Proceedings of the Sixth International Conference on Machine Learning and Cybernetics, Hong Kong, 19-22 August 2007

SEMANTIC-ENHANCED PERSONALIZED RECOMMENDER SYSTEM

RUI-QIN WANG[1,2], FAN-SHENG KONG[1]

[1] Artificial Intelligence Institute, Zhejiang University, Hangzhou 310027, China
[2] School of Computer Science and Engineering, Wenzhou University, Wenzhou, 325035, China
E-MAIL: jsj_wrq@wzu.edu.cn

Abstract:

Personalized recommender systems have emerged as a powerful method for improving both the content of customers and the profit of providers in e-business environment. Nowadays, many kinds of recommender methods have been proposed to provide personalized services. However, all these techniques have not made full use of the semantic information of objects, which leading them to an unsatisfying performance. Collaborative filter (CF) system, as the most popular personalized recommender systems, has such well-known limitations as sparsity, scalability and cold-start problem. A semantic-enhanced collaborative recommender system is proposed in this paper. The semantic information of objects is extracted to support the recommendation process. This study compares the performance of the proposed technique with the traditional CF approaches. Experimental results demonstrate the effectiveness of the proposed method.

Keywords:

Personalized recommendation; Semantic information; Collaborative filter; Clustering; Ontology

1. Introduction

In recent years, the development of WWW is so rapid that the information in the internet is overwhelming, which produces a challenge for us. How can we get the proper and usable information in such a huge volumes of data is not an easy case. The web mining based recommender system has emerged as the times require. Researchers have proposed all kinds of recommender systems, such as content-based filter, collaborative filter, and hybrid recommender systems. Among them collaborative filter is the optimum one, but we all know that it has many limitations such as sparsity, scalability and cold-start problem.

In order to solve the sparsity problem of the traditional collaborative filter method, Mukund Deshpande and George Karypis [1] proposed a model-based recommendation algorithm via computing the similarities of item-pairs instead of user-pairs in the user-item rating matrix, and then using these similarities to produce the recommendations. Deng, Zhu and Shi [2] adopted the same method and use the item-based nearest neighbor collaborative filter to predict the scores of the unrated items. Multi-thematic interest measurement method on the basis of item-based recommendation to capture users' interests and preferences was introduced in [3]. Conor Hayes and Pa´draig Cunningham proposed content-enhanced collaborative recommendation in [4], which adopted content-based filter technique to capture the current interests of the active user and used this knowledge to improve the recommendation precision. The demographical information was used in [5] to enhance the recommendation effect, which combined the demographical information of users and items via weighed average computation.

In addition, in order to overcome the limitation of individual recommendation algorithm, there are some hybrid methods in succession. A hybrid recommendation technique based on users and items was proposed in [6], in which the rating matrix was looked as a structure having missing data. The author filled the missing elements of the matrix from three aspects: the other users' ratings toward the objective item, the active user's rating toward the other items and the other similar users' rating toward the other similar items. In order to overcome the scalability problem, cluster method on the users and items are used in [7] to reduce the dimension of data. It was also a hybrid method and produced the recommendation from user and item aspects. Janusz Sobecki combined the demographical information, the content of items and the rating information to produce recommendation in [8].

The above methods aim at improving the collaborative filter technique via introducing other recommendation technique or other data sources, but they all overlooked the semantic data of the domain objects which are the most important information in recommendation system.

Recently, a new direction, intelligent recommendation, has emerged that was focused on the semantic and ontology information underlying the users or items. An item feature based recommendation technique was proposed in [9] to support the discovery of user's preference and interest via analyzing the purchase behavior model of users from the

1-4244-0973-X/07/$25.00 ©2007 IEEE

transaction records and products feature database. So it had no new-item problem. A kind of recommender method based on items' descriptors was put forward in [10], in which multiple items' descriptors were kept in the recommender system. Whenever a request comes, the system decides the recommended item via the description information of the items. This method not only had good recommendation effect but provided an easy way to understand the discovered knowledge by using the descriptors. Li and Zhong [11] offered an automated method to capture the ontology information of system via feedbacks of the users and using this ontology information to improve the recommender effect. Bezdek J C. [12] proposed a novel user preference model based on the domain ontology and used this ontology to describe the documents in a digital library and the user profiles on the documents. This model can impress the structure and semantic of user preference, at the same time it provided complicated preference operations to support the personalized retrieval and recommendation in digital library.

Although the methods above introduced certain semantic or ontology information to the recommendation process, but their methods are too complicated that are unfit to the ubiquitous application system. At the same time constructing integrated and comprehensive domain ontology is not a trivial.

In this work, we offer a novel semantic-enhanced collaborative filter recommendation method, in which the recommendation is produced using the semantic information of the category features of item as well as the user demographical data together with the usage data in form of user-item rating matrix. At the same time, we adopted the cluster method on users based on above three data sources offline to solve the scalability problem. This study compares the performance of the proposed technique with the traditional CF approach. Experimental results demonstrate the effectiveness of the proposed method in both recommendation precision and scalability.

The remainder of the paper is organized as follows: Section 2 introduces the algorithm of ontology construction of item category. Section 3 presents the approach of semantic-based user clustering via using multiple data sources. In Section 4, we introduce the whole recommendation process. We conduct experiments to compare our algorithms with the CF method in section 5 and conclude and discuss some future work in Section 6.

2. Ontology construction of item category

In order to understand and make full use of the semantic information of items, we must create the domain ontology of the referred system. Among the ontology knowledge, the most important information to the personalized recommendation system is the category of the objects. So in this section we focus on introducing the approach of ontology construction of the object category. We use multi-dimension array to organize the hierarchy structure of the ontology of object category.

The item category ontology is built upon five basic operations:

1) Making sure the total number of object categories (k) in the referred system,
2) For each object, deciding which categories it should belongs to (one object can belongs to more than one categories),
3) Constructing tree-like category hierarchy, in which every category unit is the combination of the category units in its previous layer. The number of individual categories in the units is equal in the same layer. Like the following figure 1 shows,
4) Allocating the objects which fall into these category units, at the same time using multi-dimension array to keep the objects in every category unit,
5) Computing the similarities of the object-pairs on the basis of the ratio of their common shared categories.

For a given domain, the total category number (k) and the categories that every object belongs to are usually known or can be easily found, so the work in step (1) and (2) are effortless. The computational cost of step (3) is $O(2^K)$. So constructing the complete category hierarchy ontology is time-consuming. In reality, we only construct an ontology hierarchy at a given depth for simplicity. In section 5, the experiment system on the movie domain uses a three-level ontology hierarchy.

Let t_i and t_j be two objects and they share n_{ij} common individual categories. There are N individual categories altogether in the referred domain. The similarity measure between two objects can be defined by the ratio of the shared categories of them. The larger the number is, the more similar the two objects are. Formally:

$$SIM\ (t_i, t_j) = \frac{n_{ij}}{N} \tag{1}$$

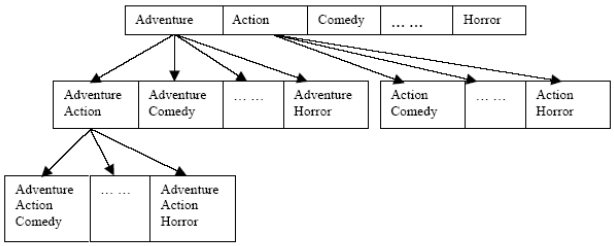

Figure 1. the hierarchy ontology of object category

Besides the category feature, the objects have other features that can be integrated into the domain ontology to provide more accurate recommendation and prediction. For example, in the movie domain there are other features such as actors, directors besides the movie category character, namely genre, can be combined into the ontologies. Even more we can extract the topic terms or key phrases from the description information about the movies as their semantic knowledge, which is our further research direction in future.

3. Semantic-enhanced user clustering

Our proposed approach is based on a collaborative filter method. The traditional collaborative filter inherits the scalability problem, namely, time-consuming user-pair similarity computation in the online recommendation stage due to the huge number of users in the system. In this work we cluster the users offline to reduce the online computational overload with the aim to overcome the scalability problem.

Previous clustering methods are almost based on the user-item matrix. Because of the sparsity of the rating matrix and no semantic information about users and items is included, the clustering effect is unsatisfied. In this study, we group the users from multiple aspects. Besides the user-item rating matrix, we also include the semantic information, user demographical data and item category features, in the computation of the user-pairs similarities.

The clustering process has the following five steps:

1) *Computing the history rating similarity of two users based on the user-item rating matrix.*

We use the updated Pearson Correlation measurement to compute the similarity between two users, u_i and u_j. Formally:

$$sim(u_i,u_j)^1=\frac{\sum_k (r_{ik}-\overline{r_i})\bullet(r_{jk}-\overline{r_j})}{\sqrt{\sum_k (r_{ik}-\overline{r_i})^2}\bullet\sqrt{\sum_k (r_{jk}-\overline{r_j})^2}} \quad (2)$$

Where r_{ik} and r_{jk} be the ratings of u_i and u_j toward item k; $\overline{r_i}$ and $\overline{r_j}$ be the average rating of u_i and u_j, respectively; k is the number of items which have been commonly rated by both u_i and u_j.

2) *Computing the demographical similarity of two users based on their demographical information.*

The demographical information of users we collected may include many dimensions. We need only the key dimensions that related to our recommendation task, such as age, occupation and so on. Much of the demographical information is nominal, but age is a numeric attribute. We can partition the age data into such subsections that people in the same subsection have the similar interest and preference. For example, we partition the age every ten years from 1 to 100 and compute their similarity in ages. Formally:

$$s_{ij}^{\ 1}=1-\frac{round(age_i/10)-round(age_j/10)}{10} \quad (3)$$

Other similarities of the nominal demographical features can be decided via the equal/unequal comparison, which result in 1 or 0.

The final demographical similarity between two users can be computed via weighed average of all the demographical component similarities as:

$$sim(u_i,u_j)^2=\sum_{k=1}^{n} s_{ij}^{\ k}*w_k \quad (4)$$

$$\sum_{k=1}^{n} w_k=1 \quad (5)$$

where $s_{ij}^{\ k}$ be the k[th] demographical feather similarity of u_i and u_j, respectively; w_k be the weightiness factor corresponding to the k[th] demographical feather in the similarity computation; n is the number of the selected demographical feathers.

3) *Computing the interest and preference similarity of two users based on the semantic similarities of items they have accessed or rated.*

$$sim(u_i,u_j)^3=\max(sim(u_i\rightarrow u_j),sim(u_j\rightarrow u_i)) \quad (6)$$

$$sim(u_i\rightarrow u_j)=\frac{\sum_{k1}\max_{k2}(SIM(t_{k1},t_{k2}'))}{k1} \quad (7)$$

$$sim(u_j \rightarrow u_i) = \frac{\sum_{k2} \max_{k1}(SIM(t_{k2}', t_{k1}))}{k2} \quad (8)$$

Where u_i has accessed/rated items t_1 to t_{k1} and u_j has accessed/rated items t_1' to t_{k2}'.

4) *Producing the final similarities of the user-pairs based on the above three similarities using weighed average method:*

$$sim(u_i, u_j) = \sum_{k=1}^{3} sim(u_i, u_j)^k * \alpha_k \quad (9)$$

$$\sum_{k=1}^{3} \alpha_k = 1 \quad (10)$$

Where α_k is the given weightiness factor.

5) Clustering the users using the k-means algorithm.

Note that any clustering algorithm can be used to generate user clusters instead of the k-means algorithm. We have chosen the k-means algorithm as the partition generation mechanism, mostly for its low computational complexity.

4. Recommendation process

When all the preparations have done, we can go into the online recommendation stage. The intelligent recommender system first analyzes the active user's preference vector and confirms the most favorite item categories of the active user through the following steps:

a. *Constructing the preference vector of the active user $p_i=(c_{i1},c_{i2},\ldots,c_{in})$, where c_{ij} is the user interest preference degree toward the j^{th} object category. Initially, all the vector elements are set to zero. Namely, $c_{ij}=0$, where i=1 to m, j=1 to n, m is the user number and n is the category number of items.*

b. *For each user u_i, analyzing the categories of every object that u_i have accessed or rated, update u_i's preference vector by adding one to the corresponding element according to the object's categories.*

c. *Reordering the vector elements in descending order, and the top-N biggest elements corresponding to the most favorite item categories of the active user.*

Online recommendation will be confined to the discovered most favorite item categories of the active user. Thus the recommendation range will be largely decreased.

Then we should make certain the cluster which the active user belongs to. In fact, the cluster he/she belongs to has already been fixed during the user clustering stage. The other users except the active user in the cluster will collaboratively predict the rating scores of the objects that the active user has never rated before.

Finally, the system recommends the top-k objects to the active user.

5. Experimental Results and Discussions

To evaluate the performance of the semantic-enhanced recommendation approach and compare the proposed approach with the conventional CF methods, we conduct experiments with artificial and real world datasets, *MovieLens*. *MovieLens* datasets were collected through the *MovieLens* web site (movielens.umn.edu) during the seven month period from September 19th, 1997 through April 22nd, 1998. This data has been cleaned up. The data set consists of 100,000 ratings (1-5) for 1682 movies by 943 users.

By using the above mentioned rating data, we compared our semantic-based recommendation method with the traditional collaborative filtering (TCF) in Pentium(R) 4, 2.41GHz CPU, 512MB Memory, Windows 2003 OS, MATLAB 7.0.1. The predicted precision is used to measure the quality of a recommendation method. Precision is the matching degree of the predicted rating ($rate\_pre_i$) of item t_i (i=1 to N) with the actual rating ($rate\_act_i$) in the testing data set. Formally:

$$precision_i = \frac{abs(rate\_pre_i - rate\_act_i)}{\max\_rating} \quad (11)$$

$$precision = \frac{\sum_{i \in test} precision_i}{|N|} \quad (12)$$

We split the data sets into five subsections by average, namely U1 to U5, and the data in each subsection was divided into a 75% training set and a 25% testing set. The experiment results are listed in table 1.

Table 1. Performance Comparison between our approach and the traditional CF approach

Data sets	Our approach (%)	Traditional CF (%)
U1	83.67	71.73
U2	82.35	70.42
U3	82.56	70.18
U4	83.03	70.02
U5	82.13	69.97

From the table we can conclude that our approach outperforms the CF method dramatically in predicted accuracy, which mainly because we used the semantic information of user demographical data and the item

category data in addition to the usage data of the user-item rating matrix in the recommendation process. Because the rating matrix is so sparsity that performance of the traditional collaborative filter method is poor.

Figure 2 shows the precision of rating prediction of our proposed approach with the traditional Collaborative Filtering (TCF) at different user/item size. From the figure we can conclude that our approach consistently outperforms the CF method dramatically in predicted accuracy.

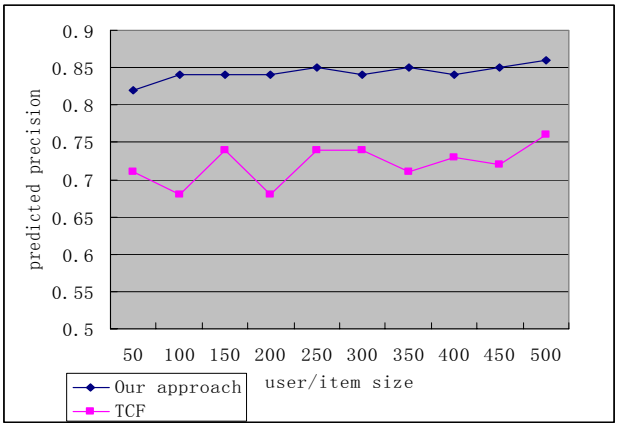

Figure 2. Precision of our approach and traditional CF

Figure 3 shows the online predict time of our approach and traditional CF. From the figure we can see our approach is stable consistently, however the predict time of CF increases linearly with the increase of user and item number. That is to say, as the scale of the recommendation system develops largely, the online recommendation time will become intolerable.

