A Preprocessing Method for Improving Effectiveness of Collaborative Filtering

Gyochang Kim
Department of Computer Engineering
College of Engineering
Myongji University
Yongin, Kyunggi-Do
kgchang@mju.ac.kr

Jonghoon Chun, Ph.D.
Department of Computer Engineering
College of Engineering
Myongji University
Yongin, Kyunggi-Do
jchun@mju.ac.kr

Sang-goo Lee, Ph.D.
School of Computer Science and Engineering
Seoul National University
Seoul
sglee@mars.snu.ac.kr

Abstract*

Collaborative filtering uses information about customers' preferences to make personal product recommendations and is achieving widespread success in e-Commerce. However, the traditional collaborative filtering algorithms do not response accurately to customers' needs. The quality of the recommendation needs to be improved in order to support personalized service to each customer. In this paper, we present novel method to improve the accuracy of the collaborative filtering algorithm. We borrow vector space model from information retrieval theory and use it to effectively discriminate the preference weights on the items for each customer. The proposed method achieves more accurate recommendations for customers who purchase similar types of products repeatedly. Our experimental evaluation on the well-known MovieLens data set shows that our method does result in a better accuracy.

1. Introduction

The explosive growth of the world-wide-web and emergence of e-Commerce has led to the development of automatic item recommendation systems. A recommendation system is based on a personalized information filtering technology, and is used to predict whether a particular user will like a particular item. Personalization service of a good quality to customer will control the success and failure of customer relationship management and e-Commerce as a whole. Many research efforts have given birth to various algorithms for such personalization services. The most representative among them is collaborative filtering. In recent years, many researches for the performance enhance of collaborative filtering have been proposed. As results, various recommendation algorithms have been proposed.

Among them, a user based recommendation system works as follows. For each user, it uses historical information to identify a neighborhood of people that in the past have exhibited similar behavior and then analyze the neighborhood to identify new pieces of information that will be liked by the user. Despite its success, GroupLens, and other similar collaborative-filtering based recommendation systems suffer from major problems [1][2][3][4].

Collaborative Filtering treats customers with different buying patterns identically. For example, customers who purchase many products evenly and others who purchase only a few products repeatedly, obviously have different buying patterns and these patterns may be used as valuable information for recommendation systems. In our work, we propose a preprocessing method in which we take into account those buying patterns and turn them into weight calculation to derive and differentiate users from one anther. We show, through experiments, that our preprocessing method indeed works better than conventional collaborative-filtering based recommendation systems.

Section 2 introduces related work about collaborative filtering. In section 3, we explain our proposed method. In section 4, we evaluate our approach on the well-known MovieLens dataset. The paper ends with a conclusion.

2. RELATED WORK

Tapestry[5] is one of the earliest implementations of collaborative filtering based recommender systems. This system relied on the explicit opinions of people from a close-knit community, such as an office workgroup. However, the case which the user does not put the comment to treat on the document is difficult be achieved for the accuracy recommendation. Later, several ratings-based automated recommender systems were developed. The GroupLens research system[1][2][6] uses a nearest- neighbor approach to find a subset of all users that have the most similar preference history as the active user. When this neighborhood is found, the similarity between the active user and a neighbor determines how much the neighbor influences the prediction for the active user. If the similarity is high, the neighbors have high influence on the final prediction. Thus, the final prediction is a weighted combination of the neighbor preferences. The above two way directly use the preference about the item of the customer. According to this reason, it's drawback that they don't recommend accuracy.

* This work was supported in part by Ministry of Health & Welfare, Korea, under the Research Project Number 02-PJI-PG6-HI03-0004.

Please address all correspondence to: Prof. Jonghoon Chun, Department of Computer Engineering, College of Engineering, Myongji University, Namdong San 38-2, Yongin, Kyunggi-Do, 449-728, Korea; jchun@mju.ac.kr; Tel:+82-31-330-6441

Our work, which uses the vector space model, a class of method to assign weights to each data object, is able to solve these problems.

3. Preprocessing Method

The traditional collaborative filtering algorithms directly use the preference about the item of the customer for recommend item to customer. But in this paper we associate weight showing the preference on item of each customer.

We use the vector space model[7][8] as follows. Let C be an m x n customer-item matrix containing historical purchasing information of m customers on n items. In this matrix, $C_{i,j}$ is one if the ith customer has purchased the jth item.

