Study on Personalized Recommendation Based on Collaborative Filtering

Taowei Wang¹, Aimin Yang¹, Yibo Ren²
¹Computer Science and Information Technology College
Zhejiang Wanli University
Ningbo, 315100
P. R. China
²Department of Computer
Zhejiang Business Technology Institute
Ningbo, 315012
P. R. China

Abstract: Collaborative filtering is the most successful technology for building personalized recommendation system and is extensively used in many fields. In the paper, a system architecture of personalized recommendation using collaborative filtering based on web log is proposed and data preparation process is detailedly described. The paper also gives an improved k-means algorithm for clustering user transactions. Experimental results show that our proposed algorithm could increase recommendation precision.

Key-Words: Collaborative filtering, Personalized recommendation, Data preparation, Cluster algorithm

1 Introduction

The World Wide Web (WWW) provides a vast source of information of almost all types and this information is often distributed among many web servers and hosts. If these chunks of information could be extracted from the WWW and integrated into a structured form, they would form an unprecedented source of information. The resulting growth in on-line information combined with the almost unstructured web data necessitates the development of powerful yet computationally efficient web data mining tools. Web data mining can be defined as the discovery and analysis of useful information from the WWW data. The Web mainly involves three types of data: data on the WWW, the web log data regarding the users who browsed the web pages and the web structure data. Thus, the WWW data mining should focus on three issues: Web Structure mining, Web Content mining and Web Usage mining[1-4]. Web usage mining includes the data from server access logs, user registration or profiles, user sessions or transactions etc. A survey of some of the emerging tools and techniques for web usage mining have been presented[3]. Current research issues in web data mining in the context of the web warehousing project called WHOWEDA (Warehouse of Web data) were discussed[5].

Web personalization recommendation is an important task from the user point of view as well as application point of view. Personalization recommendation helps the organizations to develop customer-centric Web sites. For example, web sites that display products and take orders are becoming common for many types of businesses. Organizations can thus present custom Web pages created in real time, on-the-fly, for a variety of users such as suppliers, retailers and employees. The web log data obtained from various sources such as proxy server, web server, etc. helps for web personalization recommendation according to interest and tastes of users community. Personalized recommendation content enables organizations to form lasting and loyal relationships with customers by providing individualized information, offering and services. For example, if an end user customer visits the site, he should see pricing and information that is appropriate to him, while a reseller will see a totally different price and shipping instructions. This kind of personalized approach can be effectively achieved by using web mining tools.

Personalization recommendation is any action that makes the web experience of a user personalized to the user's taste. The experience can be something as casual as browsing the web or as significant as trading stocks or purchasing a car. Existing approaches used by many web-based companies, as well as approaches based on collaborative filtering[6,7,8] rely heavily on getting human input, e.g. user profile, for determining the personalization actions.

Recently, a considerable amount of work has been carried out on Web usage mining. Mobahser et al.[9]
presented automatic personalization of a web site based on web usage mining\cite{10}. Techniques have been developed to predict HTTP requests using path profiles of users. Extractions of usage patterns from web logs using data mining techniques have been presented\cite{11,12,13}.

The rest of this paper is organized as follows: The system architecture of personalization recommendation will be presented in section 2. The process of data preparation is detailedly described in section 3. In section 4, we give an improved k-means algorithm to cluster users’ transactions. Experimental results and the discussion of the results are presented in section 5. Finally, conclusion and future work are given in section 6.

2 The system architecture of personalization recommendation using collaborative filtering based on web log

The collaborative filtering method usually requires users to explicitly input ratings about pieces of information. These ratings are then used to compute pairwise correlation coefficients among existing users. The correlation coefficient is the measure of the how similar two users are. The system can make prediction or recommendation based on the correlation coefficients.

The system architecture includes offline process and online process. The offline process includes data preparation, user visiting page matrix and clustering method. The online process includes active session, and recommendation as shown in Fig.1.

3 Data preparation

3.1 Data formatting

As the web users visit the website including all of its linked web pages, they leave some footprints behind. Link many other servers, the website saves the footprints as web server logs, which we have reformatted as following:

2008-06-06 00:08:14 218.73.20.109 -202.107.212.91 80 GET /Default.asp – 200 Mozilla/4.0+(compatible;+MSIE+5.0;+Windows+9 8;+DigExt)
2008-06-06 00:08:14 218.73.20.109 -202.107.212.91 80 GET /images/bg.gif - 200 Mozilla/4.0+(compatible;+MSIE+5.0;+Windows+9 8;+DigExt)
2008-06-06 00:08:14 218.73.20.109 -202.107.212.91 80 GET /style.gif - 200 Mozilla/4.0+(compatible;+MSIE+5.0;+Windows+9 8;+DigExt)

3.2 Data cleansing

The web log is saved to keep a record of every request made by the users. Since the log is to be used as input to organize the web pages to facilitate more effective and efficient navigation, we only want to keep the log entries that carry relevant information. Some log entries which are irrelevant to our study are deleted from the log file as follows:

1. We used a computer terminal to check the web pages for experimental purpose. Since this type of checking does not represent a normal user’s behavior, we delete all the entries with an IP address.
corresponding to the first author’s terminal.

(2) Sometimes a user request a page that does not exits. This results in error entries being recorded. Since we are organizing the existing web URLs, We are not interested in these error entries, which we have therefore deleted.

(3) A user’s request to view a particular page often results in server log entries because that page consists of other material such as graphics. We are only interested in, and only keep, what the users explicitly request because it is intended that the system should be user-oriented.

