International Journal of Artificial Intelligence in Education 14 (2004) 1-26
IOS Press

Collaborative Information Filtering: a review and an educational application

Andrew Walker, *College of Education, Lehigh University, Bethlehem, PA 18015, USA*
andy.walker@usu.edu

Mimi M. Recker, *Department of Instructional Technology, Utah State University, Logan, UT, 84322-2830, USA*
mimi.recker@usu.edu

Kimberly Lawless, *Department of Education, University of Illinois, Chicago, Chicago, IL, USA*
klawless@uic.edu

David Wiley, *Department of Instructional Technology, Utah State University, Logan, UT, 84322-2830, USA*

Abstract. This paper reviews the literature surrounding an information filtering technique, collaborative information filtering, which supports the discovery of resources in a way that is sensitive to the context of users. Moreover, via statistical clustering techniques, the system supports automated, personalized filtering and recommendation of relevant resources and like-minded users for particular user communities. The paper also describes an educational implementation of this approach, called Altered Vista, and presents results from a 3-month trial use of the system, aimed at evaluating the educational effectiveness and usefulness of the approach.

INTRODUCTION

According to recent surveys, almost 99% of full-time public school teachers in the United States have access to the Internet (National Center for Education Statistics, 1999). In tandem, teachers and students are increasingly encouraged to use Internet resources in classroom activities. For example, in 1997, the President of the United States recommended, as part of his programmatic emphasis on education, projects that used the Internet as a tool for teaching and learning (White House, 1997). Unfortunately, evidence is mounting that the Internet and its applications are not always used in productive, educationally relevant ways within instructional settings (e.g., Wallace, Kupperman, Krajcik & Soloway, 2000). For example, Rowand (2000) found that only six percent of teachers use computers or the Internet to find "model" lesson plans.

This most likely stems from a combination of socio-technical problems. First, studies show that teachers are chronically short of time, with heavy workloads (e.g., Swaim & Swaim, 1999), and using and searching the essentially unbounded Internet is time intensive, especially

1560-4292/03/$8.00

for novice users. Second, misinformation on the Web has been documented in many educational domains (Robertson, 1999). As such, Internet users need to develop robust, efficient searching and discrimination skills when accessing non-reviewed web resources.

Third, the technology itself is partly to blame. It is important to note that core Internet applications (such as Web browsers and search engines) were not built with instruction in mind. Traditional, content-indexing search engines, once the most popular means for finding Web resources, suffer from poor precision and recall. That is, for a typical keyword search, too many matches and false hits are returned to be genuinely useful for the average user. Moreover, Lawrence and Giles (1999) found that search engines display a continually decreasing level of coverage of Web content, with no more than 16% coverage by any one engine. Finally, full-text, content indexing searches have also become increasingly ineffective due to the rise in non-textual and dynamic online information (Lawrence & Giles, 1999).

More telling is the fact that typical content-indexing search engines fail to take into account the embedding context of an Internet resource. For example, a search engine cannot report how a resource was used, its juxtaposition to other resources, or its value within a particular context. A search engine also does not capture comments about a resource from its intended community of users. Such embedding context is often crucial in interpreting Web content. As argued by researchers in situated cognition, it may be impossible to separate knowledge from who knows it, and from that person's surrounding community of practice (Brown, Collins & Duguid, 1989; Brown & Duguid, 2000). As such, we believe that including such extrinsic information about a resource is critical in supporting effective discovery and re-use of resources for instructional purposes (Recker & Wiley, 2001).

In this paper, we describe an information filtering technique, collaborative information filtering, which holds the potential of addressing some of these problems. As we will describe, this approach can support the discovery of resources in a way that is sensitive to the context of users. Moreover, the approach supports automated, personalized filtering and recommendation of relevant resources for particular communities of users. Finally, the approach can provide a means for supporting community-building activities by automatically recommending like-minded users to one another for possible future collaboration.

In the next section of the paper, we provide an extensive review of the literature surrounding collaborative information filtering. We then describe an educational implementation of this approach, called Altered Vista. We report results of an empirical study, aimed at evaluating system usability and utility, as well as measuring the predictive accuracy of the recommender engine. We conclude with a discussion of our results and suggestions for future research.

COLLABORATIVE FILTERING AND RECOMMENDER SYSTEMS

Within the human-computer interaction (HCI) literature, a new paradigm of categorizing, collecting and filtering information has emerged, called collaborative information filtering. This approach is based on propagating word-of-mouth opinions and recommendations from trusted sources about the qualities of particular items (Malone, Grant, Turbak, Brobst & Cohen, 1987; Maltz & Ehrlich, 1995; Shardanand & Maes, 1995). For example, you've arrived in a brand new city, and hunger pangs have erupted. How do you make the all-important decision of where to

dine? You might consult restaurant guides, newspapers, or the phone book. More likely, you would ask friends with similar tastes in cuisine to recommend their favourite spots. In the end, you want trusted sources to provide you with information about the quality of restaurants in order to help you make the best selection.

This solution to the "restaurant problem" is the basic insight underlying research in collaborative information filtering. Collaborative filtering systems work by estimating the desirability of items under consideration. These estimates are generally made, unlike content-indexing search engines (e.g., a traditional Internet search engine), without any knowledge about the *content* of the items. Instead, systems attempt to infer the subjective "worth" or "desirability" of an item, as identified by those who use it. Once these preferences have been identified, they can be used to recommend resources to an individual user. As such, systems built on this approach are frequently called recommender systems because of their ability to provide automated, personalized recommendations of items to users (Resnick & Varian, 1997).

Collaborative information filtering systems have been implemented in a variety of domains, including recommending books, movies, research reports, and Usenet news articles (Resnick & Varian, 1997). These systems have become a staple element of e-commerce, as Internet vendors attempt to provide personalized recommendation of products to their customers. To the best of our knowledge, however, little work has been done in applying the approach within the educational domain.

It is important to note that collaborative filtering systems are most useful in situations and domains with the following characteristics:

1. The system can collect numerous data (e.g., ratings) about items, from many different users. In general, the accuracy of the predictions (called predictive accuracy) made by recommender engines increases as the data pool for estimating item desirability (either explicitly or implicitly) also increases. In short, you need lots of people with lots of preference information for lots of items.
2. Subjective or contextual information about items (e.g., tastes, preferences, opinions) are important decision aids.
3. Traditional information retrieval methods are less effective. This might be true in domains where tastes are important (e.g., restaurant selection), or where content-indexing is impractical or difficult (e.g., multimedia items).

There are several different approaches to collaborative filtering. While the specific techniques vary, all of them utilize the following steps:

1. Data gathering - through interacting with the system, a user builds a profile of his/her preferences by supplying opinions of different items. These opinions may be collected explicitly and/or implicitly.
2. Prediction/Recommendation – leveraging on the data supplied by some or all users, a user can request a prediction about the quality of an item. Alternatively, in response to a request for recommendations from a given user, predictions are made for all items not rated by the user, and the highest predicted ratings are presented as recommendations.
3. Algorithm Evaluation – as an ancillary step, the algorithm's speed, coverage, and accuracy are evaluated. It can be beneficial to pass these evaluations on to users.

In the following sections, we identify key issues within each phase of the process.

Data gathering

Collaborative filtering depends critically on gathering information about items in the domain. The more information known about people, and their preferences for various resources, the more accurate the system's predictions will be. The following discussion details the relationship between resources, people, and their preferences in the context of collaborative filtering.

Domain

A domain consists of items or resources and data about the population that uses them. The characteristics of the domain strongly influence data gathering and collaborative filtering strategies. Each domain has different characteristics (e.g., the number of items, the rate of introduction of new items, and the lifetime of items). These characteristics necessitate variations in how collaborative filtering is applied (Gupta, Digiovanni, Narita & Goldberg, 1999) and, to a certain extent, dictate whether or not it will even be a useful approach. One useful framework to assess the viability of collaborative filtering in a given domain is predictive utility.

Predictive utility, introduced by Konstan, et al. (1997) is a measure of how much influence predictions from a collaborative filtering system have on whether or not a user consumes an item. High predictive utility indicates a great deal of influence on consumption decisions and low predictive utility means the predictions will have little effect. The level of predictive utility is dependant upon the domain in which the recommender system is operating, and is a function of the value of the predictions, the cost of consuming items, and the ratio of desirable/undesirable items.

The ratio of desirable/undesirable items can have a large effect on predictive utility. If 99% of the items in a domain are desirable to the majority of the users, then making personalized predictions is generally not worthwhile—users could select an item at random and be fairly certain it will be useful. On the other hand, if a small fraction of the items are desirable, then the predictive utility should be high (assuming that the benefits of prediction outweigh the costs). For example, the domain of textbooks offers a great deal of predictive utility. There are many textbooks, and although reviewed and edited, the vast majority will not be desirable to most potential readers. The risks of consuming an undesirable book are high (the cost of the book and the time spent reading), which means that correctly rejecting a book offers a high level of benefit.

