Xiaoyi Deng

Huaqiao University, China

Utilizing User Access Sequence to Mitigate the Cold Start Problem in Collaborative Filtering Recommendation

Abstract. *Collaborative filtering (CF) is one of the most successful recommending techniques, but it suffers from the cold start problem which severely affected the quality of recommendation. To address this problem, we propose a novel hybrid approach, named UAS-CF, which incorporates user access sequence into traditional CF for improving the quality of recommendation. Experiments on three datasets were carried out to evaluate the performance of our method. Our results show that our approach outperforms other methods and improves recommendation quality effectively.*

Streszczenie. *W artykule zaprezentowano nowe podejście UAS-CF do obsługi poleceń, które włącza sekwencję dostępu użytkownika do klasycznego filtrowania uwspólnionego (ang. Collaborative Filtering), w celu polepszenia jakości rekomendacji. Badania eksperymentalne, przeprowadzone na trzech sekwencjach danych, wykazują wysoką jakość rekomendacji w porównaniu z innymi metodami. (**Pełne wykorzystanie sekwencji dostępu użytkownika do eliminacji problemu „zimnego startu" w rekomendacjach filtracji uwspólnionej**).*

Keywords: collaborative filtering, cold start, user access sequence, recommender systems
Słowa kluczowe: filtrowanie uwspólnione, zimny start, sekwencja dostępu użytkownika, systemy rekomendacji.

Introduction

Nowadays, recommender systems have become most successful application of personalized recommendation in e-commerce [1]. The core of recommender systems depends on two well-known filtering algorithms: content-based filtering (CBF) and collaborative filtering (CF) [2, 3]. Recommender systems based on CBF methods generate recommendations from items that are similar to those that the target user liked in the past, and ignore opinions or preferences of other similar users [1]. In contrast, CF only depends on historical information of user ratings, and generates recommendations to a target user based on the items that similar users liked in the past, without relying on any information about the items themselves other than their ratings [4]. Therefore, CF has an advantage over CBF in situations where it is hard to analyze the underlying content, such as music, videos and other digital products or services. Due to this reason, CF has been developed over decades and widely applied in many recommender systems and Internet-related fields [2], such as Amazon, Google News and Yahoo! Music.

Despite its advances, CF suffers from several problems, such as data sparsity. The problem of data sparsity makes CF difficult to identify similar users and items, and produce accurate predictions due to the lack of ratings. The data sparsity appears in several situations. Specially, it is also known as cold start problem, when a new user or item has just entered the system. New items can not be recommended until someone rates it, and new users are not likely provided proper recommendations because of the lack of their rating or purchase history. To solve cold start problem, many different approaches have been proposed, such as clustering, singular value decomposition (SVD), latent semantic indexing (LSI) and principle component analysis (PCA) [4]. However, useful information for recommendations related to those approaches may get lost and recommendation quality may be degraded, when certain users or items are discarded [5].

In this paper, a novel CF method named UAS-CF is proposed to improving recommendation performance under cold start conditions, which integrates user access sequence with CF. And UAS-CF is tested and evaluated with experiments on multiple datasets for its performance.

Background Review

CF generates recommendations based on the users' rating data. To provide recommendations, CF tries first to search for users who have rated the same or similar items. Once the users with common tastes are found, CF will recommend the items highly rated by those users. Generally, the more items that users have rated, the more similar the users are. The procedure of CF can be stated as follows.

It is assumed that $U=\{u_i|i=1,2,\ldots,m\}$ is a set of m users and $I=\{I_j|j=1,2,\ldots,n\}$ is a set of n distinct items. The set of user ratings is denoted by $R=\{(u_i, I_j)|u_i \in U, I_j \in I\}$ which is a $m \times n$ matrix, as shown in equation (1).

$$R = \left(r_{u_i, I_j}\right)_{m\times n}, \quad r_{u_i, I_j} = \begin{cases} S & \text{if } u_i \text{ rated } I_j \\ \varnothing & \text{if } u_i \text{ not rated } I_j \end{cases} \tag{1}$$

where $r_{u,I}$ is the rating of the item I by user u, which indicates the user's preference for different items.

After the data preparation, CF needs to select a similarity function to measure how similar two users are. Two of the most well-known similarity measures are Cosine-based similarity and Pearson correlation coefficient [1] defined in equations (2) and (3).

$$Sim\left(u_i, u_j\right) = \frac{\sum_{i\in I(u_i,u_j)} r_{u_i,I} \cdot r_{u_j,I}}{\sqrt{\sum_{i\in I(u_i,u_j)} r_{u_i,I}^{\,2}} \cdot \sqrt{\sum_{i\in I(u_i,u_j)} r_{u_j,I}^{\,2}}} \tag{2}$$

$$Sim\left(u_i, u_j\right) = \frac{\sum_{i\in I(u_i,u_j)} (r_{u_i,I} - \overline{r}_{u_i})(r_{u_j,I} - \overline{r}_{u_j})}{\sqrt{\sum_{i\in I(u_i,u_j)} (r_{u_i,I} - \overline{r}_{u_i})^2} \sqrt{\sum_{i\in I(u_i,u_j)} (r_{u_j,I} - \overline{r}_{u_j})^2}} \tag{3}$$

where $r_{u,I}$ is the rating of item I by user u; $\overline{r}_u$ is mean rating of user u, and $I(u_i, u_j)$ represents the items co-rated by users u_i and u_j

With the development of e-commerce, the magnitudes of users and commodities grow rapidly, while users' rating information is sufficient. This resulted in extreme sparsity of user rating data. To solve the sparsity problem, Anand and Bharadwaj [6] proposed various sparsity measure schemes based on local and global similarities for achieving quality predictions. Due to the extreme situation of data sparsity, i.e. cold start problem, Leung et al [7] utilized association rules to integrate domain items information into traditional CF, and introduced a preference model to comprise user-item relationships and item-item relationships. Ahn [8] applied a heuristic similarity measure method that focuses on improving the recommendation performance under the cold start conditions. Gunawardana [9] presented a novel

approach to alleviate the cold start problem using tied Boltzmann machine. Givon and Lavrenko [10] proposed a hybrid method combining social tags and CF to solve the cold start problem for improving book recommendations. Kim et al [11] designed an error-reflected model derived from explicit ratings to eliminate the effect of cold start problem for enhancing the accuracy of the prediction. Shinde and Kulkarni [12] introduced a novel centering-bunching-based clustering algorithm (CBBC) to overcome information overload for a better rating prediction. Said et al [13] investigated and evaluated the effects of weighting schemes on different types of users, and found weighting schemes can partly alleviate the cold start problem for improving precision values.

These previous researches have made several improvements on traditional CF algorithms, and they partially reduced the effect of data sparsity on the rating prediction. However, these methods significantly degraded when they were lack of information of user access sequence.

The proposed UAS-CF method

In daily life, when new users who have registered and presented few votes enter an e-commerce website, they cannot receive any personalized recommendations based on traditional CF technology. Therefore new users may feel that the recommender system does not offer the service they expected, and they may stop using it. Actually, these users may usually browse some product pages which can reflect the interests or preferences of these users. This kind of browsing behaviours will probably play an important role in producing recommendations for new users. Therefore, some researchers made analysis of the user browsing path by using clustering methods [14]. However, if the same user accesses the same product in different way (directly or indirectly), it will lead to a different browsing path which may result in different recommendations generated by these approaches. Meanwhile, operators of e-commerce websites are most concerned about which products had been purchased or browsed by the users, which can be employed to analyze the users' preferences. Apparently, user access sequence is more important than user browsing path in CF-based recommender system. However, the user access sequence was not utilized for improve the recommendation quality of recommender systems in these studies. To address the above issues, this paper introduces user access sequence to mitigate the new user cold start problem, in order to enhance prediction quality in CF-based recommendation.

