Research on Personalized Recommendation Based on Web Usage Mining Using Collaborative Filtering Technique

Taowei Wang[1], Yibo Ren[2]
[1]Computer Science and Information Technology College, Zhejiang Wanli University, Ningbo, 315100, P. R. China
[2]Department 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. This paper presents a system architecture of personalized recommendation using collaborative filtering based on web usage mining and describes detailedly data preparation process. To improve recommending quantity, a new personalized recommendaton model is proposed in which takes the good consideration of URL related analysis and combines the K-means algorithm. Experimental results show that our proposed model is effective and can enhance the performance of recommendation.

Key-Words: Collaborative filtering, Personalized recommendation, Web usage mining, Data preparation, Cluster algorithm, Similarity

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 (CF) [6,7,8] rely heavily on getting human input, e.g. user profile, for determining the personalization actions.

Currently, more and more AI technology has been applied to improve the capability of recommendation

system. To gain exact and real-time recommendation, some recommending methods have been constructed based an different theory[9~12], such as collaborative filtering algorithm, bayesian network, association rule mining, clustering, hurting graph, knowledge-based recommendation, etc.

In these recommending methods, collaborative filtering algorithm is a successful method, widely applied in many e-commerce systems, such as recommending movies or news for user. CF algorithm evaluates the current customer's near neighbours according to the rating data. Through neighbours' rating data, the current customer's evaluation for a new product can be forecasted, then, the recommendation for current customer can be obtained. So, the similarity measurement among customers and customers' classifying are the foundation of personalized recommendation system. There are many classifying methods and algorithms have been applied in many applications, but in e-commerce system, the customers' classifying has its unique features.

Recently, a considerable amount of work has been carried out on web usage mining. Mobahser et al[13]. presented automatic personalization of a web site based on web usage mining[14]. 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[15~17].

The rest of this paper is organized as follows: In section 2, we discuss web usage mining technique, basic collaborative filtering model and personalization recommendation system. A system architecture of personalization recommendation using CF technique based on web usage mining is proposed at last. The process of data preparation is detailedly described in section 3. In sction 4, we give a method of clustering user transactions combined similarity of URLs. A recommendation online algorithm is given in section 5. Experimental results and the discussion of the results are presented in section 6. Finally, conclusion and future work are given in section 7.

2 The Architecture of Personalization Recommendation System Using Collaborative Filtering Based on Web Usage Mining

2.1 Web Usage Mining

2.1.1 The Main Task of Web Usage Mining

Web is a structure of retractility. It can be neatly up against different applications. To great extent web's best using lies on the effect of web mining. Web mining usually includes three types of structure mining, usage mining and content mining.

Structure mining is a process of picking up information from linkages of web pages. Usage mining is a process of picking up information from user's how to use web sites. Content mining is a process of picking up information from texts, images and other contents. The relation of three mining type of Web mining is shown as Table 1.

Table 1 The relation of three mining type of web mining

mining type	source	form	object	collection
usage	accessing	click	behavior	logs
content	pages	texts	index	pages
structure	map	hyperlinks	map	hyperlink

Web usage mining is to find user accessing pattern by analyzing user information. The user information is usually memorized in web server log and registered by CGI recording. By web usage mining we can make sure market stratagem, improve efficiency of commercial activity, provide information for efficient buildup of web sites and offer individuation web service to special consumers.

The overall process of log mining is generally divided into two main tasks: data preprocessing and pattern discovery. Mining behavior patterns from web log data needs the data preprocessing tasks that include data cleansing, user identification, session identification and path completion. Mobasher et al.[13] presented a detailed description of data preprocessing methods for mining web browsing patterns. In this phase, web log can be transformed to transaction forms that are fit for data mining in different fields. The pattern discovery tasks involve the discovery of association rules, sequential patterns, usage clusters, page cluster, user classification or any other pattern discovery method. Lee et al.[18] provided a detailed case study of click stream analyzed from an online retail store. To measure the effectiveness of efforts in merchandising, they analyzed the shopping behavior of customers according to the following four shopping steps: product impression, click-through, basket placement, and purchase. It has been recognized that web usage mining gave better recommendation quality in the CF recommendation procedures[19,20].

2.1.2 The Application of Web Usage Mining on E-commerce

e-commerce is a activity that individual or corporation exchange commerce data and develop business by adopting digital electric manner on Internet. In modern business activities e-commerce has absolute predominance, Including no spatio-temporal limit, sending real time information, sharing world resource, falling management cost, selling straightway without shop and direct consumers' feedback.