3.1. Vector Space Model

The following matrix $C(m \times n)$ is used to express to the preference about the item n of the customer m.

$$C = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1(n-1)} & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{2(n-1)} & c_{2n} \\ \vdots & \vdots & c_{ij} & \vdots & \vdots \\ c_{(m-1)1} & c_{(m-1)1} & \cdots & c_{(m-1)(n-1)} & c_{(m-1)n} \\ c_{m1} & c_{m2} & \cdots & c_{m(n-1)} & c_{mn} \end{bmatrix}$$

- m: Customer
- m: item
- C_{ij}: The preference about the item j of the customer i

Element C_{ij} of matrix C is the preference of the customer i to item j. max[C_{i1}, C_{i2},…..., C_{ij}……,C_{in}] expresses the largest preference value among items of the customer i, and thus $cf_{i,j}$ of (1) is normalized preference value by the maximum.

$$cf_{i,j} = \frac{c_{i,j}}{\max\left[c_{i1}, c_{i2}, \ldots, c_{ij}, \ldots c_{in}\right]} \quad (1)$$

(2) gives the customers who purchase many products evenly a low value, and gives the customers who purchase a specific product many times a high value. In (2), N represents total number of items and n_i represents the distinct number of items purchase by the customer i.

$$icf_i = \left(\log \frac{N}{n_i}\right) \quad (2)$$

The weight calculation of the matrix $C(m \times n)$ is done as shown in (3).

$$w_{ij} = cf_{i,j} \times icf_i \quad (3)$$

3.2. User Similarity Computation

This general formulation of collaborative filtering first appeared in the published literature in the context of the GroupLens project. The basic idea in similarity computation between set of items rated by both customers (C_1 and C_2).

There are a number of different ways to compute the similarity between customers. Here we present two such methods. These are cosine-based similarity(4), correlation -based similarity(5).

$$sim(c_1, c_2) = \cos(\vec{c_1}, \vec{c_2}) = \frac{\vec{c_1} \bullet \vec{c_2}}{\left\|\vec{c_1}\right\| * \left\|\vec{c_2}\right\|} \quad (4)$$

$$w(c_1, c_2) = \frac{\sum_{j \in ic_1c_2} (c_{c1,j} - \bar{c}_{c1})(c_{c2,j} - \bar{c}_{c2})}{\sqrt{\sum_{j \in ic_1c_2} (c_{c1,j} - \bar{c}_{c1})^2} \sqrt{\sum_{j \in ic_1c_2} (c_{c2,j} - \bar{c}_{c2})^2}} \quad (5)$$

Ic_1c_2: set of items rated by both customers(C_1 and C_2)

3.3. Prediction Computation

The prediction $p_{c1,j}$ is the prediction for customer C_1 for item j

$$p_{c1,j} = \bar{c}_1 + \frac{\sum_{i \in Ucj} w_{c1,i} \times (c_{i,j} - \bar{c}_i)}{\sum_{i \in Ucj} \left| w_{c1,i} \right|} \quad (6)$$

Ucj: set of users who rated item j

3.3. Proposed Method

We use a vector space model for our preprocessing method. The following matrix is to add expression (3) with matrix $C(m \times n)$. It is used to computation the prediction $p_{c1,j}$, of customer C_1 on item j. ($C+W=CW$)

$$\begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1(n-1)} & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{2(n-1)} & c_{2n} \\ \vdots & \vdots & c_{ij} & \vdots & \vdots \\ c_{(m-1)1} & c_{(m-1)1} & \cdots & c_{(m-1)(n-1)} & c_{(m-1)n} \\ c_{m1} & c_{m2} & \cdots & c_{m(n-1)} & c_{mn} \end{bmatrix} +$$

$$\begin{bmatrix} w_{11} & w_{22} & \cdots & w_{1(n-1)} & w_{1n} \\ w_{12} & w_{22} & \cdots & w_{2(n-1)} & w_{2n} \\ \vdots & \vdots & w_{ij} & \vdots & \vdots \\ w_{1(m-1)} & w_{2(n-1)} & \cdots & w_{(m-1)(n-1)} & w_{(m-1)n} \\ w_{m1} & w_{m2} & \cdots & w_{m(n-1)} & w_{mn} \end{bmatrix} =$$

$$\begin{bmatrix} cw_{11} & cw_{22} & \cdots & cw_{1(n-1)} & cw_{1n} \\ cw_{12} & cw_{22} & \cdots & cw_{2(n-1)} & cw_{2n} \\ \vdots & \vdots & cw_{ij} & \vdots & \vdots \\ cw_{1(m-1)} & cw_{2(n-1)} & \cdots & cw_{(m-1)(n-1)} & cw_{(m-1)n} \\ cw_{m1} & cw_{m2} & \cdots & cw_{m(n-1)} & cw_{mn} \end{bmatrix}$$

(7) is similarity measure of customer C_1 and C_2

$$w(c1,c2) = \frac{\sum_{j \in c1c2} (cw_{1,j} - \overline{cw_1})(cw_{2,j} - \overline{cw_2})}{\sqrt{\sum_{j \in c1c2} (cw_{1,j} - \overline{cw_1})^2} \sqrt{\sum_{j \in c1c2} (cw_{2,j} - \overline{cw_2})^2}} \quad (7)$$

(8) is the prediction $p_{c1,j}$ is the prediction for customer C_1 for item j

$$p_{c1,j} = \overline{cw_1} + \frac{\sum_{i \in Ucj} cw_{c1,i} \times (cw_{i,j} - \overline{cw_i})}{\sum_{i \in Ucj} \left| w_{c1,i} \right|} \quad (8)$$

4. Experimental Evaluation

4.1 MovieLens Database

We use the MovieLens dataset[9]. The database contains ratings from 6,040 users on 3,900 movies. User ratings are recorded on a numeric five-point scale. We divide the data set into a training set and test set in the ratio of 8:2. We use MAE(Mean Absolute Error) to evaluate the precision of the forecasting.

$$|\overline{E}| = \frac{\sum_{i=0}^{N} |\varepsilon i|}{N} \qquad (7)$$

4.2 Experimental Result

We implemented two different similarity algorithms pure cosine, correlations and tested them on our data sets. For each similarity algorithms, we implemented the algorithm to compute the neighborhood and used weighted sum to generate the prediction. We ran these experiments on our training data and used test set to compute Mean Absolute Error(MAE).

