

The Pedagogical Value of Papers: a Collaborative-Filtering based Paper Recommender

Tiffany Y. Tang¹, and Gordon McCalla²

¹ *Department of Computing, Hong Kong Polytechnic University, Hong Kong*

² *Department of Computer Science, University of Saskatchewan, Saskatoon, Canada*

{cstiffany@comp.polyu.edu.hk; mccalla@cs.usask.ca}

Abstract. In this paper we discuss the pedagogical features necessary to make appropriate recommendations of papers to students in an e-learning domain. Analyzing data collected in a human subject study several characteristics of learners and of papers are found that are important to making good recommendations. These pedagogical features distinguish e-learning domains from many commercial domains where the only key factor is a user's likes and dislikes.

1. Introduction

Although there are some active research focusing on making paper recommendations (McNee *et al.* 2002, Torres *et al.* 2004), unfortunately, this research does not consider pedagogical factors when making recommendations, that is, whether or not a recommended paper will enhance learning. To deal with this issue, we proposed the notion of recommending pedagogically appropriate papers (Tang and McCalla 2005, Tang 2008). We argued that learners' overall impression towards each paper is not solely dependent on the interestingness of the paper, but also other factors, such as the degree that the paper that help to meet their 'cognitive' goals. Unlike other kinds of users, learners are willing to accept items that are not interesting, yet meet their learning goals in some way or another. This paper extends our previous work by focusing on determining the extra value of making pedagogically relevant recommendations in an e-learning system. The research is substantiated through a human-subject experiment. The full space of user evaluation is more complex than that of the most previous works, since we set our paper recommender in the context to support learners' learning pedagogically (for instance, increase their knowledge). Particularly, we split our evaluation space into two key parts: the first is to interpret the significance of the pedagogical factors in making recommendation; the second is to explore the associations among these pedagogical factors in order to understand the interactive relationships among the pedagogical factors as we believe that learner satisfaction is a complex function of learner characteristics, rather than the single topicality of a paper as matched against their interest.

The rest of the paper is organized as follows. In section 2, we will outline previous studies. The multi-dimensional paper recommender will be introduced in section 3, through a comparison between the traditional Recommender System (RS) and the proposed RS to motivate the work described in later sections. The empirical study set-up will be presented in section 4. Section 5 and 6 give out a detailed interpretation of the evaluation results as well as the proposed recommendation algorithm, while section 7 concludes this paper.

2. Related Work

2.1 Paper Recommender

There are several related works concerning tracking and recommending technical papers. Basu *et al.* (1998) defined the paper recommendation problem as: "Given a representation of my interests, find me relevant papers." They study this issue in the context of assigning conference paper submissions to reviewing committee members. (Bollacker *et al.* 1999) refined CiteSeer, NEC's digital library for scientific literature, through an automatic personalized paper-tracking module which retrieves user interests from well-maintained heterogeneous user profiles. (Woodruff *et al.* 2000) discussed an enhanced digital book with a spreading-activation-gear mechanism to make customized recommendations for readers with different type of background and knowledge. (McNee *et al.* 2002) investigated the adoption of collaborative filtering techniques to recommend papers for researchers; however, the paper did not address the issue of how to recommend a research paper, rather, how to recommend additional references for a target research paper. (Recker *et al.* 2003) studies the pedagogical characteristics of a web-based resource through Altered Vista, where teachers and learners can submit and review comments provided by learners. However, although they emphasize the importance of the pedagogical features of these educational resources, they do not consider the pedagogical features in making recommendation.

These works are different from ours in that we not only recommend papers according to learners' interests, but also pick up those not-so-interesting-yet-pedagogically-suitable papers for them. In some cases pedagogically valuable papers might not be interesting and papers with significant influence on the research community might not be pedagogically suitable for learners.

2.2 Multi-dimensional Recommendation

The majority of RSs make recommendation purely based on item-item, user-user and/or item-user correlations without considering the contextual information where the decision making happens, to name a few (Lekakos and Giaglis 2006; Konstan *et al.* 1997). For instance, a recommender system only matches the interests of a target user with other users (e.g. in terms of the correlation of their previous rating patterns to various items). Consider an example from the e-learning domain. Student Steven's job is not related to UI design, but he found out that a paper on UI design and usability engineering is useful in understanding his Software Engineering course; hence, he still rates this paper highly. His rating on the "usefulness" of this paper thus reflects the pedagogical value of it for those taking a Software Engineering course. (Adomavicius *et al.* 2005) argue that the dimensions of contextual information can include when, how and with whom the users will consume the recommended items, which, therefore, directly affect users' satisfaction towards the system performance. To deal with the multi-dimensional CF, they propose to use data warehouse and On-Line Analytic Processing (OLAP) application concepts in slicing an available database. (Manouselis and Costopoulou 2007) takes a similar look at the RS in e-commerce domain: a utility-based multi-dimensional CF. by treating each feature separately before synthesizing them to maximize the utility of the system. It is unclear though how the utility is achieved since the evaluation is conducted based on two of the most typical metrics over the performance of the RS: *accuracy* and *coverage* (Herlocker *et al.* 1999). (Recker *et al.* 2003) also studied an earlier 'version' of multi-dimensional collaborative filtering (CF) through the aggregation of users' demographic information such as their gender, age, education, address, etc. In order to make predictions to a target user, the demographic based-CF learns a relationship between each item and the type of the people who tend to like it. Then, out of 'that' type of people, the CF identifies the neighbors for the target user, and makes recommendations accordingly. The difference between traditional CF and demographic based CF is this preprocessing step of 'grouping' similar users. (Manouselis *et al.* 2007) discussed a multi-criteria CF for learning object recommendation; for each dimension of the feature, the system generates a set of user neighborhood for the target user. Then the recommendation is made in different dimension separately. (Lemire *et al.* 2005) propose another interesting work in considering the metadata of a learning object (such as title, date, author) and its rating in making recommendations. However, these works are different from ours in that we integrate the multiple features in making recommendation based on the tasks for which the RS is intended to support. In addition, our proposed multi-dimensional CF is not a utility-based CF as (Manouselis and Costopoulou 2007), since we believe that it is neither necessary nor possible to define the utility of learners, their learning experiences and thus make it impractical and inappropriate to evaluate the performance of the system (Herlocker *et al.* 2004, Winoto and Tang 2008).