Owing to fast developing e-commerce compete very intensely, web seller must be prepared for catering for requirement of on-line consumers rapidly. The refractivity of on-line selling has people in a position to supervise selling and know adaptability of price adjusting and product service in time. In addition, all important trends and modes, that can influence products, including freight, selling and stock, can be opened out by mining selling information.

Increasingly rising web visiting information and fast developing data mining technology has Web sites truly provide individuation services for their on-line consumers. Market should have Web sites put in effect to their real consumers and benefits. In a dynamic and strong competitive net environment, e-commerce must better understand frequent visiting clients and the best profitable buyers. Then they can obtain competitive advantage. If you want to know consumers' visiting behavior, you must mine your web site data by web usage mining in order that effort of web site can focus on that profitable buyers and foreground.

Now in the many companies there are vast user information brought from their own web sites. So large-scale e-commerce web sites need mining tools that are suit a great deal of data and hope to gain benefit by data mining. In addition, web is a perfect market environment. Every business is obtained and stored. By web usage mining web site can reach following aims:

(1) Identify web consumers' pivotal characteristic.
(2) Test and decide whcih market is the best influential and active.
(3) Identify consumers who are very interested in new products.
(4) Reduce wares price and improve relationship with consumers.
(5) Improve web site advertisement and selling course.

2.2 Basic Collaborative Filtering Model

Most web personalization recommendation system adopt two types of techniques: a content-based approach and a collaborative filtering (CF) approach. In the content-based approach, it recommends web objects that are similar to what the user has been interested in the past. In the collaborative filtering approach, it finds other users that have shown similar tendency to the given users and recommends what they have liked.

The collaborative filtering recommendation acts according to other users' viewpoint to produce recommendation tabulates to the goal user. Its basic thought is based on a supposition: If user grade to some product quite similarly, then they grade to other product also similar . CF recommendation model use statistics technology to search goal user's recent neighbors, then forecast goal user's grading according to neighbors' grades, thus has the recommendation. The rating data of customers are organized as a $m \times n$ matrix. In the matrix, row m represents customer m, column n represents item n. The element $R_{i,j}$ in ith row and jth column represents the rating data of customer i on item j, like Table 2 shows.

Table 2 Customer rating data matrix

	$Item_1$	$Item_2$	…	$Item_n$
$Customer_1$	$R_{1,1}$	$R_{1,2}$	…	$R_{1,n}$
$Customer_2$	$R_{2,1}$	R2,2	…	$R_{2,n}$
…				
$Customer_m$	$R_{m,1}$	$R_{m,2}$	…	$R_{m,n}$

To find user's neighbours, the similarity among users must be measured. Some measure algorithms often are applied, such as cosine similarity, correlation similarity and adjusted cosine similarity, to measure users' similarity degree. The precise computation of goal user's neighbours is the successful key of CF model. It is necessary to attempt new algorithm foundation to increase the recommendation precision through better classification. So, the similarity measurement among customers and customers' clustering are the foundation of personalized recommendation model.

In order to suit huge and sparse sample space, it is needs to reduce samples while maintaining the quality of classification. In the training process, providing the learning algorithm with some control over the inputs is an effective solution. Sometimes, it is called active learning.

2.3 Personalization Recommendation System

2.3.1 Popular Recommendation System

At present, electronic commerce has become an important strategy to integrate own resources and exterior resources effectively across the organization. The integrated research of artificial intelligence, web technology and commercial model has obtained more and more focus [21]. In the business to customer (B2C) system, the recommendation system becomes a hot research topic.

From acquiring requirements of customer [9], proves the suitable product and services for individual. Personalized recommendation system aims to solving the customer's overloading of information, and improving the performance of e-Commerce system. Currently, personalization recommendation system has been widely used in a number of different applications.The some appliation fields and popular recommendation system is shown in Table 3.

Table 3 Recommendation systems and it's application

Application f	Recommendation system
e-commerce	Amazon.com, Dietorecs, Ebay ,entrée,FAIRWIS, Ghani,Levis ,LIBRA, MIAU, RIND, Ski-europe.com
web pages	Commty, search, ass'nt, Fab, Foxtrot, ifWeb, MEMOIR, METIOREW, ProfBuilder, QuIC, Quickstep, R2P,Siteseer, SOAP, Surflen
musics	CDNOW, CoCoA, Ringo
movies	CBCF, Nakif, Moviefinder.com MovieLens, Recommends' Explorer, Reel.com, Virtual rev's
news filter	GroupLens, PHOAKS, P-Tango
e-mail filter	Tapestry
search servic	Expertise Recommender, Referral web
others	Campiello, ELFI, OWL

2.3.2 Personalization Recommendation System Based on CF

CF-based recommender system products to a target customer according to the following steps:

(1) A customer provides the system with preferenceratings of products that may be used to build a customer profile or his or her likes and dislikes.