User profile data

The items in a domain are only half of the picture; the other half consists of the population or community that consumes these resources. The primary task of a collaborative filtering system is to somehow build a repository of the community's opinions about items by storing the individual's preferences and characteristics (e.g., demographic information) of its members. Preferences can be collected in one of two general ways: explicitly or implicitly.

Explicit data is solicited directly from users. Typically this takes the form of a Likert scale ranking or "vote" (Herlocker et al., 1999), but it may also involve anything from a binary

"like/dislike" to detailed annotations (Hill, Stead, Rosenstein & Furnas, 1995). In contrast, implicit data is collected from users by leveraging information collected for other purposes (Herlocker et al., 1999), mainly as a by-product of user actions. For example, the system might infer that desirable items are used more frequently or more recently (Recker & Pitkow, 1996).

Although implicit data is more easily collected and at less cost to the user, inferences about item desirability are much less accurate than explicitly supplied ratings. Hill et al. (1995) theorized explicit and implicit data collection represent two ends of a continuum, with a "sweet spot" somewhere in the middle. This "sweet spot" or optimal point lies where users benefit from "relatively more filtering value for relatively less filtering work" (p. 195). Hybrid content and collaborative filtering techniques may represent one possible step in this direction, but more research needs to be done (e.g., Billsus & Pazzani, 1998; Lieberman, van Dyke & Vivacqua, 1999).

Sparse matrix problem

It is useful to represent user profile data as a matrix of users and items within the domain, with the matrix values specifying user preference (if available) for a given item. As an example, Table 1 contains a matrix of information about teachers explicitly rating the quality of educational web sites on a scale of 1 (not worth the bandwidth) to 5 (extremely useful). The rows contain teachers, while the columns contain web resources.

Table 1
Teacher ratings for educational web sites on a five point scale

	Astronomy for Kids	**Bagheera: In the Wild**	**Learning Network**	**The GeoNet Game**	**Leonardo Home Page**
Bob	4		1	5	
Alice		5			
Mark	1	5	4		
Beatrice	1	5	4		

In most collaborative filtering applications, the matrix has much more missing information than the one represented above. Typically, users' preferences are known for only a small fraction of the items in the domain, even if they are being gathered implicitly (Karypis, 2000). This results in what is commonly called the "sparse matrix" problem. This problem can result from several sources: cold starts, first raters, and peculiar raters.

Regardless of the collaborative filtering approach, the application needs to know something about a new user's preferences before it can start to make recommendations. This lack of information results in a cold start situation in which users will have to supply a number of ratings before the system can begin to make predictions and recommendations (Maltz & Ehrlich, 1995). As noted by Gupta et al. (1999), without an initial set of ratings, additional ratings cannot by predicted. In Table 1, Alice faces the cold start problem. Since she has rated only one item, the amount of information is still not detailed enough for accurate predictions.

A related problem to new users is new items. Recently introduced items are in need of a first rater, since the system cannot make predictions about items if user preference data is non-existent (Konstan et al., 1997). Obviously, this difficulty is pronounced in domains (like the World Wide Web) which have a constant influx of new items. In Table 1, the "Leonardo Home Page" represents a web site that has not yet been rated by any of the teachers.

A final issue that relates mostly to a class of collaborative filtering approaches (discussed in more detail below) is the problem of peculiar users. A user whose opinions are relatively unique will not have any users that agree closely. As a result, the system will produce inaccurate recommendations (Balabanovi & Shoham, 1997). In Table 1, Bob is an example of a peculiar user. Of two users who rated "Astronomy for Kids", he was the only one who liked it. Bob has not yet rated "Bagheera: In the Wild", which the rest of the community seems to love uniformly, and his favorite web site, "The GeoNet Game", has not been rated by anyone else.

User anonymity and privacy

Not surprisingly, privacy is a big concern for users of collaborative filtering systems. This is especially true in educational arenas. In order to fuel algorithms, users may be asked to provide an extremely detailed level of information about who they are and what they like—which may make users wary, especially if they are aware that such data is passed on to a third party.

As much as anonymity can be used to protect the privacy of users of a collaborative filtering system, it can also hamper its efforts. For example, people who receive recommendations may express a level of distrust unless they know the identity of users providing recommendations. A way to bridge the gap between these competing interests needs to be devised. This could include any number of possibilities, ranging from a third party that mediates and certifies pseudonymous participation, to user selectable privacy levels.

Discouraging and encouraging participation

Even if a domain offers a high degree of predictive utility, a collaborative filtering system for that domain may still fail – due to a lack of data. Avery & Zeckhauser (1997) note that collaborative filtering encourages what they call "free riding" users, who tend not to rate resources. These users have learned they can avoid the burden of reviewing an item (which involves the risk of consuming an undesirable resource, and the opportunity cost of missing one which is desirable), if they simply wait for another user to rate the resource for them. Users who do provide preference data tend to be discouraged early on. Since they have to provide several ratings before they can receive accurate predictions, they often abandon the task as they see no clear reward for their hard work (Konstan et al., 1997). Finally, systems typically require users to be both producers and consumers of information when they often only want to take on a single role (Maltz & Ehrlich, 1995).

Avery & Zeckhauser (1997) suggest three potential tactics to increase the amount of user participation. The first tactic consists of supporting the producer/consumer role distinction through subscription fees. Information consumers pay a fee to producers for their efforts in providing preference data for items. Care would have to be taken to insure the producers are representative of the population as a whole.

The second, related approach is transaction-based compensation (Avery & Zeckhauser, 1997). The system rewards users who provide early preference data, and requires payment of those who wait. Payment in this scenario could be based on the usefulness of the preference data. One downside to such an approach is the further marginalizing of peculiar users. Users who tend to be peculiar themselves will receive much less of a reward since their "subscriber" base would be much smaller than a more mainstream reviewer.

Avery & Zeckhauser's (1997) final means of encouraging participation is exclusion, a seeming contradiction that may wind up being very effective. This incentive structure would deny a user predictions unless they are among the group of early raters for a certain number of items. As long as the users perceive a reward for their efforts, they will continue to meet the minimum ratings requirement.

Prediction/Recommendation

Collaborative filtering systems can perform at least two major tasks: prediction and recommendation. In the case of prediction, systems respond to a user's request to predict how much they would like a specific item. The systems may also recommend a set of items to the user (Karypis, 2000). This usually consists of a list of the items with the highest predicted value. Alternatively, the collaborative filter may perform both tasks. At the heart of deriving these predictions and recommendations is the algorithm driving the filter. These algorithms are generally grouped into two categories. The first, most prevalent class of algorithms, as noted by Herlocker et al. (1999), is termed neighbourhood-based and includes correlation (Herlocker et al., 1999), mean squared difference (Shardanand & Maes, 1995), and personality diagnosis (Pennock, Horvitz, Lawrence & Giles, 2000). The second category of algorithms, which includes Bayesian networks (Breese et al., 1998), is non-neighbourhood.

The Altered Vista system, described below, relies on a neighbourhood-based method. Such algorithms are concerned primarily with the relationships between users, and split the prediction/recommendation task into two distinct parts.

1. Neighbourhood identification - The collaborative filtering system identifies for each user, other users with similar profiles. This is called the active user's *neighbourhood*. User similarity is often computed by correlating every user rating. As such, it is a computationally expensive task. If the system recommends people as well as resources, then the set of recommended people will come from this neighbourhood.
2. Prediction/Recommendation – Once the neighbourhood has been formed, predictions can be made on a set of items which the user supplies by using some form of a weighted average of all the preference data provided by neighbourhood members. Alternatively, predictions can be made for all items unseen by the active user, and the top predictions presented as recommendations.

Algorithm Evaluation

A key issue in collaborative filtering is determining the quality of the predictions or recommendations from a given system. Evaluation can be considered along three dimensions: accuracy, coverage, and performance. There are two main approaches to calculating accuracy:

statistical accuracy, which considers predictions, and decision support accuracy, which focuses on utility of the recommendations (Herlocker et al., 1999). Coverage is a measure of how many predictions an algorithm is able to make with the available data. Performance is a measure of how many predictions a system can generate, usually measured in predictions (or recommendations) per second (Karypis, 2000).

Finally, there appears to be benefits in providing explanations of predictions or recommendations to users. If users do not know how a recommendation or prediction is made then they will not know what level of confidence to place in the suggestion. As they encounter instances of poor predictions due to one or more errors with no corresponding explanation, users may start to reduce their trust in the system. An easy solution to this problem is to provide users with some idea of how accurate the predictions are, thus giving them an idea of how much faith to place in the system (Hill et al., 1995). Beyond accuracy, they should have some level of understanding about how the predictions are made (Herlocker et al., 2000). While some work has been done (e.g., Swearingen & Sinha, 2001; Herlocker et al., 200) regarding exactly how predictive accuracy and explanation of algorithms should be described to end-users this is still an area for further study.

