The International Review of Research in Open and Distributed Learning, Vol 13, No 5 (2012)

December – 2012

Contexts in a Paper Recommendation System with Collaborative Filtering

Pinata Winoto
Konkuk University, Korea

Tiffany Ya Tang
Kean University, USA

Gordon McCalla
University of Saskatchewan, Canada

Abstract

Making personalized paper recommendations to users in an educational domain is not a trivial task of simply matching users' interests with a paper topic. Therefore, we proposed a context-aware multidimensional paper recommendation system that considers additional user and paper features. Earlier experiments on experienced graduate students demonstrated the significance of this approach using modified collaborative filtering techniques. However, two key issues remain: (1) How would the modified filtering perform when target users are inexperienced undergraduate students who have a different pedagogical background and contextual information-seeking goals, such as task- and course-related goals, from those of graduate students?; (2) Should we combine graduates and undergraduates in the same pool, or should we separate them? We conducted two studies aimed at addressing these issues and they showed that (1) the system can be effectively used for inexperienced learners; (2) recommendations are less effective for different learning groups (with different pedagogical features and learning goals) than they are for the same learning groups. Based on the results obtained from these studies, we suggest several context-aware filtering techniques for different learning scenarios.

Keywords: E-learning; pedagogy

Introduction

A recommender system (RS) can follow the steps of its user, observe the interests of a group of similar users, and pick items that best suit the user based on either items the user liked (content-based filtering) or implicit observations of the user's followers/friends who have similar tastes (collaborative filtering, or CF; McNee et al., 2002; Herlocker, Konstan, Terveen, & Riedl, 2004; Lekakos & Giaglis, 2006). In the majority of these approaches, the successful match of the recommended item is measured by its utility, usually given a numerical rating by the user based on how much he or she liked the item (Adomavicius, Mobasher, Ricci, & Tuzhilin, 2011), a single-dimensional RS. However, users' preference for an item may be influenced by one or many contexts (Tang & McCalla, 2009; Winoto & Tang, 2010; Adomavicius et al., 2011). For instance, say a user is looking for a movie that is suitable for a fun family activity, such as a "family-friendly" movie. Contexts considered in a RS would vary depending on the applications (e.g., movies, books, music, education, etc.) and tasks the system intends to support (Gunawardana & Shani, 2009).

In the field of e-learning, a RS can help a tutor or learner to pick relevant courses, programs, or learning materials (books, articles, exams, etc.), and the contexts include the user's learning goals, background knowledge, motivation, and so on. These contextual attributes can be injected into the recommendation mechanism during either the prerecommendation or postrecommendation filtering process (Winoto & Tang, 2010; Adomavicius et al., 2011). A context-aware RS is referred to as a multidimensional RS.

Table 1 presents an example of a user rating matrix in single and multidimensional RSs for books. Here, target user John's rating of *The Da Vinci Code* can be predicted (the cell with "?") based on the ratings of those sharing similar interests with him.

Table 1
An Example of a Single-Dimensional (Top) and Multidimensional (Bottom) Book Ratings

	Jane Eyre	*Robinson Crusoe*	*Lord of the Flies*	*To Kill a Mockingbird*	*The Da Vinci Code*
Alice	8	7	6	8	5
Bob	6	7	6	6	6
Carol	6	5	7	7	6
John	5	5	8	?	?

	Jane Eyre			*Robinson Crusoe*			*Lord of the Flies*			*To Kill a Mockingbird*			*The Da Vinci Code*		
	story	style	impact	sto	sty	i	sto	sty	i	sto	sty	i	sto	sty	i
Alice	8	7	8	7	6	8	6	6	5	8	7	8	5	6	5

Alice	8	7	8	7	6	8	6	6	5	8	7	8	5	6	5
Bob	5	5	8	7	8	6	8	6	5	6	6	7	7	6	6
Carol	7	6	6	6	4	4	8	7	7	5	7	8	6	6	6
John	5	5	6	7	5	4	7	8	8	?	?	?	?	?	?

In the multidimensional RS, each book's overall rating is reflected by three sub-ratings: story, style (writing style), and impact. The overall rating, when represented by subtle and specific attributes, tends to be a more reliable predictor of whether users like the book or not. Excluding contextual considerations, Carol is the closest neighbor to John (Table 1, top); however, if we only consider the story of the books, Bob is his closest neighbor (Table 1, bottom). As such, in the first case the book *To Kill a Mockingbird* will be suggested to John (the predicted rating of the book is 7); while in the latter case, *The Da Vinci Code* will be recommended. From the situated cognition point of view, the context-aware perspective incorporated in multidimensional RSs can more appropriately capture human beings' information-seeking and cognitive behaviors (Rieh, 2002).

Tang and McCalla (2009) explored the factors that may result in a high rating for a paper in terms of its pedagogical benefits (whether the learner has gained knowledge from reading the paper will affect his or her rating). Results showed the importance of several pedagogical contexts in making a paper recommendation, particularly that learner interest is not the only dimension. Other contextual information-seeking goals, such as task- and course-related goals, are also important to learners' perceptions of the paper's value. Learners' willingness to peer-recommend a paper largely depends on how close the paper topic was to that of learners' own work. These observations can help tutors determine which items to select.