3.3 Transaction identification

3.3.1 User identification

The task of identifying unique users is greatly complicated by the existence of local caches, corporate firewalls, and proxy servers. Therefore, some heuristics are commonly used to help identify unique users. However, since the university site is mostly accessed by students and teachers in the computer laboratories without passing through proxy servers, we simply use the machines’ IP addresses to identify unique users.

3.3.2 User-session identification

For logs that span a long period of time, it is very likely that different users will use the same machine to access the server website. Therefore, we differentiate the entries into different user-sessions through a session timeout. If the time between page requests exceeds a certain limit, it is assumed that there is another user-session, even though the IP address is the same. We use a 60-min timeout because it is the one used by many commercial products. While this measure is fairly arbitrary, we find that 60 min enables us to find a balance between ensuring that the transactions are attributed to the correct user and generating enough web page accesses in one transaction set.

We assume that there is a set of n unique URLs appearing in the pre-processed web log:

\[ U = \{url_1, url_2, \ldots, url_n\} \]

and a set of m user transactions:

\[ T = \{t_1, t_2, \ldots, t_m\} \]

We represent the transactions as a bit vector

\[ t = \langle u'_1, u'_2, \ldots, u'_n \rangle \]

where

\[ u'_i = \begin{cases} 1, & \text{if } url_i \in t \\ 0, & \text{otherwise} \end{cases} \]

After the entries are grouped according to the user-sessions, the data is converted into a format as shown in Table 1, m is number of the URLs, n is the number of user transactions.

### 4 An improved k-means algorithm

After data preparation, we transpose the data so that each URL is represented as a vector of transaction group and user visiting transaction can be represented as a matrix:

<table>
<thead>
<tr>
<th>URL_1</th>
<th>RUL_2</th>
<th>…</th>
<th>URL_m</th>
</tr>
</thead>
<tbody>
<tr>
<td>User_1</td>
<td>1</td>
<td>0</td>
<td>…</td>
</tr>
<tr>
<td>User_2</td>
<td>1</td>
<td>1</td>
<td>…</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
</tr>
<tr>
<td>User_n</td>
<td>0</td>
<td>1</td>
<td>…</td>
</tr>
</tbody>
</table>

\[ M_{UserSession-Url} = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1m} \\ x_{21} & x_{22} & \ldots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{nm} \end{pmatrix} \]

where m is number of the URLs, n is the number of user transactions and \( x_{ij} \) is the amount of time user i visiting the URLj.

K-means algorithm\[^{14}\] is wide used to cluster user transactions and it is simple yet effective approach to reducing the dimensionality of problem. In order to better selecting these k centers and increasing the clustering quantigy. We improve the k-means algorithm. The algorithm description is shown as following:

**Input:** initial K cluster centers, user visiting matrix

**Output:** k cluster centers (CenterSet)

1. \( k=[K/2] \); // initialize the k=[K/2]
2. initilize(M,CenterSet,k);  // choose random initial k cluster centers
3. while \( k \leq K \) do begin
4. CenterSet=k-means(M,k,CenterSet); // get k cluster centers
5. \[ \text{max}=0; \text{newcenter}=null \]
6. for each \( m \in M \) do begin
7. \( d=0; \)
8. for each \( c \in \text{CenterSet} \) do begin
9. \( d=d+\text{distance}(M,m,c); \)
10. end
11. if \( d=\text{max} \) then begin
12. \( \text{max}=d; \)
13. \( \text{newcenter}=m; \)
(14) end
(15) end
(16) CenterSet=CenterSet ∪ {newcenter};
(17) k=k+1;
(18) end
(19) return CenterSet;

In order to improve the precision of algorithm, after every k-means, we choose a new center which is maximal sum of distance to each cluster centers. In the algorithm, the similarity ($S_{ij}$) between two users is calculated as:

$$S_{ij} = \frac{\sum_{k=1}^{m} (x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j)}{\sqrt{\sum_{k=1}^{m} (x_{ik} - \bar{x}_i)^2} \sqrt{\sum_{k=1}^{m} (x_{jk} - \bar{x}_j)^2}}$$

where $\bar{x}_i = \frac{1}{m} \sum_{k=1}^{m} x_{ik}$, $\bar{x}_j = \frac{1}{m} \sum_{k=1}^{m} x_{jk}$

5 Experimental results
We ran experiments using data from the Zhe Jiang Wanli university web server (www.zjwu.net) and select 1MB records from the web log file. After data preparation, these records were partitioned into 5 test samples as following: 200k, 400k, 600k, 800k, 1000k. We use the coverage and precision as evaluation criteria which have widely used in recommender system research. For comparisons of precision and coverage, we gave the threshold a constant value (threshold=0.5). The Fig. 2 shows the comparison of precision and the comparison of coverage is shown in Fig. 3. From the results in Fig. 2 and Fig. 3, our improved algorithm has higher precision and coverage than the k-means algorithm. Since after every k-means algorithm, we selected a new center which is maximal sum of distance to each cluster centers and that can decrease the susceptibility of edge users. When the size of test samples is 800k or 1000k, the results are more obvious.

Fig. 2 the comparison of precision

6 Conclusion and future work
In the paper, we gave a system architecture of personalization recommendation using collaborative filtering based on web log. According to the process of data preparation we gave a detailedly description. We also presented an improved k-means algorithm to cluster users’ transactions and the experimental results showed that our algorithm was effective.

Our future work in this area is to include the content-base filtering technique to our personalized recommendation system. Furthermore, we study deeply on the clustering methods to more effective for mining the web data.

References:


