Fuzzy-Clustering Web based on Mining

Liu Ying
Department of Computer and Science, Wuhan Yangtze Business University, China

Abstract—Since web is increasingly important for individuals, analyzing user behaviors on webpages is becoming essential for system designers and website providers. But current solutions cannot meet high demands on high accuracy, low time consuming and sense of privacy. Therefore, this paper designed a fuzzy clustering algorithm based web interaction system, presented multiple linear regressions to analyses user’s browsing behavior. Finally, we designed a couple of experiment to demonstrate the efficacy of this paper design system. The result shows the interaction system worked with high accuracy, low time consuming and good performance of scalability.

Index Terms—Web User; Cluster Analysis; Interest

I. INTRODUCTION

Every aspect of the Web is rapidly developing. Early Web is mainly applied to information sharing, today’s Web applications has been extended to the field of e-commerce, online games, remote login, as well as other requirement on design and functionality [1, 2]. Such as improve the organizational structure of website user-friendly to find needed information quickly and accurately; recommend users the pages may interests them; provide personalized service; discover the potential visitor groups to make accurate market positioning for different groups of visitors [3]. A traditional data mining technology used in the field of Web-Web mining thus emerged [4]. Since web information generally unstructured and lack of constraints, web information mining is with considerable difficulty [5]. Web logs have the perfect structure [6], it contains a wealth of information that can reveal the user’s browsing behavior [7], provide a good prerequisite for Web mining. We may conclude web log analysis is an important means of web mining [8].

Web user’s browsing behavior is a direct reflection of the interest of the user [9]. The Web user browsing behavior is affected by many behavior factors, so the relationship between user interest and multiple behavioral factors can be considered regression problems [10].

As the rapid development of web technology, data generated by users are increased sharply, how to capture effective information from numerous web data storage is one of the most challenges now, and of course, it will significant change people’s way of life. Practically, it would also save people time and enable them to focus on their own interesting, and save humane resource and budget as well. How to mining useful information from various web data is also important to big enterprise, makes them capture their interesting information more effectively, such as consumer prefer information, competitor’s information and so on. It will increase enterprise’s performance on market as well. Therefore, it’s important to research on web usage data mining system. But the mining performance and accuracy are among the most important part in web usage mining system. Researchers have been work a lot on how to improve these two factors.

Aibara and his colleges researched on a fuzzy neutral network to capture network data but this model needs long time and compute resource to training the network [14]. Inoue presented a clustering method by using signal neutral network, but his method cannot show a high performance in handling large scale data. The result showed acceptable accuracy in small amount of data [15] [16]. Okada researched clustering method and neutral network on finance risk management system and the data was also generated by users [17] [18]. But it highly depends on data quality and other conditions, which cannot be suited for web situation, because internet situation is more various and web user’s behavior is no obvious pattern to follow. Therefore, the predict accuracy and computing performance are the main challenges in current researches. Most of them need restricted conditions or high performance computing resources to handle the weakness of current models.

In this paper, we consider the characteristics of the above web user browsing behavior, introduce a multiple linear regression model to describe the relationship between the behavior of user interest and page views, to quantify the interest of the users of the web, directly partitions interest similarity matrix with a threshold $\lambda$, relocates the clustering results per the connection strength between a cluster and an element. Finally verify the accuracy and performance of the algorithm by an experiment. The rest of this paper is organized as follows. In section 2, Web users’ interest-based fuzzy clustering method and WFCM model are proposed. Section 3 introduces approaches and the detailed procedures of clustering methods. A study on the web users’ interest-based fuzzy clustering method is given in section 4. Finally, the concluding remarks and further works are presented in Section 5.

II. WEB USERS’ INTEREST-BASED FUZZY CLUSTERING METHOD

A. Web-related Concepts

Let $U = (x_1, x_2, ... x_n)$ For clustering Web a collection of objects that are directly related to the concept of fuzzy clustering and Web are as follows:
**Fuzzy matrix**: For any \( i = 1, 2, \ldots, n; \ h = 1, 2, \ldots, m \), which has \( r_{ik} \in [0,1] \) matrix then \( R = (r_{ik})_{nm} \) fuzzy Web.

