經營論集, 第 40 卷 1・2號 (2006年 6月)

# The Effectiveness of Collaborative Filtering-Based Recommendation Systems across Different Domains and Search Modes

Il Im*

Collaborative filtering (CF) is a personalization technology used by numerous e-commerce websites to generate recommendations for users based on others' evaluations. Although many studies have considered ways to refine CF algorithms, little is known about the effects of user and domain characteristics on the accuracy of CF systems. This study investigates the effects of two factors, domain and user search mode, on the accuracy of collaborative-filtering systems, using data collected from two different experiments — one conducted in a consumer-product domain and one in a knowledge domain. The results show that the search mode employed by users strongly influences the accuracy of recommendations. CF works better when users are looking for specific information than when they are browsing out of general interest. Accuracy drops significantly when data from different search modes are mixed. The results also show that CF is more accurate in knowledge domains than in consumer-product domains. The study implies that CF systems in either domain will provide more accurate recommendations if they identify and accommodate users' search modes.

## I. Introduction

Information overload is a serious problem for today's Internet. It is growing more and more difficult to find relevant information or suitable products [Hannabuss 2002; Ram 2001]. Electronic commerce companies searching for solutions to this problem face new opportunities as well as challenges. Personalization, one of the solutions they

*Professor, School of Business, Yonsei University

have explored, is popular enough with users that they are coming to expect and require it of Web sites they visit [Agarwal and Venkatesh 2002]. An important task that falls to Information Systems (IS) researchers is determining which personalization techniques are most effective [Straub and Watson 2001]. One approach that has gained prominence is the collaborative filtering-based recommendation system, a method that provides users with personalized recommendations based on their preferences or evaluations [Schafer et al. 1999]. Collaborative filtering (CF) is a relatively young, yet not an immature, technology [Konstan et al. 1997]. Indeed, it is a major Internet personalization method [Hirsh et al. 2000; Mulvenna et al. 2000; Perkowitz and Etzioni 2000; Yuan and Chang 2001] and one of the key technologies of electronic commerce, used by an ever-expanding group of online companies, including such industry leaders as Amazon (www.amazon.com) and CDNow (www.cdnow.com) [Collett 2002; Koufaris 2002; Schafer et al. 1999].

While many studies have investigated CF, the existing research has two major limitations. First, the existing research offers little by way of cross-domain comparisons. CF has been used mostly for consumer products like CDs, books, and movies [Wingfield 1998], and has only recently been applied to knowledge objects such as research papers [McNee et al. 2002]. One area that would likely benefit from CF is the electronic repository of information — for instance, digital libraries and Frequently Asked Question (FAQ) pages in corporate web sites. Williams [2002] argues that CF has the potential to improve information searches in digital libraries considerably. However, the task of recommending professional documents and articles may differ significantly from that of recommending consumer goods, in terms of how users' preferences are distributed and what their informational needs are. These differences may affect various aspects of CF systems, including the accuracy of the recommendations and the optimal configuration of the system. Managers of electronic commerce (EC) websites need to understand the effects of such differences when making key decisions about recommendation systems: decisions, for instance, about how to configure the system for better recommendations, what threshold to set (in

terms of user numbers) for the release of recommendations, or even whether or not to start a recommendation service at all. Thus a clear understanding of differences in CF across domains is essential for adapting CF systems to different circumstances.

Another limitation of existing CF research is the absence of studies of user-side factors. CF recommendations are based on users' evaluations. Users may evaluate items differently according to the intentions that have motivated their search, and the differences in their evaluations will likely affect the accuracy of the resulting recommendations. It is very probable, therefore, that user-side factors like search intention will affect the accuracy and usefulness of recommendations. It is also likely that if evaluations by users with different search intentions are mixed, the heterogeneity in evaluation patterns across different groups will degrade the accuracy of recommendations. Prior studies of CF systems have not adequately considered the role of user-side factors such as search intention.

To address these questions, the present study looks at the differences between CF recommendations made in two representative domains: one knowledge-object domain (research papers) and one consumer-product domain (movies). The study also examines user-side factors and their effects on the accuracy of CF systems. The primary goals of the study are to examine the effect of domain and user-side factors on the accuracy of CF systems, and to propose guidelines for designing better CF-based recommendation systems.

This paper is structured as follows. The next section explains CF systems in more detail and reviews relevant prior studies. In the third section, research issues related to domain and user-side factors are discussed and research hypotheses derived. The fourth section explains the design of the empirical study and the measurements used. Data analysis and test results are discussed in the fifth section. The paper concludes with a discussion of limitations and future research directions.

## II. Past Studies of CF

Pioneered by Goldberg et al. [1992], who applied the technology to information retrieval, "collaborative filtering" generates recommendations for users based on the evaluations of other users with similar profiles [Miller et al. 1997]. By using the ratings of an appropriate reference group rather than the average rating of all users, CF can take into account differences in taste and personal needs. A typical CF system for movies works as follows. First, a user rates a set of movies he or she has already seen. The collaborative filtering system then applies statistical techniques to identify a subset of other people who have given similar ratings to the same movies. The system uses the preferences of this subset to identify movies the user has not yet seen but would be likely to enjoy. CF algorithms are explained in more detail in section 4.

Most CF research has addressed the algorithms that generate recommendations. Studies have compared different algorithms [Breese et al. 1998; Shardanand and Maes 1995], sought variations in algorithms that would improve accuracy and security [Canny 2002; Herlocker et al. 1999; Sarwar et al. 2001] or scalability [Deshpande and Karypis 2004], and investigated combinations of CF with other methods [Ansari et al. 2000; Good et al. 1999; Melville et al. 2002]. Another stream of CF research has focused on the applications and uses of CF, investigating, for instance, the role of CF in e-mail messages [Goldberg et al. 1992], music [Shardanand and Maes 1995], movies [Good et al. 1999; Schafer et al. 2002], Usenet messages [Konstan et al. 1997; Miller et al. 1997], Internet resources [Terveen et al. 1997], TV programs [Podberezniak 1998], Web pages [Sarwar et al. 2001], research papers [McNee et al. 2002], supermarket products [Lawrence et al. 2001], and peer-to-peer (P2P) computing environments [Canny 2002]. As the variations in recommendation algorithms have increased, some studies have proposed a meta-recommendation system, one that combines results from different recommendation engines on the basis of user configurations [Schafer et al. 2002].

Given that CF has been applied to various areas (domains), it is surprising that so little research has addressed how the effectiveness of CF varies across domains. No general theoretical account exists of the conditions under which a particular collaborative recommendation application will succeed or fail [O'Mahony et al. 2004]. Most prior CF studies appear to have implicitly assumed that if CF is effective in one domain, it will be equally effective in other domains as well. For example, Mild and Natter [2002], using movie evaluation data to compare CF with other recommendation methods, concluded that regression is more accurate than CF. In fact, this result may be true in the movie domain but not in others.

At present, CF-based recommendation systems are used mostly for consumer products like CDs, movies, and novels, and rarely for knowledge objects like technical documents, manuals, engineering drawings, and news clippings. For our purposes, a "knowledge object" may be defined as a unit of codified knowledge that yields value for individuals or organizations and that is used for professional purposes. Only a very few studies have looked at the application of CF to knowledge objects (e.g., McNee et al. [2002] on research papers and Konstan et al. [1997] on Newsgroup messages). While the two domain types, consumer products and knowledge objects, are not by any means mutually exclusive, knowledge objects and consumer products *typically* have different characteristics, the most important of which are discussed further in the next section. Given these differences, it is likely that the effectiveness of CF will vary between the two domain types. We know of no studies that have directly compared the accuracy of CF in knowledge-object and consumer-product domains.

