WEB PAGE RECOMMENDATION BY URL-BASED COLLABORATIVE FILTERING

Kiyotaka Takasuka, Minoru Terada
Dept. of Information and Communication Engineering, The University of Electro-Communications
Chofugaoka 1-5-1, Chofu, Tokyo, 182-8585, Japan

Kazutaka Maruyama
Information Technology Center, The University of Electro-Communications
Chofugaoka 1-5-1, Chofu, Tokyo, 182-8585, Japan

Keywords: Web page recommendation, Collaborative filtering, Implicit user profile.

Abstract: Because the number of Web pages is very huge, and still increasing, many people have difficulty to reach pages they want. Although social bookmarking and search engines are helpful, users still have to find pages themselves.

Our goal is to recommend Web pages which are supposed to be interesting for a user, without active effort by the user. We first analyzed the http traffic data in our university collected by a sniffer, and developed a recommendation system that works on URLs and their viewers (IP address).

Our system has four features: (1) collaborative filtering, (2) implicit build of user profiles, (3) exclusion of popular Web pages (4) and use of the real activity in our university.

We evaluated the effectiveness of our system by applying it to the real http transaction data and found that there were 18 successful recommended users out of 50 users.

1 INTRODUCTION

With recent progress of network infrastructure, the Internet has come into wide use for general people. The number of web pages is getting huge, including pictures and movies as well as text pages. But the way most people use the Internet is still very limited; they only cruise around some known pages or use search engines to find some specific topics. There might be many pages in which a user would have interest. In this context, web page recommendation is an active research topics today.

Recommendation systems are classified into two classes, content-based filtering and collaborative filtering. The former uses the relationship between pages by analyzing page contents. The latter focuses the relation of users by analyzing the similarity of their browsing behaviors.

Another important point is the user profile, the representation of the interest of the user. Again there are two methods to construct the profile, the explicit approach and the implicit approach. The former requires the cooperation by the user while the latter collects informations by analyzing user activity such as browsing history.

2 PURPOSE

Our goal is to construct and evaluate a web page recommendation system, which adopts the collaborative filtering method with implicit user profiling based on the browsing history of the user.

The pages we will recommend satisfy the following conditions:

1. pages viewed by users whose interest is similar to that of the user,
2. pages NOT viewed by users whose interest is not similar
3. and pages the user has not viewed yet.

The condition 2 excludes popular pages such as portal sites and search engines which are not novel for most

Takasuka K., Terada M. and Maruyama K. (2007).
WEB PAGE RECOMMENDATION BY URL-BASED COLLABORATIVE FILTERING.
In *Proceedings of the Third International Conference on Web Information Systems and Technologies - Web Interfaces and Applications*, pages 447-450
DOI: 10.5220/0001285004470450

users. In short, the recommended page is suitable for the user and not commonly known.

3 FEATURES OF OUR PROPOSAL

We have already launched a system which gathers and provides web browsing activities of users into a single centralized server. They use our pilot web browser based on IE component or our Firefox extension to send their activities in real-time. The gathered activities are provided as the rankings of currently and heavily viewed web pages, and thus they can be regarded as ”recommendation”. We also showed the usefulness of the system through a pilot experiment by student users in our earlier study(K.Maruyama,K.Takasuka,Y.Yagihara,M.Satoshi, Y.Shirai,M.Terada, 2006).

The existing system has the following features:

1. uses only browsing activities,
2. recommends web pages without analysis of web page contents
3. and retrieves users’ activities from their web browsers directly.

The recommendation system in this paper has four features:

1. collaborative filtering,
2. implicit build of user profiles,
3. exclusion of popular web pages
4. and use of the real activity in our university.

Collaborative filtering, unlike contents based one, can recommend URLs which don’t contain any texts such as images. The implicit build of user profiles reduces users’ effort in the explicit one, because users’ interest may change in short term when they see web pages with some interesting hyper links at the upper part of them.

4 RELATED WORKS

Web page recommendation has two problems inherent in web itself: the number of web pages and the location of web servers. Most of e-commerce services provide recommendations to their customers in order to increase their sales. E-commerce services such as Amazon.com(Linden et al., 2003) provide millions of items, but all web servers in the world have tens of billions of web pages. In addition, each e-commerce service has all the items to be provided in its own servers, but web pages, in contrast, are located at so many distributed web servers in the world.

Li et al.(Jia Li and Osmar R. Zaïane, 2004) proposed a web page recommendation system with collaborative filtering. It accepts access logs of a web server as its input, analyzes the contents of the accessed web pages and the behavior of users, and then produces recommended web pages. However, it can be applied only to a particular web site because of the use of access logs.