The user access sequence is based on the assumption that two products are similar in human mind when they share similar access sequences among multiple users. For example, user access sequences on seven different items by five users are shown in Table 1.

Table 1. An example of user access sequences

	i_1	i_2	i_3	i_4	i_5	i_6	i_7
u_1	1	1	1	0	0	0	0
u_2	1	0	0	0	1	1	1
u_3	1	0	0	1	1	1	0
u_4	1	1	1	0	0	0	1
u_5	1	0	0	0	1	1	0

In Table 1, number 1 represents the item is accessed by corresponding users while number 0 means not. It is clear that i_1, i_2 and i_3 are similar from the views of u_1 and u_4. Also, i_1, i_5 and i_6 are similar because they are accessed by u_2, u_3 and u_5.

The analysis of user access sequence is stated as follows. It is assumed that m users are denoted by set $U=\{u_i|i=1,2,\ldots,m\}$, n distinct items are denoted by set $I=\{I_j|j=1,2,\ldots,n\}$ and user access sequence is marked as set $S(u)$, whose lengths is denoted by $|S(u)|$ as shown in equation (4).

$$S(u)=\left\langle u_i,\left\{I^{u_i}\middle|I_j^k\in I\right\}\right\rangle,\ \left|S(u)\right|=k \tag{4}$$

where I^{u_i} denotes the items accessed by u_i and u_j.

User access sequence is a unidirectional growing sequence, and it can be decomposed into a plurality of different lengths which is called sub-sequence. The sub-sequence is denoted by $S(u^k)$, defined in equation (5).

$$S\left(u^k\right)=I_j^1\rightarrow\cdots\rightarrow I_n^k\ ,\ 1\le k\le n \tag{5}$$

where k indicates the length of sub-sequence, and j is the order number of a item in I. When k=1, $S(u^1)$ represents a certain item accessed by u; when k=n, $S(u^n)$ is the user access sequence of u, $S(u)$.

For instant, the user access sequence of u_2 in Table 1 is $S(u_2)=I_1^1\rightarrow I_5^2\rightarrow I_6^3\rightarrow I_7^4$, and all sub-sequences of $S(u_2)$ are shown in Table 2.

Table 2. All sub-sequences of user access sequence $S(u_2)$

$S(u^1)$	$S(u^2)$	$S(u^3)$	$S(u^4)$
I_1^1	$I_1^1\rightarrow I_5^2$	$I_1^1\rightarrow I_5^2\rightarrow I_6^3$	$I_1^1\rightarrow I_5^2\rightarrow I_6^3\rightarrow I_7^4$
I_5^2	$I_5^2\rightarrow I_6^3$	$I_5^2\rightarrow I_6^3\rightarrow I_7^4$	
I_6^3	$I_6^3\rightarrow I_7^4$		
I_7^4			

When user access sequence and corresponding sub-sequences are obtain. User access sequence is employed to measure user similarity ,which is described as follows.

For two different users, the similarity measurement of is based on not only their user access sequence, but also the corresponding sub-sequences of user access sequence. In other words, both the items accessed by two users and the intersections of their sub-sequences should be considered when calculating similarity. It is assumed that the user access sequences of users u_i and u_j are denoted by $S(u_i)$ and $S(u_j)$ respectively, $|S(u_i)|=m$, $|S(u_j)|=n$; and the sub-sequences of $S(u_i)$ and $S(u_j)$ are denoted by $\bigcup_{1\le k\le n}s(u_i^k)$ and $\bigcup_{1\le k\le n}s(u_j^k)$ respectively. Therefore, the similarity should be measured by $\bigcup_{1\le k\le n}s(u^k)$ (k-length sub-sequences) and $S(u)$ (full-length sequences), respectively.