In this paper, we extend the research efforts of Tang and McCalla (2009) by performing two groups of experiments on both undergraduate and graduate students spanning a period of two years. The major reason we tested two distinct learner groups is twofold: First, it is known that RS performance is sensitive to users from different segments of the population, affected by factors such as demographic or socioeconomic status (Pazzani, 1999; Lekakos & Giaglis, 2006). Second, the two learner groups differed significantly in a number of key contexts which the recommendations should consider: pedagogical background, job-related experiences, learning goals, study practices, and so forth. Bernt and Bugbee (1993) pointed out that these contexts are good and persistent indicators of academic success in learning environments. In addition, people make judgments based on both the information and cognitive authority of the item (Rieh, 2002). The former is defined by the extent to which users think that the item is "useful, good, current and accurate," while the latter is the extent to which users think that they can trust the item (Rieh, 2002). Both of these judgment criteria motivated us to conduct two sets of studies on two pools of learners with different backgrounds. Specifically we report our findings from the comparative study of the two learner groups to address two key issues: (1) how would the recommenders perform when the target users are undergraduate students who have a different pedagogical background and contextual information-seeking goals, such as task- and course-related goals, from those of graduate students?; (2) Should we combine the pools of different learners into a single pool, combining graduate and undergraduate students into the same group, for example? Or should we separate them into different pools for the collaborative filtering? Based on these extensive studies, we further make recommendations on several context-aware techniques for different learning scenarios.

The organization of this paper is as follows: in the Related Works section we discuss the earlier efforts of researchers exploring contexts in (educational) recommendation systems. The Recommendation Techniques and Experiment Setup section provides the details of the modified CF techniques for educational paper recommendation and shows the recommendation flow of our system. The Experiment I section documents our first study (focused on experienced learners) and highlights the performance of the CF under various learning scenarios, while the Experiment II section introduces our second study (using inexperienced learners) and compares the experiment results with those in the first study. Then we provide a general discussion of our study and conclude by describing what lessons we learned from our research.

Related Works

To the best of our knowledge, no researchers have studied the cross recommendation of learning material among two or more groups of learners, and very few have studied the contextual attributes of educational RS. In this section we will discuss related work from two perspectives: context-aware RS and educational RS.

Context-Aware Recommender Systems

Adomavicius et al. (2011) argued that dimensions of contextual information can include when, how, and with whom the users will consume the recommended items, which therefore directly affects users' satisfaction about the system's performance. Pazzani (1999) studied a demographic based-CF which identified the neighbors of a target user and made recommendations accordingly. Lekakos and Giaglis (2006) considered users' lifestyles (their living and spending patterns) when making recommendations. Winoto and Tang (2010) studied a mood-aware recommendation approach that considered a user's mood to find a like-minded group for recommendation. In our study, the contexts are a learner's background knowledge and learning goals. Since undergraduate and graduate students have different background knowledge and goals, we expect their satisfaction with the recommendations will vary.

(Context-Aware) Educational Recommender Systems

Despite researchers' recent efforts to incorporate contexts into the recommendation process, the majority of early efforts in educational RS have been based on learners' interests. For example, Recker, Walker, and Lawless (2003) studied the recommendation of educational resources through Altered Vista, a system that enables teachers and learners to submit comments on the resources provided by learners who are precategorized into different "pedagogical" groups. Brusilovsky, Farzan, and Ahn (2005) reported on their user study of Knowledge Sea III, which provided "annotation-based" social navigation support for making personalized recommendations. McNee et al. (2002) investigated the adoption of CF techniques to recommend additional references for a specific research paper. A similar study conducted by Torres, McNee, Abel, Konstan, and Riedl (2004) utilized document titles and abstracts to make recommendations. Other recommendation studies made use of data mining to construct user profiles (Khribi, Jemni, & Nasraoui, 2009). These studies failed to consider whether the recommended paper is appropriate to support learning (goal-oriented RSs).

Recently, researchers have made efforts to identify and incorporate learners' pedagogical features (contexts) for recommendations. Nadolski et al. (2009) studied the effect of a learner's competence level, study time, and efforts on the performance of an educational RS. The contexts considered in Manouselis, Vuorikari, and Van Assche's study (2010) on recommending learning objects were similar to ours (learning goals, ease of use, etc.), although the target users were not students. Other similar efforts include Lemire, Boley, McGrath, and Ball (2005); Khribi et al. (2009); Gomez-Albarran and Jimenez-Diaz (2009); Manouselis et al. (2010); and Drachsler et al. (2007). Table 2 compares the contexts used in these studies with those used in ours.

