality of a document. As a result, CF systems have the potential to dramatically outperform existing systems in ranking and retrieving items relevant to a user's need.

Collaborative filtering has been a mainstream idea since 1997, when a special issue of the Communications of the ACM was published on "Recommender Systems" – an alias that is often used for CF [16]. Since that time, a considerable amount of excitement has grown around the idea of CF. Several companies

were founded to explore the idea (including FireFly, NetPerceptions, LikeMinds, TripleHop), and some continue to pursue its potential (including NetPerceptions, TripleHop). However, the most prominent proponent of CF is Amazon.com, which continues to stir everyone's imagination by pushing the limits of CF in a high-profile commercial web-site.

## 2. COLLABORATIVE FILTERING IN DIGITAL LIBRARIES

Traditional digital libraries "learn" when new documents are added to the library or when a new interface is applied. They do not evolve without the input of their steward. If we can successfully apply CF to digital libraries, we can build a digital library that learns every time a user interacts with the library. It is our goal to extend CF technologies to demonstrate that digital libraries can learn with each interaction.

We cannot simply apply existing CF technology. Current CF techniques assume that each user's tastes or information need remain mostly consistent over time. This assumption works well within entertainment domains, where an individual's taste in movies or music changes very slowly over time. However, it seems false in the realm of digital libraries; a user's information need may be entirely different every time they initiate an interaction with the digital library.

This paper presents initial results of our research to answer the question: **Can collaborative filtering be successfully applied to digital libraries in a manner that improves the effectiveness of the library?**

## 3. TESTBED

Oregon State University has received a grant to expand its tsunami wave tank facility into a nationally shared resource under the National Science Foundation Network for Earthquake Engineering Simulation (NEES) program. To complement this expansion, the Northwest Alliance for Computational Science and Engineering, an interdisciplinary research group, and the Oregon State Libraries are developing a Tsunami Digital Library (**TDL**) that provides a portal to high quality tsunami resources available on the Internet and locally at Oregon State University. We are building two separate user interfaces for the TDL, one targeted at tsunami research scientists and the other targeted at high school students. We are using a prototype of the TDL to evaluate the practicality and effectiveness of adapting CF to digital libraries.

## 4. CHALLENGES: COLLABORATIVE FILTERING FOR DIGITAL LIBRARIES

There are three primary challenges to adapting CF to web-based digital libraries.

1. **Predicting appropriate documents when the users' interest needs may change frequently and completely**. Consumers who have received Amazon.com's recommendations have experienced the first challenge when they use Amazon to purchase a gift for another person. As a result of purchasing a baby book, they may receive recommendations for baby paraphernalia for months. We need to provide recommendations that are appropriate to the user's immediate need, and only use past history that is relevant to the current need.

2. **Identifying when a document is accessed, and for what information need.** Identifying every time a document is accessed within a digital library is a challenge in a web-based environment, where we cannot continuously modify the browser software. The challenge is very apparent with the TDL because most of the content consists of web pages available at locations outside of Oregon State University's domain. When a user clicks on a hyperlink, the original site loses track of the user. Once usage can be tracked adequately, we must record the context of the use. We need to know what information need was being pursued when that document was accessed.

3. **Collecting ratings from the users**. Collaborative filtering relies on collecting ratings from users on the relevance and quality of documents. A mechanism must be provided to collect those ratings. Furthermore, since users may be unwilling to take the time to provide enough ratings explicitly, methods for observing implicit expressions of ratings must be developed. We need to know what information satisfied the user's needs and what may satisfy future users.

## 5. THE TSUNAMI DIGITAL LIBRARY PROTOTYPE USER INTERFACE

Our prototype user interface is designed to meet the challenges described above. Figure 1 shows the primary screen of the TDL. The primary goal of this screen is to collect a coherent question from the user. While a full sentence question is ideal, the system will still function adequately if the user provides a traditional short keyword query. When the user clicks the search button, two separate searches happen. The question provided by the user is compared to all questions that have been asked previously and identifies similar ones, and it is also sent as a keyword query to a traditional full-text search engine. The results are displayed in a split screen fashion that is shown in Figure 2. The top section displays the recommendations for documents that have been voted highly by previous users who asked similar questions. The bottom section shows results and summaries from the traditional full-text search engine. Note that the user's question is displayed prominently and statically near the top of the screen. The user may perform small keyword changes to the query in the query form box. This allows a single session to live across query reformulations by the user. If the user wishes to ask a question that is entirely different from the current one, he or she can click on the "New Question" button at the top. When the user clicks on a recommended document or a document returned by the search engine, the selected page is displayed inside a frame. An illustration of this is shown in Figure 3.

Every hyperlink request for a document is routed through the TDL web server, which then parses the HTML and rewrites all URLs, pointing them at the TDL instead of their original source. This ensures that no matter how deep a user clicks on hyperlinks,

Tsunami Digital Library

Home | Edit Profile | About | Log Out | New Task

OREGON STATE UNIVERSITY Libraries

**Search the Tsunami Digital Library**

**Enter your question :**

[Ask >>]

**Your recent questions:**

As you search for information using the Tsunami Digital Library, your most recent searches will appear in this area so you can easily reference them for later use.

**Examples:**

- Where can I find basic information on tsunamis?
- What can we do to prepare for a tsunami and when one strikes?
- Where can I find photographs of tsunamis?
- What is the current tool used to detect tsunamis?
- Are tsunamis measured on a scale similar to those of tornadoes and hurricanes?

Administer a Site | Site Statistics

**Browse some of our contributing sites.**

**U.S. Geological Survey**
This 18-page booklet was originally published in paper by the U.S. Geological Survey and is now available here as a PDF document. It contains lessons on how to survive a tsunami based on accounts from people who survived the tsunami generated by the largest earthquake ever measured-the magnitude 9.5 earthquake in Chile on May 22, 1960. It also contains an excellent description of what tsunamis are.

**Figure 1. The primary search screen of the Tsunami Digital Library**

Tsunami Digital Library

Home | Edit Profile | About | Log Out | New Task

OREGON STATE UNIVERSITY Libraries

Question: **How is tsunami height recorded?**     New Question

**Revise Question :** How is tsunami wave height measured     [Search]

**Users with similar questions found these pages helpful**     View by doc

Question | Page - User Rating

What modifications occur to the tsunami wave height as it travels from its source region? | [x027_iv_hm02] - 🝙🝙🝙🝙

Pacific Tsunami Museum FAQs - 🝙🝙🝙🝙
Center for Tsunami Inundation Mapping Efforts - 🝙🝙🝙

What is a tsunami? | Pacific Tsunami Museum FAQs - 🝙🝙🝙🝙
Pacific Tsunami Museum Archives - 🝙🝙🝙🝙
The Tsunami - 🝙🝙🝙

How is a tsunami generated? | How do earthquakes generate tsunamis? - 🝙🝙🝙🝙
How do landslides, volcanic eruptions, and cosmic collisions generate tsunamis? - 🝙
Pacific Tsunami Museum FAQs - 🝙🝙🝙🝙

How do undersea structures and coastline configuration affect the intensity of a tsunami wave? | Pacific Tsunami Museum FAQs - 🝙🝙🝙

**Results 1-10 of 448 for "How is tsunami wave height measured"**

Page 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21  Next >>

**Pacific Tsunami Museum FAQ's**
http://www.tsunami.org/faq.htm
... bytes) Home page What's New? What we're about FAQ's Programs Archives Member Store Links FREQUE NTLY ASKE D QU ESTIONS  The major Pacific wide **tsunami** occurred in 1964. The rare occurrence of a Pacific wide **tsunami** in recent times makes them increasingly important to understand as more and more ...

**[x026_iv_fg03]**
http://www.pmel.noaa.gov/tsunami/Faq/x026_iv_fg03
3. Are tsunamis **measured** on a scale similar to those of tornadoes and hurricanes? There is a **tsunami** intensity scale, although it is not used much anymore. Nowadays, tsunamis are usually described by their **heights** at the shore and the maximum runup of the **tsunami waves** on the land.

**Figure 2. The search results and recommendations screen of the Tsunami Digital Library. The top half of the screen shows recommendations based on previously asked similar questions. The bottom half shows results from a traditional full-text search engine.**
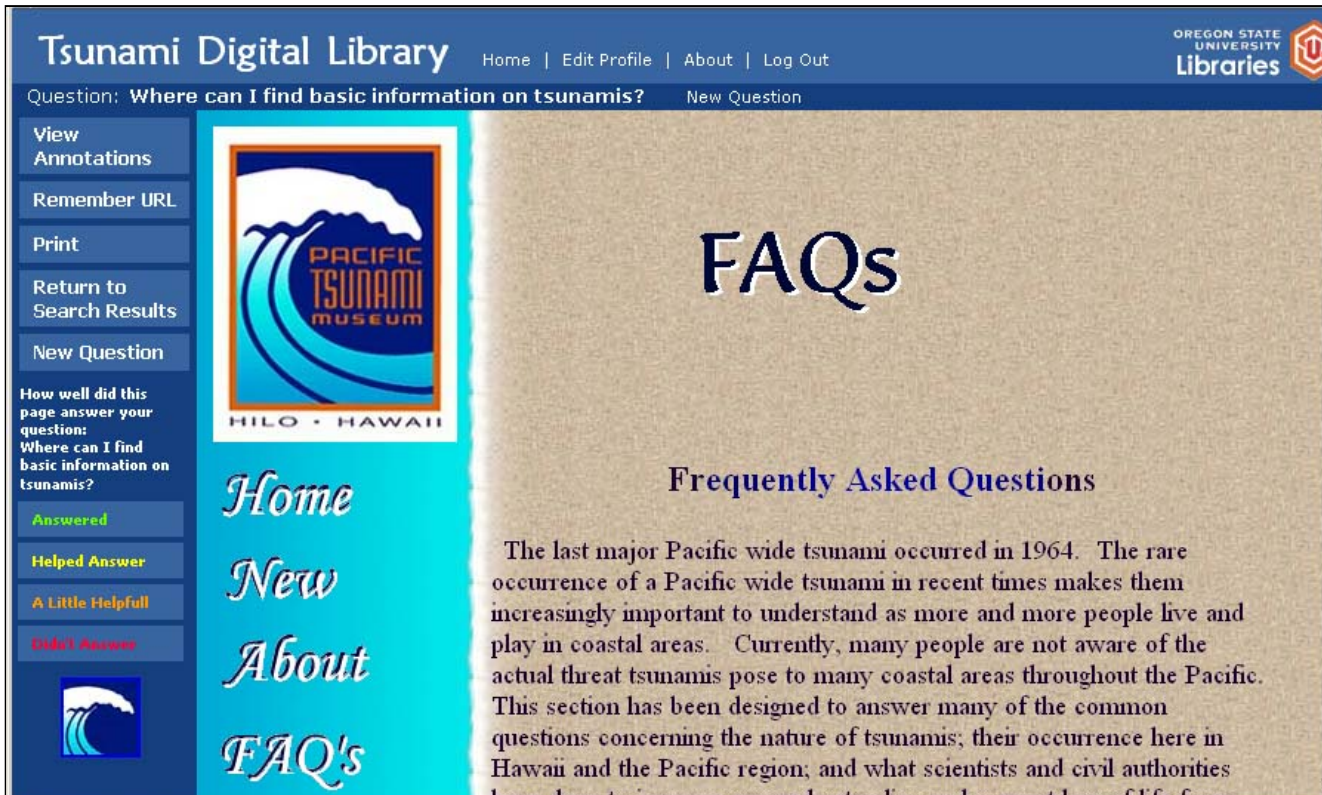
**Figure 3. The interface for viewing web pages within the Tsunami Digital Library. The top banner and the left toolbar are always present while browsing, regardless of what site the user is visiting. The left toolbar allows users to rate the current document.**

the TDL maintains control of the user interface and is able to record every event.

The left pane of the screen in Figure 3 provides buttons so the user can rate how well a document answers the current question

## 6. ADDRESSING THE CHALLENGES

### 6.1 Challenge – Dealing with Changing Information Needs

To adapt the idea of CF to the digital library, we adjusted the traditional CF model to create a model more appropriate for information retrieval tasks. In this new model, we assume that every user approaches the digital library with a question describing their information need. We assume that we have some mechanism for collecting that question. We refer to the user's interaction with the system with respect to a single question as a session. A rating for a document provided during a session is considered to be a rating of the document's relevance and quality with respect to the question that initiated the session. For example, if a user's question is "How do you measure tsunami run-up?" and the user rates page A highly, then we have evidence that page A is a good source of information on measuring tsunami run-up. The traditional CF model does not accommodate changing needs. Figure 4 illustrates the differences between the traditional collaborative filtering model and the new model we propose.

The ratings provide the information necessary to predict how well documents will answer the posed questions. Now every time a user asks a new question, we compare that question to the questions that have been previously asked. If a previously-asked question is found to be similar, the system can recommend the documents that were relevant to that previously asked question.

Obviously, we need to record the users' questions for the system to work. In our initial trials, we used a traditional keyword query as the "question" defining the information need. Taking this approach required little change to the existing user interface. However, during internal testing, we quickly determined that this approach did not work well. Keyword queries that users supply were not descriptive enough to allow another individual to identify the driving question or information need. As a result, when presented with lists of previously asked queries, test users were unable to identify if previous queries were related to their current question. The second problem we encountered trying to use keyword queries was that users frequently reformulated their queries. Users would submit a keyword query, not find any interesting search results, and then reformulate their query without rating any documents. As a result, many of the keyword queries entered had no corresponding ratings. If we are to assume that users of similar backgrounds with a similar information need are likely to issue the same test query, then what we really want are recommendations to the initial query.

We solved both of these issues by separating the concept of the question from the query. Users begin with a question that is human readable which becomes the "description" of the

information need of the session as well as the first keyword query. Users can then issue new keyword queries that are reformulations

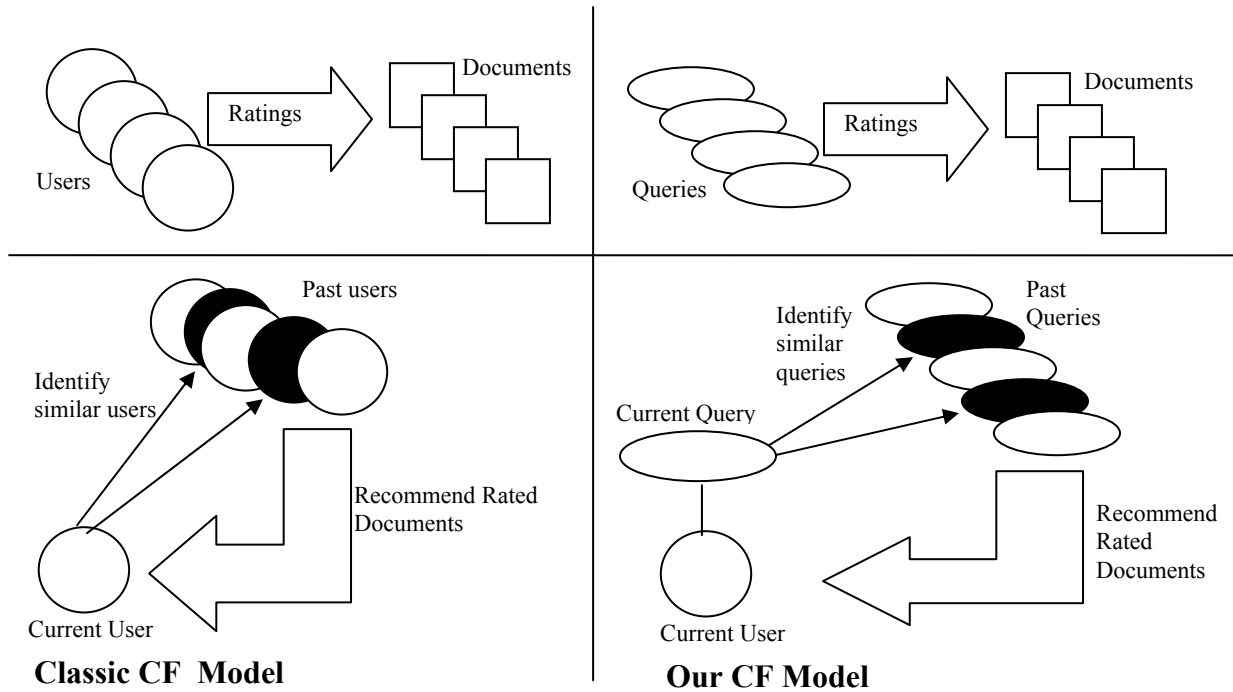without changing the context of the original question.



**Figure 4. A graphical comparison of the traditional collaborative filtering (CF) and the our proposed CF model that we empirically evaluate in this paper. In the classic CF model, users rate documents, and then those ratings are recommended to other users with similar interests. In our new model, a user's ratings establish a link between queries and documents rather than linking users with users.**

Of course, this leaves the challenge of getting users to enter full sentences as questions, when they have been trained by current search engines to only enter keyword queries. We believe that this challenge is tractable, but do not address it in this paper.

## 6.2 Challenge – Identifying Access and Question Context

As we have described in the Section 5, we designed a URL rewriting mechanism whereby our server makes all requests for all URLs on behalf of the client browser, parses the HTML and rewrites the URLs before returning the document to the client. This ensures that all requests for new documents will continue to pass through the server. It also allows the server to add the rating interface to every single page.

## 6.3 Challenge – Collecting Ratings from the Users

In earlier experiences with CF systems [6,15], we discovered that every population of users has a subgroup that willingly provides a substantial numbers of ratings. In previous systems, this prolific subgroup removed the need for all users to participate significantly in rating while maintaining the effectiveness of the system. We expect that the same phenomenon will occur with digital libraries. However, we expect that there will be many more questions in digital library systems given the variety of

information contained and the range of users. Consequently, the subgroup of prolific raters may not provide enough ratings.

We are exploring methods for inferring a user's perception of the relevance and quality of a document. The inferred recommendations will better inform the system for all users. These methods include browsing patterns (e.g. documents accessed last are more likely to have answered the question) and significant events (e.g. printing or emailing a document indicates potential relevance). However, we do not explore those possibilities in this paper.

## 7. EXPERIMENT

Keeping these three challenges in mind, we return to our key research question: can collaborative filtering be successfully applied to digital libraries in a manner that improves the effectiveness of the library? To begin to answer this question, we examined the CF-enhanced TDL empirically in a controlled context. Undergraduate students were tasked with finding the answers to questions about tsunamis, using only the TDL as a reference tool.

In our experiment we brought 53 undergraduate students, mostly computer science majors, into a computer lab in three randomly assigned groups of 15, 20, and 18. Each student was provided with a list of ten questions related to tsunamis. To avoid confusion between these ten questions and the questions that

subjects submitted to the search interface, we refer to the ten tsunami questions as the **tasks**. The ten tasks were carefully chosen from existing high school tsunami curricula [1,8]. We examined all the tasks found in the curricula, discarding tasks whose answers could not be found within the database of the TDL, and ranking the remaining tasks by estimated difficulty. We then selected a roughly equal number of easy, medium, and hard tasks for a total of ten tasks. The tasks were placed on the questionnaire in order of increasing difficulty.

The first group of students received only a traditional content-based search engine interface to the documents indexed by the TDL. We used htdig [2], a publicly available html search engine.

The second group of students received both the htdig search engine interface and the CF recommender system interface. However, when the session started, no data on previously asked questions were stored in the database, so early questions resulted in no or few recommendations. As the session continued, results from the faster students became available as recommendations for students who were not quite as fast.

Similar to Group Two, the third group of students received both the htdig search engine interface and the recommender system. However, unlike Group Two, the third group began with all the training data collected from the previous two groups. Thus subjects received recommendations for all of their questions.

The questionnaire given asked the participants, for each of the ten tsunami tasks, to rank on a scale of one to five how much they agreed with the following statement: "I was able to find correct information." They were also asked the following true/false question: "I was able to find the partial/whole answer by using the recommended page."

Once they turned in the session questionnaire, they were given a post-session survey with 12 questions that sought the subject's opinion on issues such as the user interface and system performance.

In each experiment, the system recorded every click, recommendation, search result, and query with timestamps.

# 8. RESULTS
We collected a large quantity of data as a result of this experiment, and report here some of the most significant results from initial analysis. Primary emphasis on the initial analysis is to get a general impression on whether CF adds value for users of the TDL.

## 8.1 CF Subjects Were More Likely To Find Answers
The first significant finding indicated that the subjects who had the CF recommendations felt that they were more likely to find answers to their tasks than those that did not. A summary of the responses to the statement "I was able to find correct information" is shown in Table 1, ordered by experimental group.

**Table 1. "I was able to find correct information." (On a scale of one to five, with one indicating "strongly disagree" and five indicating "strongly agree.")**

|  | No CF | CF Recommendations | |
|---|---|---|---|
|  | (Group 1) | No training data (Group 2) | Training data (Group 3) |
| N | 150 | 200 | 180 |
| Mean | 3.21 | 4.06 | 4.11 |
| Std.Dev | 1.61 | 1.19 | 1.16 |

From Table 1, we see that the participants from the groups receiving the CF recommendations strongly agreed that they were able to find correct responses to the tsunami tasks. An analysis of variance (ANOVA, $p < 0.001$) indicates that the mean response of Groups 2 and 3 is significantly higher than the mean response of Group 1. Given that the only difference between Group 1 and Groups 2 and 3 is the CF recommendation interface, this gives strong evidence that, in the view of the participants, the recommender system improved the user's ability to find answers to the given tasks.

## 8.2 CF Subjects Used the Recommendations
To cross-check that participants of Groups 2 and 3 were actually using the recommendations, we examined responses to another true/false statement: "I was able to find the partial/whole answer by using the recommended pages."

**Table 2. "I was able to find the partial/whole answer by using the recommended pages."**

|  | CF Recommendations | |
|---|---|---|
|  | No training data (Group 2) | Training data (Group 3) |
| N | 200 | 180 |
| Mean | .859 | .875 |
| Std.Dev | .349 | .331 |

Table 2 shows that 86% of Group 2 and 88% of Group 3 felt that they were able to find answers to the tsunami tasks by using the recommendations. This reinforces the belief that the recommendations improve the participants' ability to locate answers to tasks. The fact that 12-14% of the participants did not successfully use the recommendations could be explained by several factors. In Group 2, this could be explained by the "first-raters," those participants who answered the tasks before anybody else and therefore did not have access to recommendations. The remaining difference in both Group 2 and Group 3 could be due to the fact that many people in all experimental groups did not have time to satisfactorily finish the final few questions in the time allotted.

## 8.3 CF Subjects Found Answers More Efficiently

We also found that subjects who received CF recommendations found their answer more efficiently. Figure 5 compares the average number of times that subjects reformulated their keyword queries (without changing the question) for each task while searching for an answer. Figure 6 shows the average number of new questions asked for each task. Figure 7 shows the average sum of the new questions issued and the keyword reformulations for each task.

Participants in Group 1, the group that did not receive the recommendations, needed more query reformulations on almost every task. According to an ANOVA analysis, the mean number of keyword revisions for group 1 is significantly higher than the mean number of keyword revisions for Groups 2 and 3.

Tasks 1 and 2 were exceptionally easy. For example, Task 1 was "Describe what a tsunami is." The pages with correct answers to Tasks 1 and 2 appeared at the top of the full-text search engine results for any reasonable wording of the question. Having the recommendations was not necessarily an advantage for those tasks. As the tasks become less trivial, the recommendations provided a clear advantage with respect to the number of query reformulations needed. We can see this in the gap between the Group 1 line and the lines for Groups 2 and 3 in Figures 5-7

In comparing Groups 2 and 3, we do not find any significant difference. Groups 2 and 3 both received recommendations, but Group 3 began with more training data. However, apparently the first raters of Group 2 were able to provide enough data for the following users of Group 2 to receive sufficient recommendations. Thus for this experiment, the quantity of ratings data in response to a specific task does not seem to have been a factor. This is most likely because we gave each user a piece of paper with the exact same wording of the question. We expect that many of the subjects entered in similar or identical wordings of the questions.

Looking at Figure 5, we can see variance in the number of keyword revisions among tasks. In particular, we can see that Tasks 3 and 6 had larger differences between Group 1 and Groups 2 and 3. This can be explained the difficulty of the tasks. Although we ordered the tasks by what we expected to be the difficulty of each task, free form comments on the post-experiment survey indicated that many people found Tasks 3 and 6 to be the most difficult. As a result, the recommendations had a much more significant affect on the efficiency of the subject in finding an answer to those tasks.
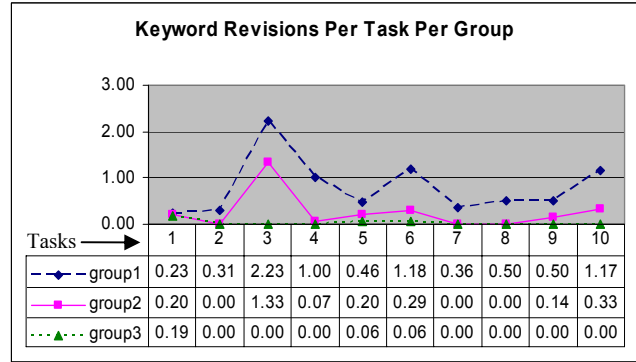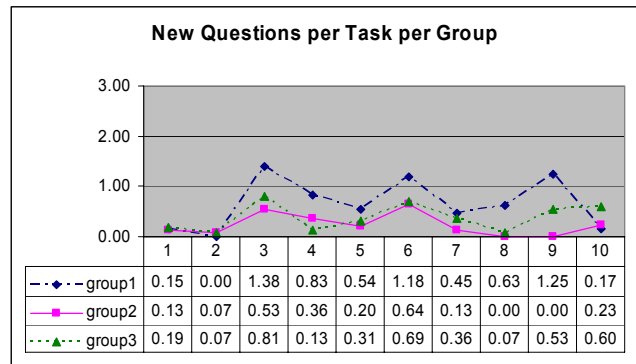


**Keyword Revisions Per Task Per Group**

| Tasks | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| group1 | 0.23 | 0.31 | 2.23 | 1.00 | 0.46 | 1.18 | 0.36 | 0.50 | 0.50 | 1.17 |
| group2 | 0.20 | 0.00 | 1.33 | 0.07 | 0.20 | 0.29 | 0.00 | 0.00 | 0.14 | 0.33 |
| group3 | 0.19 | 0.00 | 0.00 | 0.00 | 0.06 | 0.06 | 0.00 | 0.00 | 0.00 | 0.00 |

**Figure 5. Keyword Revisions per Task per Group**



**New Questions per Task per Group**

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| group1 | 0.15 | 0.00 | 1.38 | 0.83 | 0.54 | 1.18 | 0.45 | 0.63 | 1.25 | 0.17 |
| group2 | 0.13 | 0.07 | 0.53 | 0.36 | 0.20 | 0.64 | 0.13 | 0.00 | 0.00 | 0.23 |
| group3 | 0.19 | 0.07 | 0.81 | 0.13 | 0.31 | 0.69 | 0.36 | 0.07 | 0.53 | 0.60 |

**Figure 6. New Questions per Task per Group**



**Total Query Revisions (New Questions Issued + Keyword Revisions) per Task per Group**

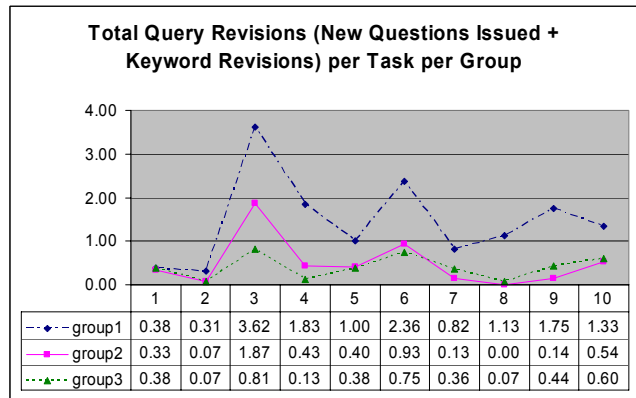| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| group1 | 0.38 | 0.31 | 3.62 | 1.83 | 1.00 | 2.36 | 0.82 | 1.13 | 1.75 | 1.33 |
| group2 | 0.33 | 0.07 | 1.87 | 0.43 | 0.40 | 0.93 | 0.13 | 0.00 | 0.14 | 0.54 |
| group3 | 0.38 | 0.07 | 0.81 | 0.13 | 0.38 | 0.75 | 0.36 | 0.07 | 0.44 | 0.60 |

**Figure 7. Query Revisions per Task per Group**

## 8.4 No Difference in the Number of Documents Viewed

We expected to find that the groups receiving collaborative filtering recommendations would need on average fewer page views to reach an answer. However, we found that these data were more inconclusive, as shown in Figure 9**Error! Reference source not found.**.
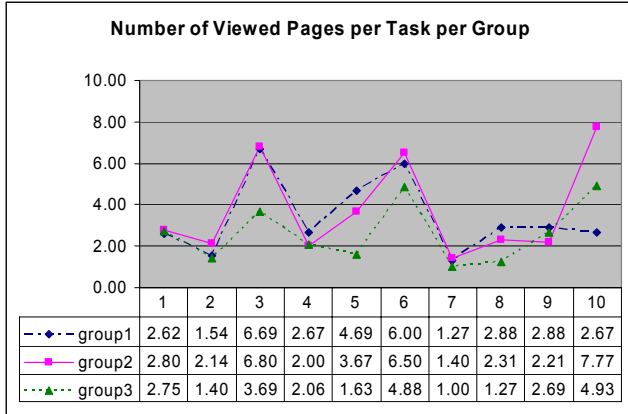
**Number of Viewed Pages per Task per Group**

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| group1 | 2.62 | 1.54 | 6.69 | 2.67 | 4.69 | 6.00 | 1.27 | 2.88 | 2.88 | 2.67 |
| group2 | 2.80 | 2.14 | 6.80 | 2.00 | 3.67 | 6.50 | 1.40 | 2.31 | 2.21 | 7.77 |
| group3 | 2.75 | 1.40 | 3.69 | 2.06 | 1.63 | 4.88 | 1.00 | 1.27 | 2.69 | 4.93 |

**Figure 8. Pages Viewed per Task per Group**

Figure 8 suggests that Group 3 might have found answers with fewer clicks, but the data are noisy and inconsistent with the other data that show Groups 2 and 3 performing at similar levels.