Firstly, the similarity measurement of sub-sequence $\bigcup_{1\le k\le n}s(u^k)$ is introduced in this part. Let $Sim(u_i,u_j)_{S(u_i^k),S(u_j^k)}$ be the similarity of k-length sub-sequences on $\bigcup_{1\le k\le n}s(u_i^k)$ and $\bigcup_{1\le k\le n}s(u_j^k)$; let the S_k be the union of $\bigcup_{1\le k\le n}s(u_i^k)$ and $\bigcup_{1\le k\le n}s(u_j^k)$. Assumed that the length of S_k is l, i.e. $|S_K|=l$, the i-th sub-sequences of S_k is denoted by $S_{k,i}$, and the ratio of sub-sequences in $\bigcup_{1\le k\le n}s(u^k)$ containing $S_{k,i}$ are denoted by $S_{k,i}(u)$. And a $l\times2$ matrix $M_{u_i,u_j}(k,2)$ is used to store the similarity between u_i and u_j on a k-length sub-sequences $\bigcup_{1\le k\le n}s(u^k)$. According to matrix $M_{u_i,u_j}(k,2)$, vectors $\overrightarrow{S_{u_i,k}}$ and $\overrightarrow{S_{u_j,k}}$ are employed to denote $\bigcup_{1\le k\le n}s(u_i^k)$ and $\bigcup_{1\le k\le n}s(u_j^k)$ respectively, as shown in equation (6).

$$\begin{aligned}\overrightarrow{S_{u_i,k}}&=\left\{S_{k,1}(u_i),S_{k,2}(u_i),\cdots,S_{k,l}(u_i)\right\}\\ \overrightarrow{S_{u_j,k}}&=\left\{S_{k,1}(u_j),S_{k,2}(u_j),\cdots,S_{k,l}(u_j)\right\}\end{aligned} \tag{6}$$

Thus, vector similarity measurement method can be utilized for the similarity measurement between $\overrightarrow{S_{u_i,k}}$ and $\overrightarrow{S_{u_j,k}}$, i.e. $Sim\left(\overrightarrow{S_{u_i,k}},\overrightarrow{S_{u_j,k}}\right)$, as defined in equation (7).

$$Sim\left(\overrightarrow{S_{u_i,k}},\overrightarrow{S_{u_j,k}}\right)=Sim\left(u_i,u_j\right)_{S\left(u_i^k\right),S\left(u_j^k\right)}=\frac{\overrightarrow{S_{u_i,k}}\cdot\overrightarrow{S_{u_j,k}}}{\left\|\overrightarrow{S_{u_i,k}}\right\|\cdot\left\|\overrightarrow{S_{u_j,k}}\right\|} \tag{7}$$

$$1\le k\le \min(m,n)$$

The similarity measurement of full-length sequence $S(u)$ is described as follows. For two different users u_i and u_j, their user access sequences are denoted by $S(u_i)$ and $S(u_j)$, and the length of $S(u_i)$ is usually not equal to that of $S(u_j)$, i.e. $|S(u_i)| \neq |S(u_j)|$. However, traditional similarity measurement method such as Manhattan and Euclidean distance can not be used to calculate the similarity between $S(u_i)$ and $S(u_j)$. In this paper, the Levenahtein distance widely applied in the field of natural language processing is introduced for the measuring similarity between $S(u_i)$ and $S(u_j)$. The similarity measuring procedure is described as follows.

Let vector $\overrightarrow{S_u}$ denote the user access sequence $S(u)$; let S_u^I denote a certain item in $\overrightarrow{S_u}$; and let $Sim\left(u_i,u_j\right)_{S(u_i),S(u_j)}$ denote the similarity measurement between $S(u_i)$ and $S(u_j)$. Then, a $(m+1)\times(n+1)$ matrix P is constructed to store the Levenahtein distances, as defined in equation (8) .

$$P=\left(P_{ij}\right)_{(m+1)\times(n+1)}=\min\begin{Bmatrix} m_{i-1,j}+1 \\ m_{i,j-1}+1 \\ m_{i-1,j-1}+d \end{Bmatrix} \tag{8}$$

where d is an integer variable; If $S_{u_i}^I=S_{u_j}^I$, d=0; else, d=1.

