International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:10, No:8, 2016

A Recommender System Fusing Collaborative Filtering and User's Review Mining

Seulbi Choi, Hyunchul Ahn

***Abstract*—**Collaborative filtering (CF) algorithm has been popularly used for recommender systems in both academic and practical applications. It basically generates recommendation results using users' numeric ratings. However, the additional use of the information other than user ratings may lead to better accuracy of CF. Considering that a lot of people are likely to share their honest opinion on the items they purchased recently due to the advent of the Web 2.0, user's review can be regarded as the new informative source for identifying user's preference with accuracy. Under this background, this study presents a hybrid recommender system that fuses CF and user's review mining. Our system adopts conventional memory-based CF, but it is designed to use both user's numeric ratings and his/her text reviews on the items when calculating similarities between users.

***Keywords*—**Recommender system, collaborative filtering, text mining, review mining.

I. Introduction

RECOMMENDER systems apply knowledge discovery techniques to the problem of making personalized recommendations for information, products or services during a live interaction [1]. Especially, CF is well known for the most successful technique for recommender systems. CF has actively been studied and applied in real-world applications [2].

A general CF system compares users based on how similar they are, and creates recommendation results with the items favored by other like-minded people. Thus, it is very important for CF to measure the similarities between users because the recommendation quality depends on it. In most cases, users' explicit numeric ratings of items (i.e. quantitative information) have only been used to calculate the similarities between users in CF. That is, traditional CF depends on users' numeric ratings as its sole source of user preference information. However, user ratings cannot fully reflect user's actual preferences from time to time. For example, Nick and Sara both like Restaurant X's pasta, so they give it five stars as their numeric ratings. But, the reasons behind their favorable ratings may be different. Nick may like the pasta because of its sauce, whereas Sara may love it due to its noodle. Consequently, it may be inappropriate to identify similar users solely based on ratings [3], [4].

Today, users have become more comfortable with expressing their opinions on the Web using text in the era of Web 2.0 [5]. Online consumer reviews have emerged as an important information source for consumers to share and acquire information about products. Such reviews have potential to provide a recommender system with more detailed, nuanced, and reliable user preference information. In other words, user text reviews can be used, in conjunction with numerical ratings, to augment the word-of-mouth recommendation process. But, there have been few studies on how those consumer reviews can be combined with conventional CF system to improve the prediction accuracy of the consumers' preferences [3], [4]. Under this background, this study presents a hybrid recommender system that integrates user text reviews and numeric ratings to each user in order to improve accuracy than general CF system.

Seulbi Choi is with the Graduate School of Business IT, Kookmin University, Seoul 02707 South Korea (e-mail: seulbimon@kookmin.ac.kr).

Hyunchul Ahn is with the Graduate School of Business IT, Kookmin University, Seoul 02707 South Korea (phone: +82-2-910-4577; fax: +82-2-910-4017; e-mail: hcahn@kookmin.ac.kr).

II. Theoretical Background

A. Recommender System

A recommender system is an automated and sophisticated decision support system that frees users from a complicated search process by providing personalized solution [6]. It typically provides a user with a list of Top-*N* recommended items he or she may be interested in, or predict how much he or she might like each item [1], [7]. Although many different approaches have been proposed to the problem of making more accurate and efficient recommender systems, two algorithms–content-based (CB) and CF –are dominantly used [6].

The CB system analyzes content of items, and creates profile that is a representation of a user's interest in terms of items. And then, it analyzes the content of items unknown to the user, and compares it with his/her profile. Finally, it constructs recommendation results with the new items that are likely to satisfy the user [6], [8]. In short, the key of CB is how to characterize items. However, it is difficult to extract contents of the items, since it usually requires manual process using human intelligence. As a result, the application of CB has been restricted to few domains such as recommendation of news articles or web pages. Due to this critical limitation, CF is more preferred as a recommendation method than CB approaches in practical applications [9]–[11].

B. CF

CF avoids the problem of CB (i.e. the necessity of collecting extensive content information about items) by recommending the items favored by other people with similar preference pattern [6]. In general, there are two major types in CF, memory-based and model-based CF [12].