One possible explanation for the lack of a significant difference between the groups with CF recommendations and the group without CF recommendations is that subjects used the data displayed on the search results screen (such as the keyword-in-context excerpts) as a primary judge for potential relevance, rather than clicking on each search result to examine the document closely. For the recommended questions and answers, it is easy to judge relevance, given a full sentence question. For the full-text search engine results, the keyword-in-context gives good clues as to the likelihood that a given result will contain the necessary information for the task.

## 8.5 CF Subjects' Perception of Quality Increases

We have shown that subjects who received the CF recommendations on average were more likely to feel that they found answers to their tasks and needed less query reformulations. We also found that subjects who received CF recommendations had a significantly more positive perception of the quality of the site, including factors that are entirely orthogonal to the recommendations.

After participants completed the tasks, they were asked to fill out a post-experiment survey. The survey sought to capture their overall impressions of the Tsunami Digital Library user interface. A total of 12 questions were asked across a variety of user interface and system issues.

Four of the questions on the post-session survey were designed to assess the usability of the prototype interface. These questions are shown in Table 3. When we initially added the questions to the survey, we did not expect to observe significant differences between the group that received CF recommendations and the group that did not. Besides the recommendations, every other factor of the user interface was exactly the same. However, the groups receiving the CF recommendations agreed more strongly with statements indicating that the user interface was good. (See Table 3 and Figure 9.)

**Table 3. Post-experiment questions related to the user interface.**

| 1 | I found the Tsunami Digital Library used vocabulary that I could understand |
|---|---|
| 2 | I found the buttons, menu items, and other controls easy to navigate |
| 3 | I found the voting menu understandable and easy to use |
| 4 | It was easy to find the recommendations and the search results |

The first question on the post-experiment survey asked if the TDL used vocabulary that they could understand. Our goal was to identify if the text messages used in our system were understandable or cryptic. However, subjects that received the CF recommendations felt more strongly that the vocabulary used was understandable than the subjects who did not receive the recommender system.

The second question of the post-experiment survey asked if the buttons, menu items, and other controls were easy to navigate. Again, we found that the subjects receiving recommendations felt that their user interface was better than the subjects who did not receive recommendations.

The topics of these questions – vocabulary and controls – are almost entirely orthogonal to the CF recommendations. Whether or not CF recommendations are present should not have any effect on the utility of the buttons and certainly not on the vocabulary of the messages in the interface. Instead, by including CF recommendations into the digital library interface, we increase the user's *perception* of the usability of all components of the system.

The remaining two user interface questions on the post-experiment survey were intended to measure the subject's response to the voting and search results interfaces. The mean response for these two questions was also higher for the groups that received CF recommendations.
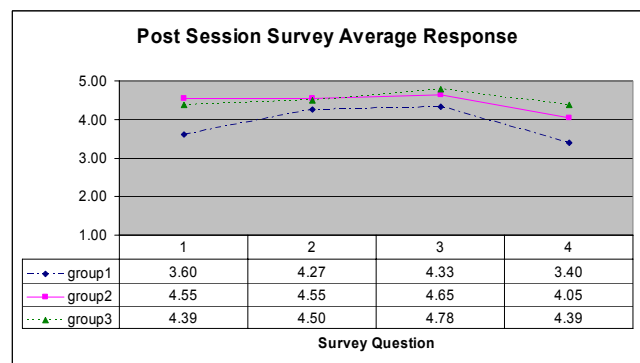
**Post Session Survey Average Response**

| | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| group1 | 3.60 | 4.27 | 4.33 | 3.40 |
| group2 | 4.55 | 4.55 | 4.65 | 4.05 |
| group3 | 4.39 | 4.50 | 4.78 | 4.39 |

**Survey Question**

**Figure 9. Post Session Survey Average Response**

## 9. SUMMARY & FUTURE WORK

Our empirical experiment with human subjects provides evidence that a CF-based recommender could significantly increase the effectiveness of a digital library search interface. In addition to decreasing the number of times that subjects had to reformulate queries, the CF recommender improved the subjects' perception of the quality of the entire user interface.

This study is only a first step – it demonstrates that a CF recommender is effective in a controlled experimental environment. Further research is necessary to determine if the user interface described can be as effective in a less controlled environment. We intend to focus our efforts initially on answering the following questions.

**Will a sufficient percentage of the population enter full questions to keep the CF recommendations valuable?** A portion of the effectiveness of the interface comes from the fact that users can view full, human readable sentences when determining if a recommendation is appropriate. We estimate that we would only need a small portion of the population of a website community to keep the effectiveness of the recommender system. We could design algorithms that would detect if a question was a "full" question, and give preference to recommendations that have "full" questions.

**How do we collect sufficient ratings to produce valuable recommendations?** We do not expect all users of a digital library to enter ratings for pages they visit, although experience from past implementations of collaborative filtering systems suggests that there is a subset of every community that is willing to rate freely. Because of the probable sparseness of the digital library ratings data (very few people will rate each question), the freely given explicit ratings are bound to not be sufficient. One approach to address this problem is to infer ratings (votes) by observing a user's navigation and actions. Certain activities initiated by a user, such as printing or emailing a document, indicate that it meets some need. We can also make inferences from the navigation pattern. A user terminates a session because they either found the information they are looking for or gave up. If many people with similar questions terminate on the same page, we have evidence that that page holds some information relevant to those questions.