ALTERED VISTA: SYSTEM DESCRIPTION

In this section, we describe our implemented recommender system, called Altered Vista. In its current implementation, Altered Vista is specifically aimed at teachers and students who review web resources targeted at education. Using Altered Vista, users submit reviews about the quality and usefulness of Web resources for education. These ratings become part of the recommendation database. Users can then access and search the recommendations of other users. The user can also request personalized recommendations from the system. In this way, a user is able to leverage the opinions of others in order to locate relevant, quality information, while avoiding less useful sites. An additional benefit of this approach is that it allows a user to locate other users (e.g., students or instructors) that share similar interests for further communication and collaboration. Moreover, by using the system to rate Web resources users may be able to improve their information literacy skills. In particular, users may have more mindful and reflective engagement with Web-based resources.

Design considerations

When developing a collaborative filtering system that gathers explicit user opinions, several design dimensions must be considered (Resnick & Varian, 1997). These are 1) the ontology of the review or rating scheme, 2) how user data are collected, 3) how user data are aggregated, 4) how user data are used, and 5) the level of user anonymity. Each of these dimensions are discussed below.

Ontology

A fundamental issue in the design of a collaborative filtering system is defining the kinds of tags or preference data that users will supply in rating the items of interest. Because they are data

about data, they are a kind of metadata (LTSC, 2000; Weibel, 1995). Together, these tags typically comprise what is called a review or rating scheme.

Devising a scheme that is both usable and useful to a potentially wide community of users, rating a wide variety of resources, is a challenging problem. For example, the scheme can allow users to provide descriptions of the resource, such as its subject. The scheme can also enable users to specify their embedding context of resource use. Finally, the scheme can enable users to rate the overall quality of the resource. At the same time, the scheme must be devised in a way that it does not impose undue cognitive load upon its users.

We have adopted an approach where the scheme is specific to the domain under consideration. We have been experimenting with a variety of content labels within this scheme. Table 2 shows our current scheme for one domain (on-line education) implemented within Altered Vista.

Table 2
Review scheme for "on-line education"

Name	Description	Format
Web Site Title	The title of the site	Text box
Internet Address	The URL of the site	Text box
Keyword(s)		multiple selection list
Added by	User name	automatically generated
Overall Rating		5 point Likert scale
Navigation Ease	How easy is it to get around the site? Can you quickly get to major topics? Are the categories for the major topics distinct and intuitive? Is there some indication as to where you are within the web site?	5 point Likert scale
Accuracy of Information	Is the information on the web site correct? Is there a clearly identified organization, group, or author who is responsible for the site's content? If so, does that person or group have or rely on a body of knowledge that is respected in their field?	5 point Likert scale
Educational Relevance	How useful the site is for educators or their students.	5 point Likert scale
Description	Any information not represented in the other review criteria, as well as justification for any extremes. If the Navigation Ease is rated as poor, there should be some indication as to what is lacking (or confusing).	text box
Grade Level	What is the target audience for this site?	multiple selection list
Would you use this web site while teaching?	Can you picture yourself using this web site as part of your own instruction? (Either in class or as something students do outside of class).	5 point Likert scale

Collection

Altered Vista currently relies on explicit, active collection of information from users. To enter their metadata, users interact with a series of interface elements, including Likert scales, text entry boxes, and multiple selection lists.

A key issue faced in the design of a collaborative filtering system for Web resources is the identifying uniqueness and granularity of resources. For example, in the domain of movies, this is a relatively easy task. There are a small number of films, their boundary is easily defined, and the title is usually sufficient for unique identification. In the case of duplicate titles, a combination of title and production year can easily be used to disambiguate. For Web resources, the situation is more complicated. Although web pages also have unique identifiers (URLs) associated with them, a single web site often has several URLs associated with it (e.g., www.yahoo.com and www.yahoo.com/index.html). Thus, three users who rate the same web site may select three different URLs to identify the same resource. The difficulty is that even though they have all rated the same web site, the collaborative filtering system will store their reviews as if they rated three different sites.

Granularity of resource is also of concern. When reviewing a URL, are users rating the page, an inline graphic, or the entire Web site? Frequently, the nature of the specific resource makes the distinction clear. Other times, this difference is ambiguous.

In order to address and reduce this possible confusion, a "find-closest" approach is used. As a first step in the review process, users supply a title and URL for the site that they wish to review. The database of existing reviews is then searched to determine whether or not the site has been previously reviewed. If no matching URLs are found, the user is presented with eight URLs that are alphabetically adjacent to where the URL they supplied would fit into the database. Users are asked to verify that one of these eight URLs does not point to the resource they are attempting to review (see Figure 1 for an example).

Aggregation

Once a rating is complete, the user submits the review form and all values are stored in a database. This database of aggregated reviews becomes a mechanism that supports search and automated recommendation of resources.

User identity

To maximize the value of contributed information, we believe it is important to recognize that contributors are members of a community. Information about who contributed is as important as the contribution itself. Hence, users must log in prior to using the system, and the data that they input is not stored anonymously, as the identity of contributing authors provides important contextual information. In our current system, the email address of the author of particular ratings is searchable, available for inspection within search results, and provided when "people" recommendations are requested.

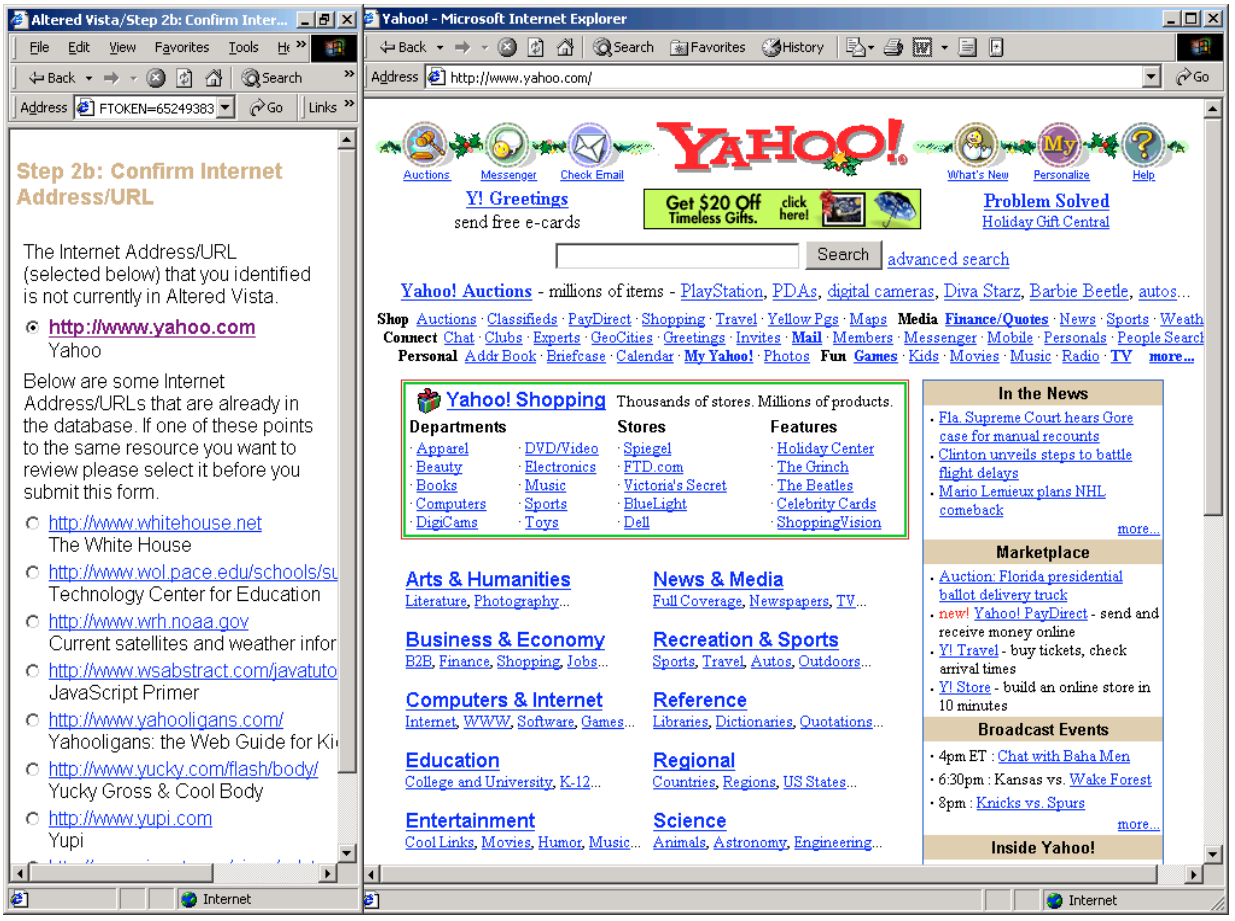

Fig. 1. "Find-closest" dialogue

Usage – searching

The scheme is used to support searching within a specific review area. A user can display all reviews, search by keyword, or search reviews by a specific contributor. Because our metadata schemes are searchable, they provide an alternate to content-indexing for discovering resources of interest.