A recent effort in incorporating context information in making recommendations is a study by Lekakos and Giaglis (2006), in which users' lifestyle is considered. Lifestyle includes users' living and spending patterns, which are in turn affected by external factors (e.g. culture and family) and internal factors (e.g. personality, emotions, and attitudes). In order to obtain their lifestyle information, users are exposed to a number of advertisements picked up from seven product categories such as food and drink, books, etc. The system will then compute the Pearson correlation of users' lifestyles to relate one user to another. After this filtering process, the system will make predictions on items for the target user based on ratings from neighbors. Most recently, (Adomavicius and YoungOk 2007) propose a multi-criteria recommendation which is capable of considering the multiple rating criteria for a movie in Yahoo!Movie. That is, the overall rating of each movie in Yahoo! Movies reflects its four main features¹: Story, Acting, Direction and Visuals. The major difference between (Adomavicius *et al.* 2005) and (Adomavicius and YoungOk 2007) is that the latter considers the numerical ratings assigned to the four dimensions of each movie.

Essentially, our approach is similar to that in (McNee *et al.* 2002; Recker *et al.* 2003 and Adomavicius and YoungOk 2007): use additional information instead of pure ratings to determine the closeness between users. However, our context is for paper recommendation where learners' pedagogical features are used to measure the similarity between them. Furthermore, we also consider paper features in the recommendation process which is different from the existing approaches that only consider users' contextual information in making recommendations such as in (Adomavicius *et al.* 2005, McNee *et al.* 2002; Recker *et al.* 2003). For instance,

¹ Interested readers can refer to <http://movies.yahoo.com/movie/1809932977/user> for more information on the rating scheme.

the popularity of each paper, denoted by \tilde{r} , is used to factor out papers that are not well received. Specifically, we consider the following factors in our proposed multi-dimensional CFs: papers' overall-ratings, popularity, value-added, degree of being peer recommended, and learners' pedagogical features such as interest and background knowledge. Due to space limits, this paper reports only some of our findings².

3. A Multi-Dimensional Paper Recommender

We begin our discussion on our proposed paper recommender by exploring the pedagogical values of a paper in order to motivate the recommendation mechanism. Specifically, we consider the following factors in making recommendations: papers' overall-ratings, popularity, value-added, degree of being peer recommended, and learners' pedagogical features such as interest and background knowledge.

3.1 The Pedagogical Values of a Paper

(Barry 1994) pointed out that *situational factors* (contextual factors) other than only the *topical content* of a selected document influence a user's judgment of document relevance as well as quality. He suggested that these situational factors include those that users bring into the reading situation including experience, background, knowledge level, beliefs, and personal preferences; and these factors should be added are the co-existence of other users' traces on the document, including the social annotations such as 'thumb up' 'thumb down' in *KnowledgeSea III* (Brusilovsky et al. 2005), the textual comments, popularity of the each article and user models annotated to the document; in other words, the social affordance of the document. When users browse a digital document space, these elements can reveal the situational factors that could influence a document's overall user acceptance.

In a paper recommendation domain, a paper's '*situated factors*' including its usefulness in helping learners gain new knowledge (referred to as *Value_addedness*) and strengthening their understandings of the course topics (referred to as *Aid_Learning*), value in practice (referred to as *Job-related*, as learners were all part-time degree students), the degree of peer-recommendation (referred to as *Peer-Rec*), and textual comments³. We think that the factors we have considered so far (e.g. interestingness, value-addedness, etc.) represent the most typical factors that need to be taken into considerations when making recommendations in the pedagogical domain. Figures 1 and 2 compare our way of making recommendations with that used in the majority of recommender systems.

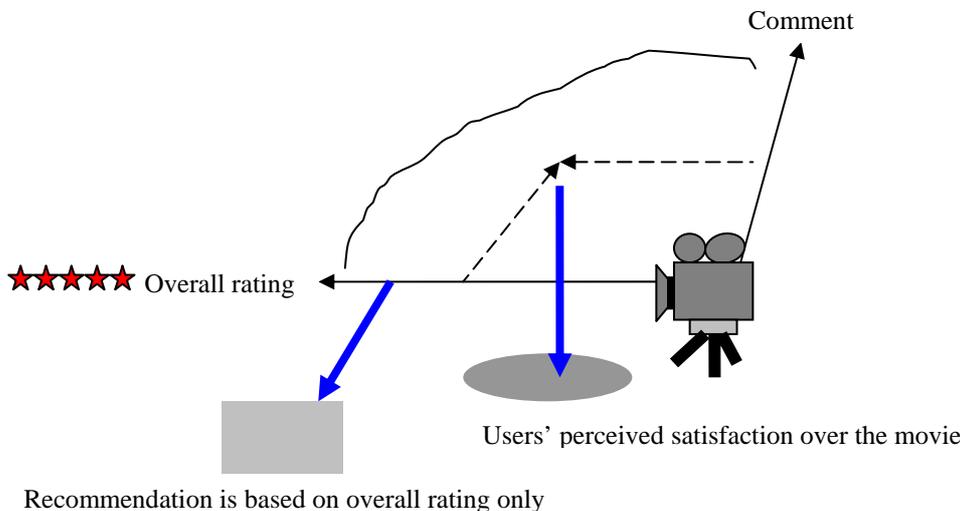


Figure 1. An illustration of user relevance evaluation on a movie (Tang and McCalla 2007)

In Figure 1, the information space allows users to review both the textual comments and the numerical rating of a movie. However, the majority of recommendation mechanisms only consider the latter in addressing users' needs and make computations on what should be recommended. Specifically, as shown in Figure 1, overall rating might reflect users' multiple feelings toward a movie, say its cast, the story etc., as studied recently in (Adomavicius and Young Ok 2007). The work propose a multi-criteria RS which incorporate users' numerical ratings on a movie's four aspect Story, Acting, Direction and Visuals in Yahoo! Movie. Results did show that combining these features can improve recommendation accuracy especially when the multi-ratings do carry meaningful information to reflect the overall rating of the item. Similarly, our pedagogical paper recommender works by incorporating

² A complete and detailed account of the study can be found at (Tang 2008).

³ A deeper discussion of this is beyond the scope of this paper. Readers can refer to (Tang 2008) for more details.

learners' additional impressions of each paper other than its overall rating. Figure 2 illustrates the recommendation mechanism.

