Clustering-based Hybrid Filtering Algorithm

Qing Li°, Byeong Man Kim, Yoon Sik Shin, En Ki Lim
Dept. of Computer Science, Kumoh National Institute of Technology
(liqing , bmkim,ysshin,eklim) @se.kumoh.ac.kr

Abstract

Recommender systems help consumers to find the useful products from the overloaded information. Researchers have developed content-based recommenders, collaborative recommenders, and a few hybrid systems. In this research, we extend the classic collaborative recommenders by clustering method to form a hybrid recommender system. Using the clustering method, we can recommend the products based on not only the user ratings but also other useful information from user profiles or attributes of items. Through our experiments on well-known MovieLens data set, we found that the information provided by the attributes of item on the item-based collaborative filter shows advantage over the information provided by user profiles on the user-based collaborative filter.

1. Introduction

Recent years we have seen the explosive growth of the sheer volume of information. Recommender system is a kind of intelligent system, which helps us prioritize information so that we can reduce the searching time and spend much more time in reading the information that we need or favor.

At the initial state, many recommender systems were fairly simple query-based information retrieval system, which can be called as content-based recommender system. Later, Goldberg and his colleagues firstly applied the collaborative filtering technology to recommender systems [1] [2]. GroupLens [3] and Ringo [4] developed independently, were the first to automate prediction. Collaborative filtering accumulates a database of consumers' product preferences, and then uses them to make recommendations for products. MovieLens system recommends movies, Jeter system recommends jokes [5], Flycasting recommends online radio [6], and GAB recommends web pages based on the bookmarks [7]. A growing number of companies, including Amazon.com, CDNow.com and Levis.com, employ or provide recommender system solutions.

Although collaborative filtering has been very successful, it can not recommend new items to users without any history and completely denies any information that can be extracted from contents of items. Further more the quality of recommendation is completely based on the user rating, instead of the information content.

For this reason, hybrid recommender systems have been provided, which can exploit both user preferences and contents. Proposed approaches to hybrid system, which combines collaborative and content-based filters together, can be categorized into two groups.

There are three main categories of hybrid recommendation systems. The first one is the linear combination of results of collaborative and content-based filters, such as systems that are described by Claypool [8] and Wasfi [9]. ProfBuilder recommends web pages using both content-based and collaborative filters, and each creates a recommendation list without combining them to make a combined prediction. Claypool describes a hybrid approach for an online newspaper domain, combining the two predictions using an adaptive weighted average: as the number of users accessing an item increases, the weight of the collaborative component tends to increase. But how to decide the weights of collaborative and content-based components is unclearly given by the author.

The second one is the sequential combination of content-based filtering and collaborative filtering. In these systems, firstly, content-based filtering algorithm is applied to find users, who share similar interests. Secondly, collaborative algorithm is applied to make predictions, such as RAAP [10] and Fab filtering systems [11]. RAAP is a content-based collaborative information filtering for helping the user to classify domain specific information found in the WWW, and also recommends these URLs to other users with similar interests. To decide the similar interests of users, scalable Pearson correlation algorithm based on the web page category is used. Fab system uses content-based techniques instead of user ratings to create profiles of users. So the quality of predictions is fully depended on the content-based techniques, inaccurate profiles result in inaccurate correlations with other users and thus make poor predictions.

The last one is the mixed combination. Both the semantic contents and ratings are applied to make recommendations, such as the probabilistic model [12] and Ripper system for recommendation [13]. Basu [13] train the Ripper machine learning system with a combination of content data and training data in an effort to produce better recommendations. Good [14] combine personal IF agents and the ratings of users to make recommendations. Popescul [12] provide a probabilistic model for unified collaborative and content-based recommendation.

In this paper, we apply clustering techniques to integrate the semantic contents of user profiles or attributes of items into the collaborative filtering to improve its recommendation performance and solve the cold start problem. We make a comparison study of these two integration methods and achieve some useful conclusion for others.

