Personalized Book Recommendation Based on Ontology and Collaborative Filtering Algorithm

Lin Cui[1,2*], Hong Li[1], Caiyin Wang[1] and Baosheng Yang[1]

[1]*Intelligent Information Processing Laboratory, Suzhou University, Suzhou, Anhui, 234000, China;* [2]*College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing Jiangsu, 210016, China*

Abstract: Information recommendation service is one of important functions of digital library, aiming at the problem that book recommendation service exists the insufficient requirement mining of service object information in the current university library, personalized book recommendation method based on ontology information and collaborative filtering algorithm (abbreviated as OI-CFA algorithm) is proposed. Firstly, this paper discusses the necessity of collaborative recommendation in digital library, introduces main methods and technology based on collaborative filtering recommendation system. However, there are several problems that are data sparse and new item forecast with collaborative filtering recommendation method based on item. In order to solve these problems, this paper introduced an integrated similarity algorithms of structured semantic information based on OI-CFA. Extracting the semantic information of items including knowledge representation based on ontology, through ontology learning, the specified domain ontology is constructed. Compared with the traditional collaborative filtering algorithm and SVM, experimental results show that this method can not only solve the problems caused by the item-based collaborative filtering algorithm, but also improve the accuracy of recommendation.

Keywords: Collaborative filtering, ontology, recommendation systems, semantic similarity.

1. INTRODUCTION

In a school, library is an information center, research center and also is an important channel by which learners acquire the learning resources. Now all levels of libraries have established information service system based on network, but most library service system only provides some simple query functions, such as access to the library information and user personal information, can not effectively provide personalized information service according to user preferences and specific needs. How to construct a personalized information service system is a very significant research content currently in the field of library. Emergency of new technology has brought many new ideas for the information service of library.

Collaborative filtering is one of the personalized recommendation technologies which is widely studied and applied successfully; its core idea is that the best way to find an interesting content by users is to firstly find other users with similar interests, and then the similar interested content of users recommended to this user [1]. Advantages of collaborative filtering recommendation system are that new interesting information of users can be discovered, and has no special requirements on the recommended resources. But collaborative filtering algorithm still has some shortcomings. The key question is that each user is generally only interested in very little item, the user score data are very sparse, which leads to the similarity between users is not accurate; the nearest neighbor users are not reliable; it is difficult to recommend or predict a new project [2]. This paper proposes a collaborative filtering algorithm based on semantic similarity, the method combined user score calculation item similarity with semantic similarity, through analyzing a large number of users borrowing records in library database, acquire the users' book preference, and provide personalized books recommendation to the user. Finally, experiments prove the validity and feasibility of the scheme.

2. RELATED WORK

Collaborative filtering technology is a main method of the current information recommendation service. In 1992, D. Goldberg *et al.* firstly put forward collaborative filtering concept and applied it to Tapestry system, in which the similarity between the current user and other users is calculated by using the user-item rating matrix of the system, finding out the most similar and the nearest neighbor sets and lastly produce the final recommendation [3]. In 1994, Resnick *et al.* extended the algorithm and designed the GroupLens system [4]. In 2000, Sarwar *et al.* proposed item-based collaborative filtering recommendation algorithm [5]. However, item-based collaborative filtering recommendation encountered many difficulties and challenges in the practical application which were the user-item score matrix is extremely sparse, cold start and recommended speed is very slow. Thus, researchers have proposed various kinds of improvement methods and experiment.

1874-110X/14 2014 Bentham Open

Table 1. User evaluation matrix based on collaborative filtering algorithm.

	Item$_1$	Item$_2$	…	Item$_N$
User$_1$	5	2	…	5
User$_2$	4	5	…	4
……	……	……	R$_{ij}$	……
User$_{N-1}$	3	4	…	2
User$_N$	4	3	…	5

Lee T. Q. *et al.* proposed a kind of uncertain collaborative filtering algorithm based on nearest neighbor [6]. Burke R. proposed a collaborative filtering recommendation based on user clustering [7]. Deshpande *et al.* proposed a top-n recommendation based on item similarity, believing that when performing recommending, the n highest similarity of items only needed to be considered, which is faster 1 to 2 orders of magnitude than recommendation algorithm and has high accuracy [8]. Rogbles *et al.* proposed a semi naive Bayesian method based on interval estimation [9]. Lee *et al.* combined collaborative filtering and association rule mining framework model, designed personalized expert model to overcome the problems in traditional recommendation algorithm [10]. Huang *et al.* adopted association rules to mine user behavior to recommend Webpage [11].

In recent years, collaborative recommendation system has applied to the digital library, and is becoming the main research topic in this field. In digital library, user preferences may change in a short time with the change of their research directions and requirements, so in order to mining user environments, collaborative filtering technology based on ontology is introduced.

3. PERSONALIZED BOOK RECOMMENDATION METHOD BASED ON ONTOLOGY AND COLLABORATIVE FILTERING ALGORITHM

3.1. Collaborative Filtering Algorithms

This paper adopts a user-based collaborative filtering algorithm, suppose that if the user scores of some item are similar, the scores on other items by them are also similar. The algorithm can find the most similar neighbor according to the target user behavior (borrowing history, evaluation, browse times, historical comparison, etc.), then in accordance with the most similar neighbor’s interests or preferences, the user interests or preferences can be predicted out and recommendation system execute information recommendation for users. The algorithm is mainly divided into three parts that are the establishment of evaluation matrix, searching the nearest neighbor and generating recommendation, respectively [12].

3.1.1. Establishing Evaluation Matrix

Through collecting user rating, evaluation behavior and so on, perform data cleaning and transformation, eventually forming the user evaluation of various item matrixes, which is as shown in Table **1**:

Where Rij denotes the j-th item evaluation from the user i, among which, Rij is usually between 0 and 5, the more Rij score, the better user evaluation of the item.

3.1.2 Searching the Nearest Neighbor

Computing the similarity between the target user and all users in the database, finding similar users to establish the nearest neighbor set, this part is the core of user-based collaborative filtering algorithm. The similarity between users is calculated by Pearson correlation similarity, which is as following [13].

$$sim(U_a,U_i)=\frac{\sum_{y\in R_a\cap R_y}(R_{i,y}-\overline{R_i})(R_{a,y}-\overline{R_a})}{\sqrt{\sum_{y\in R_a\cap R_y}(R_{a,y}-\overline{R_a})^2}\sqrt{\sum_{y\in R_a\cap R_y}(R_{i,y}-\overline{R_i})^2}} \qquad (1)$$

Among them, $sim(U_a,U_i)$ represents the similarity between $User_a$ and $User_i$, y is the common items evaluated by $User_a$ and $User_i$, $R_{a,y}$, $R_{i,y}$ respectively denote the item y evaluated by users a and user i, $\overline{R_a}$, $\overline{R_i}$ respectively denote the mean evaluation value of items. After similarity calculation is completed, needing to choose the nearest neighbor from the user group, this paper adopts the similarity threshold method to select the nearest neighbor. By setting a certain threshold, only the similarity exceeding the user threshold can be the nearest neighbor.

3.1.3. Producing the Recommendation

According to the evaluation of the nearest neighbor set, recommendation is produced by certain recommendation algorithm for the target users, by using the weighted average method. Calculation of weighted average method for target user has carried on the prediction score for item i:

$$P_{a,i}=\overline{R}_a+\frac{\sum sim(a,n)\times(R_{n,j}-\overline{R}_n)}{\sum\left|sim(u,n)\right|} \qquad (2)$$

Among them, $sim(a,n)$ denotes the similarity of user a and user n, user $R_{n,j}$ represents the recent neighbor users in n score and the item j, $\overline{R}_a$, $\overline{R}_n$ represent the users a and n mean item evaluation. The weighted average method involves that users score all items, suitable for user evaluation of more items, if the item evaluation of the user is small, individual item score would have a greater impact on the average score, and the results would be not accurate.

3.2. Constructing Domain Ontology

During the process of collaborative filtering, in order to acquire item semantic information, we must understand the hierarchy concept of structured object extraction from a large number of books information as semantic entity. Domain ontology is applied to filter book information; many problems can effectively make up for the traditional filtering technology. The domain ontology is used to describe specific domain knowledge ontology, which provides formal characteristics and rules of domain entity concept and the relationship between these domain entity concepts. Through the concept of the relationship between the description of semantic concept, which makes the interaction between a user and machine can not only based on the grammatical level, but also can be based on a complex semantic level.

If manually constructing ontology, the larger domain ontology, the bigger workload is. If we can collect enough training text, extract basic vocabulary from these texts in the field, and then use some technology obtained relationship between these vocabularies, the domain ontology construction can be realized. The specific practices are as follows:

(1) Preprocessing the training text

Execute the text segmentation and remove stop words according to stop list. the terms weighting are calculated and standardized. Normalization formula is as follows:

$$W_{ik} = \frac{f_{ik}}{\sum_{i=1}^{n_k} f_{ik}} \tag{3}$$

n_k is the total number of different terms in the document K.

(2) Constructing term-document matrix

The term-document matrix is as follows:

$$W = \begin{vmatrix} w_{11} & \cdots & w_{1n} \\ \vdots & \ddots & \vdots \\ w_{m1} & \cdots & w_{mn} \end{vmatrix} \tag{4}$$

where each row represents a term weights in each document, each column represents a document vector, w_{ik} represent the *i-th* weight terms in the K document.

(3) Performing SVD decomposition for W.

$$W = u^k \times s^k \times v^k \tag{5}$$

u^k is the term concept matrix, v^k is a document-concept matrix, s^k can be a set of concepts and describe a set of terms for each concept, finally, the relationships between these concepts can be clustered. When clustering, each concept is regarded as a vector consisting of m terms and is clustered. Ontology database in book fields is constructed by using this method.

3.3. Personalized Book Recommendation Algorithm based on Ontology and Collaborative Filtering

Combined ontology and collaborative filtering algorithm, the process of personalized book recommendation algorithm based on ontology and collaborative filtering is showed in Fig. (**1**):

The architecture of OI-CFA is mainly divided into the following modules:

(1) Pre-processing data and user feature model is established

Data preprocessing is the process of data optimization, user model is established through this process, the model is

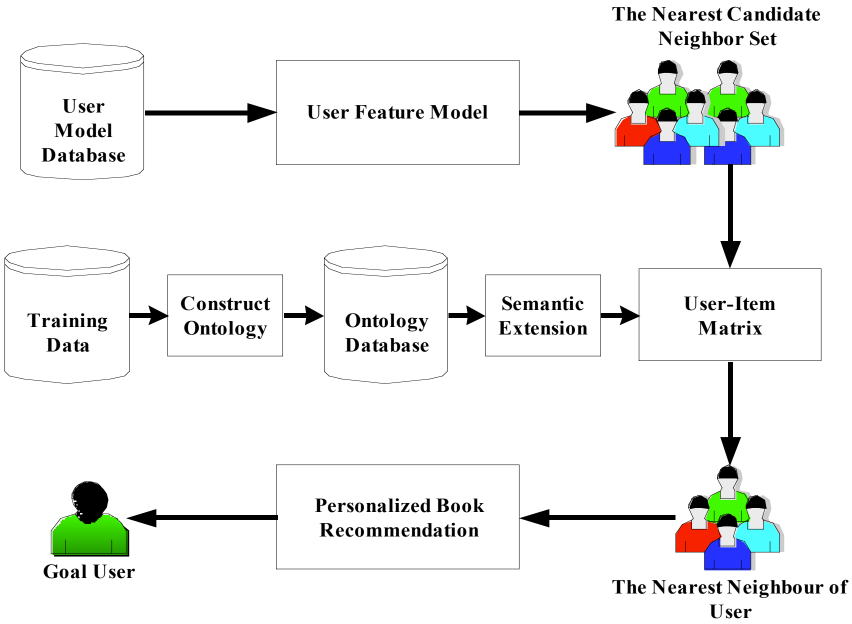

Fig. (1). Personalized book recommendation based on ontology and collaborative filtering algorithm.