Another shortcoming of prior studies is that they do not question the implicit assumption by CF systems that users' needs and goals are invariable. Suppose a user named Tom loved *Star Wars* when he saw it and rated it highly when making his initial evaluations in a CF system. The system assumes that he would like the movie under any and all circumstances, and groups him accordingly with other likeminded users — about whom it makes the same assumption. No allowance is made for the possibility that Tom would like the movie less on a different day or in a different

mood. Some CF systems do allow users to modify the initial evaluations that establish their profile, but even in those systems, it is assumed that a new evaluation replaces the old one permanently. In reality, however, people look for products and information for many different reasons, and their responses to the material they find may vary according to their goals. Their perception of the accuracy and usefulness of a recommendation may therefore be affected by the goals that led them to search in the first place. In other words, users have multiple usage scenarios, and the performance of CF may vary depending on which scenario motivates their search, as suggested in McNee et al. [2002].

The accuracy of recommendations should be understood in the context of multiple usage scenarios, in which different search goals obtain under different circumstances. We need, therefore, to identify the most common goals of an information search, and then to investigate how the accuracy of CF varies across those goals. Though a few studies (e.g. Miller et al. [1997]) have indirectly investigated the effects of product category on accuracy, no research, to the best of our knowledge, has examined the impact of search intention on the accuracy of CF recommendations.

## III. Research Issues and Hypotheses

There are many factors that may affect the accuracy of CF. One is the number of users: as previous studies have shown [Konstan et al. 1997; Shardanand and Maes 1995], when the number of users increases, the subset of people with similar preferences grows as well. As discussed above, it is also likely that the recommendation accuracy of CF will be affected by search domain and by user search intention.

## 1. Number of Users and Critical Mass

An increase in the number of users raises the probability of finding people with similar preferences, which should in turn increase recommendation accuracy. Several studies (e.g., Shardanand and Maes [1995]) have shown that the accuracy of CF increases as the number of users increases. Our first hypothesis (H1) is a kind of "control hypothesis," in that it is obvious and has already demonstrated in prior studies. It will simply be a basis for other hypotheses.

**H1.** The accuracy of a CF system increases as the total number of users increases.

It is very likely that a certain number of users is required for a certain level of recommendation accuracy. We call this minimum number of users the "critical mass." As the relationship between the number of users and performance may differ in different situations, so may the critical mass vary according to circumstances. Figure 1 shows three hypothetical patterns. In one, pattern C, accuracy increases sluggishly at

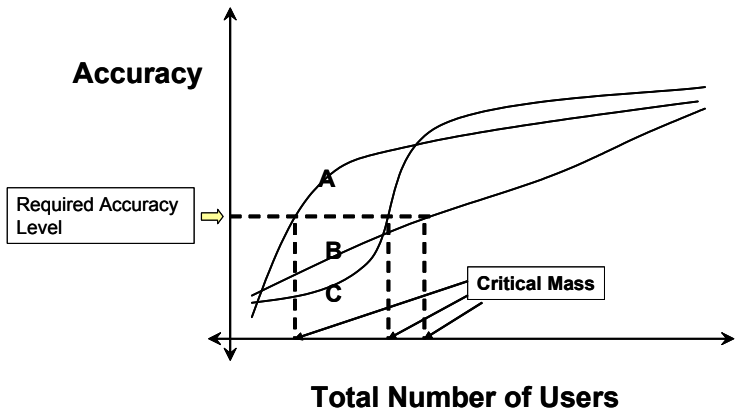



Figure 1. Hypothetical Relationships between Total Number of Users and CF System Accuracy

first, then begins to rise rapidly, and then at a later point slackens off. Of course, these three hypothetical patterns do not cover all possible patterns, and are offered for purposes of illustration. An entirely different pattern may emerge in the actual data.

The point is that accuracy may increase in different patterns depending on domain and other factors. Knowing the typical pattern of accuracy increase in given circumstances would help CF system managers make key decisions: whether a CF system should be implemented in a certain domain (can the desired accuracy level be achieved?), when a new system should be released (how many users will yield the desired accuracy level?), and so forth.

## 2. Domain Type and Preference Heterogeneity

"Domain" refers to the area in which recommendations are given. As explained above, there are different types of domains, some of them consumer-product domains, others knowledge-object domains. Movie recommendation systems belong to the movie domain — a consumer-product domain. A system that recommends newsgroup messages belongs to the newsgroup domain — a knowledge-object domain. Miller et al.'s [1997] study comparing the characteristics of evaluations (e.g. correlations of the evaluations across users) and the accuracy of CF across different newsgroups (e.g. food, humor) found that the characteristics of evaluations and the accuracy of CF varied across different newsgroups.

"Preference heterogeneity" is a term used to represent the pattern of consumer preference in a specific domain. High preference heterogeneity means there is substantial variation in consumer preferences — a low consensus in evaluation, and many distinct clusters of ideal points in attribute space [Feick and Higie 1992]. In domains characterized by low preference heterogeneity, consumers give similar attribute weightings, share similar preference attributes, and appear to rely upon more objective standards of evaluation [Feick and Higie 1992].

Figure 2 shows examples of heterogeneous and homogeneous domains. Suppose

there are two items in each domain, and how much a person likes or dislikes each item can be measured. The first domain is heterogeneous because people express different likes and dislikes, and no single consensus emerges. The second domain is less heterogeneous (more homogeneous) because most people express similar preferences: they dislike item 1 and like item 2. It has been shown that the degree of heterogeneity in consumer preferences varies across product categories and product attributes [Allenby et al. 1998; Bapna et al. 2004]. Since CF generates recommendations based on the evaluations of a group of people with similar preferences, different preference distribution patterns will likely result in different recommendation accuracies.

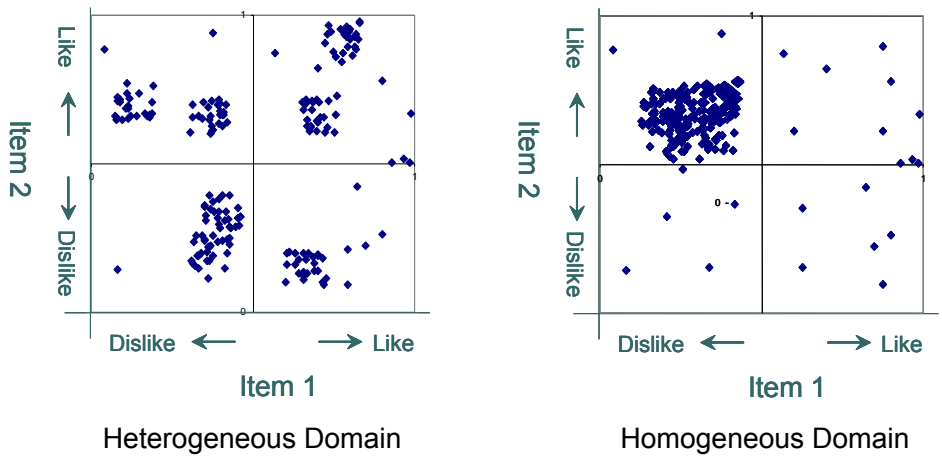



Figure 2. Hypothetical High- and Low-Heterogeneity Domains

It is likely that the pattern of accuracy increase (as in Figure 1) will be different across domains that have different levels of preference heterogeneity (as in Figure 2). For example, if a domain is relatively homogeneous, accurate recommendations may be expected with a small number of users, because most will have similar preferences, and the pattern of accuracy increase will likely resemble line A in Figure 1. If a domain is heterogeneous, however, the pattern of increase might be closer to line C.

In this case, while more users would be needed to find a subset with similar preferences (i.e. to attain critical mass), eventually, as the number of users increased, each cluster would hold sufficient users for a CF system to make accurate recommendations. Because clusters in heterogeneous domains are more distinctive from each other than those in homogeneous domains [Feick and Higie 1992], recommendation accuracy might be expected at this point to increase sharply, producing the steep rise typical of the S-curve shown as line C. Thus, with a large user set, recommendations for a heterogeneous domain may actually be more accurate and useful than those for homogeneous domains.