(2) Then, these system apply statistical techniques or machine learning techniques to find a set of customers, known as neighbors, which in the past have exhibited similar behavior (they either rated similiarly or purchased similar set of products), Usually, a neighborhood is formed by the degree of similarity between the customers.

(3) Once a neighborhood of similar customer is formed, these systems predict whether the target customer will like a particular product by calculating a weighted composite of the neighbor's rating of that product(prediction problem), or generate a set of products that the target customers is most likely to purchase by analyzing the products the neighbors purchased (top-N recommendation problem).These system, also known as the nearest neighbor CF-based recommender system have been widely used in practice. However, its widespread use has exposed some limitations, such as sparsity, scalability, and black box.

2.4 A System Architecture of Personalization Recommendation Using Collaborative Filtering Based on Web Usage Mining

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 uses. 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 Fig.1 shows an system architecture of personalization recommendation. The architecture mainly includes two parts: offline process and online process. The offline process includes data preparation and web usage mining and the online process is made up of recommendation engine. The offline process is called model acquisition phase and online process is called model application phase.

3 Data preparation

We uses the access logs form the Wanli University web server (www.zjwu.net). The web log indicating the date, time, and address of the requested web pages, as well as the IP address of user's machine.

There are serveral pre-processing tasks to be done befor data mining algorithms can be performed on the web server logs. These include data cleansing, user identification, session identification, path completion and formatting. These pre-processing tasks are the same for any web usage mining problem

and disscussed by Coolye et al. [16]. The original server logs are formatted, cleansed, and the grouped into meaningful transactions before the user session clustering algorithm being performed.

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

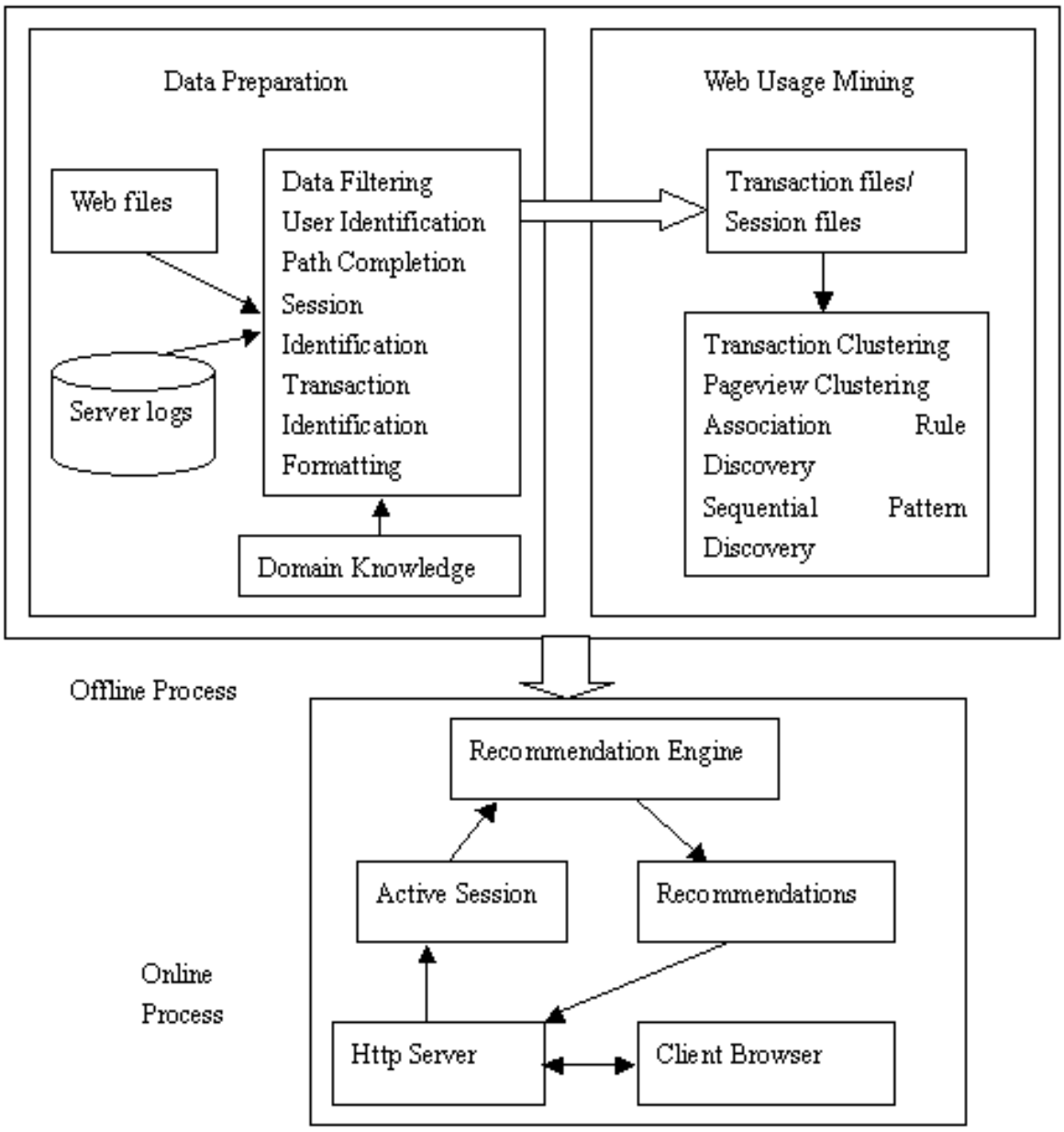


Fig. 1 A system architecture of personalized recommendation using CF based on web usage mining

3.1 Data Formatting

As the web users browse 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 shown Table 4.

3.2 Data Cleansing

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.

Table 4 A web server access log

date	time	c-ip	cs-method	cs-uri-stem	cs-status	cs(user-agent)
2008-06-06	00:08:14	220.184.17.121	80 POST	/sosu/sosu.asp	200	Mozilla/4.0+(compatible; +MSIE+5.0; +Windows+98;+DigExt)
2008-06-06	00:08:14	220.184.17.121	80 GET	/sosu/img/bg-g.gif	200	Mozilla/4.0+(compatible; +MSIE+5.0; +Windows+98;+DigExt)
2008-06-06	00:08:14	220.184.17.121	80 GET	/sosu/img/zik1.gif	200	Mozilla/4.0+(compatible; +MSIE+5.0; +Windows+98;+DigExt)
…	…	…	…	…	…	…
2008-06-06	00:12:10	203.175.243.70	80 GET	/images/bg.gif	304	Mozilla/4.0+(compatible; +MSIE+5.0; +Windows+98;+DigExt)
2008-06-06	00:12:10	203.175.243.70	80 GET	/pop/040520/logo.gif	304	Mozilla/4.0+(compatible; +MSIE+5.0; +Windows+98;+DigExt)

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.

Table 5 Format of data grouped by transactions

	URL_1	RUL_2	…	URL_m
$User_1$	1	0	…	0
$User_2$	1	1	…	1
…	…	…	…	…
$User_n$	0	1	…	0

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

$$\vec{t} = <u_1^t, u_2^t, \ldots, u_n^t>$$

where

$$u_i^t = \begin{cases} 1, if\ url_i \in t \\ 0, otherwise \end{cases}$$

After the entries are grouped according to the user-sessions, the data is converted into a format as shown in Table 5, m is number of the URLs, n is the number of user transactions.

4 Clustering User Transactions

4.1 Clustering Transactions into User Groups

Even after the above pre-processing procedure, the web log data is still not ready for effective application in personalization recommendation system based on collaborative filtering, because the number of transactions is very large. The number of inputs of the personalization recommendation system will need to be equivalent to the number of transactions, and because this number is so large, it will not be feasible with this data. The size of the web log data set not only consumes large amounts of processing time, but also limits the applicability of the system to data in the real world (for example, web logs collected by a search engine in a month).

To solve this problem, Ref.[13] combines hypergraph partitioning and association rules, to cluster the web documents. However, we can not use the same technique because we are trying to achiver more than just clustering of documents: we want to map the web documents into a two-dimensional space, where the locations will indicate the similarity between documents, as indicated by the navigation patterns. The pages are deemed similar if they are accessed by similar kinds of people. That is, rather than deeming two pages to be similar because the web-master has identified their relationship based on content or keywords, we are employing a data-driven approach where the similarity of two documents or web pages depends on them being accessed by the same users. If the same collection of web pages is repeatedly accessed by a group of users, then these pages are defined as being similar (at the very least, they are similar in term of the interest level of these users at that time).

Due to our focus on user navigation patterns, we have devised another simple yet effective approach to reducing the dimensionality of the problem. By using the K-means clustering algorithm [22]. The K-means algorithm is outlined as follows:

The K-means algorithm used in this study comprises the following four steps:

1. Choose K initial cluster centres (representing the K transaction groups) randomly from the centre of hypercube.

2. Assing all data points (representing the transactions) to their closest cluster (measuring from the cluster centre). This is done by presenting a data point x and calculate the similarity (distance) d of this input to weight w of each cluster centre j. The closest cluster centre to a data point x is the cluster centre with minimum distance to the data point x.

$$d_j = \left\| x - w_j \right\| = \sqrt{\sum_{i=1}^{n} (x_i - w_{ij})}$$

3. Recalculate the centre of each cluster as the centroid of all the data points in each cluster. The centroid c is calculate as follows:

$$\vec{C} = < w_1^c, w_2^c, \ldots, w_n^c >$$

where

$$w_i^c = \frac{\sum_{j \in c} u_i^j}{N^c}$$

4. If the new centres are different from the previous ones, repeat Step2, 3 and 4. Otherwise terminate the algorithm.

4.2 Clustering User Transactions Combined Similarity of URLs

Generally, for standardization management, web site puts the files into different hierarchy directories. So that, we can define the similarity of URLs according to the page hierarchy structure.