Table 2

Various Studies in (Context-aware) Education RS

	Object to be recommended	Contexts	Evaluators (size)
Manouselis *et al.*, 2010	learning resources	Ease of use, facilitate learning, and topical relevance	Teachers
Khribi *et al.*, *2009*	Learning objects	Learner goals and learner needs	Simulations
Nadolski *et al.*, *2009*	Learning programs	Competence level, study time, and efforts	Simulation
Drachsler *et al.* 2007	Learning resources	Learner features, learner preference, demographic data, prior knowledge	Prototype only
Lemire *et al.*, 2005	Learning objects	Title, date, and author	Prototype only
Our work	Reading materials	Facilitate learning, topical relevance, popularity, and ease of understanding	Graduate (25) Undergraduate (45)

Recommendation Techniques and Experiment Setup

Pedagogical Paper Recommendation: Techniques

Since providing recommendation in a pedagogical context differs from doing so in other settings, we have modified the traditional techniques according to the characteristics of the learning domain. Broadly speaking, these characteristics are (1) the limited number of users, (2) the large number of unrated or new items, (3) the likelihood of the learners having difficulty understanding the items, and (4) the numerous purposes of the recommendation.

When there are a limited number of users and large numbers of unrated/new items, our RS cannot rely solely on rating-based CF (a cold start problem). Therefore, we considered a user model-based CF that does not need many learner ratings. We also took into consideration paper popularity (the average overall ratings of a paper) in an attempt to start the recommendation process when there were not many ratings in the system (this technique is largely considered to be nonpersonalized). Factors considered in our multidimensional CF were mainly used to correlate one user with another. Specifically, we considered the following factors: a paper's overall ratings, popularity, value-added, frequency of peer recommendation (or peer_rec), and learners' pedagogical features, such as interest and background knowledge. Overall rating represents the total rating given to a paper by a user (using a Likert scale of 1 to 4). Value-added represents the knowledge the user learned from a paper, and peer recommendation is defined as the user's willingness to recommend a paper to other learners.

Regarding the numerous purposes for recommendation, a tutor may aim for overall learner satisfaction (the highest possible overall rating), to stimulate learner interest only (the highest possible interest rating), or to help the learner gain new information only (the highest possible value-added rating), and so on. It is thus both appealing and imperative to collect multiple ratings and study multidimensional CFs that can utilize them. Table 3 categorizes the various recommendation techniques used in our studies, which generally fall into three main categories: content-based, CF-based, and hybrid recommendation.[1]

Table 3

A Summary of the Various Recommendation Techniques

Category	Name	Remarks
Content-based	ContentF	Content-based filtering
CF-based	1D-CF	Uni-dimensional rating-based CF
	3D-CF	Multidimensional rating-based CF
	UM-CF (2D-CF)	User model-based CF
	PopUMCF	A combination of non-personalized and UM-CF
Hybrid	PopCon2D	A combination of non-personalized, user item content filtering and 2D-CF

Nonpersonalized Recommendation (Benchmark)

Note here that we regard the inclusion of paper popularity as a nonpersonalized method. That is, this type of recommendation technique generates items based on a group of users tastes. We treat all of the students in the same class as a group. For this equation, the average rating of each paper, k, among all same-grade learners (denoted as rk), will represent the paper's popularity.

1D-CF

Unidimensional, rating-based CF is the traditional CF that has been used in the literature (Herlocker et al., 2004; Adomavicius et al., 2011). First, we calculated the Pearson correlation between users a and b, using the formula

$$P(a,b)=\frac{\sum_{k\in K}(r_{a,k}-\bar{r}_a)(r_{b,k}-\bar{r}_b)}{\sqrt{\sum_{k\in K}(r_{a,k}-\bar{r}_a)^2\sum_{k\in K}(r_{b,k}-\bar{r}_b)^2}} \quad (1)$$

where k is the rating by user i on item r, $\bar{r}_i$ is the mean rating by user i for all items, and k is the set of items co-rated by both a and b. The estimated rating target user a gave for a paper, j $r^*_{a,j}$ is then calculated with the target user's neighbors, denoted as B, using the following formula:

$$r^*_{a,j}=\frac{\sum_{B}P(a,b)\times r_{b,j}}{\sum_{B}P(a,b)} \quad (2)$$

In the learning domain, not many papers (less than 30 for each student in one semester) are commonly assigned as part of the learning activities in a course. Thus, our research focus is on a limited number of co-rated papers, so $|K| \leq 5$, and the number of neighbors |B| ranges from 2 to 15.

User-Model-Based Collaborative Filtering (UM-CF)

In rating-based CF, a target user needs to rate a few papers before we can find his or her neighbors, which is a major drawback especially at the beginning of each semester when ratings have not yet been provided by students. This issue has long been known as one type of a cold start problem, known as the new user problem (Schein, Popescul, Unger, & Pennock 2002). Fortunately, a user model-based CF (UM-CF) can be employed for users who previously have not rated any papers. The UM-CF extracts a user's interest and his or her background knowledge and injects these features into *Equation 1* in order to compare the similarity between each user. In other words, UM-CF compares users based on their interests and background knowledge, and makes recommendations accordingly.

