Advanced Science and Technology Letters
Vol.66 (Networking and Communication 2014), pp.78-80
http://dx.doi.org/10.14257/astl.2014.66.19

# A Study on Predicting Preference for Recommendation by using Data Mining Technique

Jong Seok Um[1], Ji Hoon Choi [1], Min Woo Yang [1], and Byong Su Choi [1],

[1] Dept. of Mulltimedia Engineering, Hansung University,
Samsungyuro 16 gil, Seoul, Korea
{jsum, mp5678, bytesize, cbs}@hansung.ac.kr,

**Abstract.** We propose a method for predicting user's preference on the item serviced on the web pages. Here we use collaborative filtering combining user-base and item-base prediction results. To resolve data sparsity, we include similar user's rating on similar items. The result shows that proposed method gives reliable prediction on to the test user.

**Keywords:** Collaborative Filtering, Predicting Preference, Recommendation

## 1 Introduction

Now a day, most of us get information through web searching and browsing. Web search engines give a list of information on what we are searching. Users are also finding information by browsing Web directories which are well organized in hierarchical order of subjects. Conceptual model of the Web directory which is called domain ontology will improve the efficiency of information searching and browsing. Domain ontology is constructed to express the knowledge about application domain which is easy to understand by human.

After building domain ontology, we preprocess the access logs for applying collaborative filtering and usage analysis. A Web server log file contains Web request made to Web Server, recorded in chronological order. Based on the result of the collaborative filtering, we propose the method to serve personalized web pages. Also, domain ontology is updated according to the result of the usage analysis.

## 2 Proposed Methodology

Our method is divided into four steps. Figure 1 shows our method. The first step consists of constructing the domain ontology of the application area. We follow the method TODE proposed at [1]. We use three existing ontological resources: suggest upper merged ontology (SUMO, www.ontologypotal.org), WordNet2.1 (wordnet.princeton.edu), and MultiWordNet Domain(wndomains.itc.it). It consists of three hierarchical layers.

ISSN: 2287-1233 ASTL

The second step consists of preprocessing the access log in a Web server. During preprocessing, we remove images and unnecessary files having extension like js and make table according to users and resources(items).

The third step consists of applying the combined collaborative filtering and usage analysis. User-based collaborative filtering predicts a test user's interest on a test item based on the similarity between users. Cosine similarity or correlation is used for similarity metric between users. Item-based approach also predicts a test user's interest on a test item based on the similarity between items. Unifying these two similarities, we predict a test user's interest on a test item [2]. Last step consists of recommending web page items based on the collaborative filtering results.

Let assume that there are K users and M items in the web server. Then user-item matrix V = $(v_{i,j})$, where $v_{i,j}$ is a rating (or vote) of user i on item j. Let $v_{k,m}$ is a rating of the test user k on item m, which need to be estimated. Predicted rating for $v_{k,m}$ is obtained from 2 sources, user-based and item-base collaborative filtering. Let us say user-base result as $u\hat{v}_{k,m}$ and item-base result as $i\hat{v}_{k,m}$. Then each result is obtained as follows.

$$u\hat{v}_{k,m} = \overline{v}_k + \kappa \sum_{i=1}^{n} w(k,i)(v_{i,m} - \overline{v}_i) \quad (1)$$

$$i\hat{v}_{k,m} = \frac{\sum_{i=1}^{n} s(m,i)v_{k,i}}{\sum_{i=1}^{n} |s(m,i)|} \quad (2)$$

Here $\overline{v}_i$ is the mean of the user *i*, $w(k,i)$ is the similarity between user *k* and user *i* and $s(m,i)$ is the similarity between item *m* and item *i*. Pearson correlation and cosine distance are used for similarity measure. Combining these two predictions, we have a following prediction on $v_{k,m}$ where $\lambda$ is the mixture ratio. Since there is a difference between users and also difference between items, ratings are normalized. Here $\tilde{v}_m$ is the mean of the item *m*.

$$v_{i,j} = v_{i,j} - (\overline{v}_i - \overline{v}_k) - (\tilde{v}_j - \tilde{v}_m) \quad (3)$$

$$ui\hat{v}_{k,m} = u\hat{v}_{k,m} * \lambda + i\hat{v}_{k,m} * (1 - \lambda) \quad (4)$$

## 3 Experiment

We applied the proposed method for predicting preference to the job searching web site. Users for this site should join the membership to access any item on the web pages. Table 1 shows user-item matrix. Test user is user6000 and test item jobOfferGubun. For similarity measure, we use cosine distance and we choose top-6 similar users and top-6 similar item. Item similarities between test item and similar item are given at bottom row and user similarities are given at the right most column. Using this result, we get $u\hat{v}_{k,m}$= 3,955 and $i\hat{v}_{k,m}$= 4,002 and with $\lambda$=0.7 combined predicted rating is $ui\hat{v}_{k,m}$ = 3989. Comparing with actual rating which is 5078, it is

underestimated. When actual rating is high, predicted rating is underestimated and actual rating is low predicted rating is overestimated.

**Table 1.** Use-item matrix with user similarity and item similarity.

| ITEM / USER | jobOffer Gubun | selectJob Offer InfoView | getJo Manage | selectJo Manage ReqstList | selectJo Manage List | getTnMn grInfo | getSimple SysMain | USER similarity |
|---|---|---|---|---|---|---|---|---|
| user6000 | ? | 1,011 | 560 | 343 | 62 | 22 | 9 | 1 |
| user5000 | 4,076 | 767 | 391 | 504 | 7 | 14 | 11 | 0.970 |
| user486 | 3,711 | 272 | 51 | 15 | 49 | 14 | 29 | 0.913 |
| user1234 | 4,423 | 269 | 12 | 14 | 65 | 6 | 69 | 0.835 |
| user7000 | 4,112 | 409 | 201 | 294 | 2 | 12 | 21 | 0.957 |
| user1400 | 3,236 | 128 | 2 | 0 | 9 | 3 | 4 | 0.846 |
| user7346 | 3,020 | 214 | 7 | 0 | 95 | 3 | 14 | 0.799 |
| ITEM similarity | 1 | 0.884 | 0.653 | 0.623 | 0.704 | 0.902 | 0.807 | |

## 4 Discussion

Here we propose a method to combine item-base and user-based collaborative filtering. The result shows that when actual rating is high it has tendency of underestimating and vice versa. Even we use more rating data, still there is a sparsity problem. By combining two sources, we have more reliable prediction. Also recommending items could be confined to a certain hierarchical topic based on predicted ratings.

## References

1. Stamou, S., Ntoulas, A., Christodoulakis, D.: TODE: An Ontology-Based Model for the Dynamic Population of Web Directories. Information Science Reference, pp1--17 (2008)
2. Wang,J., Vries, A.P., Reinders, M.J.T.: Unifying user-based and item-based collaborative filtering approaches by similarity fusion. SIGIR '06 Proceedings of the 29th annual International ACM, pp. 501--508. ACM, New York (2006)

 