Each URL is represented as a vector of transaction group and users' browsing transaction can be represented as a matrix.

Defination 1. The users' browsing matrix is defined as

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

where m is number of the URLs, n is the number of user transactions and x_{ij} is the amount of time user i visting the URL_j.

Defination 2. The similarity of two URLs U_i and U_j is defined as

$$S(U_i; U_j) = \frac{|URL_i \cap URL_j| - 1}{\max(|URL_i|, |URL_j|)} \quad (1)$$

where URL_i is represented the path from root of website to current node (for example, www.zjwu.net ~ graduate/student/doctor99.htm). $|URL_i|$ is represented the path length.

From the above defination, it can be seen that the similarity of two URLs U_i and U_j is a number between 0 and 1. When the similarity of two U_i and U_j is 0, then U_i and U_j are completely dissimilar. On the other hand, is the similarity is 1, then the two URLs are completely similar. Thus, for any two URLs U_i and U_j,

$S(Ui; Uj) \in [0, 1]$

S (Ui; Ui) = 1
S (Ui; Uj) = S (Uj; Ui).

So that, the matrix M formed by similarity of URLs can be represented as follows:

		URL_1	URL_2	URL_3	…	URL_m
	URL_1	1	S_{12}	S_{13}	…	S_{1m}
	URL_2	S_{21}	1	S_{23}	…	S_{2m}
M =	URL_3	S_{31}	S_{32}	1	…	S_{3m}
	…	…	…	…	…	…
	URL_m	S_{m1}	S_{m2}	S_{m3}	…	1

where $S_{ij} = S(U_i; U_j)$.

For example, we give one group URLs {www.zjwu.net/graduate/student/doctor98} and {www.zjwu.net/graduate/student/doctor99}. From the defination 2, the similarity of the group URLs is: $\frac{3-1}{\max(4,4)} = 0.5$.

4.3 An User Session Clustering Algorithm

K-means algorithm is wide used to cluster user transactions and it is simple yet effective approach to reducing the dimensionality of problem.Based on this algorithm,foremore, we takes good considerations of URL related anaylsis in our recommendation system and then an user session clustering algorithm is proposed as follows:

(1) function k-means ()
(2) initialize k prototypes (w_1, …, w_k) such that $w_j = us_i$, $j \in \{1, …, k\}$, $i \in \{1,…, n\}$
(3) each cluster C_j is associated with prototype w_j repeat
(4) for each user session us_i, where $i \in \{1,…,n\}$
(5) assign us_i to the cluster C_j with most similarity prototype w_j
(6) end for
(7) for each cluster C_j, where $j \in \{1, …, k\}$
(8) update the prototype w_j to be the centroid of all sessions currently in C_j

$$W_j = \sum_{us_i \in c_j} \frac{us_i}{|C_j|}$$

(9) end for
(10) compute the equation

$$E = \sum_{i=1}^{k} \sum_{us \in C_j} (S'_{ij})^2 > \partial$$

(11) until E is true

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. Taking the considerations of URL related analysis, the similarity (S'_{ij}) between two users is calculated as following equation:

$$S'_{ij} = \frac{\sum_{k=1}^{m} \sum_{l=1}^{m} (x_{ik} - \overline{x_i})(x_{jk} - \overline{x_j}) \cdot S_{kl}}{\sum_{k=1}^{m} (x_{ik} - \overline{x_i})^2 \cdot \sum_{k=1}^{m} (x_{jk} - \overline{x_j})^2} \quad (1)$$

where

$\overline{x_i} = \frac{1}{m} \sum_{k=1}^{m} x_{ik}$, $\overline{x_j} = \frac{1}{m} \sum_{k=1}^{m} x_{jk}$ and the S_{kl} is calculated with equation 1 (see Defination 1).