Combinations of Nonpersonalized Recommendation and User Model-Based Collaborative Filtering (PopUM-CF)

PopUM-CF is a combination of UM-CF with the nonpersonalized recommendation method. It is used to overcome rating-based CF's reliance on co-rated papers.

Combinations of Content-Based Filtering, Nonpersonalized Recommendation, and User Model-Based Collaborative Filtering (PopCon2D)

Another hybrid method combines content-based filtering with nonpersonalized recommendation and user model-based CF, namely PopCon2D (it stands for popularity + content-based filtering + 2D user model-based CF). However, we normalized the closeness value by dividing each value with maxB (|closenessb|) so that our closeness value is always between -1 and 1.

The Recommendation System at a Glance

The proposed paper recommendation is achieved through a careful assessment and comparison of both learner and paper characteristics. In other words, each individual learner model will first be analyzed in terms not only of learner interest but also pedagogical features. Paper models will also be analyzed based on the topic, degree of peer recommendation, and so on. The recommendation is carried out by matching learner interest with the paper topics, with the goal of ensuring that the technical level of the paper should not impede the learner from understanding it. Therefore, the suitability of a paper for a learner is calculated by whether this paper is appropriate to help the learner in general.

When the tutor first initiates the system, she or he will be requested to fill in briefly the learner model, including learning goals, interest, knowledge background, and so forth. The tutor will then see the user interface listing an initial set of recommended articles that matches the learner's profile. Figure 1 illustrates the overall architecture of the system.

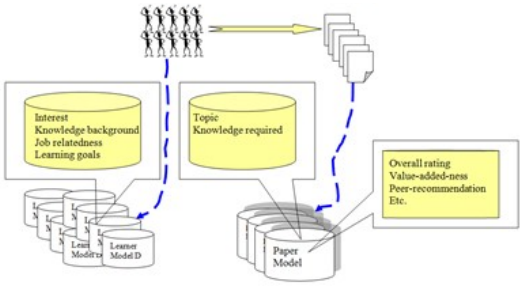


Figure 1. A closer look at the recommendation in our system.

The system consists of four main panes, showing the user model, paper model, paper rating, and recommendation (see Figure 2).

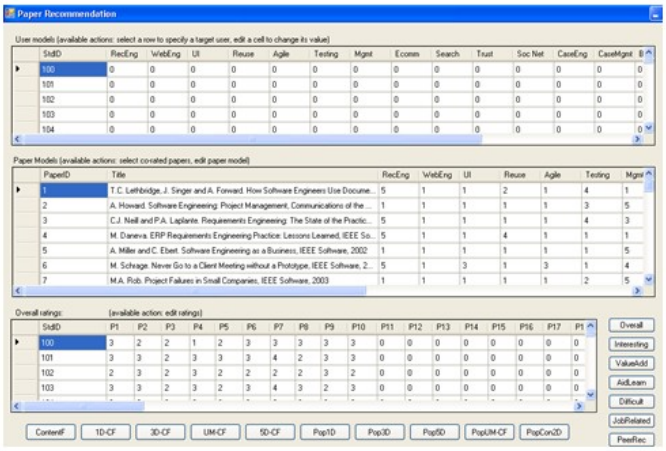
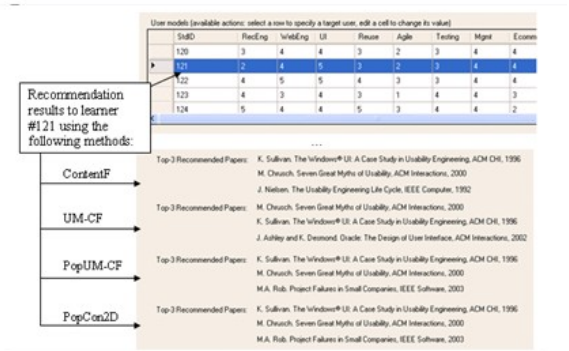


Figure 2. The initial system with ratings from previous users and paper models.

Figure 3 illustrates the results of an ensemble method, where we applied a weighted voting mechanism. The best three recommended papers from each applicable method (ContentF, UM-CF, PopUM-CF, and PopCon2D) are shown at the top of the figure, while at the bottom the calculation of users' weighted voting is visible. As a class proceeds, more papers will be rated by the learners, and therefore more methods can be used to make recommendations.



Paper by:	ContentF	UM-CF	PopUM-CF	PopCon2D	Total Vote
Sullivan	3	2	3	3	**11**
Chrusch	2	3	2	2	9
Nielsen	1	0	0	0	1
Ashley & Desmond	0	1	0	0	1
Rob	0	0	1	1	2

Figure 3. An illustration of an ensemble method.

As Figure 3 illustrates, the Sullivan paper is recommended to learner #121 after obtaining the highest votes from four applicable recommendation methods.

Experiment I

Data Collection

The first experimental study (hereafter Experiment I) was conducted in an introductory software engineering course for master's-level students. In total, 40 part-time students attended the course. Since in this pool students have at least

one year of working experience in the IT industry or an IT-related field, we considered them experienced learners. During the class, 22 papers were selected and assigned to students as their reading assignments according to the curriculum of the course, without considering the implications the choices had for our research. The number of papers assigned each week varied according to their length. In total, 24 students agreed to participate in this experiment.