When matrix P is estiblished, it can be found that the value of $P_{m+1,n+1}$ is equal to the Levenahtein distance between $\overrightarrow{S_{u_i}}$ and $\overrightarrow{S_{u_j}}$. Thus, the similarity measurement of user access sequence can be calculated by equation (9).

$$Sim\left(u_i,u_j\right)_{S(u_i),S(u_j)}=\left|\frac{P_{m,n}}{\max(m,n)}-1\right| \tag{9}$$

After both k-length sequence similarity and full-length sequence similarity are obtain, the final the similarity measurement can be figured out by equation (10).

$$\begin{aligned} Sim\left(u_i,u_j\right)&=\sqrt{Sim\left(u_i,u_j\right)_{S\left(u_i^k\right),S\left(u_j^k\right)}\cdot Sim\left(u_i,u_j\right)_{S(u_i),S(u_j)}} \\ &=\sqrt{\frac{\overrightarrow{S_{u_i,k}}\cdot\overrightarrow{S_{u_j,k}}}{\left\|\overrightarrow{S_{u_i,k}}\right\|\cdot\left\|\overrightarrow{S_{u_j,k}}\right\|}\cdot\left|\frac{P_{m,n}}{\max(m,n)}-1\right|} \end{aligned} \tag{10}$$

Experimental Results

In this section, a numerical experiment is designed to test and evaluate UAS-CF. The experiment on three real-world datasets is carried out on a computer with Intel Xeon E3-1230 3.3GHz CPU, 16GB RAM and Windows 2003 operation system. And the other three CF algorithms are used as the benchmarks in this experiment.

All the experiments are carried out on three real world datasets for completeness and generalization of results, as shown in Table 1. These three datasets are publicly open for research purpose and provided by GroupLens Research Group at University of Minnesota and NetFlix Company. The sizes of the three datasets are given in Table 3. MovieLens and NetFlix datasets provide ratings on movies in the scale of 1 to 5.

Table 3. Characteristics of three real world datasets

Dataset	User	Movie	Rating	Sparsity
MovieLens-100K	943	1682	100K	6.30%
MovieLens-10M	71567	10681	10M	1.31%
NetFlix-100M	117000	8500	190M	1.92%

For all the experiments, all datasets are randomly divided into two groups: 80% of the data is used as a training set and 20% of the data is used as a test set. In the other word, 80% of the users are utilized as the reference for similarity calculation, and actual recommendation is conducted to 20%; similarly, 80% of the movies are used for similarity calculation, while 20% are actually recommended to users.

In order to evaluate the performance of our approach, the following metrics are selected: Mean Absolute Error (MAE), Root Mean Square Error (RMSE) and Coverage. Metrics are define in equations (11)~(13).

$$MAE=\sum_{i=1}^{N}\left|P_i-Q_i\right|\Big/N \tag{11}$$

where P_i is the rating prediction, Q_i is corresponding real rating and N is the number of user rating in rating matrix. The lower MAE is, the better prediction performance is.

$$RMSE=\sqrt{\sum_{i=1}^{N}\left(P_i-Q_i\right)^2\Big/N} \tag{12}$$

where P_i is the rating prediction, Q_i is corresponding real rating and N is the number of user rating in rating matrix. And the lower RMSE is, the better prediction performance is.

$$Coverage=\sum_{u\in U}\left|IP(u)\cap IR(u)\right|\Big/\sum_{u\in U}\left|IR(u)\right| \tag{13}$$

where set $IP(u)$ is rating prediction for user u, set $IR(u)$ is corresponding real ratings of user u and users are denote by set U. If Coverage is higher, the recommendation performance is better.