This study looks at one consumer-product domain — movies — and one knowledge-object domain — research papers. The set of users searching for movie recommendations will be larger and more various (in terms of tastes, interests, income, education, etc.) than the set of users searching for an appropriate research paper. However, movie-goersjudge movies by broadly shared standards — the quality of the acting, the plot, the visual effects, and so forth — while a scholar evaluates a research paper by very narrow criteria, namely, its relevance to his or her specific research. If Tom finds a particular movie excellent, it is likely that most people he knows will also like it. However, even if Jane finds a paper very relevant to her research, most likely only a few others in her research community will admire it as much, because they are active in different research areas. Thus, while *people* in the movie domain may be more various than users in the research paper domain, their *preferences* will be more homogeneous, because they will share common standards and overlap far more in their evaluations. Consequently, after a certain threshold has been crossed, the accuracy of CF should increase faster in the research paper domain. Our second hypothesis can therefore be formulated as follows:

**H2.** The accuracy of CF will increase faster for research papers than for movies.

## 3. Mode of Search

One challenge in designing information retrieval systems for a knowledge-object domain, such as a knowledge base or a digital library, is accommodating users' varying information needs [Korfhage 1997]. Researchers using a digital library, for example, will have various research projects in a variety of subject areas. Differences of subject area can be resolved to a certain extent by using CF systems: due to the nature of their algorithms, CF systems will find people who evaluated common items, and who thus probably share some research interests with the searcher. However, even among users interested in the same items, researchers may seek information for different reasons, and thus evaluate its worth differently.

Studies have categorized information search behavior in various ways. In an ethnographic study with professional intermediaries, O'Day and Jeffries [1993] categorized information searches into three categories: *monitoring, information search following a plan*, and *exploring*. Monitoring is a search for information search on a well-known topic. An information search following a plan is searching for information on a specific topic, following a typical plan. Exploring is an undirected information search. Vandenbosch and Huff [1997] identified two search modes: *scanning* and *focused search*. A user performing a focused search is driven by a need to answer a specific question. A user who is scanning browses through data in order to understand trends or to sharpen his or her general understanding of a field. Similarly, El Sawy [1985] categorized the information retrieval behaviors of managers as *scanning* and *problemistic searches*. A problemistic search is stimulated by a problem and directed towards finding a solution; scanning is not directed towards any particular problem.

Although the categories and the associated terms differ across studies, it is clear that there are two major information search modes: one with specific questions and one without. We will use the terms *problemistic search* and *scanning* to distinguish them. In the problemistic search, users already have very specific ideas about what they are

looking for; users engaged in a scanning search do not. In scanning mode, users apply broad and diverse criteria when they evaluate materials, because they have no specific goals. In this mode, users' evaluations will likely be more homogeneous, because there will be more overlaps in their interests. In contrast, users engaged in a problemistic search will apply narrower criteria when they evaluate materials. As evaluation criteria grow narrower, they overlap less and less, producing greater heterogeneity.

Miller et al. [1997] have shown that a CF system is effective for domains that have high overall correlations among users. If people evaluate items using similar criteria, their evaluations will have higher correlations, which suggests that CF accuracy will be higher for users in scanning mode than for users in problemistic search mode. However, it is also possible to argue that heterogeneity will actually increase CF accuracy. Users engaged in a problemistic search are more heterogeneous than users in scanning mode. As shown in Figure 2, heterogeneous domains will show a greater number of tight clusters, while homogeneous domains will show fewer clusters that are more loosely grouped. It is possible that users in the problemistic search mode can be divided into clearly delineated clusters, and that users within each cluster will have high correlations, thus improving CF accuracy. If this is so, CF will actually work better for users engaged in a problemistic search, because higher correlation within a group means higher CF accuracy.

These conflicting predictions may be resolved by the concept of critical mass. As discussed above, users in scanning mode evaluate items using broader criteria than users in problemistic search mode. Therefore, evaluations in scanning mode will have less variability across users than in problemistic search mode. Less variability in evaluations means fewer users (a lower critical mass) are required to achieve the same accuracy level. With a small number of users, therefore, a CF system will perform better when the users are in scanning mode, because the accuracy threshold will be lower. When there are sufficient number of users in problemistic search mode, however, they can be grouped into clusters large enough to achieve critical mass but relatively homogeneous internally.

**H3a.** With a small number of users, the accuracy of CF will be higher for users in scanning mode than for users in problemistic search mode.

**H3b.** The accuracy of CF will increase for users in problemistic search mode faster than for users in scanning mode.

In most CF systems, users in different search modes are mixed together. Recommendations are generated without consideration either for the search mode of the current user or for the search modes of the users whose evaluations are being used to generate recommendations. However, users can be in different search modes at different times, and different search modes entail different mindsets [El Sawy 1985; Vandenbosch and Huff 1997]. Users may have different evaluation criteria according to which search mode they are in. The recommendations a CF system produces will likely be less accurate when users in different search modes are grouped together. Conversely, when all evaluations put into a CF algorithm come from users in a single search mode, the resulting recommendations are likely to be more accurate. Thus our final hypothesis can be expressed as follows:

**H4.** The accuracy of a CF system will be greater for users in a single search mode than for users in mixed search modes.

## IV. Empirical Study

In order to test our hypotheses, we collected and analyzed data from two different domains, movies and research papers. The accuracy of CF systems was calculated using simulations — the most widely used method in CF research [Ansari et al. 2000; Breese et al. 1998; Herlocker et al. 1999].

## 1. CF Algorithms

The general recommendation process is similar for most CF applications, though the specific algorithms differ according to application [Breese et al. 1998; Herlocker et al. 1999; Shardanand and Maes 1995]. A CF system begins by soliciting a certain number of initial evaluations from the user for whom it is to generate recommendations. Once these initial evaluations have been entered, the system can identify people with similar preferences by calculating the degree of similarity (referred to here as the *similarity index*) between the user and every other user in the system. The result is a set of people (referred to here as the *reference group*) with preferences similar to those of the user in question. The system then looks up the items that have not yet been seen by that user and, using the evaluations of the other users in the reference group, predicts the user's evaluation of each item. It recommends the items with the highest predicted evaluation scores.

## 2. Selection of Domains and Similarity Measures

One of the goals of this study was to compare the effectiveness of CF in a consumer-product domain to its effectiveness in a knowledge-object domain. Movies are a representative consumer product widely used in CF research [Ansari et al. 2000; Breese et al. 1998; Herlocker et al. 1999]. Research papers were chosen as a typical knowledge-object domain.

Although it is not our purpose to compare different recommendation algorithms, selection of the type of CF algorithm to be used is a critical issue, for the choice of algorithm will affect the accuracy of the recommendations. Prior studies [Breese et al. 1998; Herlocker et al. 1999; Shardanand and Maes 1995]have shown that correlation is one of the best similarity indices for CF. A comparison of correlation and vector similarity showed that the two measures achieved equal performance [Huang et al.

2004]. Thus a correlation coefficient was used as the similarity index in this study. As for reference group selection, the two most common selection methods are the "thresholding" and the "best-n-neighbor" methods. The latter is reasonably accurate, widely used, and easily analyzed [O'Mahony et al. 2004]. Moreover, it has higher coverage than the thresholding method; that is, recommendations can be generated for a greater percentage of items and users [Herlocker et al. 1999]. Although the accuracy of the thresholding method is slightly higher, its coverage is sometimes unacceptable — 19% in one study, as opposed to 99% for the best-n-neighbor method [Herlocker et al. 1999]. Further, one of the variables that affects the accuracy of CF is reference group size. This variable, which must be controlled for precise analysis, cannot be controlled with the thresholding method [Herlocker et al. 1999]. For all these reasons, the best-n-neighbor method was selected for this study.

## 3. Experiments

The systems for the two experiments — one for movies and one for research papers — were developed as Web applications using Microsoft Access, ASP, Oracle, Borland's Delphi, JavaScript, and HTML. The system for the first experiment contained about 490 movies in various genres. The subjects for this experiment were recruited from undergraduate and graduate classes at a major university on the east coast of the US. They participated in the experiment as a class assignment. Once they had registered, the experiment was explained to them, and they were shown a consent form, to which they assented by clicking on "Agree." The subjects were then instructed to think of a specific occasion, for instance a party or a family gathering, for which they might need to find a good movie. They were then shown a randomly selected set of 10 movie titles. If they had seen a movie, they were asked to evaluate it twice, on the basis of two different criteria: *in general and for the specific occasion chosen*. (The former was an evaluation in scanning mode, the latter an evaluation in problemistic search mode.) Once subjects finished evaluating the 10 movies, they were

given two options: get ten movie recommendations for each search mode ("in general" and "for the specific occasion"), or evaluate more movies to get better recommendations. Of the 168 subjects, 102 chose to evaluate more movies before they received recommendations.

The procedure for the second experiment was similar. The system contained abstracts of about 2,000 academic articles from recent issues (1991~2000) of five leading IS journals: *Communications of the ACM, Information Systems Research, Journal of MIS, Management Science, and MIS Quarterly*. Academics in the IS field were selected as the subject pool because the experiment required substantial knowledge of and experience in IS research. E-mails soliciting participants were sent to IS faculty members whose e-mail addresses were listed, as of April 2000, in the IS Faculty Directory of ISWorld (www.isworld.org). After registering and agreeing to the consent form, subjects could search for papers by keyword. This search function was provided because, unlike in the first experiment, the length and number of papers in the system would require subjects to spend an excessive amount of time with papers if they were presented randomly. As in the first experiment, subjects were asked to evaluate each paper according to two criteria: *overall usefulness/relevance of the paper for general IS research and usefulness/relevance of the paper for the subject's specific research project*. Other procedures were similar to those in the first experiment. Of 259 subjects, 47 chose to evaluate more than the minimum 10 papers before receiving recommendations.

The first experiment lasted about four weeks; 168 subjects participated. The second experiment lasted about six weeks; about 480 people visited the experimental site, of whom 259 participated.

## 4. Simulations

Using data from the experiment, we ran simulations to calculate the accuracy measures; no generated data were used. This is a method commonly used in time

series analysis [Griffiths et al. 1993] and other CF studies [Ansari et al. 2000; Breese et al. 1998; Herlocker et al. 1999] to calculate the accuracy of estimations.

For some hypotheses, the accuracy needed to be calculated using subsets of the full set of users. For example, for H1, the CF accuracy had to be calculated with varying number of users — 100, 110, 120, and so on — to see the effect of the increase of user numbers. Thus, for the calculation of accuracy with 100 users, that many users needed to be selected from the whole pool (259 users) as the sample. However, the accuracy of the recommendation might be affected by *which* 100 users were selected. In order to eliminate this selection bias, the simulation was repeated a certain number of times (the details are discussed below) with different sets of 100 users, and the average was used as the final accuracy measure. A similar process was used for all accuracy calculations that involved less than the full pool of users.

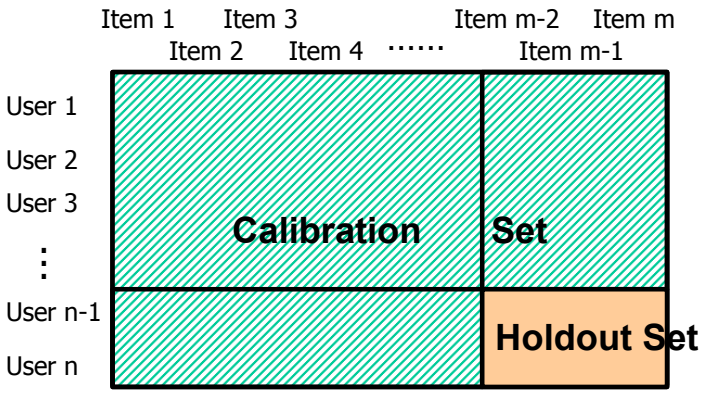



Figure 3. Holdout Set and Calibration Set in the Simulation

Once the sample was selected (100 users in the above example), it was divided into two sets, a *holdout set* and a *calibration set*, as shown in Figure 3. The estimated evaluations for the items in the holdout set were calculated using the evaluations in the calibration set. Since the actual evaluations in the holdout set were known, the estimation errors (i.e. actual evaluation minus predicted evaluation) for the holdout set

could be calculated. The following table summarizes the procedure for calculating the estimated evaluations and the final accuracy measures when two users and three items from each user are in the holdout set (as depicted in Figure 3).

**Step 1:** Select two users who will belong to the holdout set.

**Step 2:** Select three items for each user to put in the holdout set.

**Step 3:** Calculate estimated evaluations for the three items selected in Step 2. The estimated evaluations are calculated as if the data in holdout set were not known.

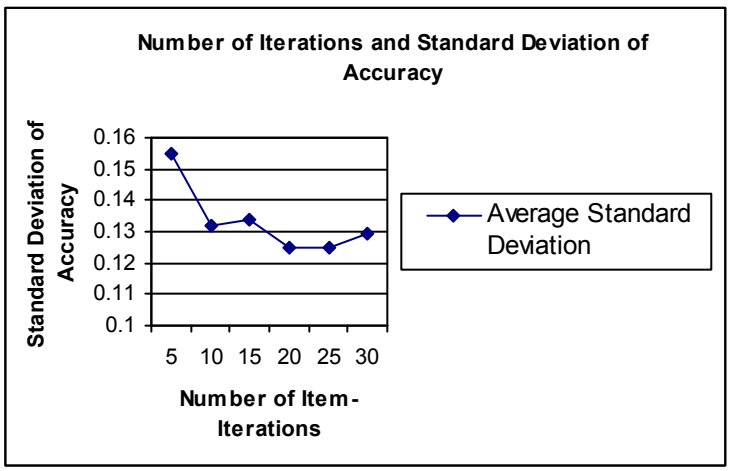
**Step 4:** Calculate accuracies by comparing the actual evaluations to the estimated evaluations.

**Step 5:** Repeat Steps 2 through 4 for all other users in the holdout set.

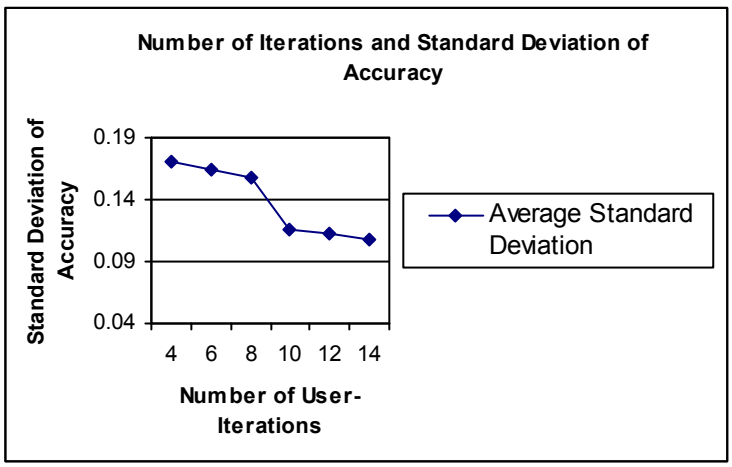
**Step 6:** Average the accuracies of all the items of all users in the holdout set. The average is the final accuracy measure.

The problem of selection bias described above also applies to the selection of the holdout set and calibration set. For example, if there are to be two users in the holdout set, two users need to be selected from the subject pool. No matter how those two users are selected, there will likely be a selection bias. Similarly, for each user in holdout set, choosing the items to set aside as holdout items will produce a selection bias. In theory, these selection biases can be eliminated by trying all possible combinations of users and items and averaging the results. However, since the number of possible combinations is staggeringly large, it is more practical to carry out a limited number of iterations with randomly selected combinations of users and items. In order to determine an optimal number of iterations, a pilot simulation was run to see how the variance or standard deviation of the accuracy changed as the number of iterations increased. Figure 4 shows how the standard deviation of accuracy measures changed as the number of iterations increased. The figure shows that the standard deviation of the accuracies stabilized after a certain number of iterations — approximately 20 item iterations and 10 user iterations. Thus the procedure described

above was repeated a total of 200 times, once for each simulation condition: 10 different random combinations of users, and 20 different random combinations of items for each combination of users.




a. Number of Item Iterations and Standard Deviation of Accuracy




b. Number of User Iterations and Standard Deviation of Accuracy

Figure 4. Numbers of Iterations and Standard Deviations of Accuracy

Another set of pilot simulations was run to check whether there was an optimal reference group size. For movies, accuracy was highest when the reference group size was around 70. For research papers, however, there seemed to be no optimal reference group size; accuracy did not change much as reference group size varied, though accuracy did appear to be slightly higher at around 60-70users. In order to eliminate the effect of reference group size in the movie domain, the simulation in this research was iterated with different reference group sizes. Reference groups from 41 to 100 (70 ± 30) for movies and from 31 to 100 (65 ± 35) for research papers were used. The average accuracies of these different reference groups were taken as the final accuracy measures for the two experiments.

Changes in the number of users in the holdout set (holdout user size) and the number of items in the holdout set (holdout item size) might also shift the final accuracy measures; thus we needed to use consistent numbers. The holdout user size for both movies and research papers was set at five. After considering the average number of evaluations per user (14.7 for movies and 7.01 for research papers, as shown in Table 1), we set the holdout item size at four. Thus, unless otherwise specified, all analyses in this paper are based on simulations with five holdout users and four holdout items. In the simulation, those users who did not have enough evaluations (fewer than three after holdout items were taken out) were excluded from accuracy calculations, so that the small numbers of items could not bias the results.

## 5. Accuracy Measures

The most commonly used measures in investigations of CF system accuracy are mean absolute deviation (MAD) [Ansari et al. 2000; Goldberg et al. 1992; Herlocker et al. 1999; Sarwar et al. 2001; Shardanand and Maes 1995], mean squared error (MSE) [Miller et al. 1997], root mean squared error (RMSE) [Sarwar et al. 2001], and correlation between actual and predicted evaluations [Hill et al. 1995]. MAD, MSE, and RMSE are variations of the same measure (the difference between actual and

estimated evaluation), and there would not be substantial differences between them if they were all used. Thus MAD was chosen as one of our two primary accuracy measures and MSE and RMSE omitted.

In fact, however, MAD, MSE, and RMSE have several common limitationsas accuracy measures for CF. First, they do not consider how CF performs compared to non-CF recommendation methods. One simple non-CF method is to recommend items based on the average of all users' evaluations ("average evaluations of everybody"). A CF system is not necessarily more accurate than this simple alternative, even if its recommendations have good MAD, MSE, or RMSE measures; the non-CF system might score just as well. Second, these measures cannot be used to compare domains with different evaluation scales; for example, measures on a 1-5 scale and measures on a 1-7 scale cannot be directly compared.

The use of a rank-based accuracy measure can eliminate these problems. Rank is a proxy for user utility, since users prefer to find relevant results earlier [McNee et al. 2002]. A rank-based measure developed in this study calculated the sum of the rank differences between actual and predicted evaluations. If the system did not have the CF feature, the "average evaluations of everybody" would be used to generate recommendations. Thus "average evaluations of everybody" becomes a baseline against which the accuracy of CF is measured.

**Accuracy** = estimation error by "average evaluations of everybody" — estimation error by CF

**Estimation error by "average evaluations of everybody" and estimation error by CF =**

$$\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\left|r(a_{ij})-r(e_{ij})\right|}{nm}$$

where $r(a_{ij})$ = rank of actual evaluation by user $i$ of item $j$

$r(e_{ij})$ = rank of estimated evaluation (by CF or "average evaluations of everybody") by user $i$ of item $j$

$n$ = number of users in the holdout set

$m$ = number of items in the holdout set

A positive value for the rank-based measure meant that CF was more accurate than the “average evaluations of everybody”; a negative value meant that CF was less accurate. The rank-based measure was the second of the two primary accuracy measures used in this study.

## V. Results and Discussion

A total of 168 subjects participated in the movie experiment; the total number of evaluations was 4,946 (2,473 for each search mode). The number of movies evaluated by at least one user was 475.Approximately 480 people visited the experimental site in the second experiment, of whom 259 participated; the total numbers of evaluations were 3,634 (1,817 for each mode). In all 1,063 papers were evaluated by one or more users. Other basic evaluation statistics are summarized in Table 1.

Table 1. Basic Evaluation Statistics

| | | Average of Evaluations | Std. Dev. | Average Number of Evaluations |
|---|---|---|---|---|
| Movies (0 to 6 scale)* | Scanning Mod | 3.89 | 1.60 | 29.4 per user (14.7 per mode) |
| | Problemistic Search Mod | 3.50 | 1.83 | |
| Research Papers (0 to 6 scale)** | Scanning Mod | 3.33 | 1.72 | 14.0 per user (7.0 per mode) |
| | Problemistic Search Mod | 2.88 | 1.94 | |

*How much do you like the movie? (0 = awful, 6 = wonderful)

**How relevant is the paper to your research? (0 = negligible, 6 = extraordinary)

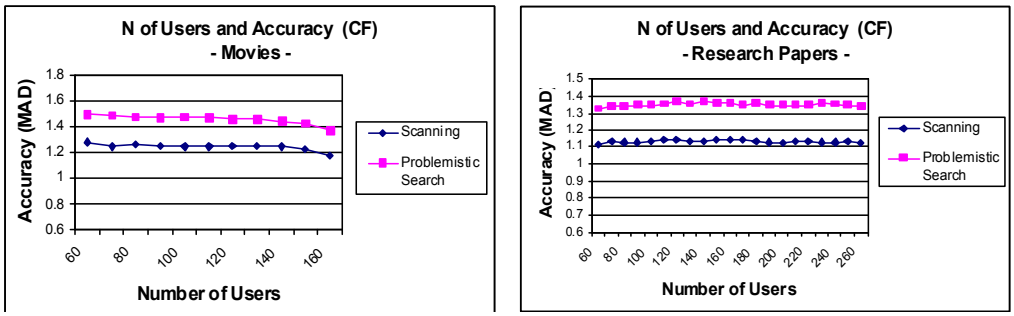
Interestingly, the domains share two important features: the average evaluation is higher in scanning mode than in problemistic search mode (with differences significant at the = 0.01 level for both research papers and movies), and the standard deviations in evaluations are smaller in scanning mode than in problemistic search mode. These results indirectly support an argument made earlier: people evaluate items with broader (higher average) but more similar (smaller standard deviation) criteria in scanning mode and with narrower (lower average) but more diverse (higher standard deviation) criteria in problemistic search mode.

Table 1 also reveals that research papers received lower ratings than movies regardless of search mode (with differences significant at the = 0.01 level for both modes). This suggests that research papers comprise a more heterogeneous domain than movies do. However, further research incorporating a more operational definition of "preference heterogeneity" would be needed to corroborate this finding.

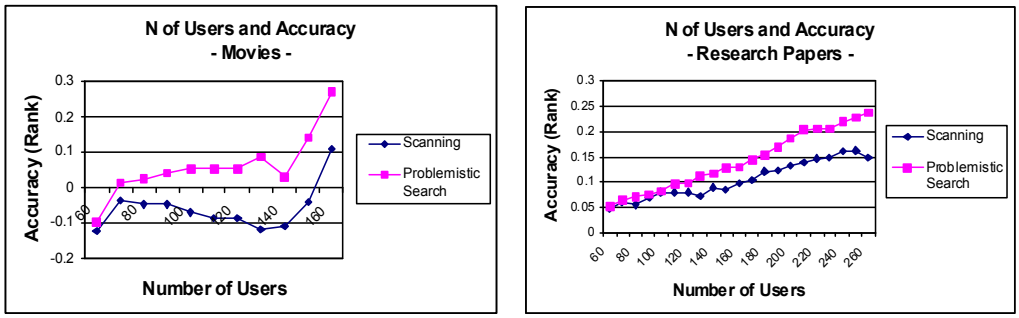
## 1. Number of Users and the Accuracy of CF Systems

In order to test H1, we conducted simulations with different numbers of users and calculated the recommendation accuracies, as shown in Figure 5. In the figure, the horizontal axis represents the simulated total number of users. For example, 60 on the horizontal axis means that the accuracy was calculated with 60 randomly selected subjects.

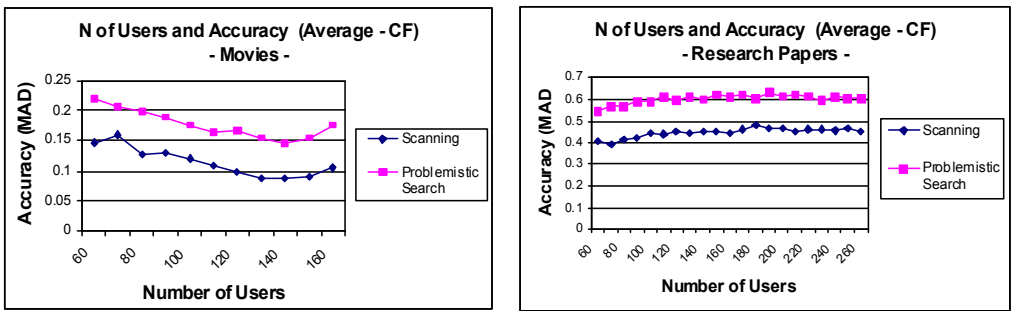
In the MAD accuracy measure (Figure 5a) smaller numbers represent more accurate recommendations. In the movie domain, the accuracy of a CF system measured by MAD increases as the number of users increases; in research paper domain, however, the accuracy of a CF system seems to remain almost unchanged. The rank-based measure (Figure 5b) presents a somewhat different picture: accuracy in the movie domain seems generally to increase, but with a moderate number of users (in the middle of the graph) it stabilizes or even decreases. In contrast, accuracy increases almost monotonically in the research paper domain.




a. MAD (CF)




b. Rank-Based Accuracy




c. MAD (Average Evaluations of Everybody CF)

Figure 5. Number of Users and CF Accuracy

The differences between Figure 5a and Figure 5b can be explained by Figure 5c, which shows the differences in MAD between CF and "average evaluations of everybody." This difference keeps increasing in the research paper domain, which indicates that as the number of users increases, recommendations by CF become more accurate than "average evaluations of everybody" recommendations. In the movie domain, however, the difference decreases first and then begins to increase.

## 2. Mode of Search and the Accuracy of CF Systems

More rigorous tests were conducted to determine whether the increase in the rank-based measures was statistically significant and whether there were differences across the movie and research paper domains. A test of normality (histogram and Q-Q plot) showed that the distribution of the data did not follow normal distribution. Therefore, a non-parametric test method had to be used for data analysis; a parametric method is not appropriate when the probabilistic distribution of the data is not normal [Conover 1999]. One way of testing whether the increase was statistically significant would be to check whether the data deviated significantly from the non-increasing (horizontal) line. If the increase in accuracy was statistically significant, the differences between accuracy measures in Figure 5and an increasing line would be significantly smaller than the differences between the accuracy measures and the best horizontal line. The best horizontal line to compare with would be the average of the measures. This logic is similar to that used in regression analysis in the t-test on the coefficients of independent variables.

A Wilcoxon Signed Rank Test [Conover 1999] was conducted to test whether the accuracy increase was statistically significant. A simple linear regression-fitting line was used as the increasing line. The results (summarized in Table 2) show that the accuracy of CF increases significantly in the research paper domain but not in the movie domain.

The accuracy of CF for research papers increases almost linearly, while the accuracy

of CF for movies increases following an inverse S-shape curve. This suggests that for consumer products such as movies, there is a "take-off point" at which the accuracy of CF recommendations, compared to those based on "average evaluations of everybody," begins to increase dramatically. In contrast, the accuracy of CF systems for knowledge objects such as research papers appears to increase relatively steadily. However, it is possible that here too the accuracy of CF for research papers is in its initial stage, and that some larger number of users would constitute a "take-off point," after which accuracy would increase more rapidly. Further research with a larger dataset would provide a better understanding of this phenomenon. Nonetheless, it is useful to know that the pattern of increase for our sample knowledge-object domain (research papers) differs from that for our sample consumer-product domain (movies).

Table 2. Results of Pattern of Accuracy Increase Test

| | | Willcoxon Signed Rank Test Z value (Horizontal Line vs. Regression Fit) | |
|---|---|---|---|
| | | MAD | Rank-based |
| Movie | Scanning Mod | -0.525 | -0.657 |
| | Problemistic Search Mod | -0.919 | -0.098 |
| Research Paper | Scanning Mod | -0.226 | -4.880** |
| | Problemistic Search Mod | -0.201 | -5.031** |

**significant at = 0.01 level

Figure 5b shows that when gauged by rank-based measures, with a given number of users CF recommendations for research papers are more accurate than CF recommendations for movies. To the degree that research papers and movies are representative of their domain types, the implication is that CF can provide more effective recommendations for knowledge objects than for consumer products.

A post-experiment questionnaire measured the subjects' perceptions of the performance of the recommendations they were given. The relevance and novelty of

recommendations have been the two most important perception measures of CF performance in prior studies [McNee et al. 2002]. Users were asked two questions to measure perceived performance, one concerning "relevance" and one concerning "provision of a new perspective." The analysis results are summarized in Table 3.

Table 3. Perceived Effectiveness of Recommendations

| | | Relevance | | Recommending New Item | |
|---|---|---|---|---|---|
| | | Average** | z | Average** | z |
| Movies (n = 122) | Scannin | 3.78 | -0.264 | 3.50 | -2.217 |
| | Problemistic Searc | 3.83 | | 3.73 | |
| Research Papers (n = 41) | Scannin | 2.80 | -1.244 | 3.07 | -0.645 |
| | Problemistic Searc | 2.59 | | 2.97 | |

*Significant at = 0.05 level

**0 = Poor, 6 = Excellent

The differences of accuracy in the two search modes were tested statistically using a non-parametric test method. One of the four tests was statistically significant: "provision of a new perspective" in the movie domain. This shows that users perceived that CF recommendations for movies provided a significantly new perspective in problemistic search mode than scanning mode. The insignificant results for other cases are probably due to the small sample size.

To test H2, H3a, and H3b, we compared the accuracies of recommendations in the two search modes. As shown in Figure 6, the accuracies in scanning and problemistic search modes are similar with small numbers of users. However, as the number of users increases, the accuracy of the problemistic search mode improves faster than the accuracy of the scanning mode in both domains.

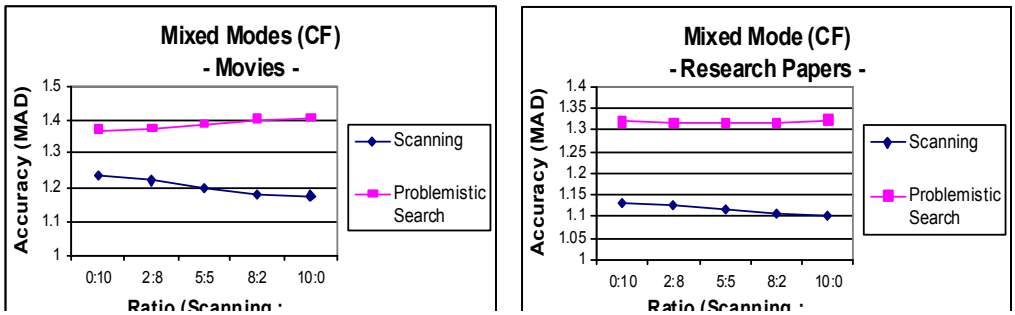
In order to test whether the increase in accuracy in problemistic search mode was significantly faster than that in scanning mode — in other words, to determine whether the slopes of the two lines in Figure 5 were significantly different — we

conducted a non-parametric test. If the two slopes in the figure were significantly different, the differences between the two modes would keep increasing. Following the rationale outlined in section 5.2, we performed the Wilcoxon Signed Rank Test. The z-value for research papers was -3.492 and was significant at the $\alpha = 0.01$ level. The results for movies were not statistically significant. These results indicate that CF systems will work better for users needing help with a specific problem. Interestingly, the accuracy of CF in problemistic search mode is higher than in scanning mode even with small numbers of users, disproving H3a.

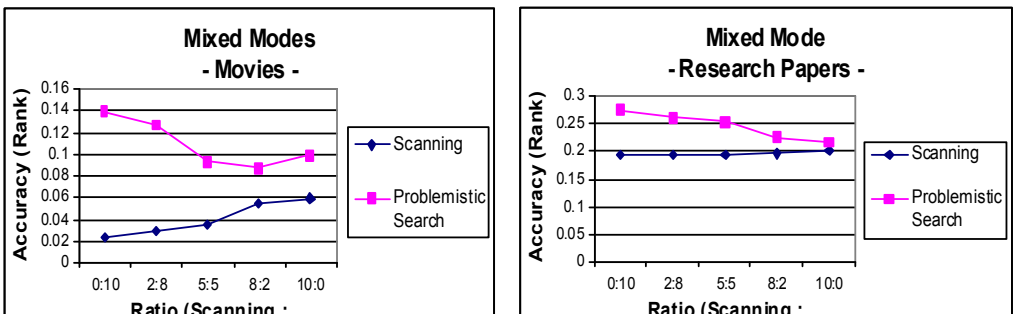
The accuracies in Figure 5 were calculated using data from unmixed search modes. To test H4, we randomly mixed users in the two search modes in five different proportions (0:10, 2:8, 5:5, 8:2, and 10:0) and calculated accuracies. The recommendation accuracies for the two domains were calculated for each proportion. In Figure 6, the five points on the horizontal axis represent the proportions of two search modes. For example, the point on the left, 0:10, means that the ratio of users in scanning and problemistic search mode was 0:10.

The accuracy of CF for scanning mode increases as the proportion of users in scanning mode increases (moves to the right), while the accuracy for problemistic search increases as the proportion of users in problemistic search mode increases (moves to the left). Two non-parametric tests were conducted to determine whether the changes in CF accuracy across different proportions were statistically significant. Kendall's W-test was used to test mean difference across different proportions [Conover 1999].

Table 4 shows the results of two the tests. The changes in accuracy were statistically significant, which implies that the accuracy of CF decreases as more evaluations from different modes are mixed into the database. Table 5 summarizes the results for our hypotheses.




a. MAD




b. Rank-Based Accuracy

Figure 6. Mixed Modes and Accuracy

Table 4. Accuracy with Users in Mixed Modes

| | | Kendall's W-test (Differences across Different Mixes | |
|---|---|---|---|
| | | MA | Rank-base |
| Movie | Scanning Mod | 0.886*** | 0.805*** |
| | Problemistic Search Mod | 0.836*** | 0.620*** |
| Research Paper | Scanning Mod | 0.886*** | 0.287* |
| | Problemistic Search Mod | 0.437** | 0.769** |

* Significant at α = 0.1 level

** Significant at α = 0.5 level

*** Significant at α = 0.01 level

Table 5. Summary of Hypothesis Test Results

| Hypothesis | Test Result |
|---|---|
| H1. The accuracy of a CF system increases as the total number of users increases. | ✓ |
| H2. The accuracy of CF will increase faster for research papers than for movies. | ✓ |
| H3a. With a small number of users, the accuracy of CF will be higher for users in scanning mode than for users in problemistic search mode. | |
| H3b. The accuracy of CF will increase for users in problemistic search mode faster than for users in scanning mode. | ✓ |
| H4. The accuracy of a CF system will be greater for users in a single search mode than for users in mixed search modes. | ✓ |

✓ - Supported
× - Not supported

# VI. Conclusionsand Implications

This study investigated behavioral aspects of CF systems, an area neglected in the existing research. Based on prior studies in CF and related areas, this study has identified key factors that affect the accuracy of CF systems and examined the nature of that impact.

## 1. Limitations

The study's sample and research methods produce several limitations. First, only two domains — movies and research papers — were examined. Although the two domains are good representatives of distinct domain types, consumer products and knowledge objects, it would be desirable for purposes of generalization to investigate more domains. Second, the sample size is a limitation. Large-scale experiments or large data sets from actual CF systems would improve our understanding of CF accuracy with large numbers of users. Third, the study operationalized two typical

search modes, scanning and problemistic search, by means of two typical search domains, movies and research papers. While the existing research validates this approach as theoretically sound, these choices may nonetheless have introduced unintended biases. Fourth, this study only investigated users' explicit evaluations. CF can also utilize users' implicit evaluations by monitoring, for instance, clickstream data. With those implicit evaluations, results may be different.

## 2. Implications

The study results have direct implications for the development of CF-based recommendation systems. The study shows that the performance of CF systems is domain-dependent. The domains typically used for CF research and for commercial applications — movies and consumer products — are in fact less suited to CF than knowledge-intensive domains, where CF algorithms demonstrate greater accuracy. This may be encouraging news for repositories such as FAQs and corporate knowledge bases, where collaborative filtering has rarely been applied. Another implication of the study is that designers implementing a new CF system may need to conduct a pilot test to assess the suitability of CF for the intended domain.

Most prior research into CF accuracy has ignored the goals and intentions of users. This study shows that a user's search mode strongly influences the accuracy of the results. CF works better when users are looking for specific information with a specific goal than when they browse information out of general undirected interest. Developers of collaborative filtering systems also need to separate these search modes, for CF accuracy drops significantly when the algorithms indiscriminately combine results from both modes. Finally, the study shows that the effect of critical mass needs to be considered when developing a CF system. Critical mass appears to be higher for consumer products than for knowledge objects.

### 3. Future Research Directions

Extending this research would improve our understanding of issues surrounding CF systems. We suggest three major research directions. First, further research in other domainsis needed if our findings are to be generalized. Second, more research is needed to investigate how the patterns of evaluations affect the accuracy of CF systems. As shown in Table 1, the average evaluations and standard deviations differ across search modes, which implies that the evaluation patterns are different. Research on the evaluation patterns in various domains (e.g. what kind of evaluation patterns exist, and how patterns of evaluation should be measured) will provide a better understanding of their impact on CF accuracy. Finally, further research is needed to examine how search modes can be identified with minimal intrusion for users, stored effectively in a CF database, and incorporated into CF algorithms.