5 Recommendations Online

The main work online process is recommendations for active user sessions. It mainly includes: active user session conversion, similarity calculation and real-time recommendations. In the actual system, recommended threshold (α) is an important parameter. α represents minimal recommendation factor. When the recommendation fact of a web page is lager than α , we will put this page into recommended set. So that, α value is large, the recommended pages are few. On the contrary, the α value is small, the recommended pages are more. The recommendations online algorithm is given as follows:

Input: session cluster set (C), active user session (s), α value
Output: recommended set (RecSet)

(1) RecSet= Φ ;
(2) compare each cluster $c_j \in C$ with active session s with the equation 2.
(3) get the most similarity of the cluster centroid
(4) for each itemset I of the cluster centroid
(5) if average(support(I) $\geq \alpha$) then
(6) if I not appear in s then
(7) RecSet $\leftarrow$ {I}
(8) end if
(9) end if
(10) endfor

6 Experiments and Its Analysis

We ran experiments using data from the Zhe Jiang Wanli university web server (http://www.zjwu.net) and date is from the 6th of June 2008 to 16th of June 2008. After data preparation, the experimental data set contains 327 users, 385 URLs.

We use the COVERAGE and PRECISON as

evaluation criteria which have widely used in recommender system research[23,24]. In the experiment, we compare the traditional CF method with our proposed method.

COVERAGE is percent of pages that user likes which correctly recommends.

$$\text{COVERAGE}=\frac{|US \cap \text{Re}cSet|}{US}$$

PRECISON is percent of recommended pages liked by user.

$$\text{PRECISON}=\frac{|US \cap \text{Re}cSet|}{\text{Re}cSet}$$

where US represents web page set liked by user, RecSet represents recommended page set.

Firstly, we fixed the α values from 0.1 to 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9 respectively and the size of test sample was 600k. The Fig. 2 shows the comparison of PRECISON and the Fig. 3 shows the comparison of COVERAGE while α values are different.The X-coordinate expresses the different threshold (α) values, the Y-coordinate in Fig. 2 expresses PRECISION and the Y-coordinate in Fig. 3 expresses COVERAGE.

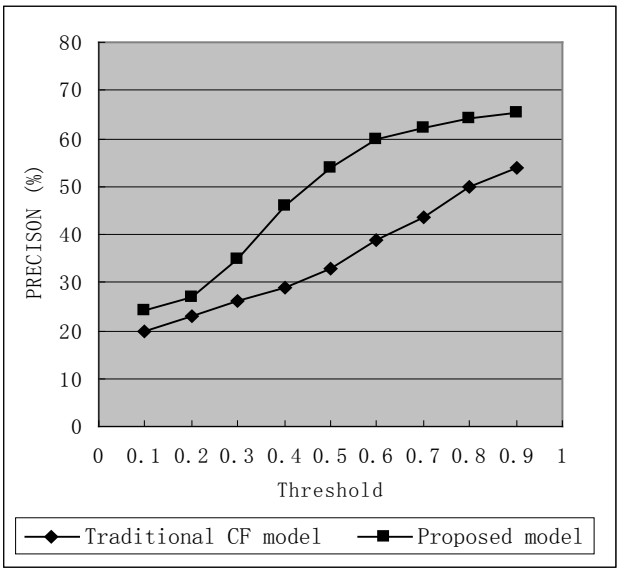
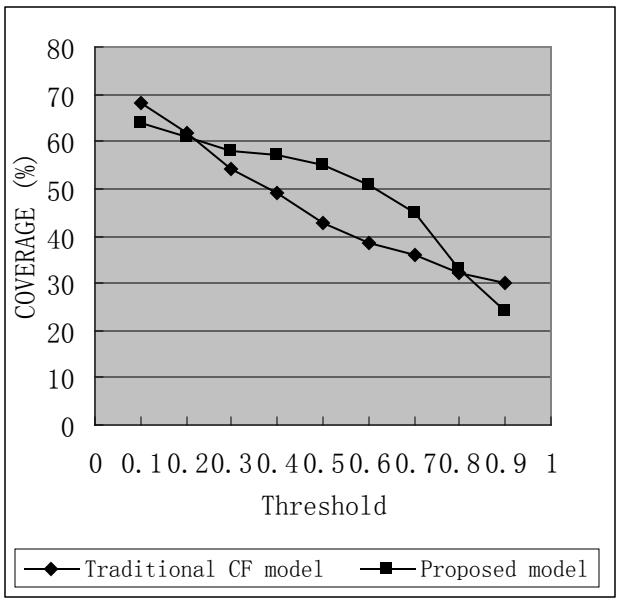


Fig. 2 The comparison of PRECISON

Then, to comparisions of PRECISON and COVERAGE, we gave the threshold (α) a constant value(α =0.5) and selected 1MB records from the web log file. These records were partitioned into 5 test samples as following: 200k, 400k, 600k, 800k,1000k. The Fig. 4 shows the comparison of PRECISON and the comparison of COVERAGE is shown in Fig. 5.