2. Our approach

Up to now, the dominant paradigm for performing collaborative filtering in recommender systems has been based on nearest neighbor regression. It has reached a high level of popularity, because they are simple and intuitive on a conceptual level. It uses a general two-step approach. First users or items are identified that are similar to some active user or items for which a recommendation has to be made. Then recommendations are computed based on the preferences of

This work was supported by Korea Research Foundation Grant (KRF-2002-041-D00459).

similar users or items.

However, as we know, the classic collaborative filtering makes recommendations only based on the user ratings disregarding some useful information, and it is hard to make recommendations when the new user or item comes, because we lack of historical information to find the nearest neighbors for this new user or item. To cope with these problems and achieve better performance, we extend the classic collaborative filtering algorithm.

The basic idea of our approach is that first we apply clustering algorithm to group the users or items, then use the result, which is represented by the fuzzy set, to create a group-rating matrix. Second normalize the group-rating matrix and combine it with original user-item matrix to form a new rating matrix. At last, using the classic collaborative to make recommendations or predications for users.

In our approach, if we group the items and apply item-based collaborative filtering algorithm [15] to make predictions, we call *ICHM* (Item-based Clustering Hybrid Method). If we group the user profiles and apply user-based collaborative filtering algorithm to make predictions, we call it *UCHM* (User-based Clustering Hybrid Method).

As Figure 1 shows, as for *UCHM*, clustering is based on the attributes of user profiles and clustering result is treated as items. However, as for *ICHM*, clustering is based on the attributes of items and clustering result is treated as users.

UCHM

	Movie 1	Movie 2	Movie 3	Cluster 1	Cluster 2
User 1	...	...	...	...	...
User 2	...	...	...	...	...
User 3	...	...	...	...	...

Item-rating Matrix Group-rating Matrix

ICHM

ICHM	User 1	User 2	User 3	Cluster 1	Cluster 2
Movie 1	...	...	...	...	...
Movie 2	...	...	...	...	...
Movie 3	...	...	...	...	...

User-rating Matrix Group-rating Matrix

Figure 1. UCHM & ICHM

In the following subsections, we will describe the detail algorithms we applied in our approach.

2.1 Clustering Algorithm

K-means Clustering Algorithm is a simple and fast clustering method, which has been popularly used [16]. So we apply it with some modifications. The difference is that we apply the fuzzy set theory to represent the affiliation between an object and a cluster. As shown in Figure 2, firstly, user profiles are grouped into a given number of clusters. After completion of grouping, the possibility of one object (here one object means one user profile or item) belonging to a certain cluster is calculated as follows.

$$Pro(j,k) = 1 - \frac{CS(j,k)}{MaxCS(i,k)} \qquad (1)$$

where *Pro(j,k)* means the possibility of object *j* belonging to the cluster *k*; The *CS(j,k)* means the counter-similarity between the object these
j and the cluster *k*, *MaxCS(i,k)* means the maximum counter-similarity between an object and the cluster *k*.

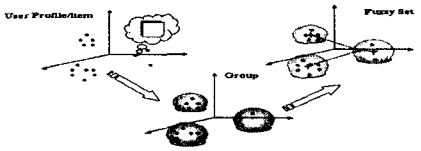

Figure 2. Adjusted K- means Clustering Algorithms

However, in our adjusted k-means algorithm, the fuzzy membership in a cluster is only assigned at the last step. It seems to represent the fuzzy memberships of objects unessentially. So the fuzzy k-means algorithm [17] is also applied, in which a fuzzy membership is assigned to each object during each iteration as Figure 3 shows.

