Clustering for Probabilistic Model Estimation for CF

Qing Li[1, 2], Byeong Man Kim[2], Sung Hyon Myaeng[1]
[1]Information & Communications University, Daejeon, 305-732, Republic of Korea
(liqing, myaeng)@icu.ac.kr
[2]Kumoh National Institute of Technology,Kumi, kyungpook, 730-701, Republic of Korea
bmkim@se.kumoh.ac.kr

ABSTRACT

Based on the type of collaborative objects, a collaborative filtering (CF) system falls into one of two categories: item-based CF and user-based CF. Clustering is the basic idea in both cases, where users or items are classified into user groups where users share similar preference or item groups where items have similar attributes or characteristics. Observing the fact that in user-based CF each user community is characterized by a Gaussian distribution on the ratings for each item and the fact that in item-based CF the ratings of each user in item community satisfy a Gaussian distribution, we propose a method of probabilistic model estimation for CF, where objects (user or items) are classified into groups based on the content information and ratings at the same time and predictions are made considering the Gaussian distribution of ratings. Experiments on a real-world data set illustrate that our approach is favorable.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval –*information filtering.*

General Terms

Algorithms, Experimentation.

Keywords

Information filtering, Probabilistic model, Collaborative filtering

1. INTRODUCTION

Collaborative filtering (CF) refers to the technique wherein peer opinions are employed to predict the interests of others. A target user is matched against the database to discover neighbors, which are other users who have similar historical tastes. Items that neighbors like are then recommended to the target user. Sarwar [1] extended the concept to items, where a target item is matched against the database to discover similar items which share similar ratings from users and proved that item-based CF is better than user-based CF on precision and computation complexity. However, in order to capture accurate similarities among items or users, both have to face the three challenges: non-association, user bias and cold start problem [3].

In our previous work, we proposed a clustering method to integrate the content information into the collaborative filtering framework to alleviate those challenges [3]. Traditional CF methods make recommendation only based on the historical user ratings; more specifically, they calculate user similarities based on user ratings to find out neighbors. Our previous method extended the user rating matrix by adding a group rating matrix which is obtained from content information, as shown in Table 1. With the information from content, this method alleviates the three challenges and results in a good recommendation performance.

Observing the fact that in user-based CF each user community where users share similar preferences is characterized by a Gaussian distribution on the ratings for each item, we propose a probabilistic model to predict the user rating in this paper. The similar observation can be obtained item-based CF. Therefore, our approach is realized with two models: user-based and item-based probabilistic model. When computing the similarities of user to build a community, we are not only based on ratings but also on the content information of users or items.

Table 1. Example for item-based case

User rating matrix						Group rating matrix		
item / user	1	2	3	4		Group / user	G 1	G 2
1	5		1			1	98%	4%
2		4				2	95%	5%
3						3	15%	96

2. PROBABILISTIC MODEL

The domains we consider consists of a set of users $U = \{u_1, ...u_n\}$, and a set of items $Y = \{y_1, ...y_m\}$ and a set of possible rating $V = \{v_1, ...v_k\}$. In CF, we are interested in the condition probability $P(v|u,y)$ that user u will rate item y with rating v. If rating v possesses a numerical scale, then it is appropriate to define the deterministic prediction function $g(u,y)$ which indicates the user u's rating on item y, as

$$g(u,y) = \int_v v p(v|u,y) dv$$

where $p(v|u,y)$ denotes a probability mass function (discrete case) or a conditional probability density function (continuous case) dependent on the context. As for user-based model, we introduce a variable z which can be treated as the interest group of users or user community. Therefore, $p(v|u,y)$ can be calculated as

$$p(v|u,y) = \sum_z p(z|u) p(v|z,y)$$

Since users in the same community share similar preferences for items, most of user ratings on a certain item will fall into the same rating range. We assume that the ratings for a certain item y made by users in community z satisfy a Gaussian distribution:

$$p(v|z,y) = p(v;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp[-\frac{(v-\mu)^2}{2\sigma^2}]$$

Therefore, the deterministic prediction function $g(u,y)$ that predicts the user u's rating on item y can be computed as:

$$g(u,y) = \int_v v p(v|u,y) dv = \int_v \sum_z p(z|u) p(v|z,y) v dv$$
$$= \sum_z p(z|u) \int_v p(v|z,y) v dv = \sum_z p(z|u) \mu_{y,z}$$