To compare the performance of our algorithm, three other typical CF algorithms are implemented: a item-based CF algorithm proposed by Sarwar and et al (denoted by kNN-20) [3], an item-based CF approach based on error-reflected (IErrorCF) [11], and a popular item-based CF method named weighted SlopeOne (W-SlopeOne)[15]. kNN-20 applies Cosine-based similarity to predict rating; IErrorCF employs Cosine-based similarity to perform rating prediction; and W-SlopeOne utilizes rating deviation to predict user rating. Our proposed UAS-CF is evaluated by comparing with the three benchmark algorithms.

The experimental results from comparisons of MAE, RMSE and Coverage of four algorithms on three dataset are shown in Figures 1~3, respectively.

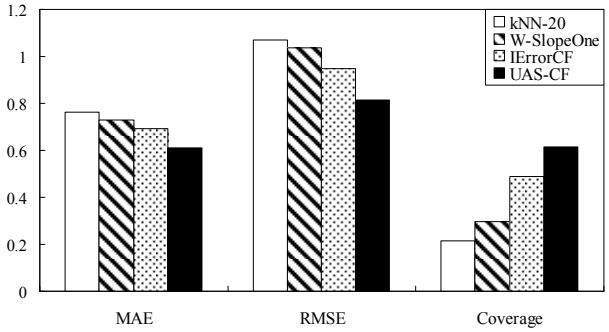

Fig.1. Comparisons of four algorithms' results on MovieLens-100K

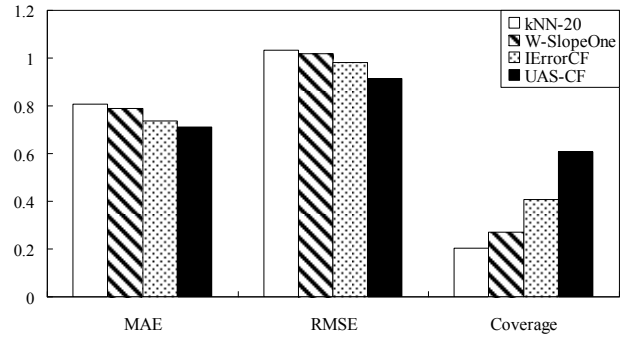

Fig.2. Comparisons of four algorithms' results on MovieLens-10M

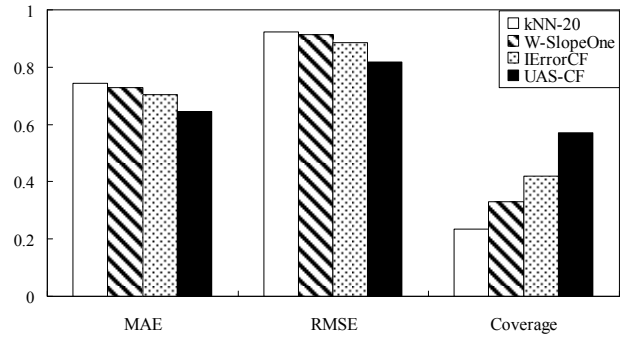

Fig.3. Comparisons of four algorithms' results on NetFlix-100M

From Figures 1 to 3, it is clear that our UAS-CF outperforms the other three typical CF approaches, and it can effectively improve the quality of collaborative recommendation.

Conclusion

This paper presents an improved collaborative filtering method UAS-CF to enhance the prediction quality of collaborative recommendation. UAS-CF employs user access sequence for similarity measurement to search target users' nearest neighbourhoods and reduce the impact of cold start problem on prediction quality. The experimental results have shown that UAS-CF succeeds in advancing the quality of rating prediction. Compared with the other algorithms, UAS-CF has both the minimum MAE and the minimum RSME. Moreover, UAS-CF has the maximum value of Coverage. This means UAS-CF outperforms the other three typical CF approaches in terms of quality. This indicates that UAS-CF is more applicable in situations where context and relationship are critical to the success of the application.