The global cost function, membership between an object and a clust er, and the mean value of one cluster are calculated as follows.

$$GCF_{fuz} = \sum_{i=1}^{c} \left(\sum_{j=1}^{n} ((Pro_{i,j})^b \times Dis_{i,j}) \right)$$

$$Mean_j = \frac{\sum_{j=1}^{n} (Pro_{i,j})^b X_j}{\sum_{j=1}^{n} (Pro_{i,j})^b}$$

$$Pro_{i,j} = \frac{(1/Dis_{i,j})^{\frac{2}{b-1}}}{\sum_{r=1}^{c} (1/Dis_{i,j})^{\frac{2}{b-1}}}$$

where, GCF_{fuz} means the fuzzy global cost function; c means the cluster number; b is a free parameter chosen to adjust the blending of different clusters; Dis_{ij} is the Euclidean distance between the mean value of cluster i and the object j; X_j is the vector of object j; Pro_{ij} means the membership between the cluster i and the object j.

However, no matter what kind of clustering algorithms is used, how to choose the initial cluster center is a critical problem. We recommend the refinement algorithm suggested by Bradley [18].

Algorithm : Fuzzy K-means Clustering

Input: the number of clusters *k* and items attribute features.

(1) Initialize the parameters, and membership between objects and clusters;

(2) Repeat (a) and (b) until global cost function has small change;

  a) Recompute the mean value of each cluster.

  b) Recompute the membership of each object.

(3) Return the membership.

Figure 3. Two Clustering Algorithms

2.2 Similarity Computation and Collaborative prediction

Due to difference in value range between item-rating matrix (or user-rating matrix) and group-rating matrix, we should normalize them to the same level. As for item-ratings (or user-rating) matrix, the rating value is integer; As for group-rating matrix, it is the fuzzy set value ranging from 0 to 1. In our approach, we transform the discrete data range from [1 5] to [0 1] and then apply Pearson correlation-based algorithm [19] to calculate similarity.

Prediction for an item is then computed by performing a weighted average of deviations from the neighbor's mean. Here we use top *N*

or items according to which method we apply - *UCHM or ICHM*. Details can be referred to [19].

3. Experimental Evaluation

Currently, we perform experiment on a subset of real movie rating data collected from the MovieLens web site. The data subset contained 100,000 ratings from 943 users and 1,682 movies, with each user rating at least 20 items. The ratings in the MovieLens data are explicitly entered by users, and are integers ranging from 1 to 5. We divide data set into a training set and a test data set. 20 percent of MovieLens data are used as a training data set; the other 80 percent are used as a test data set.

Since the MovieLens data set do not contain any other information of movies except the genre information, as for UCHM, we only use the genre information of movie to create the user profiles. Details can be referred to our former work [19]. As for ICHM, we group the items, only based on one attribute – movie genre. So we can make a fair comparison between ICHM and UCHM.

3.1 Clustering Algorithm Effection

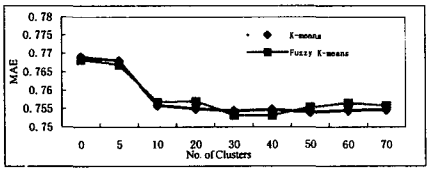

Figure 4 Clustering Algorithm

As for UCHM, we implement grouping rating method described in section 2.1 and test them on the MovieLens data with the different number of clusters. Figure 4 shows the experimental results. It can be observed that the number of clusters does affect the quality of prediction. As we have discussed before, the fuzzy k-means algorithm seems more essentially represent the fuzzy membership than the adjusted k-means algorithm. However, in our experiment, it does not show obvious advantages, in addition, as for ICHM we get the similar result. Since the computation complexity of fuzzy k-means algorithm is heavier than the adjusted k-means algorithm, we choose our adjusted k-means algorithm in following parts.

3.2 Comparison

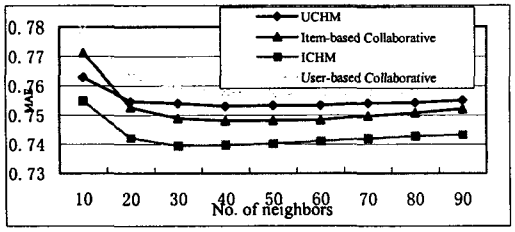

Figure 5 Comparison

The size of the neighborhood has significant effect on the prediction quality [15]. It can be observed from Figure 5 that the size of neighborhood does affect the quality of prediction. When the number of neighbors changes from 30 to 50 in our approach, it arrives at the optimal *MAE* value.