From the results in Fig. 2, our proposed model has higher PRECISON than the traditional CF model. However, the PRECISON of the proposed model is a little higher than the traditional CF model with α =0.1 or 0.2. When the α value is between 0.4 to 0.7, the advantage of the proposed model is obvious.



From the Fig.3, we can see that when the α value

Fig. 3 The comparison of COVERAGE

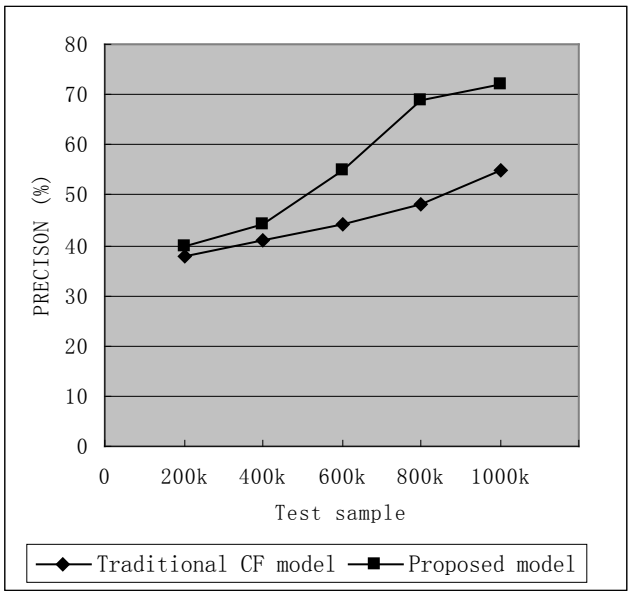


Fig. 4 The comparison of PRECISON

is between 0.3 to 0.7, the COVERAGE of the proposed model is higher than the traditional CF model. However, the COVERAGE is lower than the traditional CF model with α =0.1, α =0.2, α =0.9.

Therefore, from the above analysis, the PRECISON and COVERAGE of the proposed medel with α =0.4, α =0.5, α = 0.6 has much better advantage than with α setted other values.

We can see from Fig. 3 and Fig. 4 that when the

size of test sample is smaller (for example: 200k, 400k), the PRECISON and COVERAGE of the proposed model is a little higher than the traditional CF model. When the test sample is 600k or above , the PRECISON and COVERAGE of our proposed model is more higher than the traditional CF model.

Since we take the good consideration of URL related analysis in our proposed model, the experimental results show the proposed model is effective. Thus, it can enhance the performance of recommendation.

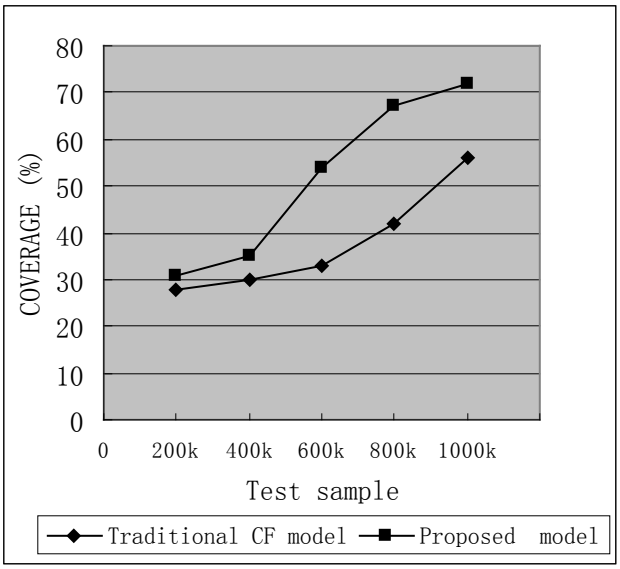


Fig. 5 The comparison of COVERAGE

7 Conclusion and Future Work

In the paper, we firstly discussed the web usage mining method, collaborative clustering technique and personalization recommendation system. Then, we presented a system architecture of personalization recommendation using collaborative filtering technique based on web usage mining. According to the process of data prepariation and cluster users' transactions, we also gave a detailedly description. Furthermore, combining with the similarity of URLs and K-means algorithm, we proposed a new presonalization recmmendation model. Finally, we gave the PRECISON and COVERAGE comparison results, experimental results showed that our proposed model had higher PRECISON and COVERAGE than the traditional CF model.

Our future work in this area is to include the content-base filtering technique to our personalized recommendation system and more experiment will be done to examine reliabilty of the proposed model and speed. And the class uncertainty and sampling affection those affect quality of recommendation are needed to disscuss. Furthermore, we study deeply on the clustering methods to more effective for mining the web data.