In contrast, the web page recommendation system proposed by Zhu et al.(R.Greiner, T.Zhu, G.Haubl, K.Jewell, 2005) can recommend web pages at all web servers in the world. A special web browser for the system enables it, but requires user to evaluate web pages explicitly. As mentioned above, implicit evaluation of web pages is expected.

Another approach to web page recommendation is to use bookmarks of users because a bookmark of an user shows his or her interests (Rucker and Polanco, 1997) (Jung et al., 2001). If an user kept his or her bookmark up to date, the bookmark would reflect his or her short term interests. Updating bookmarks corresponds explicit evaluation.

5 OUTLINE OF ALGORITHM

The algorithm for generating recommendation is explained in 4 steps.

step 1: 5 URLs are extracted from the history of the target user.

step 2: A group of users who have viewed one or more of the 5 URLs in step 1 is extracted out of all users.

step 3: The similarities between target user and each member of the group extracted in step 2 are computed and the group is ordered by the result. From the history of the most similar user we get candidates for the recommendation.

step 4: Computing the page score of each candidates, we determine the recommendation.

The outline of the algorithm is illustrated in Figure 1.

5.1 Recommender Group

In step 1, our system gets the latest five URLs from the history of the target user. We use latest URLs in order to reflect the short-term interest of the target user.

In step 2, a set of users who share one or more of the 5 URLs is extracted, which we refer *the recommender group* in this paper.

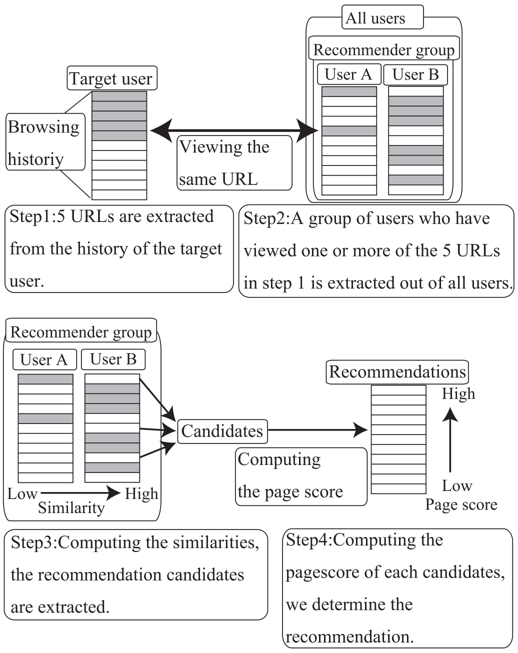
All users
Recommender group
User A
User B
Target user
Browsing historiy
Viewing the same URL
Step1:5 URLs are extracted from the history of the target user.
Step2:A group of users who have viewed one or more of the 5 URLs in step 1 is extracted out of all users.
Recommender group
User A
User B
Recommendations
High
Candidates
Computing the page score
Low
Page score
Low
High
Similarity
Step3:Computing the similarities, the recommendation candidates are extracted.
Step4:Computing the pagescore of each candidates, we determine the recommendation.

Figure 1: The outline of the algorithm.

5.2 The Similarity Between Users

In step 3, we compute the similarity between two users. To compute the similarity, we consider these two factors:

1. the number of pages viewed in common between the users
2. and the number of unique users who have viewed the page.

The second factor is introduced to make popular pages less influential. Even if two users view a popular site in common, it doesn't necessarily mean that the interest of them is similar. We use the number of unique users as the denominator in the calculation. The browsing history of user *a* is represented as $hist(a)$. The number of unique users viewing a page x is represented as $uusr(x)$. The similarity of user *a* and *b*, that is $similarity(a,b)$, is computed as follows.

$$similarity(a,b) = \sum_{x \in hist(a) \cap hist(b)} \frac{1}{uusr(x)}$$

5.3 Page Score

We computes the page score to determine the recommendations. First, we organize another group – *the control group* – containing the same number of users as the recommender group, chosen randomly from users not in the recommender group. The control group is sorted by the similarity to the target user in ascending order. We give one-to-one correspondence between the member of each group, such as the most similar user in the recommender group and the least similar user in the control group.

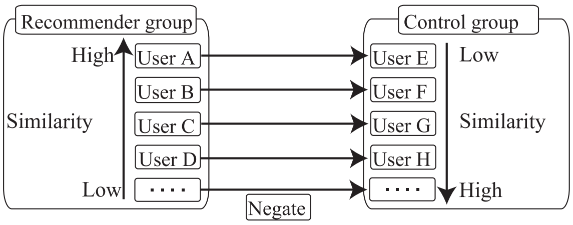

Figure 2: How to determine the vote value of the control group.

Member of each group votes for a page. The value of the vote is determined as follows. Member of the recommender group has positive value which is the similarity between the member and the target user. But, the most similar user has 1.0. Member of the control group has negative value, obtained by negating the value of the corresponding recommender member (Figure 2). But, if the user hasn't viewed the page to be scored, the vote value is 0. The most similar user in the recommender group has the highest vote value, while the least similar user in the control group has the lowest.