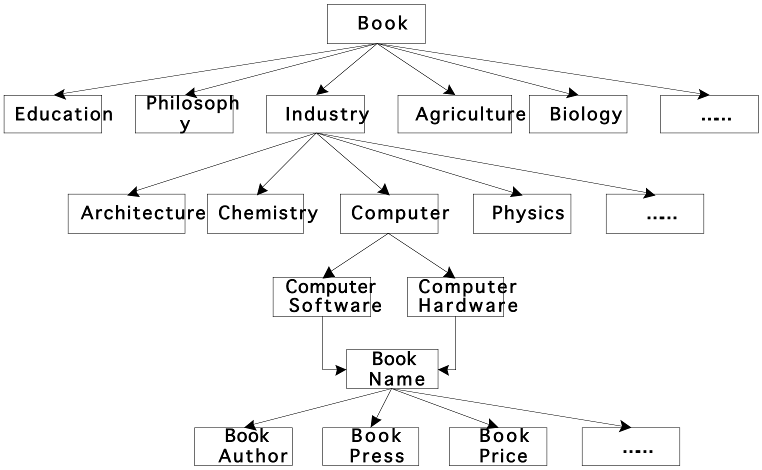

Fig. (2). Book ontology framework.

able to represent the user's needs and interests. This system uses the vector space model to represent users, as to facilitate the user clustering and similarity calculation. Source data is the library user's lending records, pretreatment process mainly including data cleaning, data conversion, normalization processing. Data cleaning is a data optimization process, to remove those who do not meet the requirements of the lending records, can be more effective for the recommended books user access to high quality. With the change of time, user learning content and users' interests will change, demand for books may have very big difference, delete the early user lending records, can better grasp the needs of users.

(2) Determining the top-n nearest neighbor set

The core of collaborative filtering recommendation system is to find the nearest neighbor set for the target users. The system determines the top-n nearest neighbor set of target users according to the similarity of users, namely the main operation of this step is to calculate the similarity between users. There are several methods of measuring similarity; this system uses the correlation similarity. Classification that users i and user j commonly interested in is expressed by Uij, Ri and Rj respectively represent average interest in all of the classification from users i and user j, the similarity sin (i, j) between user i and user j is as shown in formula (5): Through user similarity, the similarity between the target user and other users is calculated, and then select n users with greatest similarity, composed of top-n nearest neighbor set of the target user.

(3) Construction of ontology and semantic expansion

Ontology provides some important concepts and semantic relationship between these concepts, the goal of ontology is to capture the domain knowledge, provide the common understanding of the domain knowledge, determine the commonly recognized vocabulary of domain, and giving these terms in formal models from different levels and a clear definition of the correlation between vocabularies. Based on ontology, item semantic information is extracted and after establishing domain ontology, using domain ontology, the semantic information of project is extracted. The relationship between classes and class in ontology database is regarded as a relational database, build a specific domain semantic classification trees. Fig. (**2**) is the construction of relevant books classification ontology, from the books the extracted entities are instances of these classes, these examples include semantic information.

(4) Integration of semantic similarity and collaborative filtering

In order to conveniently calculate the item semantic similarity, the extracted instance attributes are converted to be vectors, using vector space model to represent the attributes of the item, the item can be expressed as $t_i = \{(t_1, w_1), (t_2, w_2), \cdots (t_i, w_i)\}$, where t_i is an attribute of item t, w_i denotes the weight of attribute t_i corresponding to the given item, describing the importance degree of attribute in the item. Similarity between items is measured through using the distance between vectors, calculation formula is as follows:

$$S(i,j) = (\sum_{k=1}^{M} W_{ik} \times W_{jk}) \Big/ \sqrt{(\sum_{k=1}^{M} W_{ik}^{\ 2}) \times (\sum_{k=1}^{M} W_{jk}^{\ 2})} \tag{6}$$