At the beginning, learner profiles were drawn from a questionnaire consisting of four basic categories: interest, background knowledge, job nature, and learning expectation. After reading each paper, students were asked to fill in a paper feedback form evaluating several features of each paper, including how difficult it was to understand, its degree of job-relatedness for the user, how interesting it was, its degree of usefulness, its ability to expand the user's knowledge (value-added), and its overall rating (on a 4-point Likert scale). Using the collected ratings, we applied the recommendation techniques explained in the previous section to find the best one (top 1), the best three (top 3), and the best five (top 5) recommended papers for each target learner. Then we recorded the ratings given by the target learners to these recommended papers for our analysis.

Evaluation Metric

Evaluation protocols and methodologies should be designed to appropriately reflect the tasks that the RS supports and the users of the system (Herlocker et al., 2004; Gunawardana & Shani, 2009; Winoto & Tang, 2010). Our system is not intended for accurately predicting user ratings; instead, it helps learners understand the materials well in order to recommend the most appropriate papers. Therefore, our evaluation stresses learner acceptance rather than pure prediction accuracy (similar to those evaluations done by researchers in earlier works, including McNee et al., 2002, Torres et al., 2004, Brusilovsky et al., 2005, and Recker et al., 2003), which aligns with baseline measurement of a RS's quality (Herlocker et al., 2004).

Results

In some experiments involving CF methods, we chose 10 neighbors, selected after our analysis of a one-dimensional CF. Figure 4 shows some of the average overall ratings in 1D-CF with the number of neighbors, N, for the number of co-rated papers |K| equal to 2, 4, and 8. It also shows the average overall ratings for all combinations of |K| (from 2– to 15) and (top 1, top 3, top 5), labeled as "Total Average."

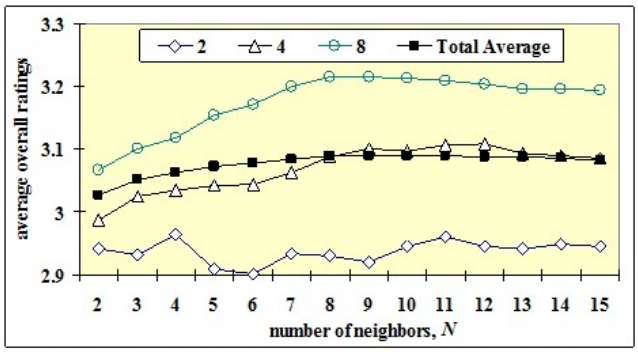


Figure 4. The average overall ratings with the number of neighbors used in 1D-CF.

When the number of co-rated papers is two, the performance is very unstable. The average ratings slightly decrease when the number of neighbors increases. This phenomenon indicates that pure CF-based methods rely on the quality of the neighbors and the denseness of the rating database. Therefore, the key issue is the similarity between the ratings. In our data, the total average reaches its maximum value when N = 10, which we used in our rating-based CF methods.

Discussion

The experiment's results are encouraging, especially because they confirm that making recommendations to learners is not the same as making recommendations to users in commercial environments such as Amazon.com. In learning environments, users are willing to accept items that are not interesting if they meet their learning goals in some way or another. For instance, our experimental results suggest that a user model-based CF works well with content-based filtering and nonpersonalized methods (such as paper popularity), such as PopCon2D. Although the computation required for PopCon2D is more sophisticated than other CF-based approaches, under certain circumstances it helps inform the recommender and therefore improve the recommendations.

Our findings illuminate learner satisfaction as a complicated element of learner characteristics, rather than a single issue of whether the paper topics matched learner interests. The results lead us to speculate that if there are a limited number of both papers and learners in the domain, considering other features, rather than relying on an overall rating and user interest, can help inform the recommendation. Table 4 summarizes our recommendations for adopting appropriate mechanisms based on various learning scenarios. Here, PopCon2D performs very well in three typical learning contexts for picking the single best paper, and the more complex PopUM-CF works well for making the best three recommendations. Due to its characteristics, PopCon2D can not only be used to start the recommendation but also to inform the recommendation (since it contains information such as paper popularity, paper content, user model of learner interest, and knowledge background, all of which can be used to generate recommendations without paper ratings). In dimensions such as this, with a limited number of both papers and learners (and other constraints, such as the course syllabus), we conclude that considering features other than just overall ratings and user interest can help inform the recommendation.

When the system does not have enough data on paper and user models, a content-based filtering method is appropriate because it matches the new user model and existing user and paper models. However, when there are not

enough papers to perform the matching, some other features, such as popularity, rj, need to be injected to inform the RS, as in PopCon2D and PopUM-CF. These methods define the features of pedagogical paper recommendation and reflect the reality that human judgments about scientific articles are influenced by a variety of factors, including a paper's topical content, its content appropriateness, and its value in helping users achieve their task (Custard & Sumner, 2005). They also highlight the importance of appropriately incorporating such factors into the recommendation process.