Usage – recommendation

Upon user request, the aggregated database of user reviews can be used to provide automated, personalized recommendations of both resources and people. The recommendation algorithm relies upon a specific implementation of the neighbourhood-based approach to collaborative filtering (Herlocker et al., 1999).

When a user requests recommendations, the system first determines the *neighbourhood* for the current, active user. In a pair-wise fashion, the overall rating for resources provided by the

active user is correlated with all other users. To be considered, users must have mutually reviewed at least two resources. The set of users with high correlations (at least .50) between their set of overall ratings comprises the active user's neighbourhood.

Resources rated highly by users within a neighbourhood but unseen by the active user form the basis for automated recommendations. The system calculates a predicted rating for the unseen resource for the active user. This predicted rating is a weighted average of the ratings of users in the neighbourhood, and their correlation of agreement with the active user. The current system only recommends resources with a predicted rating greater than or equal to 4.00 (on a 5-point scale).

Members (and not just resources) of the neighbourhood can also be recommended to the active user. In this way, the active user can locate other users that share similar interests for further communication and collaboration.

The specific algorithm used for recommendation is as follows (bold text represents algorithm parameters; italic text represents output from the system).

1. Every morning at 3am, the system runs a batch process:
    a. For each user pair, determine if the pair has reviewed **2 or more** of the same resource.
    b. If they have:
        i. Run a Pearson correlation on the overall ratings supplied by each user for each resource to determine their level of agreement.
        ii. Store the correlation value in the database.
2. Upon a request for recommendation:
    a. Find all of the users who correlate with the "active" user at a level of **>= .50** (this represent this user's "neighbourhood").
        i. *Rank the neighbourhood in order of correlation and present as a list of recommended people.*
    b. Obtain a list of unique resources rated by the neighbourhood.
    c. Subtract any resources already rated by the active user from the list.
    d. For each remaining item:
        i. Retrieve the overall rating supplied by each of the neighbourhood members who reviewed it.
        ii. Use list of relevant neighbours to compute weighted average for prediction.

$$p_{a,i} = \bar{r}_a + \frac{\sum_{u=1}^{n} (r_{u,i} - \bar{r}_u) * \left(w_{a,u} * \left(o_{a,u} / 50 \right) \right)}{\sum_{u=1}^{n} w_{a,u} * \left(o_{a,u} / 50 \right)}$$

Equation 1 – weighted average formula.

1. $P_{a,i}$ = prediction for user a on item i
2. r_a = average ratings of user a
3. n = number of neighbours
4. $r_{u,i}$ = rating by neighbour u on item i
5. r_u = average ratings of neighbour u
6. $w_{a,u}$ = correlation between user a and neighbour u

7. $o_{a,u}$ = number of times user a and neighbour u have rated the same resource.

*e.*Order the list of resources by predicted value*, and display as a list of recommended web sites to the end user. Only display sites with a predicted value >= **4.00**.*

System interactions

Figure 2 shows a sketch of the system's architecture. All users must first log in to interact with the system. When users log in for the first time, they must complete a questionnaire form.

Altered Vista supports two primary user modes. First, the system administrator or teacher can define any number of review areas, and, for each area, any number of review tags in the review scheme. In our current system, we have defined one area, online education.

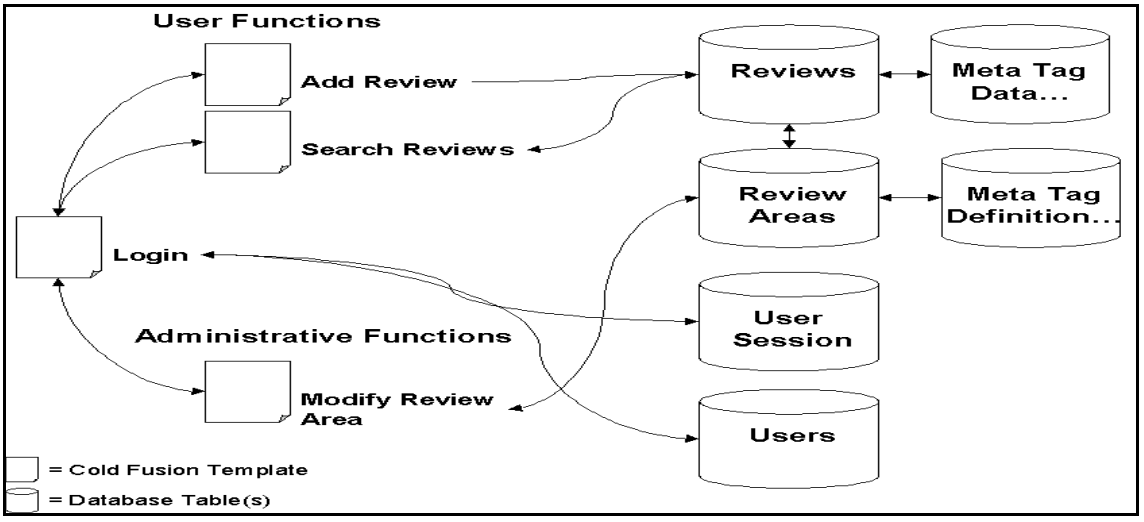

Fig. 2. System architecture

Second, once a review area has been defined, users log into the system, and select an area in which they will contribute reviews for particular Web resources. Figure 3 shows an example screen shot for entering a review. As can be seen, on one side of the screen, the user views the target Web site, while on the other side of the screen, the review of the site is entered using the pre-defined review tags.

System specifications

Altered Vista is implemented on a Linux server, running Apache for http services. Reviews are stored in a database, and communication between it and the server is accomplished using PHP. Users may access the system using any browser supporting Javascript (or VB Script) and Cascading Style Sheets. This system can be accessed at http://alteredvista.usu.edu.

These reviews are then stored in the Altered Vista database. Users can then search the reviews submitted by other users. Alternatively, as previously described, they can request

personalized recommendations of unseen Web resources. Figure 4 shows a screen shot of a composite review, based upon several user ratings for one resource.

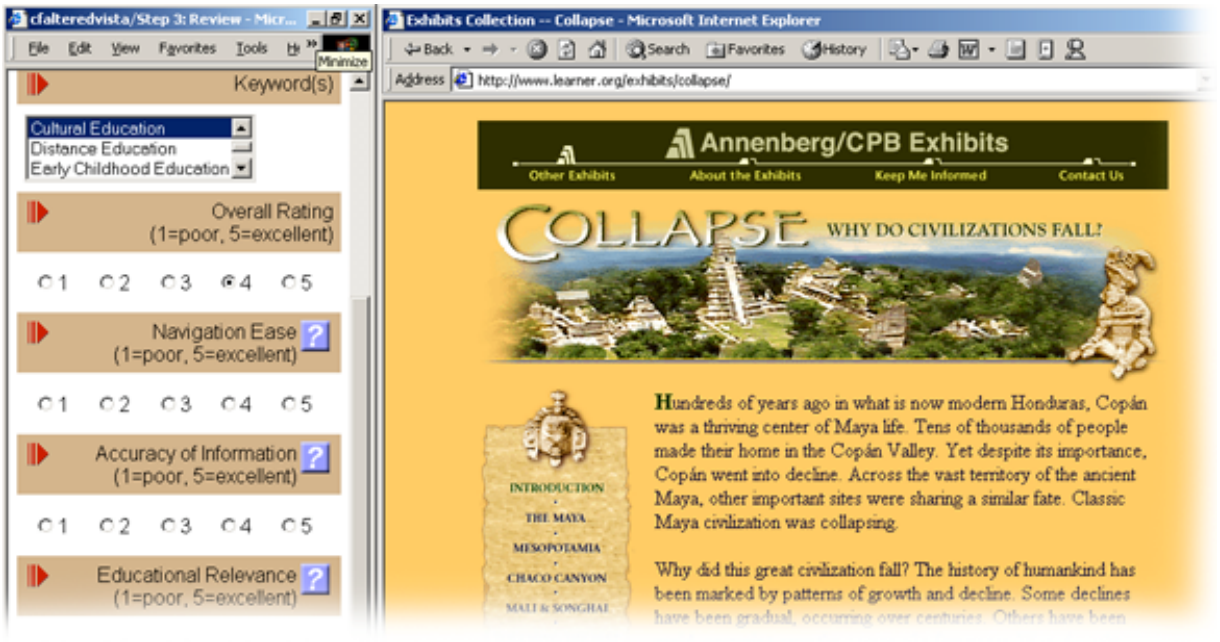

Fig. 3. Entering a review into Altered Vista

EMPIRICAL EVALUATION OF ALTERED VISTA

To evaluate the effectiveness and usefulness of the approach embodied in the Altered Vista system, we conducted a 3-month trial involving 85 students. The purpose of the study was to:

1. Evaluate system usability and performance. Specifically, are users able to use the system to submit reviews and to receive recommendations?
2. Evaluate predictive accuracy of the recommender engine. Are automated, personalized recommendations of resources and people useful and, if so, in what way?
3. Evaluate the extent that reviewing Web resources within a community of users supports and promotes collaborative and community-building activities.
4. Evaluate the extent that critical review of Web resources leads to improvements in user's information literacy skills. In particular, are users more mindful and reflective when engaging with Internet-based resources?