The memory-based CF approach repeatedly scans the preference database (so-called user-item matrix) to identify the peer group for a user. The people in the peer group are chosen based on their similarity in tastes. A prediction is then

computed by weighting the votes of the users in the peer group. This approach is also called the correlation-based or nearest-neighbor CF [6].

The model-based CF approach formulates a statistical or data mining model from user ratings, and uses it for predictions. This approach generally requires a shorter time for generating recommendation list in comparison to the memory-based approach. However, it needs extra time to train the data set for building a model. Moreover, it is not appropriate for the circumstances in which user preference models must be continuously updated in real time [13].

C. Text Mining

In general, the text is a collection of unstructured documents. To discover meaningful underlying information from the text, people use text mining. Text mining denotes all the tasks that try to extract probably useful information by analyzing large quantities of text and detecting patterns [14]. Text mining is similar to data mining, but the former deals with unstructured or semi-structured data sets (such as email, full-text documents, and HTML files), whereas the latter handles structured data from databases or XML files [15]. Online consumer reviews are a kind of text, so they can also be analyzed using text mining.

From a technical perspective, text mining is based on natural language processing (NLP) techniques. NLP techniques imply the methods that enables computers to understand human language. In specific, they intelligently parse the text to locate the topic. In NLP, the text can be expressed in various ways, but VSM (Vector Space Model) is generally used to express the text. VSM can summarize the topics of the text according to the frequencies of the terms used in each document of the text.

III. Proposed System

In this research, we propose a novel recommender system that generates recommendation list using numeric ratings well as text reviews. Our proposed system is basically based on memory-based CF, but it is designed to consider both ratings and text reviews when calculating similarities between users.

Fig. 1 presents six stages of the recommendation process for our proposed system. The detailed explanation on each phase of proposed model is as follows.

A. Step 1: User-Item Rating Matrix Creation

In Step 1, the system constructs user-item rating matrix by extracting data regarding users and their ratings on the products (i.e. items) from the user rating database. The user-item rating matrix makes it easier to calculate the similarity between two users since each row of the matrix represents a user's rating pattern on the items.

B. Step 2: Term Extraction and Analysis

Since user reviews written in text are unstructured in nature, they cannot be used without preprocessing in CF. Thus, in this step, the system first extracts meaningful terms from the reviews stored in the user review database, and store them as a set of keywords. In our proposed system, only nouns are considered as the candidates of keywords. Then, it counts the frequency of each keyword for each review.

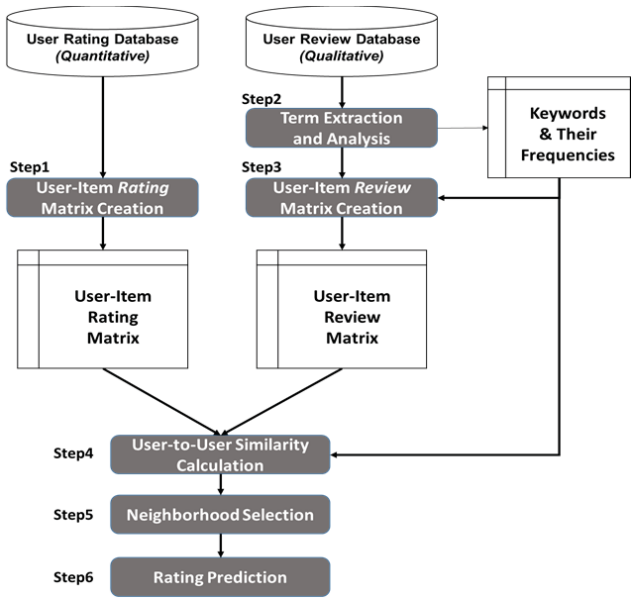

Fig. 1 Procedure of the proposed system's recommendation

C. Step 3: User-Item Review Matrix Creation

Each user review can be transformed into a set of keywords during the process in the previous step. Based on the transformation, user-item review matrix as shown in Fig. 2 is created in this step.