References:

[1] D. Backman and J. Rubbin, Web log analysis: Finding a recipe for success, 1997, *http://techweb.comp.com/nc/811/811cn2.html*

[2] R. Cooley, Web usage mining: Discovery and application of interesting patterns from Web Data, *SIGKDD Exploration*, 2000, pp.12-23.

[3] R. Cooley, B. Mobasher and J. Srivastava, Web mining: Information and pattern discovery on the world wide web, *Proc. 9th IEEE Int. Conf. Tools AI (ICTAI'97)*, 1997.

[4] J. Pitkow, In Search of Reliable Usage Data on the WWW, *Proc. 6th Int. World Wide Web Conf., Santa Clara, CA*, 1997.

[5] S. K. Madria, S. S. Bhowmick, W.-K. Ng and E. P. Lim, Research issues in web data mining, *Proc. 1st Int. Conf. Data Warehousing Knowledge Discovery, Florence, Italy*, 1999.

[6] J. Herlocker, J. Konstan, A. Borcher and J. Riedl, An algorithmic framework for performing collaborative filtering, *Proc. 1999 Conf. Res. Dev. Inf. Retrieval*, 1999.

[7] J. Konstan, B. Miller, D. Maltz, J. Herlocker, L. Gordon and J. Riedl, GroupLens: Applying collaborative filtering to usenet news, *IEEE Bull. Tech. Committee Data Eng*. Vol.21, No.1, 1998.

[8] U. Shardanand and P. Maes, Social information filtering: Algorithms for automating "Word of Mouth", *Proc. ACM CHI Conf.*, 1995.

[9] Ralph Bergmann, Padraig Cunningham, Acquiring customer's requirement in electronic commerce, *Artificial Intelligence Review*, Vol.18, 2002, pp.163-169.

[10] Gediminas Adomavicius, Toward the next generation of recommender systems: a survey of the state-of–the-art and possible extensions, *IEEE Transactions on Knowledge and Data Engineering*, 17(16): 734-749, 2005.

[11] Gao Linqi, Li Longzhu, Production recommender method based on customer's behavior, *Computer Engineering and Application*, 3, 188-190, 2005.

[12] Jae Kyeong Kim, A personalized recommendation procedure for Internet shopping support, *Electronic Commerce Research and Application*, 1(4): 301-313, 2002.

[13] B. Mobasher, R. Colley and J. Srivastava, Automatic personalization based on web usage mining, *Communications of ACM*, Vol.43, No.8, 2000, pp.142-151.

[14] S. Schechter, M. Krishnan and M. D. Smith, Using path profiles to predict HTTP requests, *Proc. 7th Int. World Wide Web Conf., Brisbane, Australia*, 1998.

[15] A. Buchner and M. D. Mulvenna, Discovering internet marketing intelligence through online analytical web usage mining, *SIGMOD Rec.*Vol. 27, No.4, 1999.

[16] R. Cooley, B. Mobasher and J. Srivastava, Data preparation for mining world wide web browsing patterns, *Knowledge and Information Systems*, Vol. 1, No.1, 1999.

[17] M. Spiliopoulou and L. C. Faulstich, WUM: A web utilization miner, *Proc. EDBT Workshop WebDB98, Valencia, Spain, Lecture Notes in Computer Science 1590* (Springer Verlag, 1999).

[18] Lee, J., M. Podlaseck, E. Schonbery, R. Hoch, Visualization and analysis of clickstream data of online stores for understanding web merchandising, *Data Mining and Knowledge Discovery*, Vol.5, No.1, 2001, pp.59-84.

[19] Ahn, D.H., J. K. K. Kim, Y. H. Cho, A comparative evaluation of hybrid product recommendation procedures for web retailers, *9th Asia-Pacific Decision Science Institute Conference* 2004

[20] Cho, Y. H., J. K. Kim, Application of web usage mining and product taxonomy to collaborative recommendations in e-commerce, *Expert System with Applications*, Vol.26, 2004, pp.233-246

[21] Robin Burke, Hybrid recommender systems: survey and experiments, *User Modeling and User-Adapted Interaction*, Vol. 12, No.4, 2002, pp.331-370,.

[22] J. A. Hartigan, Clustering Algorithm, *Wiley, New York*, 1975.

[23] Cho, Y.H., J.K. Kim, Application of Web usage mining and product taxonomy to collaborative recommendations in e-commerce,*Expert Systems with Applications*, Vol.26, 2004, pp.233-246.

[24] Sarwar, B., G. Karypis, J. Konstan, J. Riedl, Analysis of recommendation algorithms for e-commerce,*In Proc. ACM E-Commerce*, 2002, pp.158-167.