We first describe the study's methods and participants. We then present usage and user questionnaire results, followed by results summarizing performance of the collaborative filtering

algorithm. We close with an analysis of the predictive accuracy of the recommendations provided by the system.

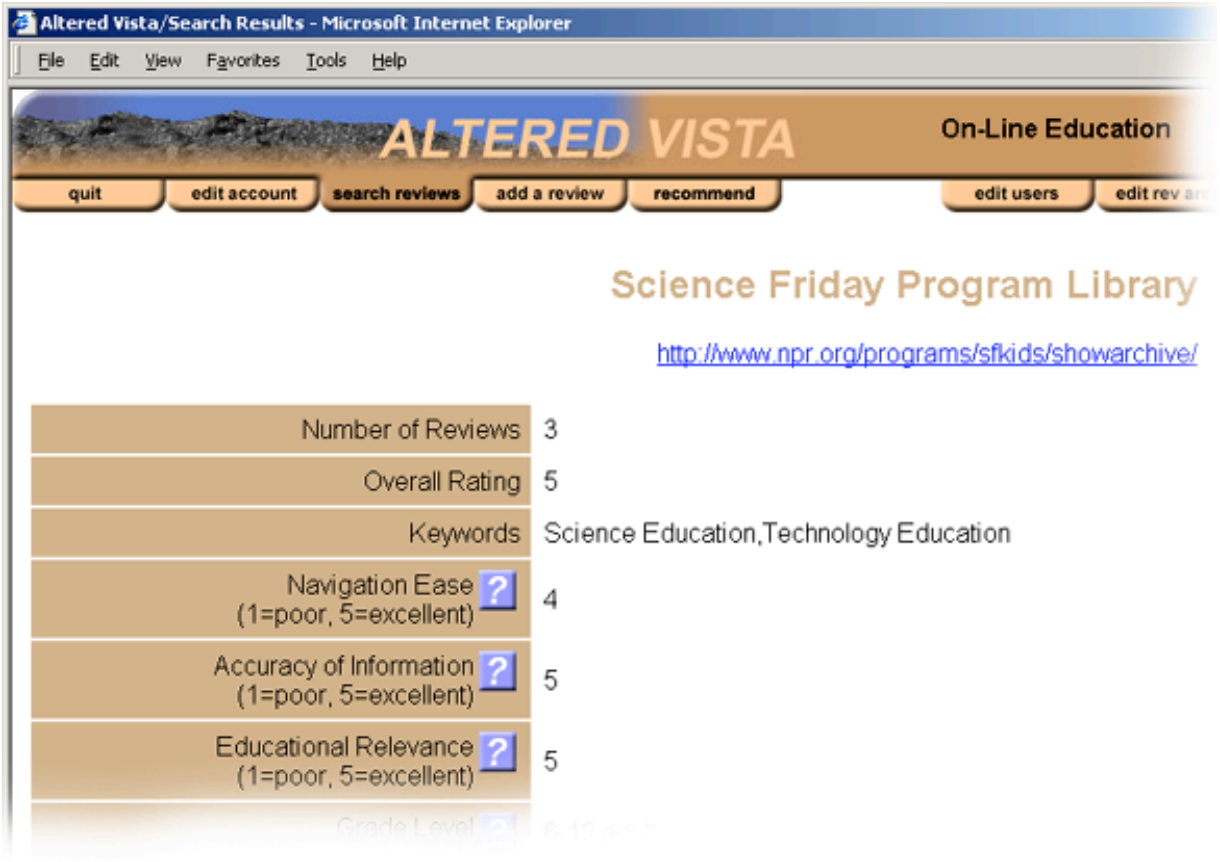

Fig. 4. A composite review

Participants and methods

Sixty three students (41% male and 59% female) from two U.S. universities participated in the study as part of course credit. As shown in Table 3, most participants comprised a mix of current classroom teachers taking additional professional development classes, and students preparing to become teachers.

In the context of the educational technology courses in which they were enrolled at their respective institutions, students were asked to use Altered Vista to review web resources related to "online education". Initially, students were asked to review five sites from a pre-selected list of Web resources. These sites were selected by an expert in online learning, and then reviewed by two additional experts. The list represented a broad, cross section of the type of resources that

teachers would typically encounter, and were intended to run the gamut of quality in terms of content, design, and overall utility.

Table 3
Participant background

Participant descriptor	Frequency
In-service	22 (35%)
Pre-service	19 (30%)
Other	9 (14%)
Religious education	10 (16%)
University instructor	3 (5%)

Participants were also asked to review five sites of their own choice. Finally, they were asked to review five sites reviewed by other users in the Altered Vista database. Thus, at a minimum, they were asked to contribute fifteen reviews during the course of the trial evaluation period. The goal was to ensure a critical mass of overlapping reviews in order to provide data to the recommendation algorithm.

Prior to using Altered Vista, all sixty three participants completed an online survey, which asked background information. At the end of the trial, students completed an exit survey that asked participants to rate the usability, usefulness, and accuracy of Altered Vista. Fifty two (82%) of the participants completed this exit survey. The surveys consisted of several 5-point Likert scale (1=strongly disagree; 5=strongly agree) and short answer questions. Comments were also collected from an online course bulletin board used by some participants.

Usage Results

Table 4
Usage results

Number of participants	63
Total number of reviews submitted	934
Mean time spent entering a review (seconds)	228
Total number of resources reviewed	242
Mean number of reviews per resource (St.Dev.)	3.86 (7.17)
Mean number of reviews submitted per user (St.Dev.)	14.83 (2.32)

As shown in Table 4, almost 1000 reviews were submitted for 242 unique Web resources. Of note is that the mean number of reviews submitted per user (14.83) is less than the required number of fifteen from the class assignment. Seven users did not meet the minimum requirements and were dropped from the remainder of the analysis. Resources received a mean

of 3.86 reviews, but their distribution is skewed. Figure 5 shows that most resources had a small total number of reviews, while a small number of resources received a high number. Indeed, as with many Internet datasets, the distribution follows Zipf's law (Heaps, 1978).