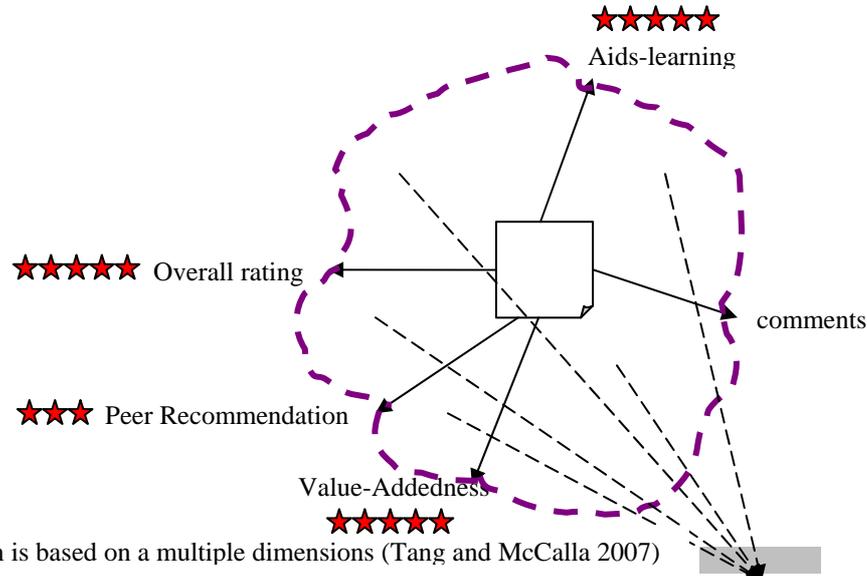


Figure 2 An illustration of users' relevance evaluation of a paper in the pedagogical paper recommender (Tang and McCalla 2007)

In the pedagogical paper recommender as shown in Figure 2, a paper is recommended based on a variety of dimensional *factors* that a learner has provided in terms of not only its overall rating of the topical appropriateness, but also some pedagogical values (situated factors), including its usefulness in helping learners gain new knowledge (referred to as *Value_addedness*) and strengthening their understandings of the course topics (referred to as *Aid_learning*), value in practice (referred to as *Job-related*, as learners were all part-time degree students) and the degree of peer-recommendation (referred to as *Peer-Rec*) as illustrated in Figure 2. Our goal is to understand the many factors driving learners to judge the 'goodness' of a paper. When the system allows users to unfold these aspects of a paper, it actually creates a rich space for learners to interact with the system and other learners. For instance, it can help raise the awareness of a learner towards the candidate papers or provide an opportunity for the learner to socialize with others through initiating discussions. Our experimental studies confirm our speculations that making recommendations to learners in social learning environments is not the same as that to users in Amazon.com etc. Learners are willing to accept those items that are not interesting, yet meet their learning goals in some way or another; learners' overall impression towards each paper is not solely dependent on the interestingness of the paper, but also other factors, such as the degree that the paper that help to meet their 'cognitive' goals (Tang 2008).

4. An Empirical Study

4.1 Study Goals

The goal of our study is to measure the multi-dimensionality of paper recommendation in an attempt to understand the effect of the incorporation of the pedagogical elements in boosting the recommendations made, because these factors will be projected in the algorithm to reflect the pedagogical values of each paper, and therefore, how they in turn can help push up the quality of the recommendation is of our first concern. In other words:

- Will learners be happy with papers that can expand their knowledge (i.e. they feel that after reading them, they learned something 'new')? In the e-learning domain, this is very important, as to fulfil their knowledge needs is the ultimate goal.
- How important is learner interest is in our domain? For instance, how far will learner be willing to tolerate a paper that is not so interesting?
- If a paper is too technical, will learners be comfortable with it, even if it matches their interest or from educators' perspective, it is a required reading (for instance, a seminal paper on a topic)?

These questions can provide us insights on the importance of the elements in making recommendations, therefore, in turn, guide us in tune the variables and weights used in recommendation algorithm.

4.2 Data Collection

We conducted a study to investigate the importance of various pedagogical factors in papers being read by students in a graduate course. The study was carried out with postgraduate students enrolled in a master program at the Hong Kong Polytechnic University. They were all registered in a course entitled Software Engineering (SE), with curriculum designed primarily for mature/working students with various backgrounds. In total 40 part-time students attended the course, offered as an evening class in the fall semester 2005. 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 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 by releasing their data (ratings and student model) after their final mark has been finalized (after the term ends).

4.3 Learners, Learner Profiles and Learner Feedback

At the beginning, learner profiles are drawn from a questionnaire consisting of four basic categories: interest, background knowledge, job nature, and learning expectation. Students represent a pool of learners with working experience related to information technology, but do not necessarily have background in computer science. After reading each paper, students were asked to fill in a paper feedback form (Figure 3).

1. Is the paper difficult to understand?
4. very difficult 3. difficult 2. easy 1. very easy
2. Is the content of paper related to your job?
4. very much 3. relatively 2. not really 1. not at all
3. Is the paper interesting?
4. very much 3. relatively 2. not really 1. not at all
4. Is the paper useful to aid your understanding of the SE concepts and techniques learned in class?
4. very much 3. relatively 2. not really 1. not at all
5. Do you learn something "new" after reading this paper?
4. absolutely 3. relatively 2. not really 1. not at all
6. What is your overall rating towards this paper?
4. very good 3. good 2. relatively 1. bad
7. Will you recommend this paper to your fellow classmates?
3. absolutely yes 2. maybe 1. no

Figure 3. Learner feedback form where several pedagogy-related questions were asked.

Several features of the papers were to be evaluated by each student, including its degree of difficulty to understand, its degree of job-relatedness with the user, its interestingness, its degree of usefulness, its ability to expand the user's knowledge (value-added), and its overall rating. We used a Likert 4-scale rating for the answer, except for Q7 in Figure 3. Several of the above questions are related to the pedagogical value of each paper, i.e. the degree of difficulty to understand (Q1), the degree of job-relatedness with the user (Q2), the value-added-ness (Q5), and degree of peer-recommendation (Q7). Since basically, the collections of candidate papers are mainly from popular technical magazines, therefore, it is not difficult to understand even for learners without much mathematical background knowledge. Indeed, almost all learners admitted that it is not difficult to read the recommended papers. In view of this, in our paper recommender, two of the pedagogical factors were focused in the analysis presented here: the value-added-ness(Q5) and the degree of peer-recommendation (Q7), along with overall rating (Q6). Section 5 and 6 document two studies to examine: 1) the characteristics of the paper recommender (a statistical analysis); and 2) our proposed multi-dimensional pedagogical paper recommender and evaluations.