	Item 1	Item 2	...	Item *m*
User 1		{Convenience, Speed...}		{Discomfort, Power...}
User 2	{Design, Power, ...}			{Art, Slow, Function...}
User 3	{Price, Function...}	{Update, Friend, Power...}		{Color, Position...}
...				
User *n*-1	{Function, Speed...}			{Company, Show, Time...}
User *n*	{Family, Interest...}	{Money, Safe, Process...}		{Guest, Fail, Shock...}

Review Words (Vector)

Fig. 2 Example of User-Item Review Matrix

D.Step 4: User-to-User Similarity Calculation

Typical CF estimates the similarities between users using evaluation scores from user-item matrix. Pearson's Correlation Coefficient (PCC) is generally used as the similarity measure between two users. Equation (1) represents PCC between user *X* and *Y*.

$$S(X,Y) = \frac{\sum_{i \subseteq I_{x,y}}^{n} (R_{x,i} - \bar{R}_x)(R_{y,i} - \bar{R}_y)}{\sqrt{\sum_{i \subseteq I_{x,y}} (R_{x,i} - \bar{R}_x)^2} \sqrt{\sum_{i \subseteq I_{x,y}} (R_{y,i} - \bar{R}_y)^2}} \quad (1)$$

where $S(X,Y)$ denotes PCC between the active user *X* and each of the other user *Y* who have the co-rated items with the active user *X*, *i* is the index of each item that both user *X* and user *Y* have rated, $R_{x,i}$ is the rating of user *X* for item *i*, $R_{y,i}$ is the rating of user *Y* for item *i*, $\bar{R}_x$ is the average rating of user *X*, and $\bar{R}_y$ is the average rating of user *Y*.

As shown above, conventional PCC only considers ratings when calculating similarities between users. Therefore, it cannot be adopted in our proposed system without modification since our system should simultaneously consider the similarities of both ratings and reviews. Thus, we adopts modified PCC in this research [3]. The modified PCC can be measured using (2):

$$S^*(X,Y) = \frac{\sum_{i \subseteq I_{x,y}} \mu_{x,y,i} (R_{x,i} - \bar{R}_x)(R_{y,i} - \bar{R}_y)}{\sqrt{\sum_{i \subseteq I_{x,y}} \mu_{x,y,i} (R_{x,i} - \bar{R}_x)^2} \sqrt{\sum_{i \subseteq I_{x,y}} \mu_{x,y,i} (R_{y,i} - \bar{R}_y)^2}} \quad (2)$$

where $S^*(X,Y)$ denotes modified PCC between the active user *X* and each of the other user *Y* who have the co-rated items with the active user *X*, $\bar{R}_x = \frac{\sum_{i \subseteq I_{x,y}} \mu_{x,y,i} \times R_{x,i}}{\sum_{i \subseteq I_{x,y}} \mu_{x,y,i}}$, and $\mu_{x,y,i}$ implies the review similarity between the active user *X* and the other user *Y* on the item *i*.

In our study, $\mu_{x,y,i}$ represents the number of the keywords that are used in both user X's review and user Y's review on item i. For example, $\mu_{x,y,i}$ becomes $n(\{\text{Price, Satisfaction}\}) = 2$ when the word vector of user *A*'s review on item *i* is {Price, Power, Satisfaction, Purchase} and the word vector of user *B*'s review on the same item is {Price, Convenience, Color, Time, Satisfaction}.

E. Step 5: Neighborhood Selection

In Step 5, the system constructs the set of the nearest neighbors, *N*. The neighborhood set *N* consists of *n* neighbors who have the highest similarity to the active user. The similarities calculated in the previous step are used as a criterion for choosing nearest neighbors.

F. Step 6: Rating Prediction

The last step predicts the active user's unanswered rating from a combination of the ratings of selected neighbors. The predicted rating ($\hat{R}_{x,i}$) of the active user *X* for a target item *i* can be calculated using (3):

$$\hat{R}_{x,i} = \bar{R}_x + \frac{\sum_{v \in N} (R_{v,i} - \bar{R}_v) \cdot S^*(X,v)}{\sum_{v \in N} |S^*(X,v)|} \quad (3)$$

where $\bar{R}_x$ is the average rating of the active user *X*, v is the index of each nearest neighbor in *N*, and $S^*(X,v)$ is the similarity (modified PCC) between the user *X* and v.