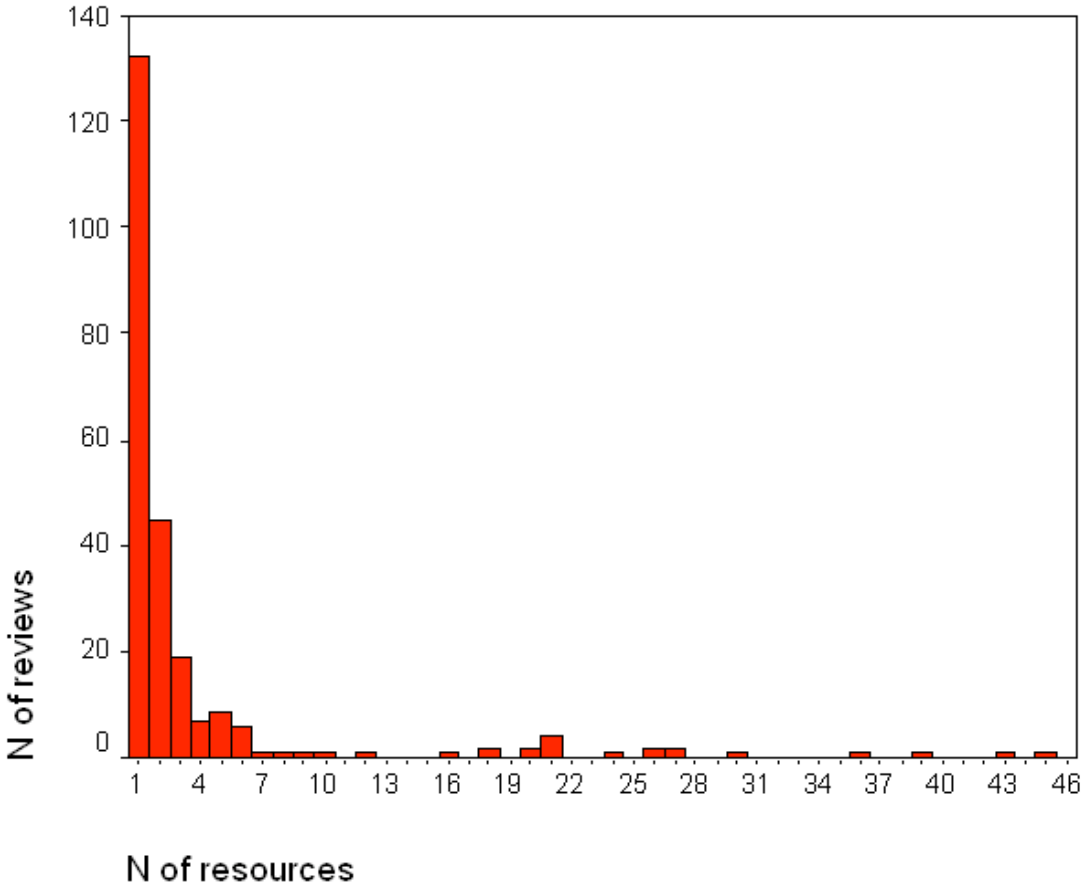

Fig. 5. Frequency of reviews per resource

Recall that users were asked to review resources using a number of Likert-scale dimensions, including the resources' "overall rating." Figure 6 shows the frequency of Likert-scale ratings for the "overall rating" category. Note the strong trend toward positive ratings of sites. This same trend is exhibited in both the initial list of pre-selected sites and the user-selected sites. Users appeared to have a positive opinion of the Web resources that they reviewed.

Recommender system performance

As previously described, the recommender algorithm employed by Altered Vista relies upon a neighbourhood-based method (Herlocker et al., 1999). This algorithm involves finding pairs of

users who agree by finding high correlations in the overall ratings that they provided for mutually reviewed resources.

The number of correlations computed between user pairs who reviewed two or more of the same Web resources is 1,473. This is reduced from the higher total number of user pairs (2,015). Figure 7 shows the frequency distribution of different correlation values between pairs of users. Results show high agreement about quality within this user group. This is probably due to the fact that, overall, resources received high ratings.

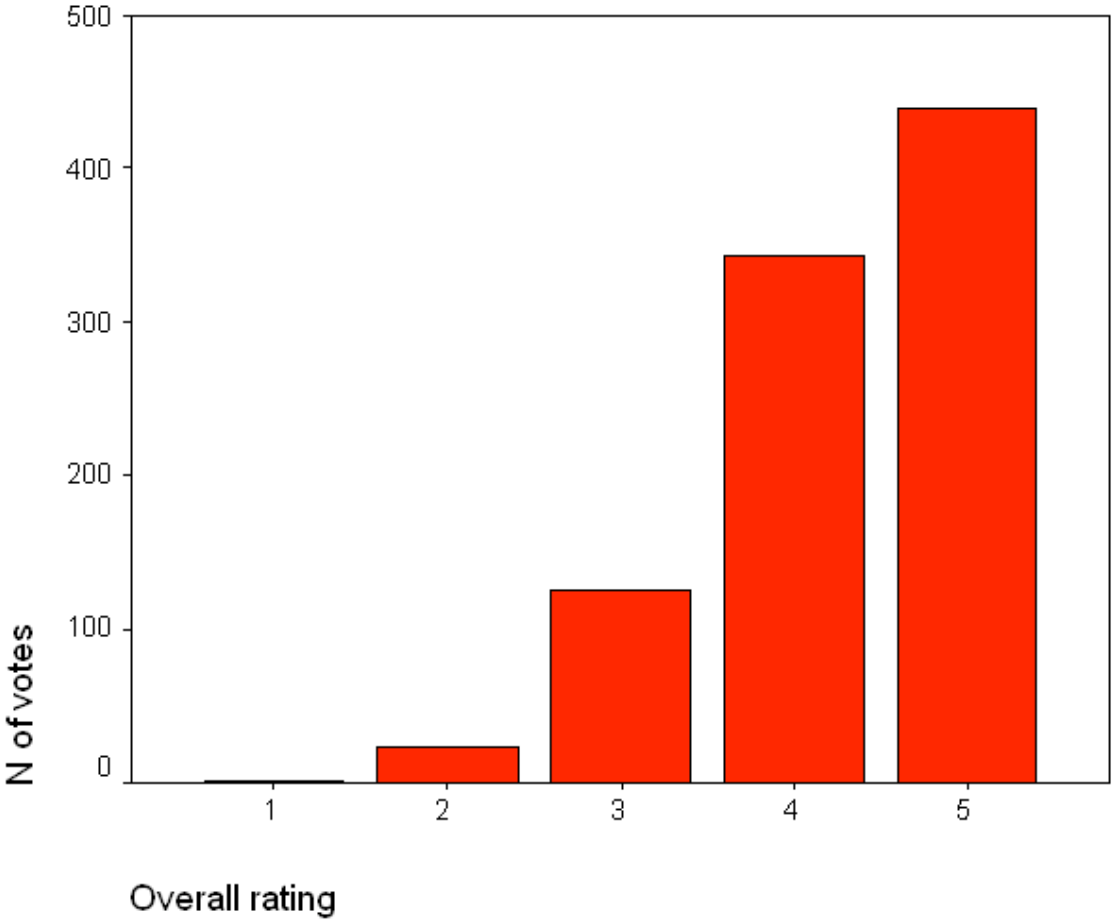

Fig. 6. Frequency of overall ratings

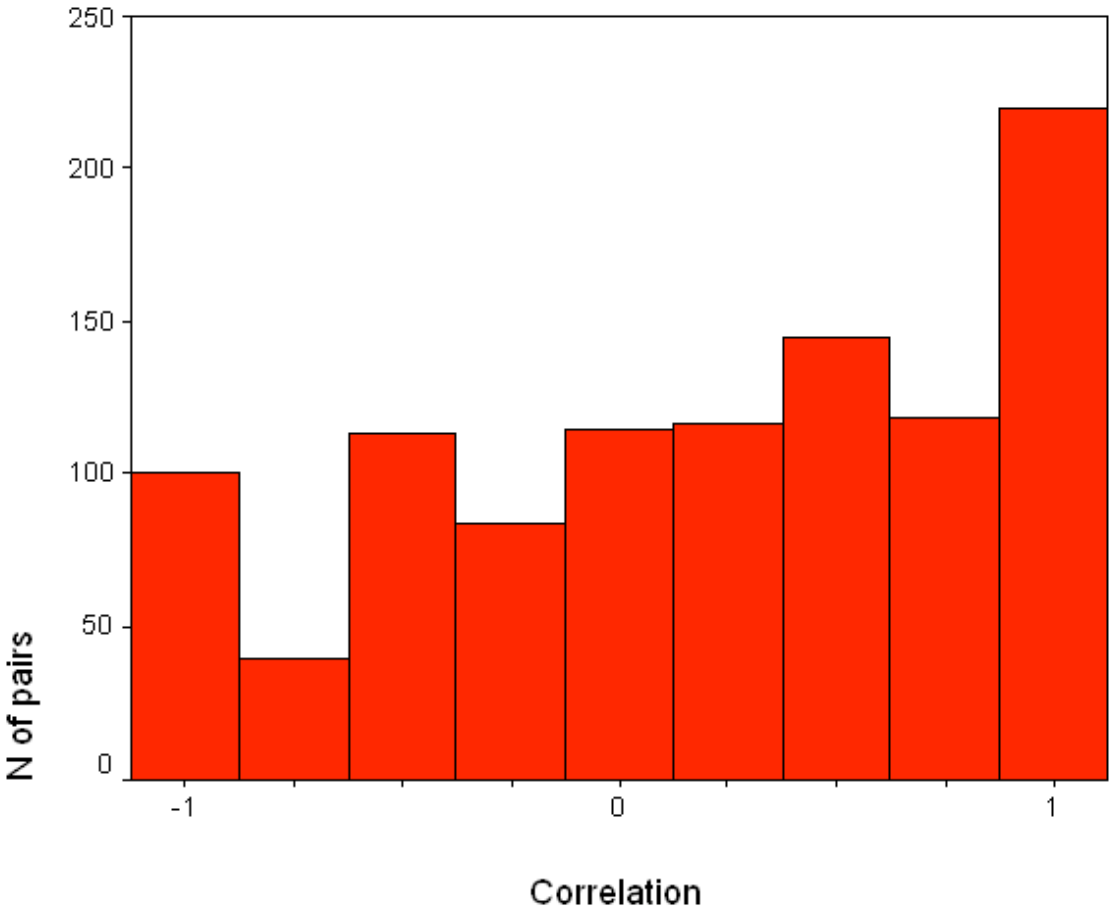

Fig. 7. Frequency of user pair correlation values

Table 5
Performance of the recommender engine

Mean number of recommended resources per user (St.Dev.)	46.56 (28.00)
Minimum	0
Maximum	76
Mean number of recommended people per user (St.Dev.)	16.57 (6.79)
Minimum	0
Maximum	31

In order to make recommendations for the active user, the user's neighbourhood must first be determined. The algorithm searches through the database to find users whose overall rating values correlate with the active user at least at the .50 level and have also reviewed at least two of the same resources. This neighbourhood becomes the list of recommended people, and a

weighted average of their favourite resources becomes the list of recommended items. As shown in Table 5, users received a mean number of approximately 46 recommended resource and 16 recommended people.