Table 4

A Summary of Suggested Recommendation Methods

Learning Scenario		Appropriate recommendation method(s)	
		Top one	Top three
When there are enough ratings and papers	The learner is new to the course	PopCon2D	PopUM-CF
	The learner is half way in the course	PopCon2D	PopUM-CF
When there are not enough ratings and papers	The learner is new to the course	PopCon2D	PopUM-CF

Experiment II

Experiment Setup and Evaluation Metrics

In this study (hereafter Experiment II) we collected data from 45 undergraduates who took a junior-level software engineering course. It is a mandatory course for full-time undergraduates in three majors: computer science, information technology, and a double-degree program of computer science and management. Most students claimed that they were inexperienced in practical software engineering when taking the course. Evaluation metrics were the same as those used in Experiment I in that we assessed recommendation performance in terms of three ratings: overall, aid_learning and value_added; however, we were more interested in ratings for both aid_learning and value_added because most undergraduates were inexperienced and this was their first software engineering course. Intuitively, undergraduates with less background knowledge may find more "new" information compared to the graduates. Hence, both aid_learning and value_added become more important recommendation goals in an undergraduate course.

Comparative Results and Discussion

The recommendation algorithms we used included PopUM-CF, PopCon2D, and 1D-CF, as suggested by Experiment I (Table 4).

Tables 5 to 7 compare the performance of PopUM-CF, PopCon2D, and 1D-CF for recommending the top 1 and top 3 papers to graduates and undergraduates. The values reported in the tables are the average ratings. The column "Grads" consists of the average ratings given by experienced learners (grads) to the recommended papers. The column "Undergrads" is divided into two columns: "From Grads" and "From UG", which denote whether the recommended papers were rated by experienced learners or inexperienced ones (undergraduates). We also included the best-case benchmark (recommending the most popular papers) after the solidus mark to compare whether the particular personalized recommendation is better (shown in bold font) or not. As we show in Table 7, all results from 1D-CF are significantly higher than those from the best-case benchmark, suggesting that when we have enough co-rated papers (at least eight), the recommendations for undergraduates work well. However, this is not always true for PopUM-CF (Table 5) or PopCon2D (Table 6), which means a pure, nonpersonalized recommendation may be useful in a cold start situation. It is not clear why both PopUM-CF and PopCon2D fail to provide a better outcome. One possible explanation is that most undergraduates, due to their lack of background knowledge, cannot specify their interests accurately; hence, adding personalized recommendations by matching less-accurate user models to papers or users could not improve the results.

Table 5

Average Ratings from PopUM-CF: Popularity Only

	Top 1			Top 3		
	Grads	Undergrads		Grads	Undergrads	
		From Grads	From UG		From Grads	From UG
Aid_learning	3.160	**3.178**/3.111	3.111/3.178	3.000	3.030/3.089	3.067/3.089
Value_added	3.280	3.244/3.267	**3.356**/3.267	3.240	3.170/3.185	**3.296**/3.244
Overall	3.160	2.933/3.022	3.044/3.044	3.093	2.933/2.970	**3.022**/2.978

Table 6

Average Ratings from PopCon2D: Popularity Only

	Top 1			Top 3		
	Grads	Undergrads		Grads	Undergrads	
		From Grads	From UG		From Grads	From UG
Aid_learning	3.160	2.978/3.111	2.933/3.178	3.040	2.963/3.089	3.089/3.089
Value_added	3.240	3.244/3.267	**3.311**/3.267	3.173	3.178/3.185	**3.259**/3.244
Overall	3.240	3.022/3.022	2.956/3.044	3.027	2.933/2.970	2.911/2.978

Table 7

Average Ratings from 1D-CF (Co-Rated Papers = 8): Popularity Only

	Top 1			Top 3		
	Grads	Undergrads		Grads	Undergrads	
		From Grads	From UG		From Grads	From UG
Aid_learning	3.170	**3.291**/3.111 (p = .015)	**3.344**/3.178 (p = 0.027)	3.075	**3.153**/3.089 (p = 0.086)	**3.184**/3.089 (p = 0.027)
Value_added	3.359	**3.428**/3.267 (p = .034)	**3.550**/3.267 (p = 0.001)	3.295	**3.328**/3.185 (p = 0.003)	**3.402**/3.244 (p = 0.001)
Overall	3.214	**3.252**/3.022 (p = .011)	**3.302**/3.044 (p = 0.002)	3.090	**3.109**/2.970 (p = 0.007)	**3.086**/2.978 (p = 0.028)

Another result indicated that the performances of PopUM-CF from graduates to undergraduates (column "From Grads") are always worse than those from undergraduates to other undergraduates (column "From UG") for both value_added and overall ratings (Table 5). The results suggest that collaborative filtering within the same group (from undergraduates to other undergraduates) works better than it does across groups (from graduates to undergraduates). This conclusion is supported by the results of 1D-CF in Table 7 where the recommendation across each group is mostly lower than those within each group (the only exceptional case is for the top 3 overall rating, where 3.086 < 3.109).