5. The Pedagogical Factors Exhibiting in Learner Ratings: The Correlation Analysis

The major goal of our study described in this section is to explore the characteristics of pedagogical paper recommendation, which differentiates our study from other paper recommendation approaches designed for non-learning environments: we are interested in what makes a paper a high overall rating in terms of the pedagogical benefits it brings to the learner. Statistically, we are interested in the interaction among the pedagogical variables used in the recommendation mechanism. To achieve it, we abandon traditional approaches such as MAE, ROC; instead, we took several steps further in using some statistical analysis methods to uncover the associations among these factors. These objective evaluation methodologies are normally used to sort out alternative explanations for relations between variables and therefore essential to support our claims and to motivate the research studies. To further substantiate our understanding, we make the following conjectures:

Conjecture 1. The overall rating given by learners to a paper may not only depend on the interestingness of the paper from their perspective, but also on the richness of knowledge that has been gained by them from reading the paper and/or the usefulness of the paper in helping them to understand the course subject.

Conjecture 2. The intent of learners in recommending a paper to others may not only depend on the interestingness of the paper from their perspective, but also on the richness of knowledge that has been gained by them from reading the paper and/or the usefulness of the paper in helping them understand the course material.

Conjecture 3. The closeness of learners' jobs to a paper's topics may also affect their overall ratings of that paper and their likelihood of recommending it to others.

In order to validate our conjectures, it is necessary to show that ratings on Value_added (Q5) or Aid_learning (Q4) indeed affect Overall (Q6) or Peer_rec (Q7) ratings independently from Interesting (Q3). A total of four analyses were formed, and two of them will be described in this paper: partial correlation and Principal Components Regression and Partial Least Squares Regression Analysis (Tang 2008). These statistical analysis methods are of particular value to social scientists to probe into the interactivity among variables.

5.1 Partial Correlation and Results

Partial correlation is commonly used in modeling causality of models with 3 or 4 variables. Let $r_{AB.C}$ be the Pearson correlation of variables A and B, controlling for variable C, and r_{AB} be the Pearson correlation of variables A and B. If $r_{AB.C} = r_{AB}$, the inference is that the control variable C has no effect. If $r_{AB.C}$ approaches 0, then r_{AB} is spurious (the correlation is spurious), i.e. there is no direct causal link between A and B (see Figure 4(a)). It is either C affects A and B (antecedent), or A affects C which affects B (intervening). If $r_{AB} > r_{AB.C} > 0$, then we have partial explanation (see Figure 4(b)). In this case, A partially affects B regardless of whether it affects (or is affected by) C. Our computations on partial correlation are shown in Table 1.

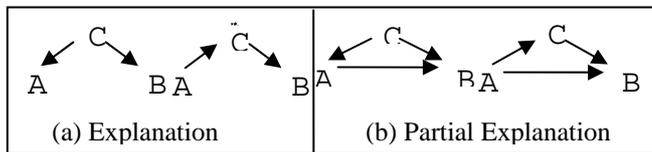


Figure 4. Causal inference with partial correlation when (a) $r_{AB.C} = 0$, and (b) $r_{AB} > r_{AB.C} > 0$.

Three groups (I, II, and III) are used as the comparison. In Group I, we check r_{AB} and $r_{AB.C}$ for $C=Interest$, $A=Value_added$, and B is either Overall or Peer-rec. The results of r_{AB} are 0.4798 and 0.4335 for $B=Overall$ and $B=Peer-rec$, respectively. After introducing Interest as a control, the correlations decrease to 0.3539 and 0.3017, respectively; hence, $r_{AB} > r_{AB.C} > 0$ in this group. In Group II, we check r_{AB} and $r_{AB.C}$ for $A=Aid_learning$. In Group III, we check r_{AB} and $r_{AB.C}$ for the reverse causality, i.e. Interest is affected by Value_added or Aid_learning. In fact, $r_{AB} > r_{AB.C} > 0$ for all groups, or the results favor a partial explanation model. In other words, in some degree Value_added and Aid_learning affect Overall and Peer_rec ratings independently from Interest. However, the presence of multicollinearity among variables in partial correlation analysis may diminish the validity of the claim. In addition, it is not clear whether the model is still valid in the presence of other variables (e.g. Difficulty or Job_related).

Table 1. Results of partial correlations

Variables:	C (control)	A	Pearson partial correlation	
			B:	B:
			Overall	Peer_rec
Group I	-	Value_added	0.4798	0.4335
	Interest	Value-added	0.3539	0.3017
Group II	-	Aid_learning	0.4242	0.3740
	Interest	Aid_learning	0.3038	0.2469
Group III	-	Interest	0.6046	0.5574
	Value_added	Interest	0.5282	0.4780
	Aid_learning	Interest	0.5456	0.4975

5.2 Principal Components Regression and Partial Least Squares Regression Analysis and Results

Principal components regression (PCR) combines principal components analysis (PCA) and linear regression. PCA transforms observations from a p -dimensional space to a q -dimensional space, $q \leq p$, while conserving as much information as possible (in terms of the total variance) from the original dimensions. The resulting dimensions are non-correlated weighted components which are linear combinations of the original variables. The weights are usually represented by eigenvalues produced during transformation. High-eigenvalue

components are principal components which contain the most information of the original data (Kelloway 1998), providing a window of opportunity for researchers to analyze the associations between the original variables. Meanwhile, by removing low-eigenvalue components, we can also simplify the regression model. If an explanatory variable is redundant (e.g. collinear with other variables), then it will vanish during dimensional reduction by PCR. In our test, we will check if Value_added and/or Aid_learning will vanish when we reduce the dimensionality of explanatory variables to two components only. In other words, we will see whether these two variables have any impact on the Overall rating or Peer-rec or both. Partial least squares regression (PLS) also uses PCA in building non-correlated components but differs from PCR in the sense it considers the accuracy of regression during the selection of components in regression. The components selected are not necessarily those with the highest eigenvalues, but those which explain as many independent variables as possible. As such, PLS performs a simultaneous decomposition of explanatory variables (components) and dependent variables with the constraint that these components explain as much as possible of the covariance between explanatory and dependent variables.