System usability and usefulness

After using the system, participants were asked to complete an online exit survey. Table 6 shows summary results from these surveys.

Table 6
Summary of exit survey results

Results from exit survey (Lickert scale: 1= strongly disagree; 5=strongly agree)	% rating 4 or 5	Mean	StDev
1.I found Altered Vista (AV) easy to use	80	3.89	1.14
2.AV usage made me think more deeply about Web design	74	3.96	.90
3.AV useful tool for finding quality resources	87	4.27	.89
4.AV useful tool for finding people with shared opinions	74	3.98	.85
5.AV provided me with useful recommendations of resources	65	3.85	1.02
6.AV provided me with useful recommendations of people with similar opinions	54	3.56	1.07
7.I would use AV even if it weren't a course requirement.	45	3.30	1.16
8.AV helped me find resources that I would otherwise not have found.	73	3.87	1.02
9.AV allows me to find and communicate with other professionals in my field that I would not normally have access to.	54	3.53	1.01
10.AV allowed me to see opinions about the quality of resources from people with different expertise.	36	3.51	1.05

As can be seen, a high percentage of respondents mostly or strongly agreed with the statement that Altered Vista was easy to use (Q1 in Table 6). Most respondents also thought that it was a useful tool for finding quality resources or resources that they would not have found (Q3 and Q8 in Table 6). Three quarters of the respondents also felt that the system allowed them to find people with similar opinions (Q4 in Table 6). As noted by one participant:

> *It takes a lot of time to evaluate websites and if there were a place where teachers could go to see evaluations already completed (and have a list of other sites that they may be interested in) it would save a lot of time in the long run. In this way teachers will find encouragement and resources that will help them integrate technology into their curriculum without having to reinvent the wheel. If teachers had a place to share there impressions about sites they had looked at, and all teachers had access to the data, just think of the work and time that could be saved. Especially if the data was searchable by grade level and subject.*

In terms of personalized recommendations, responses were more mixed (Q5 & 6). Respondents appeared to like the resource recommendation, but seemed unsure about the people recommender. However, respondents thought that this was an interesting feature:

> *This is a COOL feature!! Like [person x] mentioned, What a time saver this could be if all the teachers could have access to this kind of a system. What was fascinating to see how the top few people on my list responded almost exactly as I had done. Knowing this kind of a trend, I could then search through the sites they rated high in order to find some thing of interest to me, with very few exceptions.*

In addition, it appeared that many respondents felt that using Altered Vista helped them think more critically about Web design (Q2). Thus, weak evidence suggests that critical review of Web resources can lead to improvements in user's information literacy skills.

Three questions brought a greater diversity of opinion. Slightly less that a half of the respondents thought that they would use the system if it weren't a course requirement (Q7). This result highlights difficult issues relating to incentives for using the system, and sustained use. Lastly, respondents were less clear about the value of using Altered Vista to find and communicate with other users (Q9 & Q10). In the words of one respondent:

> *Usually recommendations are more valuable if the credentials of the recommender are known. Is there a way (besides guessing from what they say) to display expertise level of the recommender?*

Recommender system predictive accuracy

In this section, we report analyses evaluating the *predictive accuracy* of automated, personalized resource recommendation. In other words, are the recommendations provided to users worthwhile? A common way to measure this is called average absolute deviation (Breese et al., 1998) or mean average error (MAE) (Herlocker et al., 1999).

In this approach, users' actual ratings for resources are withheld from analyses. These withheld ratings are instead computed from the recommender data set, then compared to the actual user rating. Specifically, to calculate an MAE, the algorithm withholds one randomly selected rating from each user for the prediction set. The remaining items are given to the recommendation algorithm as a training set of the user's preferences. For each of the withheld ratings, the algorithm computes a prediction and the difference (or error) between the predicted rating and the user supplied rating is recorded. All of the error values are averaged across all predictions to come up with an MAE value. With good recommender system performance, the MAE is small. In addition, the MAE should be smaller than the difference between the overall average rating provided by users (called the popular average) and the actual rating. If this is not the case, the recommender engine should simply recommend on the basis of average user ratings of resources, and not bother with a computationally expensive collaborative filtering algorithm.

Breese et al. (1998) also advocate varying the size of the "training set" data to ascertain how well the algorithm performs with extensive (or limited) information. In particular, a varying number of actual ratings (e.g., 1, 5, or 10) can be withheld for prediction, and all remaining data

is reserved for the training set. In our analyses, only a single rating is withheld for prediction and all of the remaining data is given to the training set. In addition, only users who submitted ten or more reviews were used in this analysis (comprising 63 users).

Recall that the deployed system required a minimum of 2 overlapping reviews and a minimum correlation of .50 in overall rating for membership to the active user's neighbourhood. In the following analyses, we varied these thresholds and calculated the resulting MAEs. Specifically we varied the correlation threshold (from .50 to .90) and the minimum required number of overlapping reviews (from 2 to 5). As thresholds are increased, we expect better performance of the recommender engine, but a smaller number of possible predictions.

Table 7
MAE and coverage for various correlation and overlap thresholds

Overlap Threshold	Correlation threshold				
	0.50	0.60	0.70	0.80	0.90
2	.61 (37)	.70 (36)	.71 (36)	.73 (35)	.89 (33)
3	.60 (36)	.69 (33)	.69 (31)	.73 (30)	.91 (20)
4	.67 (32)	.74 (24)	.74 (21)	.81 (20)	1.23 (9)
5	.75 (23)	.77 (19)	.80 (12)	1.12 (11)	1.53 (5)

For each of these different thresholds, we ran the recommender engine to compute the overall MAE and the number of possible predictions made (called the coverage). Table 7 shows the resulting MAEs for various recommendation thresholds. The following results can be compared to baseline results used by computing the overall mean rating (called the popular average) provided by users. In our dataset, the popular MAE is .56, with 43 predictions out of a possible total of 54 predictions.

Not surprisingly, as the two thresholds are increased, the number of predictions that can be made decreases (reducing the coverage of the recommender engine). For all thresholds, the number of predictions (or coverage) for the correlation algorithms is always less than the number made using the simple popular ratings.

What is surprising is the uniform trend of increasing error as the correlation and overlap thresholds are increased. In theory, limiting the neighbourhood size to members with high levels of agreement, based on more data should increase the accuracy of predictions. This may be due to the strong positive bias towards all of the user supplied ratings.

Summary and discussion

Results from our exit surveys suggest that students found Altered Vista easy to use and a useful tool for finding quality resources and like-minded people. Respondents also felt that use of the system made them think more critically about website content and design. Its role in fostering community-building and collaboration is much less clear. Our users reported mixed opinions about the value of a system that automatically recommends potential (and usually unknown) collaborators.

Although the current study involved a fairly large number of people, two shortcomings affected the system's performance and results. First, we observed a strong trend toward positive ratings and thus high agreement among users about the quality of sites. Although other collaborative filtering datasets of user ratings occasionally exhibit a trend towards positive ratings (Konstan et al., 1997), ours was particularly marked. While it is unclear what caused this phenomenon, it certainly impacted the performance of the recommender engine: if everyone agrees, personalized ratings will in general add little extra benefit.

Because we collected user preference data along a number of dimensions (see Table 2), the recommender algorithm could use a multidimensional analysis to help improve the accuracy of personalized recommendations (and compensate for the ceiling effect in users' "overall" ratings). However, the Cronbach's alpha coefficient for our questionnaire data was high (.8124), suggesting that our scale measures a single unidimensional latent construct. As such, using other user preference data would not improve recommender accuracy.

Finally, because the Web has an essentially unbounded set of resources, our database suffered from a very low ratio of reviewers to resources. In the literature, this is called the "sparse-matrix" problem. As a result, the distribution of reviews for resources is severely skewed, and many resources had few reviews. Unfortunately, the recommendation algorithm relies upon a critical mass of both reviewers and resources for effective performance.

To verify that our lack of predictive power is a result of a skewed dataset, we obtained a publicly available dataset of user ratings of movie ratings, called MovieCritic (see http://research.compaq.com/SRC/eachmovie/). The ratings within this dataset were more normally distributed and less sparse. This dataset was run through the same algorithm used by Altered Vista, and its predictive accuracy was computed. This trial resulted in personalized predictions that outperformed the "popular" average – showing that personalized recommendations can be accurate with a suitable set of data.