## References

Agarwal, R., and Venkatesh V. 2002. Assessing a firm's Web presence: A heuristic evaluation procedure for the measurement of usability. *Info. Syst. Research* 13, 2, 168-188.

Allenby, G.M., Arora, N., and Ginter, J.L. 1998. On the heterogeneity of demand. *J. Market. Research* 35, 384-389.

Ansari, A., Essengaier, S., and Kohli, R. 2000. Internet recommender system. *J. Market. Research* 37, 363-375.

Bapna, R., Goes, P., Gupta, A., and Jin, Y. 2004. User heterogeneity and its impact on electronic auction market design: An empirical exploration. *MIS Quart.* 28, 1, 21-43.

Breese, J.S., Heckerman, D., and Kadie, C. 1998. Empirical analysis of predictive

algorithms for collaborative filtering. In proceedings of Fourteenth Conference on Uncertainty in Artificial Intelligence, Madison, WI,

Canny, J. 2002. Collaborative filtering with privacy. In proceedings of IEEE Conference on Security and Privacy, Oakland, CA,

Collett, S. 2002. 35 years of IT leadership: The Web's best-seller. *Computerworld*, 36, 40 (September 30), 40-42.

Conover, W.J. 1999. *Practical Nonparametric Statistics*. John Wiley & Sons, New York.

Deshpande, M. and Karypis, G. 2004. Item-based top-N recommendation algorithms. *ACM Trans. Internet Tech*. 22, 1, 143-177.

El Sawy, O.A. 1985. Personal information systems for strategic scanning in turbulent environments: Can the CEO go on-line? *MIS Quart*. 9, 1, 53-60.

Feick, L. and Higie, R.A. 1992. The effect of preference heterogeneity and source characteristics on ad processing and judgments about endorsers. *J. Advert*. 21, 2, 9-24.

Goldberg, D., Nicolas, D., Oki, B.M., and Terry, D. 1992. Using collaborative filtering to weave an information tapestry. *Comm. ACM* 35, 12, 61-70.

Good, N., Schafer, B., Konstan, J.A., Borchers, A., Sarwar, B.M., Herlocker, J.L., and Riedl, J. 1999. Combining collaborative filtering with personal agents for better recommendations. In proceedings of Conference of the American Association of Artificial Intelligence, Orlando, FL,

Griffiths, W.E., Hill, R.C., and Judge, G.G. 1993. *Learning and Practicing Econometrics*. John Wiley & Sons, New York.

Hannabuss, S. 2002. Internet editorial. *Lib. Manage.* 23, 4/5, 254-258.

Herlocker, J.L., Konstan, J.A., Borchers, A., and Riedl, J. 1999. An algorithmic framework for performing collaborative filtering. In proceedings of Conference on Research and Development in Information Retrieval, New York,

Hill, W., Stead, L., Rosenstein, M., and Furnas, G. 1995. Recommending and evaluating choices in a virtual community of use. In proceedings of ACM

Computer Human Interactions, Denver, CO, 194-201.

Hirsh, H., Basu, C., and Davison, B.D. 2000. Learning to personalize. *Comm. of the ACM* 43, 8, 102-106.

Huang, Z., Chen, H., and Zeng, D. 2004. Applying associative retrieval techniques to alleviate the sparsity problem in collaborative filtering. *ACM Transact. Info. Syst*. 22, 1, 116-142.

Konstan, J.A., Miller, B.N., Maltz, D., Herlocker, J.L., Good, N., and Riedl, J. 1997. GroupLens: Applying collaborative filtering to Usenet news. *Comm. ACM* 40, 3, 77-87.

Korfhage, R.R. 1997. *Information Storage and Retrieval.* John Wiley & Sons, New York.

Koufaris, M. 2002. Applying the technology acceptance model and flow theory to online consumer behavior. *Info. Syst. Research* 13, 2, 205-214.

Lawrence, R.D., Almasi, G.S., Korlyar, V., Viveros, M.S., and Duri, S.S. 2001. Personalization of supermarket product recommendations, Data Mining and Knowl. *Disc.* 5, 1, 11-32.

McNee, S.M., Albert, I., Cosley, D., Gopalkrishnan, P., Lam, S.K., Rashid, A.M., Konstan, J.A., and Riedl, J. 2002. On the recommending of citations for research papers. In proceedings of Computer Supported Collaborative Work (CSCW), New Orleans, LA, 116-125.

Melville, P., Mooney, R.J., and Nagarajan, R. 2002. Content-boosted collaborative filtering for improved recommendations. In proceedings of National Conference on Artificial Intelligence, Edmonton, AB,

Mild, A. and Natter, M. 2002. Collaborative filtering or regression models for Internet recommendation systems? *J. Target. Meas. Anal. Market*. 10, 4, 304-313.

Miller, B.N., Riedl, J., and Konstan, J.A. 1997. Experiences with GroupLens: Making Usenet useful again. In proceedings of Usenix Winter Technical Conference, Anaheim, CA, 1-17.

Mulvenna, M.D., Anand, S.S., and Büchner, A.G. 2000. Personalization on the net

using web mining, *Comm. ACM* 43, 8, 122-125.

O'Day, V. and Jeffries, R. 1993. Orienteering in an information landscape: How information seekers get from here to there. In proceedings of International Computer-Human Interaction (INTERCHI), Amsterdam, 438-445.

O'Mahony, M., Hurley, N., Kushmerick, N., and Silvestre, G. 2004. Collaborative recommendation: A robustness analysis. *ACM Transact. Internet Tech.* 4, 4, 344-377.

Perkowitz, M. and Etzioni, O. 2000. Adaptive web sites. *Comm. ACM* 43, 8, 152-158.

Podberezniak, E. 1998.Collaborative filtering in TV recommender. Unpublished Master's thesis, New Jersey Institute of Technology.

Ram, S. 2001. Intelligent agents and the World Wide Web: Fact or fiction? *J. Data. Manage*. 12, 1, 46-47.

Sarwar, B.M., Karypis, G., Konstan, J.A., and Riedl, J. 2001. Item-based collaborative filtering recommendation algorithms. In proceedings of WWW Conference, Hong Kong,

Schafer, J.B., Konstan, J.A., and Riedl, J. 1999. Recommender systems in E-commerce. In proceedings of ACM Conference on Electronic Commerce, New York,

Schafer, J.B., Konstan, J.A., and Riedl, J. 2002. User-controlled integration of diverse recommendations. In proceedings of Conference on Information and Knowledge Management (CIKM), McLean, VA, 43-51.

3Shardanand, U. and Maes, P. 1995. Social information filtering: Algorithms for automating "word of mouth." In proceedings of ACM Computer Human Interaction, Denver, CO, 210-217.

Straub, D.W. and Watson, R.T. 2001. Research commentary: Transformational issues in researching IS and net-enabled organizations. *Info. Syst. Research* 12, 4, 337-345.

Terveen, L., Hill, W., Amento, B., McDonald, D., and Creter, J. 1997. PHOAKS: A system for sharing recommendations. *Comm. ACM* 14, 3, 59-62.

Vandenbosch, B. and Huff, S.L. 1997. Searching and scanning: How executives obtain information from executive information systems, *MIS Quart.* 21, 1, 81-107.

Williams, J.F. 2002. Hot technologies with a purpose. *Lib. J.* 51,

Wingfield, N. 1998. Unraveling the mysteries inside web shoppers' minds. *Wall St. J.* 18 June 1998.

Yuan, S. and Chang, W. 2001. Mixed-initiative synthesized learning approach for web-based CRM. *Exp. Syst. Applic.* 20, 2, 187-200.