In our test here, we set the stopping criteria for both PCR and PLS as when they found at most two components. Thus, other components, if any, will be excluded from the regression. We use XLSTAT 2007 to perform both PCR and PLS, with Difficulty, Job_related, Interesting, Aid_learning, and Value_added as explanatory variables, and Overall and Peer_rec as dependent variables. The PCR model uses 61.1% variability of original explanatory data, i.e. the amount of information retained by the first two components of PCA. This value is low (the suggested variability in PCR is at least 80%, i.e. the default setting of XLSTAT). We restrict our model to a low variability in order to verify the “survivability” of Value_added and Aid_learning as explanatory variables in the model. We found that the parameters of Value_added and Aid_learning are between 0.170 and 0.297, while the parameters of Interest are between 0.235 and 0.314. Here, the parameters of Value_added and Aid_learning are relatively big with respect to that of Interest; hence, the result supports the survivability of these two variables in explaining Overall and Peer_rec ratings. In fact, the variable-importance-in-the-projection index (VIPs) of both Value_added and Aid_learning from PLS are above the critical value 0.8, which lead us to strongly believe that they contribute significantly to the model (Wold 1995) (see Figure 5).

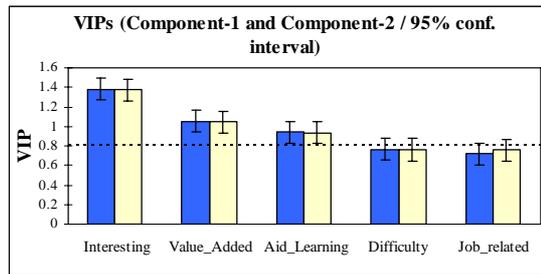


Figure 5. The variable-importance-in-the-projection index (VIPs) of both component 1 and component 2 from PLS.

6. The Proposed Paper Recommender and Evaluation

6.1 The Algorithm

Convinced that Overall rating is affected by other than Interest ratings, we believe that a multi-dimensional recommendation is more suitable in paper recommendation. This is of important, especially in the cold start recommendation, i.e. when we do not have enough co-rated papers in performing collaborative filtering (CF). We first consider three factors as a basis to measure the closeness of a pair of users, i.e. the Overall, Value_add and Peer_rec. Since we have three different ratings (3 dimensions) for each paper, we may obtain three different Pearson correlations for each pair of users. Suppose $P_d(a, b)$ is the Pearson correlation based on the rating r_d on dimension d , then, we can combine those three correlations into a weighted sum Pearson correlation as:

$$P_{3D}(a, b) = w_{overall} P_{overall}(a, b) + w_{valueadd} P_{valueadd}(a, b) + w_{peer_rec} P_{peer_rec}(a, b) \quad (1)$$

where $w_{overall} + w_{valueadd} + w_{peer_rec} = 1$.

Through this computation, we find a group of similar users to a given target user. Our approach is similar to that in (Lekakos and Giaglis 2006) which adopts users’ ‘life style’ to measure the closeness of each pair of users in order to identify its neighbors.

Suppose all learners have also provided their student models (e.g. their interests in various topics and background knowledge such as programming skill, etc.), also on a Likert scale. Next, we can compute the

2D-Pearson correlation between learners based on their student models; that is, we compute the aggregated Pearson correlation between student interest and their knowledge background as follows.

$$P_{2DStdModel}(a, b) = w_{interest} P_{interest}(a, b) + w_{bkgrKnowledge} P_{bkgrKnowledge}(a, b) \quad (2)$$

Since we have various weights on combining Pearson correlations, we may tune them to study the relative importance of each factor in making recommendations. We then combine this with a 3D-Pearson correlation from co-rated papers:

$$P_{5D}(a, b) = P_{3D}(a, b) + w_{2D} P_{2DStdModel}(a, b) \quad (3)$$

From $P_{5D}(a, b)$ we can identify the best N neighbors for a target user. After that, we use the following formula to calculate the aggregate rating of each paper:

$$r_k^{5D} = \sum_B P_{5D}(a, b) r_{b,k} \quad (4)$$

In the end, we combine this rating with the average rating of each paper (i.e. paper's popularity \tilde{r}) to obtain a 6D-CF based rating for the papers:

$$r_k^{6D} = r_k^{5D} + w_{\tilde{r}} n \tilde{r} \quad (5)$$

where n is the number of neighbors = $|B|$ and $w_{\tilde{r}}$ is the weight of a paper's popularity \tilde{r} .

Based on the ranking of r_k^{6D} , we can find the best papers to recommend to a target user. To summarize, the six elements we have used are Overall, Value_add, Peer-rec, \tilde{r} , *learner interest*, *learner background knowledge*. Although the 6D-CF computation is more complex than the other CF-based recommendation techniques, we speculate, under certain circumstances, that it is necessary to improve recommendations in e-learning applications¹.

6.1 Evaluation and Results

6.1.1 Experiment Setup

The weighting combinations (overall-rating, value-addedness, peer recommendations) are not chosen randomly; in fact, a lot other sets have been tested. However, we found out that only when the sum of the second and third weights is less than 0.1 can the benefits of recommendation performance. We conjecture that in recommender systems, the overall rating is still the major factor in determining recommended items. It might be possible that human users are more consistent decision-makers when they agree or like items. The number of neighbours is set to be 5 and 10 respectively, although in our experiments other values were also evaluated, and here, we report two of the best results we obtained. Last but not the least, we would also be eager to see, among the two key pedagogical features (value-addedness and peer-recommendation) and each paper's popularity \tilde{r} , which factors can boost the recommendation performance. In other words, the weight reflects the value of the corresponding variable in making recommendation.