Our future will be carried out on two aspects. On one aspect, our work is related to the social tag. Recently, tagging technology has been used to describe contents of products/services and users' interests on Internet. It would be very functional to use tags for providing navigation among all the profiles of users who share a particular tag to describe their likes/dislikes. Thus, we would like to investigate the usage of social tagging in our method for facilitating rating predictions. On the other aspect, we have to face the problem of having a huge amount of data which is highly increased, so we will need to make our computations with more extensibility. To deal with this, we plan to study the possibility of parallelizing our algorithms.

Acknowledgments

This research is supported by the Scientific Research Foundation of Huaqiao University. The author would like to express his gratitude to those colleagues who provided the constructive comments and suggestions.

REFERENCES

[1] Su X.Y., Khoshgoftaar T.M., A Survey of Collaborative Filtering Techniques, *Advances in Artificial Intelligence*, 2009(2009), No.4, 1-19

[2] Linden G., Smith B. and York J., Amazon.com recommendations: item-to-item collaborative filtering, *IEEE Internet Comput.*, 7(2003), No.1, 76-80

[3] Sarwar B., Karypis G., Konstan J. et al, Item based collaborative filtering recommendation algorithms, *Proc. of the 10th international conference on World Wide Web*, Hong Kong, (2001), 285-295

[4] Park D.H., Kim H.K., Choi I.Y., Kim, J.K., A literature review and classification of recommender systems research, *Expert Systems with Applications*, 39(2012), No.11, 10059-10072

[5] Adomavicius G., Toward the next generation of recommender systems: a survey of the state-of the-art and possible extensions, *IEEE Tran. Knowl. Data En.*, 17(2005), No.6, 734-7497

[6] Anand D., Bharadwaj K.K., Utilizing various sparsity measures for enhancing accuracy of collaborative recommender systems based on local and global similarities, *Expert Systems with Applications*, 38(2011), No.5, 5101-5109

[7] Leung C.W., Chan S.C., Chung F., An empirical study of a cross-level association rule mining approach to cold start recommendations, *Knowledge-Based Systems*, 21(2008), No.7, 515-529

[8] Ahn H.J., A new similarity measure for collaborative filtering to alleviate the new user cold starting problem, *Information Sciences*, 178(2008), No.1, 37-51

[9] Gunawardana A., Meek C., Tied boltzmann machines for cold start recommendations, *Proceedings of the 2008 ACM Conference on Recommender Systems*, 2008, 19-26

[10] Givon S., Lavrenko V., Predicting social-tags for cold start book recommendations, *Proceedings of the 3rd ACM Conference on Recommender Systems*, 2009, 333-336

[11] Kim H.N., El-Saddik A., Jo G.S., Collaborative error-reflected models for cold start recommender systems, *Decision Support Systems*, 51(2011), No.3, 519-531

[12] Shinde S.K., Kulkarni U., Hybrid personalized recommender system using centering-bunching based clustering algorithm, *Expert Systems with Applications*, 39(2012), No.1, 1381-1387

[13] Said A., Jain B.J., Albayrak S., Analyzing weighting schemes in collaborative filtering: Cold start, post cold start and power users, *Proceedings of the 27th Annual ACM Symposium on Applied Computing*, 2012, 2035-2040

[14] Kuo R.J., Liao J.L., Tu C., Integration of ART2 neural network and genetic K-means algorithm for analyzing Web browsing paths in electronic commerce, *Decision Support Systems*, 40(2005), No.2, 355-374

[15] Lemire D., Maclachlan A., Slope One Predictors for Online Rating-Based Collaborative Filtering, *Proceedings of the SIAM Data Mining Conference*, 2005, 471-475

***Authors**: dr. Xiaoyi Deng, College of Business Administration, Huaqiao University, Chenghuabei Road No.269, Fengze District, 362021 Quanzhou, China, E-mail: Londonbell@hqu.edu.cn*