Let *R* be the recommender group, *C* be the control group and the vote value for user *b* be $vote(b)$, the score for a page *x* is

$$score(x) = \frac{\sum_{b \in (R \cup C)} vote(b)}{uusr(x)}$$

6 EXPERIMENT

We implemented the recommendation system for the experiment and tried to generate web page recommendations to several users picked up randomly based on the real activities of web users in our university captured by the sniffer. By examining the recommendations, which consist of multiple recommended URLs, we found some successful and unsuccessful results. A recommendation is classified as a successful one if it contains URL with page score that is higher than 1.0. If only the most similar user has viewed the URL, the URL's page score would be 1.0. If the users with high similarity have viewed the URL, the URL's page score would be higher than 1.0. If the users with low similarity have viewed the URL, the

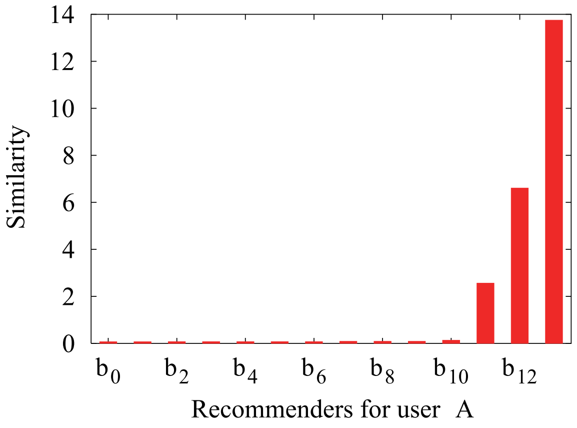

Figure 3: The distribution of the recommender group's similarity to the user A.

URL's page score would be lower than 1.0. In other cases, recommendation is classified as an unsuccessful one.

The criterion seems to be appropriate. The successful recommendations tend to contain web pages of which the category is the same as the target user's browsing history, and which seem to be viewed only by users interested in their category.

There are 18 successful recommended users of 50 users.

User A received the recommendations of some blog entries about a case taken up by mass media at that time. Actually, the user had been interested in the case because his browsing history contains the search engine result pages and the news articles of the case. The recommender group to him consists of 14 users and the distribution of their similarities is shown in figure 3.

The group contains both the users with high similarity, b_{11} to b_{13} in figure 4, and users with low one, b_0 to b_{10}, although the former is less than the latter. Therefore,the recommended web pages satisfy the point mentioned in section 2. They are viewed by similar users and not by dissimilar ones. For user A, as described above, our recommendation system provides successful recommendations which are strongly related to his interests.

In the successful cases, the distribution of the similarities of the recommender group tends to be similar to the case of user A. In the unsuccessful cases, there are no users with high similarity or are the users with only high or low similarity.

7 CONCLUSION

We examined the feasibility of the recommendation only by URLs using real data captured by sniffer.

As a result, we succeeded to give recommendations for a user only using browsing URL histories. We conclude the recommendation by only URLs is possible enough.

In addition, we recognized that the highest similarity score and the distribution of the recommender group's similarity affect the recommendation success.

8 FUTURE WORKS

We have following several future works:

- computing similarity in advance to increase in speed of generating recommendations,
- correcting the bias in the recommender group to let recommendation succeed,
- Introduction of new parameters for calculation,
- excluding irrelevant URLs
- and evaluation of recommendation.

ACKNOWLEDGEMENTS

We would like to thank Information Technology Center of the University of Electro-Communications for providing us with user data.

REFERENCES

Jia Li and Osmar R. Zaïane (2004). Combining Usage, Content, and Structure Data to Improve Web Site Recommendation. In *EC-Web*, pages 305–315.

Jung, J. J., Yoon, J.-S., and Jo, G. (2001). Collaborative information filtering by using categorized bookmarks on the web. In *INAP*, pages 343–357.

K.Maruyama,K.Takasuka,Y.Yagihara,M.Satoshi, Y.Shirai,M.Terada (2006). Real-Time Discovery of Currently and Heavily Viewed Web Pages. In *proc. of WEBIST 2006*, pages 352–359.

Linden, G., Smith, B., and York, J. (2003). Amazon.com recommendations: Item-to-item collaborative filtering. *IEEE Internet Computing*, 07(1):76–80.

R.Greiner, T.Zhu, G.Haubl, K.Jewell (2005). A Trustable Recommender System for Web Content. *Beyond Personalization 2005*, pages 83–88.

Rucker, J. and Polanco, M. J. (1997). Siteseer: Personalized navigation for the web. *Commun. ACM*, 40(3):73–75.