6.1.2 Evaluation Protocol

Different from that in classical literature, we do not use the "all-but-one" protocol for the evaluation of our recommendation techniques. In the "all-but-one" protocol (mainly in the CF-based recommendation), all but one of the ratings given by a target user are used to find neighbors who eventually used to predict the rating of the single item that is held out. Instead, our CF-based methods allow neighbors to pick and recommend the best items (papers), regardless the user has rated them or not. Then based on the recommended paper list(s), we then examine the rating(s) the target user provided. If more than one paper is returned, an average of paper ratings is reported. The major reason of applying this type of the evaluation is due to the fact that our paper recommended is not intended to accurately *predict* user ratings. The way that previous works have put too much focus on the prediction accuracy has been heavily debated and criticized recently (Herlocker et al. 2004, McNee *et al.* 2006, Winoto and Tang 2008). In the domain we are studying, it is more important to recommend pedagogically useful papers than to only suggest papers matching learner interest.

¹ In our experiments, in addition to the 6D-CF discussed here, we have also studied 3D- and 4D-CFs and compared their performances. Due to the space limitations, we focus on the performance of 6D-CF.

For each target learner, we randomly assign 30 combinations of co-rated papers in finding his/her neighbours, who later recommend one or five papers to the target learner. Then, we record the average ratings of the recommended papers by the target learner. Therefore, for each treatment we have collected $24 \text{ learners} \times 30 \text{ combinations} = 720 \text{ ratings}$. In each treatment, we tune in the weights of those pedagogical elements to identify how important and useful they contribute to the overall recommendation. The average ratings are reported in next section and in Figures 6, 7, and 8. However, we do not perform statistical test in comparing two average ratings. Instead, our analysis is based on the pattern of those average ratings on each dimension of our control variables.

6.1.3 Results

In this section, we will keep our focus on three key findings aiming at interpreting the significance of the pedagogical factors through comparisons of the recommendation approaches.

- *The effect of value-added-ness and peer-recommendation on overall rating*

Results suggest that incorporating ratings from value-added-ness and peer-recommendation can improve the performance of CF-based recommender systems. Specifically, the average overall rating increases from 3.055 to 3.06 when the recommendation considers the paper's value-added-ness and peer-recommendation (weight = 0.02).

- *The effect of incorporating learner knowledge background on overall rating*

We also look at whether or not the incorporation of learners' knowledge background would have either negative or positive effects on the performance of the recommender.

Figure 6 captures the experimental results when $(w_{interest}, w_{bkgKnowledge}) = \{(1, 0), (1, 0.5), (1, 1)\}$, $w_{2D} = 0.5$, $(w_{overall}, w_{valueadd}, w_{peer_rec}) = (100, 0, 0)$, and the number of co-rated papers are 2 and 4.

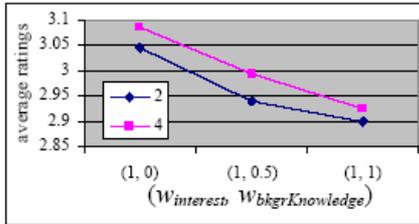


Figure 6. 6D-CF (100, 0, 0) with the number of co-rated papers as 2 and 4 in the case of recommending the best paper.

Similar patterns were obtained when the number of co-rated papers is 8 and 15. One observation is that when the weight on knowledge background, $w_{bkgKnowledge}$, increases from 0, to 1. That is, when we begin to consider knowledge background in computing the Pearson correlation between two learners, the performance of the recommendations decreases. In fact, when we look at other treatments of this group of the experiments, exactly same results are obtained which lead us to strongly believe that the incorporation of knowledge background does not bring advantage into the recommendation process. That is, the understanding of papers does not strongly depend on their background knowledge. The results are not surprised to us, since almost all the papers come from popular magazines such as *CACM*, *IEEE Software*, etc., which aims at general readers. Hence, the papers are more understandable, compared with those more technical papers from, say, *IEEE Trans. on Software Engineering*. It is noted that although we did not establish a strong relationship between learners' background knowledge and the ratings, it does not necessarily mean that this relationship does not exist. In fact, in one of the papers, there are a few mathematical formulas; and a few students felt that the paper is relatively difficult to understand, though the content of it is interesting.

- *The effect of incorporating learner knowledge background on overall rating*

Would adding learner interest into CF increase the quality of a recommendation? Here, we present our analysis to answer this question when $(w_{interest}, w_{bkgKnowledge}) = (1, 0)$. When we first compared the performance of the 6D-CF for recommending the best single paper, the results are shown in Figure 7. Now, the effect of $w_{2DStdModel}$ (horizontal line) represents the weight of learner interest only.

When $w_r = 1$ (the left 6 data elements in each diagram) and we introduce learner interest ($w_{2DStdModel}$ increases from 0 to 0.5), the quality of recommendations drops in both top diagrams but increases in both bottom diagrams, which means that *incorporating learner interest in CF has a small positive impact when we have a relatively larger size of co-rated papers (15 in this case)*.

However, the benefit is not persistent when we increase $w_{2DStdModel}$ more, because it eventually drops after $w_{2DStdModel} > 1$. When $w_r = 5$ (see the right 6 data elements in each diagram), the performance is quite steady with respect to $w_{2DStdModel}$, showing recommendations are independent of the weights on learner interest

($w_{2DStdModel}$ increases from 0 till 10), except when $w_{2DStdModel}$ equals to 10 in both bottom diagrams. From both top diagrams we can see that the recommendations made are more satisfying when $w_{\bar{r}} = 5$. In other words, the effect of *learner interest* on the outcome of recommendations made is less important than that of papers' *popularity*. The primary reason is that we cannot accurately identify "similar" neighbors using a small number of co-rated papers. For other combinations of ($w_{overall}$, $w_{valueadd}$, w_{peer_rec}), similar results are observed. When the recommender is choosing the top five papers (instead of recommending the best single paper to a target user), its performance is different, as shown in Figure 8. Things fail to change for the better even when we increase the weight of *popularity* (from $w_{\bar{r}} = 1$ to $w_{\bar{r}} = 5$), and *learner interest*. The overall performance shows a downward trend. When the number of co-rated papers is low (top diagrams), the performance is even worse when we increase the value of $w_{\bar{r}}$. This is just the opposite of what we obtained when the recommender makes the best single recommendation (top diagrams in Figure 7). For other treatments of this group of experiments, similar results are obtained. The results suggest that care should be taken when the recommender is required to pick up the top 5 papers in which more information is needed if the recommender is expected to maintain its performance stability.