CONCLUSION

In this paper, we described a system based on collaborative filtering applied in an educational setting, and presented results from an empirical study. In future system development, we envision several improvements. For example, the current recommendation algorithm relies exclusively on the "overall rating" category. However, we collect review data along a number of rating dimensions (or categories), and multivariate analyses of such data may help performance of the recommender engine. Similarly, we hypothesize that using elements of the user profile may also help recommender engine performance. For example, it seems possible that users with similar backgrounds (e.g., teachers of similar subject matter or grade level) may intrinsically have higher agreement in their ratings. Thus, we can calculate predicted ratings using a weighted average of profile agreement.

To help reduce the cognitive load of explicitly entering reviews, we wish to explore implicit rating methods. These methods will be based on collecting metrics about prior usage of the resource. In particular, in previous research, we showed that object desirability is strongly correlated to recency and frequency of prior object usage (Recker & Pitkow, 1996).

Our research also raises a number of important issues concerning the use of collaborative filtering in education, which are worthy of further study. The first issue is best exemplified in the following comment from a participant:

> *Reviewing web sights (sic) is not something I would do without some kind of motivation.*

People are loath to explicitly contribute reviews without some kind of incentive; hence it is difficult to seed and grow a review database. As a result, the review database can be sparse, impacting its reliability when searching for and recommending resources (Konstan et al., 1997). This is especially true if the user is an early contributor. In future trials, we must pay close attention to incentives in using the system, in order to have sustained and meaningful use within a community of users. Put more colloquially, how do we make Altered Vista a "sticky" site? One method for encouraging user participation might be through the use of profiles. If the "right" information can be gathered from short user surveys, then user neighbourhoods could be formed on the basis of this information—without new users having to take on the arduous task of rating several resources before seeing any benefits in the form of recommendations. A more extensive and normally distributed dataset will have to be collected before this hypothesis can be investigated reliably.

Second, our system raises important issues surrounding user privacy. Again, this is best summarized by a user comment

> *I found that the recommender listed 18 email addresses of people with my common ratings. It was interesting to see what others had researched, but I don't know if I would agree to having this information widely available on the web - would this be an additional open invitation for the invasion of my privacy - if there is such a thing on the web?*

While privacy issues are hardly unique to our system, it does point to a pressing concern in online environments. Specifically, we will need to examine the extent that anonymity of participation (or even pseudo-anonymity via a proxy) impacts user acceptance of the system and the recommendations it provides.

In the end, it may be that the Web is not an ideal environment for a collaborative filtering system. Because of its essentially unbounded and heterogeneous nature, it is difficult to overcome the sparse matrix problem. In addition, because of the wide variety resources on the Web (from large Web sites to small applets), resource "granularity" is hard to define. Instead, such systems may work best in a more constrained environment. Indeed, in current research, we are exploring the application of our approach within a digital library of educational resources (Recker & Wiley, 2001). Ultimately, this may prove to be a more suitable domain, because items in a digital library are easily itemized (and catalogued), and (hopefully) are used by a large number of people.

ACKNOWLEDGEMENTS

We thank the students who participated in our study. We also thank Jen Walker and Richard Cutler for helpful advice. This work reported in this paper was partially supported by a grant

from the National Science Foundation (NSF NSDL DUE-0085855) and a Utah State University Grant. Portions of this paper were presented at the 2001 Meeting of the American Education Research Association, Seattle, WA, USA.

REFERENCES

Avery, C., & Zeckhauser, R. (1997). Recommender systems for evaluating computer messages. *Communications of the ACM,* 40(3), 88-89.

Balabanovi, M., & Shoham, Y. (1997). Content-based collaborative recommendation. *Communications of the ACM,* 40(3), 66-72.

Billsus, D., & Pazzani, M. (1998). Learning collaborative information filters. *Proceedings of the Fifteenth International Conference on Machine Learning,* (pp. 46-54). Madison, WI: Morgan Kaufmann.

Breese, J., Heckerman, D., & Kadie, C. (1998). Empirical analysis of predictive algorithms for collaborative filtering. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*. Madison, WI: Morgan Kaufmann.

Brown, J.S., & Duguid, P. (2000). *The social life of information*. Cambridge, MA: Harvard Business School Press.

Brown, J.S., Collins, A., & Duguid, P. (1989). Situated cognition and the culture of learning. *Educational Researcher,* 18, 32-42.

Heaps, H. S. (1978). *Information Retrieval: Computational and Theoretical Aspects*.

Herlocker, J., Konstan, J., Borchers, A., & Riedl, J. (1999). An algorithmic framework for performing collaborative filtering. In *Proceedings of SIGIR'99* (pp. 230-237). ACM.

Gupta, D., Digiovanni, M., Narita, H., & Goldberg, K. (1999). Jester 2.0: Evaluation of an new linear time collaborative filtering algorithm. *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 291-292). New York, NY: ACM.

Hill, W., Stead, L., Rosenstein, M., & Furnas, G. (1995). Recommending and evaluating choices in a virtual community of use. *Conference Proceedings on Human Factors in Computing Systems* (pp. 194-201). New York, NY: ACM.

Karypis, G. (2000). Evaluation of Item-Based Top-N Recommendation Algorithms (Tech. Rep. No. 00-046). Minneapolis, MN: University of Minnesota, Department of Computer Science/Army HPC Research Center.

Konstan, J., Miller, B., Maltz, D., Herlocker, J., Gordon, L., and Riedl, J. (1997). GroupLens. *Communications of the ACM*, 40(3).

Lawrence, S., & Giles, L (1999). Accessibility and Distribution of Information on the Web. *Nature*, 400, 107-109.

Lieberman, H., van Dyke, N., & Vivacqua, A. (1999). Let's browse: A collaborative browsing agent. *Knowledge-Based Systems,* 12, 427-431.

LTSC. (2000). IEEE P1484.12 Learning Objects Metadata Working Group homepage [On-line]. Available: http://ltsc.ieee.org/wg12/index.html.

Malone, T., Grant, K., Turbak, F., Brobst, S., & Cohen, M. (1987). Intelligent information sharing systems. *Communications of the ACM,* 30(5).

Maltz, D., & Ehrlich, K. (1995). Pointing the way: Active collaborative filtering. *ACM Conference on Human Factors in Computing Systems* (pp. 202-209). New York, NY: ACM.

National Center for Education Statistics (2000). *Common Core of Data (CCD): School Years 1993-94 through 1997-98*. Washington D.C.: U.S. Department of Education.

Pennock, D., Horvitz, E., Lawrence, S., & Giles, C. (2000). Collaborative filtering by personality diagnosis: A hybrid memory and model-based approach. *Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence* (pp. 473-480). San Francisco, CA: Morgan Kaufmann.

Recker, M., & Pitkow, J. (1996). Predicting Document Access in Large, Multimedia Repositories. In *ACM Transactions on Computer-Human Interaction (ToCHI)*, 3(4), 352-375.

Recker, M., & Wiley, D. (2001). A non-authoritative educational metadata ontology for filtering and recommending learning objects. *Interactive Learning Environments.* 1, 1-17.

Resnick, P. & Varian, H. (Eds.) (1997). Recommender systems, Special Issue. *Communications of the ACM*, 40(3).

Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., & Riedl, J. (1994). GroupLens: An Open Architecture for Collaborative Filtering of Netnews. In *Proceedings of ACM 1994 Conference on Computer Supported Cooperative Work* (pp. 175-186). Chapel Hill, NC..

Robertson, J.S. (1999). The curse of plenty: Mathematics and the Internet. *Journal of Computers in Mathematics and Science Teaching*, 18(1), 3-5.

Rowand, C. (2000). Teacher use of computers and the Internet in public schools. *Education Statistics Quarterly* [http://nces.ed.gov/pubs2000/quarterly/summer/3elem].

Shardanand, U., & Maes, P. (1995). Social information filtering: Algorithms for automating word-of-mouth. *ACM Conference on Human Factors in Computing Systems* (pp. 210-215). New York, NY: ACM.

Swaim, M., & Swaim, S. (1999). Teacher time (or rather, the lack of it). *American Educator*, 23(3), 20-26.

Swearingen, K., & Sinha, R. (2001). Beyond algorithms: An HCI perspective on recommender systems. ACM SIGIR Workshop on Recommender Systems. New Orleans, LA: September 9-13.

Wallace, R., Kupperman, J., Krajcik, J., and Soloway, E. (2000). Science on the Web: Students online in a sixth-grade classroom. *Journal of the Learning Sciences*, 9(1), 75-104.

Weibel, S. (1995). Metadata: The foundations of resource description. *D-Lib Magazine*, July 1995.

White House Office of the Press Secretary (1997, April 19). Expanding access to Internet-based educational resources for children, teachers, and parents.