WWW 2005, May 10-14, 2005, Chiba, Japan.
ACM 1-59593-051-5/05/0005

The user u 's rating on item y can be regarded as the sum of the product of $\mu_{y,z}$ (the average rating on a certain item y in a user community z) and the posteriori probability $p(z|u)$ which dependents on the relationship between user u and user community z represented by $p(u|z)$:

$$p(z|u) = \frac{p(u|z)p(z)}{\sum_{z'=1}^{k} p(u|z')p(z')} = \frac{ED'(V_u,V_z)^{-1}\frac{|C_z|}{|C_u|}}{\sum_{z'=1}^{k} ED'(V_u,V_{z'})^{-1}\frac{|C_{z'}|}{|C_u|}} = \frac{ED'(V_u,V_z)^{-1}|C_z|}{\sum_{z'=1}^{k} ED'(V_u,V_{z'})^{-1}|C_{z'}|}$$

where V_u denotes the rating vector of user u , V_z denotes the centre vector of variable z , C_z, C_u are the set of community z and all users respectively, $ED'(.)$ is the function of adjusted Euclidean distance and is applied to measure the relationship of user u and user community z . $p(z)$ is computed as the ratio of users falling into the community z over all users. We assume that the relationship between user u and user community z is determined by the Euclidean distance. A shorter distance represents a close relationship. We found it helpful to adjust the Euclidean distance of two users based on the number of items in common:

$$ED'(V_u,V_z) = \frac{\max(|V_u \cap V_z|,\beta)}{\beta} ED(V_u,V_z)$$

where $ED(.)$ denotes the Euclidean distance function, and $|V_u \cap V_z|$ denotes the number of items in common.

The model introduced above groups users into similar interesting communities and makes recommendation based on the relationship of users. An alternate approach is to classify items into similar item communities and predict by the relationships of items. In this item-based probabilistic model, we also introduce a variable z , which refers to the similar group of items instead of user community in user-based case.

As we can see, how to construct the user or item community z in both models is a key point. In both models, we create the community z by a K-means clustering algorithm based on both user ratings and content information extracted from user profiles or item attributes. From our previous work [2], content information extracted from user profiles or item attributes contribute to alleviate the three challenges aforementioned. Thus, it helps us capture more accurate similarities between users or items. Subsequently, we can build more accurate communities to put similar users or items into groups. We adopted our previous approach [3] to create a group rating matrix based on content information for similarity computation, as shown in Table 1.

3. EXPERIMENTAL EVALUATION

We carried out experiments based on the EachMovie data. There are 1,623 movies (items) in this data set and 61,265 users with a total of over 2.8 million ratings. This data set is to our knowledge the largest publicly available data set for collaborative filtering testing. MAE (mean absolute error) and Score for Ranked List [4] are evaluation metrics.

Comparison: We have implemented a baseline method to evaluate the achieved results where users are randomly grouped into classes. In addition, we have implemented a standard memory-based CF method - Simple Pearson method [4]. Our previous method [3] realized in user-based case called UCHM and in item-based case called ICHM is also presented as a comparison. As shown in Table 2, our approach realized in user-based probability model (*UPM*) and item-based probability model (*IPM*) show a better performance than others.

Table 2. Comparison

Method	Rank Scoring	Rel. improv.	MAE	Rel. improv.
Baseline	13.46	0	1.472	0
User-based Pearson	20.46	52%	1.03	30%
Item-based Pearson	21.50	59.7%	0.984	33.15%
UCHM	21.10	56.76%	0.99	32.74%
ICHM	22.78	69.24%	0.974	33.83%
UPM	23.16	72.1%	0.952	35.33%
IPM	23.56	75%	0.946	35.73%

4. Future Work

Our current work only focuses on the numerical ratings, however, binary ratings is widely used on the internet and studied by many researchers [5]. We will extend our approach to the binary ratings based on the Bernoulli or Poisson models, which are more accurate than Gaussian model for binary data.

5. ACKNOWLEDGEMENTS

This work was supported by Korea Research Foundation Grant (No. R05-2004-000-10190-0).

6. REFERENCES

[1] Sarwar, B. M., Karypis, G., Konstan, J. A. and Riedl, J(2001).Item-based Collaborative Filtering Recommendation Algorithms, In Proc. of the Tenth Int. WWW Conf. 2001, pp. 285-295

[2] Li, Qing, B.M. Kim, G. D. Hai, D.H Oh (2004). A Music Recommender based on Audio Features, In Proc. Of the SIGIR-04, Sheffield, UK

[3] Q. Li, B.M. Kim. Clustering Approach for Hybrid Recommender System, In Proc. of IEEE/WI, Canada. pp. 31-37, 2003.

[4] Breese, J. S., Heckerman, D. and Kardie, C. (1998). Empirical Analysis of Predictive Algorithms for Collaborative Filtering. In Proc. Of the 14th UAI, pp.43-52.

[5] Nasraoui,O. and M. Pavuluri (2004). Accurate Web Recommendations Based on Profile-Specific URL-Predictor Neural Networks. In Proc. of the WWW04, NY.