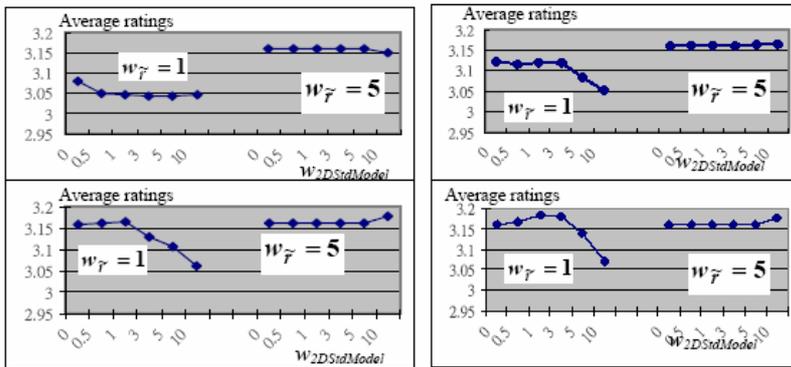


Figure 7. Performance comparison between (6D-CF (100, 0, 0)) with the number of co-rated papers as 2 (top-left diagram), 4 (top-right diagram), 8 (bottom-left diagram) and 15 (bottom-right diagram) to find the best recommended paper.

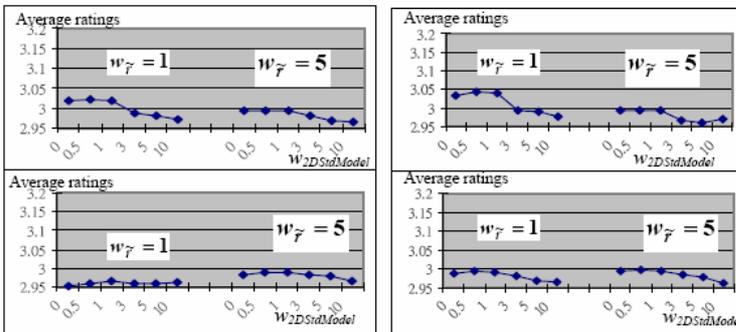


Figure 8. Performance comparison between (6D-CF (100, 0, 0)) with the number of co-rated papers as 2 (top-left), 4 (topright), 8 (bottom-left), and 15 (bottom-right) to make the best five recommendations.

6.2 Discussions

Our results indicate that for a recommender system in such a domain individual learner models should track learners' interests, their goals, and their background knowledge in specific topics. Papers should also be analyzed based on the topic, degree of peer recommendation, etc. The recommendation is carried out by matching the learner interest with the paper topics where the technical level of the paper should not impede the learner in understanding it. Therefore, the suitability of a paper toward a learner is calculated as the summation of the fitness of learner interest toward the paper and the appropriateness of it to help the learner in general. Experimental results support one fundamental conclusion: making recommendation to learners in learning environments is not the same as it is in many commercial domains where user likes are all that matters. Learners are willing to accept those items that are not interesting, yet meet their learning goals in some way or another; learners' overall impression towards each paper is not solely dependent on the interestingness of the paper, but also other factors, such as the degree that the paper that help to meet their 'cognitive' goals, which is consistent with human user's information-seeking behaviors as Rieh (2002) summarized: 'people make judgment of information quality and cognitive authority' on consumed items (p. 146).

Unique to the tasks that our paper recommender intends to support in satisfying learners pedagogically, some of our evaluations are therefore performed to provide us insights on the learner satisfaction towards the recommended items (Herlocker *et al.* 2004, Pazzani 1999, Terveen and McDonald 2005). Our findings suggest learner satisfaction as a complicated function of learner characteristics, rather than, the single topicality of a paper as matched against learner interest. In fact, we conducted interviews with some students. Both the interview results and student feedbacks on the course are both overwhelmingly good: they all claimed that the readings opens a new horizon for them in that although some of them are software engineers themselves, they are not aware of some terms that used in their field; as such, they gained a lot of knowledge from these extra reading materials, although they admitted that it is difficult to balance their time and energy between their heavy work and study loads.

We realized that one of the biggest challenges is the difficulty to test the effectiveness or appropriateness of a recommendation method due to a low number of available ratings. Testing the method with more students, say, in two or three more semesters, may not be helpful, because the results are still not enough to draw conclusions as strong as those from other domains where the ratings can be as many as millions.

In additions, in our study, the papers are related to software engineering (including user interface design and usability engineering); hence, it is hard to generalize the results to make recommendation to students in other classes. Since in some subjects, papers may exhibit more technical difficulties due to their inherent features (e.g. in artificial intelligence or data mining), so are students who may also be different when they begin to take this course, which in turn affect on the effect of those pedagogical factors considered on the performance of the recommender system. Hence, we are eager to see the collaborations from different institutions in using the system in a more distributed and larger scale fashion (as it is very difficult to achieve it in using one class each time and in one institution). Through it, our future work includes the design of a MovieLens-like benchmark database as a test bed on which more algorithms can be tested (including ours).

7. Concluding Remarks

In this paper, we discussed a multi-dimensional paper recommendation approach for e-learning domains. The experiments suggest that in the e-learning domain, it is imperative for us to inject other factors, among them, the popularity of each paper, learner knowledge background, learner interest, learner knowledge background and job experience, although these factors are less important for making recommendations on movies, books, CDs. Another interesting observation is that user interest isn't the number one key factor to boost the performance of recommendations; instead, users are willing to accept 'risky' recommendations that are not matched to their interest during the learning. As such, the focus is more on how to find more pedagogically appropriate papers. Currently, we are studying the degree of effects that other learner features such as their job relatedness and learning goals might have on the overall performance of recommendations.

Acknowledgement

We would like to thank three anonymous reviewers for their constructive comments.