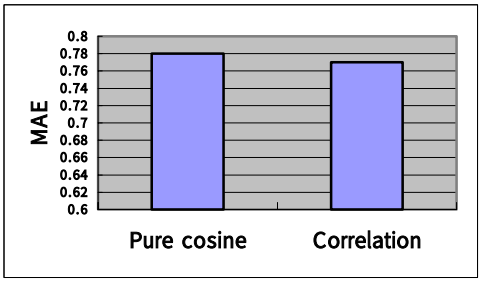

Figure 1 Impact of the similarity computation measure on user-based collaborative filtering algorithm. (No preprocessing data)

Figure 1 shows the experimental results with no preprocessing dataset. Figure 2 shows the experimental result with preprocessing dataset. It can be observed from the results that correlation computation for preprocessing method has advantage, as the MAE is significantly lower in this case. Hence, we select the adjusted preprocessing method for the rest of our experiments. Figure 3 shows experimental result that MAE of customers who purchased only a few categories of products repeatedly is lower than customers who purchased many different types of products evenly. Many customers in e-Commerce environment have tendencies to repeatedly purchase similar group of products over and over again.

5. Conclusion

Recommender systems provide powerful new technology for extracting additional value for a business. These systems benefit users by enabling them to find items they like. Conversely, they help the business by generating more sales. Recommender systems are rapidly becoming an indispensable tool in e-Commerce and need for innovative technologies are overwhelming especially, for improvement of accuracy of recommender systems.

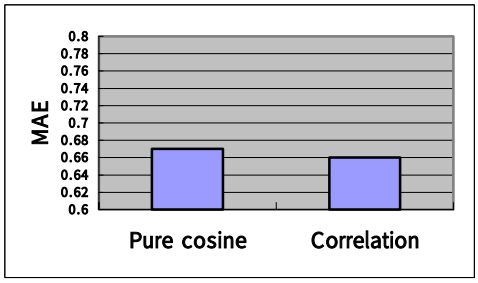

Figure 2 Impact of the similarity computation measure on user-based collaborative filtering algorithm (preprocessing data)

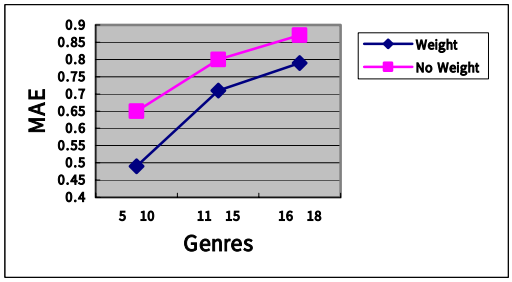

Figure 3 Sensitivity of the genres size on user-based collaborative filtering algorithm

In this paper we presented and experimentally evaluated a new preprocessing method for a user-based recommendation system. Experimental result showed that our weight calculation method provides more accurate recommendation than those provided by traditional user-based collaborative filtering techniques.

6. References

[1] B Konstan, J., Miller, B., Maltz, D., Herlocker, J., Gordon, L., and Riedl, J. (1997). GroupLens: Applying Collaborative Filtering to Usenet News.*Communications of the* ACM.

[2] Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., and Riedl, J, (1994). GroupLens: An Open Architecture for Collaborative Filtering of Netnews. *In Proceedings of CSCW '94 Chapel Hill, NC*

[3] Shardanand, U., and Maes, P. (1995). Social Information Filtering: Algorithms for Automating 'Word of Mouth'. *In Proceedings of CHI '95. Denver, CO.*

[4] Sarwar, B., Karypis, G., Konstan, J., Riedl, J.,(2001) Item-Based Collaborative Filtering Recommendation Algorithms

[5] Goldberg, D., Nichols, D., Oki, B. M., and Terry, D.(1992). Using Collaborative Filtering to Weave an Information Tapestry. *Communications of the* ACM. December.

[6] P. Resnick, N. Iacovau, M. Sushak, P. Bergstrom, and J. Riedl, "GroupLens: An Open Architecture for Collaborative Filtering of Netnews? *In Proceedings of the 1994 Computer Supported Collaborative Work Conference.*

[7] R. Baeza-Yayes, and B. Ribeiro_Neto . Modern Information Retrieval. Addison Wesley 1998.

[8] G. Salton and M. J. McGill. Introduction to Modern Information Retrieval. McGraw-Hill, 1983

[9] http://www.grouplens.org

"**Note**: The full paper is available from the CD of conference proceedings"