The attribute weight values are calculated by TF-IDF formula, where t_i is an attribute of item T_j.

$$W(t_i, T_j) = \frac{tf(t_i, T_j) \times \log(N / n_1 + 0.01)}{\sqrt{\sum_{t_i \in T_j} \left[tf(t_i, T_j) \times \log(N / n_1 + 0.01) \right]^2}} \tag{7}$$

where $W(t_i, T_j)$ denotes the weight of t_i in T_j, $tf(t_i, T_j)$ is the frequency item t_i appear in item T_j, N is the total number of item, n_1 is the number of item t_i appeared in item set. Finally, for a item, integrated the semantic similarity and collaborative filtering, combined to a linear way to measure the similarity project.

$$S_{Inte}(i,j) = \alpha S_{Sem}(T_i, T_j) + (1-\alpha) S_{Rat}(i,j) \tag{8}$$

where α is the weight parameters, when $\alpha = 0$, the combination similarity $S_{Inte}(i,j) = S_{Rat}(i,j)$; $\alpha = 1$, combination similarity $S_{Inte}(i,j)=S_{Sem}(T_i,T_j)$. From the formula (5), we can see, OI-CFA similarity algorithm has two advantages that are the combination of similarity algorithm can further explain whether the user is interested in a particular project; under the situation of the sparse rating data or no score, still can use semantic similarity to provide reasonable recommended products. Combined with the similarity algorithm, can obtain the corresponding prediction score or recommendation, $P_{a,i}$ is the predicting value the user scored the target item, which is as shown:

$$P_{a,i}=\sum_{j=1}^{k}(P_{a,j}\times S_{Inte}(i,j))/\sum_{j=1}^{k}S_{Inte}(i,j) \tag{9}$$

(5) Produce book recommendation

After determine top-n nearest neighbor set of the target user, according to the book resource access condition of the nearest neighbor set, recommended books to the target user are generate. This system produce recommended books of the target user from the following several aspects:

(a) Recommending some classic books on some book category. Based on top-N nearest neighbor set to borrow books, you can get the book category that user is interested in, these books category may be the target users are interested in or will be interested in the books category, Recommendation system can put these classic books in the book category to recommend the user.

(b) Recommending new shelves of books on some book category. After getting books category that the target users are interested in or will be interested in, also can put these new shelves of books in the category recommended to the user;

(c) Recommending borrowing books of top-n nearest neighbor set. In determining the target users of the top-N nearest neighbor set, extracted the borrowing book by these users, in these borrowing books, the target users not to borrow the book may be the books which people are interested in, the recommendation system recommend these not borrowing books to the users.

4. EXPERIMENTAL ANALYSES

4.1. Experimental Data Sets

The adopted data set came from Book-Ranting data in an open source data sharing website Data tang (http://www.datatang.com/data/44305), which includes 271,380 books information and 1,149,781 pairs scoring data from users, the number range that each user scored on each book is from 1 to 5, randomly selected 4000 of the data as the experimental data. In order to realize the semantic similarity algorithm based on the proposed OI-CFA method, using the extracted ontology from data set to wrapper agent the extracted book instance, every book instance includes semantics information.

4.2. Evaluation Index

Establishing a reasonable evaluation index system would greatly promote the development of the recommendation algorithm research and recommendation system, this paper adopts the most widely used recommendation algorithm MAE (mean absolute error). MAE can measure the prediction accuracy by calculating deviation between user prediction scores and user actual scores. The smaller the MEA, the higher the quality of recommendation is. The average absolute error is calculated as the formula (10) [13]:

$$MAE=\frac{\sum_{\alpha=1}^{n}\left|r_\alpha-v_\alpha\right|}{n} \tag{10}$$

In the formula (9), n is the number of items scored by target users, r_α is the actual score of users, v_α denotes the predicting score of the recommendation system.

4.3. Experimental Results Analysis

When performing the proposed OI-CFA algorithm under considering influence of weight parameter α on MAE, the optimal range of α value could be found out. Experimental results are shown in Fig. (**3**), it can be found out that the range of α is the best in the range [0.3, 0.5].