References

- [1] Adomavicius, G., Sankaranarayanan, R., Sen, S. and Tuzhilin, A. (2005). "Incorporating contextual information in recommender systems using a multidimensional approach". *ACM TOIS*, 23 (1): 103-145. January 2005.
- [2] Adomavicius, G. and Young Ok Kwon. (2007). "New Recommendation Techniques for Multi-criteria Rating Systems". *IEEE Intelligent Systems*, Volume 22, Issue 3 May-June 2007, pp. 48-55.
- [3] Barry, C. (1994). "User-defined relevance criteria: an exploratory study". *Journal of the American Society for Information Science (JASIS)*, pp.149-159. April 1994.
- [4] Basu, C., Hirsh, H. and Cohen, W.W. (1998). "Recommendation as classification: using social and content-based information in recommendation". *AAAI'98 (Madison, Wisconsin, USA: Morgan Kaufman)*, pp.714-720.
- [5] Bollacker, K., Lawrence, S. and C. Lee Giles, C. L. (1999) "A system for automatic personalized tracking of scientific literature on the web". *ACM/IEEE JCDL'1999 (Berkeley, CA, USA, pp.105-113.*
- [6] Browne, M. W. and Cudeck, R. Alternative ways of assessing model fit. In K. A. Bollen and J. S. Long (Eds.) *Testing Structural Equation Models*. Sage: Newbury Park, CA. 1993.
- [7] Brusilovsky, P., Farzan, R. and Ahn, J. (2005). "Comprehensive personalized information access in an educational digital library". *ACM/IEEE JCDL'2005 (Denver, CA, USA: ACM Press)*. pp.9-18.
- [8] Herlocker, J., Konstan, J., Borchers, A., and Riedl, J. (1999). "An algorithmic framework for performing collaborative filtering." In *Proc. of the 22nd annual International ACM SIGIR Conference on Research and Development in Information Retrieval SIGIR'99 (Berkeley, CA, USA: ACM Press)*, pp. 230-237.
- [9] Herlocker, J., Konstan, J., Terveen, L. And Riedl, J. (2004). "Evaluating collaborative filtering recommender systems". *ACM TOIS*. 22(1):5-53, January 2004.
- [10] Kelloway, E. K. (1998) *Using LISREL for Structural Equation Modeling: A Researcher's Guide*, Sage. 1998.

- [11] Konstan, J., Miller, B.N., Maltz, D., Herlocker, J., Gordon, L.R. and Riedl, J. (1997) "GroupLens: applying collaborative filtering to Usenet news". *CACM*, 40(3): 77-87, March 1997.
- [12] Lekakos, G. and Giaglis, G. (2006) "Improving the prediction accuracy of recommendation algorithms: approaches anchored on human factors". *Interacting with Computers*, 18(3): 410-431, 2006.
- [13] Lemire, D., Boley, H., McGrath, S. and Ball, M. (2005) "Collaborative filtering and inference rules for context-aware learning object recommendation". *International Journal of Interactive Technology & Smart Education*, 2 (3), 2005. <http://www.daniel-lemire.com/fr/documents/publications/itse2005.pdf>
- [14] Manouselis, N., Vuorikarim, R. and Van Assche, F. (2007). "Simulated analysis of MAUT collaborative filtering for learning object recommendation". SIRTEL Workshop, EC-TEL 2007. http://infolab-dev.aua.gr/sirtel2007/papers/Manouselis_et_al.pdf
- [15] Manouselis, N. and Costopoulou, C. (2007). "Experimental analysis of design choices in multi-attribute utility collaborative filtering." *International Journal of Pattern Recognition and Artificial Intelligence*, 21(2), pp.311-331, April 2007. <http://infolab-dev.aua.gr/files/publications/en/1169683003.pdf>
- [16]McNee, S., Albert, I., Cosley, D., Gopalkrishnan, P., Lam, S., Rashid, A., Konstan, J. and Riedl, J. (2002). "On the recommending of citations for research papers." *ACM CSCW'02* (Orleans, Louisiana, USA: ACM Press), pp.116-125.
- [17] McNee, S., Riedl, J. and J.A. Konstan. (2006). "Being accurate is not enough: how accuracy metrics have hurt recommender systems." *In the Extended Abstracts of the 2006 ACM Conference on Human Factors in Computing Systems 'CHI 2006* (Montreal, Canada: ACM Press), pp.1097-1101.
- [18] Pazzani, M. (1999). "A Framework for collaborative, content-based, and demographic filtering." *AI Rev.*, pp.393-408, Dec.1999.
- [19] Recker, M., Walker, A. and Lawless K. (2003). "What do you recommend? Implementation and analyses of collaborative information filtering of web resources for education." *Instructional Science* 31:299-316, 2003.
- [20] 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. January 2002.
- [21] Tang, T. Y., and McCalla, G. I. (2004). "Utilizing artificial learners to help overcome the cold-start problem in a pedagogically-oriented paper recommendation system". *AH 2004* (Eindhoven, The Netherlands: Springer) pp.245-254.
- [22] Tang, T. Y., and McCalla, G.I. (2005). "Paper annotations with learner models." *AIED 2005*: pp.654-661.
- [23] Tang, T.Y., and McCalla, G. I. (2007). "The social affordance of a paper". In *Workshop of Assessment of Group and Individual Learning Through Intelligent Visualization (AGILEeViz), 13th International Conference on Artificial Intelligence in Education (AIED 2007)*. Marina Del Rey, CA, USA, 34-42. <http://aied.inf.ed.ac.uk/AIED2007/AgileVizWorkshopAIED2007.pdf>
- [24] Tang, T. Y. (2008). "The Design and Study of Pedagogical Paper Recommendation." PhD Thesis. Department of Computer Science, University of Saskatchewan, Canada. April 2008. <http://library2.usask.ca/theses/available/etd-03262008-002314/>
- [25] Terveen, L. and McDonald, D. (2005). "Social matching: a framework and research agenda." *ACM TOCHI*, 12(3):401-434. September 2005.
- [26] Torres, R., McNee, S. M., Abel, M., Konstan, J.A. and Riedl, J. (2004). "Enhancing digital libraries with TechLens." *ACM/IEEE JCDL'2004* (Tuscon, AZ, USA: ACM Press), pp.228-236.
- [27]Winoto, P and Tang, T. Y. (2008) "If you like the Devil Wears Prada the book, will you also enjoy the Devil Wears Prada the movie? A study of cross-domain recommendations". Special Issue of Web-based Recommendation Systems Technologies and Applications, *New Generation Computing* (Ed. Janusz Sobecki), Vol. 26 No. 3 2008.
- [28] Woodruff, A., Gossweiler, R., Pitkow, J., Chi, E. and Card, S. (2000). "Enhancing a digital book with a reading recommender". *CHI'00* (The Hague, The Netherlands: ACM Press), pp.153-160.
- [29] Wold, S. (1995). "PLS for multivariate linear modelling". In van de Waterbeemd H. (ed.), *QSAR: Chemometric Methods in Molecular Design*, Vol 2. Wiley-VCH, Weinheim, Germany. pp.195-218.