IV. Conclusion

We have proposed a new kind of recommender system in order to improve the accuracy and performance than conventional CF-based recommender system. Specifically, we proposed to create user-item rating matrix as well as user-item review matrix. Based on these matrices, our proposed system can measure the similarity between users better since it calculates PCC using both user's ratings and reviews.

Previous studies such as [3] and [4] have already tried to fuse CF and user's review mining, but their approaches have some critical limitations. The method proposed by [3] requires feature extraction, which should be done by human experts. Moreover, it uses complicated refinement process since it adopts opinion mining rather than simple text mining. The method of [4] is more similar to one of our study, but it contains a critical error in its algorithm. Since the review similarity between users can vary according to items, it should be measured using all of the reviews on the co-rated items. However, the study of [4] just considered the review similarity of the target item, which is an unrealistic presupposition.

Future research plan of this study is as follows: to validate the usefulness and the applicability of the proposed system, we need to collect the experimental data set. So, we are going to build and operate a Web-based data collection system first. Then, using the data set, we will compare the performance of our proposed method with one of conventional CF system to examine the effectiveness of the proposed system.

References

[1] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Item-Based Collaborative Filtering Recommendation Algorithms," in *Proc. of the 10th international conference on World Wide Web*, pp. 285-295, 2001.

[2] K.-j. Kim, and Y. Kim, "Recommender System using Implicit Trust-enhanced Collaborative Filtering," *Journal of Intelligence and Information Systems*, vol. 19, no. 4, pp. 1-10, 2013.

[3] Z. Zhang, D. Zhang, and J. Lai, "urCF: User Review Enhanced Collaborative Filtering," in *Proc. of 20th Americas Conference on Information Systems, Savannah*, pp. 1-11, 2014.

[4] B. Jeon, and H. Ahn, "A Collaborative Filtering System Combined with Users' Review Mining: Application to the Recommendation of Smartphone Apps." *Journal of Intelligence and Information Systems,* vol. 21, no. 2, pp. 1-18, 2015.

[5] S. Dhanasobhon, P.-y. Chen, and M. D. Smith, "An Analysis of the Differential Impact of Reviews and Reviewers at Amazon.com," in *Proc. of International Conference on Information Systems*, pp. 1-17, 2007.

[6] K.-j. Kim, and H. Ahn, "Collaborative Filtering with a User-Item Matrix Reduction Technique," *International Journal of Electronic Commerce*, vol. 16, no.1, pp. 107-128, 2011.

[7] X. Yang, Y. Guo, Y. Liu, and H. Steck, "A survey of collaborative filtering based social recommender systems," *Computer Communications*, vol. 41, pp. 1-10, 2014.

[8] M. Balabanovic and Y. Shoham, "Fab: Content-based, collaborative recommendation," *Communications of the ACM*, vol. 40, no. 3, pp. 66-72, 1997.

[9] D. Billsus and M.J. Pazzani, "Learning Collaborative Information Filters," in *Proc. of the 15th International conference on Machine Learning*, pp. 46-54, 1998.

[10] Y.H. Cho and J.K. Kim, "Application of Web usage mining and product taxonomy to collaborative recommendations in e-commerce," *Expert Systems with Applications*, vol. 23, no. 2, pp. 233-246, 2004.
[11] Y.H. Cho, J.K. Kim, and S.H. Kim, "A personalized recommender system based on Web usage mining and decision tree induction," *Expert Systems with Application*, vol. 23, no. 3, pp. 329-342, 2002.
[12] J.S. Breese, D. Heckerman, and C. Kadie, "Empirical Analysis of Predictive Algorithms for Collaborative Filtering," in *Proc. of 14th Conference on Uncertainty in Artificial Intelligence*, pp, 43-52, 1998.
[13] J.B. Schafer, J. Konstan, and J. Riedl, "Electronic commerce recommender applications," *Journal of Data Mining and Knowledge Discovery*, vol. 5, no. 1-2, pp. 115-152, 2001.
[14] I.H. Witten, *Text Mining*, 2005.
[15] W. Fan, L. Wallace, S. Rich, and Z. Zhang, "Tapping the power of text mining," *Communications of the ACM*, vol. 49, no. 9, pp. 76-82, 2006.