Another approach is to "seed" the digital library with recommendations. For example, if high school students are a target audience for the TDL, we could work with teachers to collect recommendations prior to the students using the TDL. This approach could develop a dynamic digital library with great potential as a powerful teaching and learning resource. However, this approach does not fully address the challenge of changing information needs of a wide variety of users.

Yet another approach is to devise incentives for ratings. One of the most effective incentives in past collaborative filtering systems was to make the accuracy of the recommendations depend on the quantity of ratings that the user has provided in the past. Our current algorithm does not have this feature, but we may explore artificially adjusting the recommendation algorithm to have such a property.

Integrating the human evaluation of information into digital libraries may add another means for them to "learn" and expand. As our preliminary research shows, collaborative filtering has potential that should be further explored.

## 10. RELATED WORK
Articles by Konstan et al [12], Shardanand and Maes [18], and Hill et al [10] introduce the concept of collaborative filtering in more detail, including descriptions of algorithms. You can read about research collaborative filtering systems that have been built for movies [6], music [18], and jokes [7], among others. Examples of detailed analyses of collaborative filtering algorithms can be found in [5,9]. Previous articles on CF that have appeared in the ACM Digital Libraries conference or JCDL include [4,11,14,17]. The proceedings of the NSF/DELOS Workshop on Personalization and Recommender Systems in Digital Libraries [3] contain some articles that are relevant to the intersection of CF and digital libraries.

The AntWorld project at Rutgers [13] has the most similarity to our current work. AntWorld is 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 in time 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, the AntWorld has never been evaluated in an empirical user study.

## 11. ACKNOWLEDGEMENTS

## 12. REFERENCES

1. Servicio Hidrográfico y Oceanográfico de la Armada de Chile, Departamento de Oceanografía, Programa de Geofísica Marina, Intergovernmental Oceanographic Commission & International Tsunami Information Center 2002. Earthquakes and Tsunamis: High School Teacher's Guidebook. http://www.shoa.cl/oceano/itic/pdf-docs/hsteacher.pdf

2. ht://Dig web site. http://www.htdig.org/ . 2003.

3. Proceedings of the NSF/DELOS Workshop on Personalization and Recommender Systems in DigitalLibraries. http://www.ercim.org/publication/ws-proceedings/DelNoe02/index.html . 2003.

4. Alspector, J., Kolcz, A., Karunanithi, N., 1998. Comparing feature-based and clique-based user models for movie selection. *Proceedings of the third ACM Conference on Digital Libraries*.

5. Breese, J. S., Heckerman, D., Kadie, C., 1998. Empirical analysis of predictive algorithms for collaborative filtering. *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence (UAI-98)*. San Francisco, (pp. 43-52).

6. Dahlen, B. J., Konstan, J. A., Herlocker, J. L., Good, N., Borchers, A., Riedl, J., 1998. Jump-starting movielens: User benefits of starting a collaborative filtering system with "dead data". University of Minnesota TR 98-017.

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

8. Goodrich, M., Atwill, T., 2000. Oregon Earthquake and Tsunami Curriculum: Grades Seven Through Twelve, Revised edition. National Tsunami Hazards Mitigation Program, Oregon Department of Geology and Mineral Industries.

9. Herlocker, J. L., Konstan, J. A., Riedl, J., 2002. Empirical Analysis of Design Choices in Neighborhood-based Collaborative Filtering Algorithms. *Information Retrieval* 5 287-310.

10. Hill, W., Stead, L., Rosenstein, M., Furnas, G. W., 1995. Recommending and Evaluating Choices in a Virtual Community of Use. *Proceedings of ACM CHI'95 Conference on Human Factors in Computing Systems*. (pp. 194-201).

11. Huang, Z., Chung, W., Ong, T.-H., Chen, H., 2002. A Graph-based Recommender System for Digital Library. *Proceedings of the Second ACM/IEEE-CS Joint Conference on Digital Libraries*. New York, NY, (pp. 65-73).

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

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

14. Mooney, R. J., Roy, L., 2000. Content-based book recommending using learning for text categorization. *Proceedings of the Fifth ACM Conference on Digital Libraries*. New York, NY, (pp. 195-204).

15. Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., Riedl, J., 1994. GroupLens: An open architecture for collaborative filtering of netnews. *Proceedings of the 1994 Conference on Computer Supported Collaborative Work*. New York, (pp. 175-186).

16. Resnick, P., Varian, H. R., 1997. Recommender Systems. *Communications of the ACM* 40 (3), 56-58.

17. Riggs, T., Wilensky, R., 2001. An Algorithm for Automated Rating of Reviewers. *Proceedings of the First ACM/IEEE-CS Joint Conference on Digital Libraries*. New York, NY, (pp. 381-387).

18. Shardanand, U., Maes, P., 1995. Social Information Filtering: Algorithms for Automating "Word of Mouth". *Proceedings of ACM CHI'95 Conference on Human Factors in Computing Systems*. New York, (pp. 210-217).