With respect to 1D-CF, we observed that most results were significantly higher than the best-case benchmark as shown by the low p-value (< 0.05), which means 1D-CF can provide effective recommendations. In fact, the results here are better than those for graduates in terms of the overall ratings, which had less significant gains (p = .33 and p = .29 for top 1 and top 3 respectively, see also Table 5).

General Discussion

Due to the limited number of students, papers, and other learning restrictions, a tutor cannot simply require students to read many papers in order to stock the database. As a result, the majority of typical recommender systems cannot work well in the pedagogical domain. Our study attempts to bridge the gap by proposing a set of recommendation mechanisms that do work well in this domain. Through extensive experimental studies, we discovered three key findings to answer our three broad research questions:

1. It is worth the trouble of complicating the traditional single-dimensional recommendation by incorporating contextual information to inform the recommendations. This can be achieved by adopting approaches such as PopUM-CF for a number of learning contexts.
2. The simple 1D-CF performs equally well for both graduates and undergraduates. Hence, paper recommendation systems can be effectively used for inexperienced learners along with experienced ones.
3. Gathering recommendations from across different learning groups (with different pedagogical features and learning goals) is less effective than it is to gather them from within the same learning groups, especially with collaborative filtering.

Our two studies suggest that learners make judgments based on the information and cognitive authority of a paper (Rieh, 2002). Hence, appropriately designing a RS and evaluating its performance is key to improving system performance. Experiment results suggest that a user model-based CF works well with some nonpersonalized methods, including PopCon2D and PopUM-CF. Although the computations in PopUM-CF and PopCon2D are more sophisticated than other CF-based approaches, under certain circumstances they help improve system performance. Our experiments and evaluation also highlight the importance of appropriately designing a RS and evaluating its performance. As Herlocker et al. (2004) declared, "accurate recommendations alone do not guarantee users of recommender systems an effective and satisfying experience. Instead, systems are useful to the extent that they help users complete their tasks."

This study attempts to bridge the gap between commercial recommendation systems and educational ones by proposing a set of recommendation mechanisms that work well in the learning domain. Through our experiments and prototypical analysis, we were able to draw a number of important conclusions regarding the design and evaluation of these techniques. In spite of this, more work needs to be done to further our understanding of this complex issue.

For instance, one of our biggest challenges was the difficulty of testing the effectiveness or appropriateness of a recommendation method due to a low number of available ratings. Testing the method with more students in two or three more semesters may not be helpful because the results would still not be enough to draw conclusions as strong as those found in other domains where there can be millions of ratings.

Hence, we are eager to see different institutions collaborate to use the system in a more distributed manner and on a larger scale (as it is very difficult to achieve accurate recommendations using only one class in one institution each time). Through this broader collaboration, our future work will include the design of a MovieLens-like benchmark

database to use as a test bed on which more algorithms can be investigated, including ours.

In addition, as shown in the analysis, the papers are related to software engineering; hence, it is not appropriate to generalize the results to make recommendations to students in other classes. Papers may exhibit more technical difficulties due to their inherent features in some subjects (e.g., in artificial intelligence or data mining), and students may also be different when they begin to take a course, which in turn affects those pedagogical factors considered in the performance of the recommender system. For instance, in user model-based CF, we intend to match user interest and background knowledge to a paper's topics, and the difficulty lies in making recommendations that ensure students have enough background knowledge to understand the paper. If, for example, the recommendation is made to students taking an artificial intelligence course, due to the nature of papers in this topic tutors should consider the technical difficulty of each paper carefully by giving more significant weight to paper difficulty (in the SE course, and due to the overall nature of the papers, the weight was suggested to be 0 after performance comparisons) to reflect the importance of that variable.

In this study, we only investigated making recommendations for research articles. Nevertheless, we believe that the study can be extended to other educational resources, such as learning objects, chapters with different topics in a digital book, tutorial materials, and so on. In fact, almost all educational resources can be regarded as learning objects with different granularity, situated environments, and purposes. Hence, the various recommendation mechanisms can be extended to make personalized recommendations for learning objects to individual learners with different needs.

Conclusion

Obviously, finding a "good" paper is not trivial: It is not as simple as finding out whether the user will either accept the recommended items or not; rather, it is a multiple step process that typically entails users navigating the paper collection, understanding the recommended items, seeing what others like/dislike, and making decisions. Therefore, a future research goal we derived from this study is to design RSs for different kinds of social navigation in order to study their impact on user behavior and how over time user behavior gives feedback to influence the system's performance. Additionally, we realized that one of the biggest challenges is the difficulty of testing the effectiveness or appropriateness of a recommendation method due to a low number of available ratings. Testing the method with more students in two or three more semesters may not be helpful because the results would still not be enough to draw conclusions as strong as those from other domains where there can be millions of ratings. Hence, we are eager to see the collaborations between different institutions to use the system in a more distributed and large-scale fashion. Through this broader collaboration, our ambition is to create a MovieLens-like benchmark database for the education domain to use as a test bed on which more algorithms can be investigated.