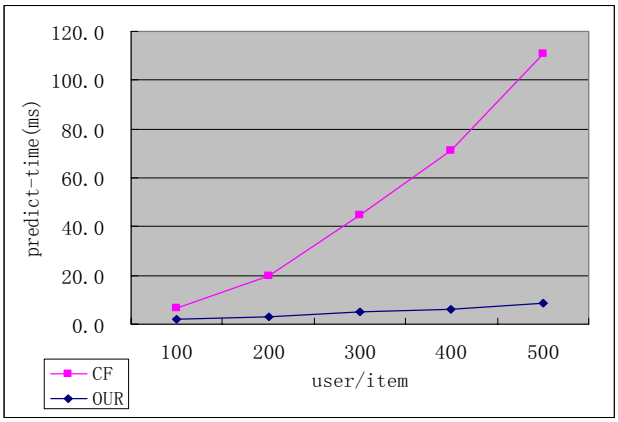

Figure 3. online predict time of our approach and CF

6. Conclusions and future works

We have proposed a semantic-enhanced technique to the personalized recommender system in this paper. The recommendation and prediction is produced via using the usage data and semantic knowledge. At the same time, users are grouped offline according to their similarities from multiple aspects. Experiments in real world data sets indicate the good performance of the proposed approach comparing with the traditional collaborative filter recommendation method. Among the advantages of the approach are its steady high precision in prediction and recommendation and the short online recommendation time

Besides the category feature, the objects have other features that can be integrated into the domain ontology to provide more accurate recommendation and prediction. Our further research direction is to construct more comprehensive domain ontology of the application system. Furthermore, we will apply our proposed semantic-based recommendation method to the real applications to test its performance.

References:

[1] Item-Based Top-N Recommendation Algorithms, MUKUND DESHPANDE and GEORGE KARYPIS, ACM Transactions on Information Systems, Vol. 22, No. 1, January 2004, Pages 143–177.

[2] DENG Ai-Lin, ZHU Yang-Yong, SHI Bai-Le. A Collaborative Filtering Recomm endation Algorithm Based on Item Rating Prediction. Journal of Soflware, 1000．9825 / 2003 / l4(09)l62

[3] Improving Recommendation Lists Through Topic Diversification, CaiNicolas Ziegler,Sean M. McNee,Joseph A. Konstan,Georg Lausen, WWW 2005, May 1014,2005, Chiba, Japan

[4] Context boosting collaborative recommendations, Conor Hayes, Pa´draig Cunningham, Knowledge-Based Systems 17 (2004) 131–138

[5] Enhancing Collaborative Filtering with Demographic Data: The case of Item-based Filtering, Manolis Vozalis and Konstantinos G. Margaritis,the fourth International Conference of Intelligent Systems Design and Applications (ISDA04)

[6] Unifying User-based and Item-based Collaborative Filtering Approaches by Similarity Fusion, Jun Wang, Arjen P. de Vries, Marcel J.T. Reinders1SIGIR'06, August 6–11, 2006, Seattle, Washington, USA

[7] A hybrid collaborative filtering method for multiple-interests and multiple-content

recommendation in E-Commerce,Yu li, Liu lu, Li xuefeng, Expert Systems with Applications 28 (2005) 67–77

[8] Implementations of web-based Recommender Systems Using Hybrid Methods, Janusz Sobecki, International Journal of Computer Science & Applications vol.3 Issue 3, pp52-64

[9] Feature-based recommendations for one-to-one marketing, Sung-Shun Weng, Mei-Ju Liu, Expert Systems with Applications 26 (2004) 493–508

[10] Using Item Descriptors in Recommender Systems, Eliseo Reategui, John A. Campbell, Roberto Torres, American Association for Artificial Intelligence

[11] capturing evolving patterns for ontology-based web mining, yuefeng li,ning zhong, Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence (WI'04)

[12] Bezdek J C. Pattern Recognition with Fuzzy Objective Function Algorithm s. New York: Plenum Press, 1981.