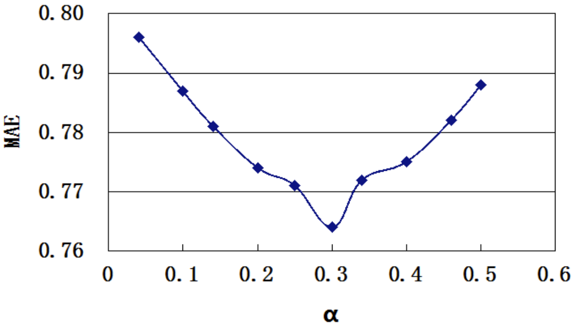

Fig. (3). Influence of α on MAE.

In order to illustrate that the proposed OI-CFA algorithm can solve two disadvantages of standard project collaborative algorithm (problems of data sparseness and cold start), we also compared the proposed OI-CFA method with item-based CFA and SVM. As α is in [0.3, 0.5] through the above experiment, the value of α is set to 0.3. Fig. (**4**) shows MAE comparative results of OI-CFA method, item-based CFA and SVM, as can be seen from Fig. (**4**), the similarity computed by OI-CFA algorithm is more accurate than the standard item-based collaborative algorithm and SVM.

Fig. (**5**) illustrates that the combined similarity algorithm OI-CFA can solve the new item predicting problem. In the case of forecasting items not scored, the collaborative filtering algorithm based on items cannot be predicted as the "cold start" problem, but the proposed OI-CFA can predict the new produced item because OI-CFA algorithm avoids the cold start problem. From the above analysis, through combining the semantic similarity with standard project collaborative algorithm, a semantic relationship between items is mined, extracting semantic information of items not only can solve the item rating sparse problem of the standard cooperative filtering algorithm and prediction problem of new items, which finally improves the recommended accuracy and further explains whether users are interested in the specified items or not.

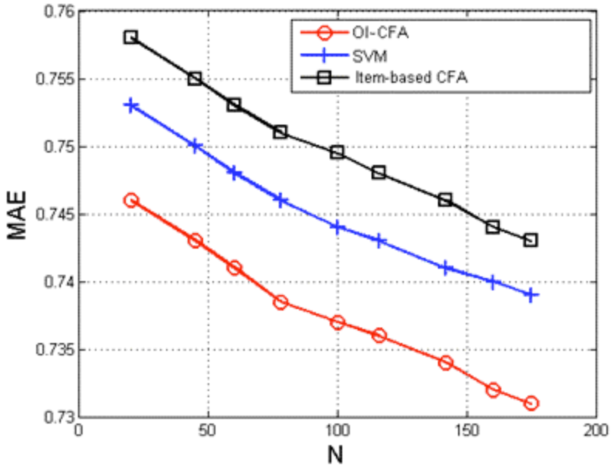

Fig. (4). Comparisons between recommendation algorithms.

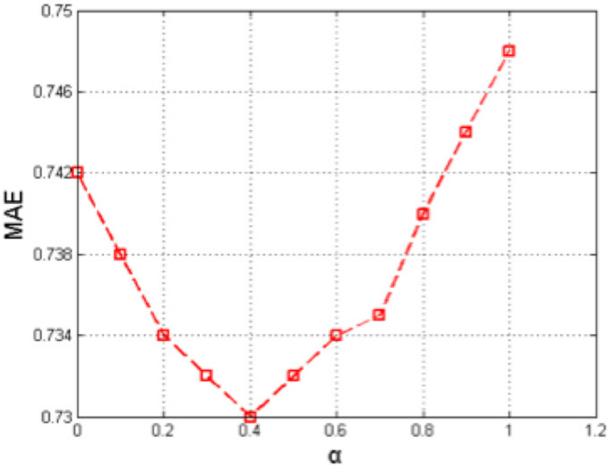

Fig. (5). New item prediction based on OI-CFA.