As Figure 5 shows, the ICHM and UCHM compare favorably with rule to select the nearest *N* neighbors based on the similarities of users item-based collaborative algorithm and user-based collaborative algorithm respectively. Furthermore, the ICHM shows the best performance among all of them.

4. Conclusion

In this paper, we extend our former work [19] to item-based collaborative filtering framework. Our comparison study shows that the correct application of the item information can further improve the recommendation performance.

REFERENCES

[1] Goldberg, D., Nichols, D., Oki, B. M. and Terry, D. (1992). Using Collaborative Filtering to Weave an Information Tapestry. Communication *of the ACM Science*, Vol. 35, pp.61-70.

[2] Douglas B. Terry (1993) A tour through tapestry. *Proceedings of the ACM Conference on COOCS.*

[3] Resnick, P., Iacovou, N., Suchak, M., Bergstorm, P. and Riedl, J. (1994). GroupLens :An open architecture for collaborative filtering of Netnews. *the Conference on Computer Supported Cooperative Work*

[4] Upendra, S. and Patti, M.(1995). Social Information Filtering: Algorithms for Automating "Word of Mouth" . *ACM CHI'95 Conference on Human Factors in Computing System.*

[5] Gupta, D., Digiovanni, M., Narita, H. and Goldberg, K. (1999). Jester 2.0: A New Linear-Time Collaborative Filtering Algorithm Applied to Jokes. *Proceedings of Workshop on Recommender Systems: Algorithms and Evaluation.*

[6] Hauver, D. B. and French,J. C. (2001). Flycasting: Using Collaborative Filtering to Generate a Play list for Online Radio. *International Conference on Web Delivery of Music.* Italy.

[7] Wittenburg, K., Das, D., Hill, W. and Stead, L. (1995). Group Asynchronous Browsing on the World Wide Web. *In Proc. of Fourth International World Wide Web Conference.*

[8] Claypool, M., Gokhale, A., Miranda, and Sartin, M.(1999). Combining content-based and collaborative filters in an online newspaper. *ACM SIGIR '99 Workshop on Recommender Systems*

[9] Wasfi, A. M. A. (1999). Collecting User Access Patterns for Building user Profiles and Collaborative Filtering. *International Conference on Intelligent User Interface.*

[10] Delgado, J., Ishii, N. and Ura, T. (1998). Content-based Collaborative Information Filtering: Actively Learning to Classify and Recommend Documents. *CIA' 98.*

[11] Balabanovic, M., Shoham, Y. (1997).Fab: Content-Based, Collaborative Recommendation. *Communications of the ACM.*

[12] Popescul, A., Ungar, L. H., and Lawrence,S. (2001). Probabilistic Models for United Collaborative and Content-Based Recommendation in Sparse-Data Environments. *UAI 2001.*

[13] Basu C., and Cohen (1998). Using Social and Content-based information in Recommendation. *In Proc. of the AAAI-98.*

[14] Good N., Borchers, A., Sarwar, B., Herlocker, J., and Riedl, J. (1999). Combining Collaborative Filtering with Personal Agents for Better Recommendations. *In Proc. of the AAAI-99.*

[15] Sarwar,B., Karypis,G., Konstan, J. and Riedl,J. (2001). Item-based Collaborative Filtering Recommendation Algorithms. *Proceedings of the 10th WWW Conference.* Hong Kong.

[16] Han, J., and Kamber, M.(2000). Data mining: Concepts and Techniques. New York: Morgan-Kaufman.

[17] Duda, Richard O., Hart,Peter E. and Stork,David G. (2000) Pattern Classification. *Wiley-Interscience Publicatio.*

[18] Bradley, P. S. and Fayyad U.M. (1998) Refining Initial Points for K-Means Clustering. *ICML '98.*

[19] Qing Li, B.M. Kim. "An Approach for Combining Content-based and Collaborative Filters", IRAL2003, workshop of ACL2003, Sapporo, Japan.