Acknowledgments

We would like to thank three anonymous reviewers for their constructive comments and the editors for their time and valuable remarks. This work was supported by Konkuk University in 2012.

References

Adomavicius, G., Mobasher, B., Ricci, F., & Tuzhilin, A. (2011). Context-aware recommender systems. *AI Magazine, 32*(3), 67–80.

Bernt, F. L., & Bugbee, A. C. (1993). Study practices and attitudes related to academic success in a distance learning programme. *Distance Education, 4*(1), 97–112.

Brusilovsky, P., Farzan, R., & Ahn, J. (2005). Comprehensive personalized information access in an educational digital library. *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries* (pp. 9–18). New York, NY: ACM.

Custard, M., & Sumner, T. (2005). Using machine learning to support quality judgment. *D-Lib Magazine, 11*(10). Retrieved from http://www.dlib.org/dlib/october05/custard/10custard.html

Drachsler, H., Hummel, H. G. K., & Koper, R. (2007). Recommendations for learners are different: Applying memory-based recommender system techniques to lifelong learning. In E. Duval, R. Klamma, & M. Wolper (Eds.), *Creating new learning experiences on a global scale: Second European Conference on Enhanced Technology Learning, EC-TEL* (pp.1-9). Crete, Greece.

Gomez-Albarran, M., & Jimenez-Diaz, G. (2009). Recommendation and students' authoring in repositories of learning objects: A case-based reasoning approach. *International Journal of Emerging Technologies in Learning, 4*, 35–40.

Gunawardana, A., & Shani, G. (2009). A survey of accuracy evaluation metrics of recommendation tasks. *Journal of Machine Learning Research, 10*, 2935–2962.

Herlocker, J., Konstan, J., Terveen, L., & Riedl, J. (2004). Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems (TOIS), 22*(1), 5–53.

Khribi, M. K., Jemni, M., & Nasraoui, O. (2009). Automatic recommendations for e-learning personalization based on web usage mining techniques and information retrieval. *Educational Technology & Society, 12*(4), 30–42.

Lekakos, G., & Giaglis, G. (2006). Improving the prediction accuracy of recommendation algorithms: Approaches anchored on human factors. *Interacting with Computers, 18*(3), 410–431.

Lemire, D., Boley, H., McGrath, S., & Ball, M. (2005). Collaborative filtering and inference rules for context-aware learning object recommendation. *International Journal of Interactive Technology and Smart Education, 2*, 179–188.

Manouselis, N., Vuorikari, R., & Van Assche, F. (2010). Collaborative recommendation of e-learning resources: An experimental investigation. *Journal of Computer Assisted Learning, 26*(4), 227–242.

McNee, S., Albert, I., Cosley, D., Gopalkrishnan, P., Lam, S., Rashid, A., Konstan, J., & Riedl, J. (2002). On the recommending of citations for research papers. *Proceedings of the 2002 ACM Conference on Computer Supported Cooperative Work, New Orleans, Louisiana* (pp. 116–125). New York, NY: ACM.

Nadolski, R., Van den Berg, B., Berlanga, A., Drachsler, H., Hummel, H., Koper, R., & Sloep, P. (2009). Simulating lightweight personalised recommender systems in learning networks: A case for pedagogy-oriented and rating based hybrid recommendation strategies. *Journal of Artificial Societies and Social Simulation, 12*(14). Retrieved from http://jasss.soc.surrey.ac.uk/12/1/4.html

Pazzani, M. (1999). A framework for collaborative, content-based, and demographic filtering. *Artificial Intelligence Review, 1*(5–6), 393–408.

Recker, M., Walker, A., & Lawless, K. (2003). What do you recommend? Implementation and analyses of collaborative information filtering of web resources for education. *Instructional Science, 31*, 299–316.

Rieh, S. Y. (2002). Judgment of information quality and cognitive authority in the web. *Journal of the American Society for Information Science and Technology, 53*(2), 145–161.

Schein, A., Popescul, A., Unger, L. H., & Pennock, D. (2002). Methods and metrics for cold-start recommendations. *SIGIR '02, Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 253–260). New York, NY: ACM.

Tang, T. Y. (2008). *The design and study of pedagogical paper recommendation* (Doctoral dissertation). University of Saskatchewan, Saskatoon.

Tang, T. Y., & McCalla, G. (2009). A multi-dimensional paper recommender: Experiments and evaluation. *IEEE Internet Computing, 13*(4), 34–41.

Torres, R., McNee, M., Abel, M., Konstan, J. A., & Riedl, J. (2004). Enhancing digital libraries with TechLens. *The Fourth ACM/IEEE Joint Conference on Digital Libraries (JCDL 2004)* (pp. 228–236).

Winoto, P., & Tang, T. Y. (2010). The role of user mood in movie recommendations. *Expert Systems with Applications, 37*(8), 6086–6092.

---

[1] Interested readers can refer to Tang (2008) for details on these algorithms.