5. CONCLUSION AND FUTURE WORK

In this paper, personalized book recommendation method based on ontology and collaborative filtering algorithm OI-CFA is proposed, which provides a new idea for the research and practice of library personalized service. This paper extends the item-based collaborative filtering algorithm through structured semantic information to calculate the similarity of items. A domain specific ontology is constructed by ontology learning. The similarity measurement method is the combination of semantics similarity based on domain and user-item mapping similarity. Experimental results showed that advantages of the proposed OI-CFA method compared with the traditional item-based CFA and SVM, which shows that "OI-CFA" algorithm can improve the prediction accuracy of the traditional item-based CFA and also can produce reasonably accurate recommendation for new items or not scored items. In the case where data is very sparse, a higher prediction quality is provided. Our next research is that using domain features and machine learning techniques realizes the automatic determination of semantic parameters combination value and deeply study the automatic extraction and measurement of semantic similarity.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

This work was supported by the key project of Anhui Province Colleges and Universities Natural Science Foundation of China (No. KJ2014A250, No. KJ2012A263), the teaching research project at Suzhou University of China (No. szxyjyxm201413), the open project of Intelligent Information Processing Laboratory at Suzhou University of China (No. 2013YKF14, No. 2014YKF40, No. 2014YKF43, and No. 2011YKF10).

REFERENCES

[1] J. Bobadilla, F. Ortega, A. Hernando, J. Barnal, "A collaborative filtering approach to mitigate the new user cold start problem," *Knowledge-Based Systems*, vol. 26, pp. 225-238, 2012.

[2] S. J. Gong, "A collaborative filtering recommendation algorithm based on user clustering and item clustering," *Journal of Software*, vol. 5, no. 7, pp. 745-752, 2010.

[3] D. Goldberg, D. Nichols, B.M. Oki, and D. Terry, "Using collaborative filtering to weave an information tapestry," *Communications of the ACM*, vol. 35, no. 12, pp. 61-70, 1992.

[4] R. Burke, "Hybrid recommender systems: survey and experiments," *User Modeling and User-Adapted Interaction*, vol. 12, no. 4, pp. 331-370, 2002.

[5] B. M. Sarwar, G. Karypis, J.A. Konstan, and J. Riedl, "Application of dimensionality reduction in recommender system,", In: *ACM WebKDD Workshop*, 2000.

[6] T.Q. Lee, Y. Park, and Y.T. Park, "A time-based approach to effective recommender systems using implicit feedback," *Expert Systems with Applications*, vol. 34, no. 4, pp. 2396-2495, 2008.

[7] G. Adomavicius, and A. Tuzhilin, "Towards the next generation of recommender systems: a survey of the state-of-the-art and possible extensions," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, pp. 734-739, June. 2005.

[8] M. Deshpande. and G. Karypis, "Item-based top-N recommendation," *ACM Transactions on Information System*, vol. 22, no. 1, pp. 143-177, 2004.

[9] J. Bobadilla, F. Ortega, and A. Hernando, "A collaborative filtering similarity measure based on singularities," *Information Processing & Management*, vol. 42, no. 2, pp. 204-217, 2012.

[10] R. Jin, J.Y. Chai, and L. Si, "An automatic weighting scheme for collaborative filtering", In: *Proceedings of 27th International ACM SIGIR Conference Research and Development in Information Retrieval (SIGIR '04)*, 2004, pp. 337-344.

[11] C.B. Huang, and S. J. Gong, "Employing rough set theory to alleviate the sparsity issue recommender system,", In: *Proceedings of the 7th International Conference on Machine Learning and Cybernetics*, pp. 1610-1614, 2008.

[12] M.H. Hsu, A personalized english learning recommender system for ESL students, *Expert Systems with Applications*, vol. 34, pp. 683-688, 2008.

[13] C. Porcel, and E. Herrera-Viedma, "Dealing with incomplete information in a fuzzy linguistic recommender system to disseminate information in university digital libraries," *Knowledge-Based Systems*, vol. 23, no. 1, pp. 32-39, 2010.

Received: September 22, 2014 Revised: November 30, 2014 Accepted: December 02, 2014