**Web fuzzy similar matrix**: \( R \) Order \( n \) Web fuzzy matrix, \( I \) is a unit matrix, and the reflexivity: is satisfied (1) \( 1 \leq R; \) (2) Symmetry: wherein \( R^T \) is a transposed matrix of \( R \). \( R \) is denoted as n-order Web fuzzy similarity matrix.

**Fuzzy equivalent matrices**: \( R \) Fuzzy similarity matrix of order \( n \), if \( R \) is transitive, that is, \( R \cdot R \leq R \), where \( \circ \) Web fuzzy matrix composition operation, called to order the Web fuzzy equivalent matrix.

**Transitive closures**: Let \( R \) fuzzy similar matrix of order \( n \) Web, there is a smallest natural number \( k(k \leq n) \), making the transfer the closure \( t(R) = R^k \) for all natural number \( l \) greater than \( k \) constant \( R' = R^k \), \( t(R) = R^k \) for \( n \)-order Web matrix mold equivalence.

Typical web architecture is showed as below; it is consisted of five components: cache, load balancer, application server, database, and memory cache.

![Web Pages components](image)

**B. Clustering Process Model**

Web of clustering objects given web object collection that the object properties of the source data, Web fuzzy clustering direct role for web fuzzy similar matrix or web fuzzy equivalent matrix, so we must first abstract source data obtained represent web object properties of web data matrix, then suitable web data matrix into fuzzy clustering method of operation of the web, fuzzy similar matrix or web fuzzy equivalent matrix basis last in the web on the use of certain web fuzzy clustering methods, clustering results [11]. Web fuzzy clustering process model WFCM (Web Fuzzy Clustering Model), shown in Figure 2.

The Clustering pattern technology has been coped with the increasingly grim security situation; researchers have proposed many security programs to ensure the network. Currently, Mainstream method includes firewall, intrusion detection system, and anti-virus gateway; Spam filtering, malware detection and so on. Firewall simply detects the single data packet which is currently passing by using the method of setting access control policy. They view source IP/destination IP address, Source port/destination port and a simple protocol type combined with access control rules to achieve selected pass of packet. This technology is simple, fast processing speed, and transparency for the upper application, but there are many problems: such as the inability to determine the contents of the packet, which means powerless to commence at the application layer attacks and intrusions. However, due to the application layer protocol up to the most complex, greater space, therefore, using application systems and programs vulnerability to cyber-attacks has gradually evolved into a trend, the harm is also growing. As a result, the researchers begin to apply fuzzy clustering technology to a variety of security systems.

![Cluster model](image)

**C. Challenges for Current Solutions**

Fuzzy clustering technology scan and analysis the content of package’s payload, and then determine the treatment strategy by the results of analyzing. It is different with traditional methods which only scan packet header. Deep packet inspection technology to discover hidden specific content during detected network packet and given known “features” clustering pattern technology is necessary to complete such operations. Therefore as the supporting of packet inspection technology, Clustering pattern has been common concern hot issues in academia and industry, and have a pivotal position in the field of network security. Further research and improving of clustering pattern technology have important practical significance to improve the performance of the application system and the security of the network. However, there are still challenges for researchers in the following fields:

As the increasing amount of elements in web pages, how to track this large quantity orderly would be a big challenge for system designers. What more, as web is becoming continuously open, but users are showing strong sense of protecting private information; this will be also a hot point for researchers. Collecting data for analyzing is important but we cannot get user’s private information, especially, some credit card numbers, their birthday, personal contact and so on.

It’s a clear trend that more and more folks transfer their traditional works to web, which means the data of
web is becoming larger and larger, how to analyze these large amount of data effectively is very important. As we know, there are a lot of valuable information in this data and also some useless data as well. How to analyzing and determining them effectively is still a challenge for researchers.

People are familiar with web technologies now, that means they would act more quickly during do business on internet. How to track the quick behavior on web is increasingly attracted by researchers.

III. CLUSTERING WEB USERAGE MINING SYSTEM

Web log contains extremely rich user access information, but the data form of Web log is a far cry from the similarity matrix data formats [13]. Thus the transformation from the original user access data information into structured user similarity data is needed. The process includes: data cleaning, user identification, session identification, filtering information - lack of data, users interest identifying and structure similarity matrix.

Overview of web pages requesting and rendering structure is showed as below figure. Usage model is consisted of data, user request, device detection, web sharing, devices validation and screen components. Web pages module is consisted of JavaScript, cache, HTML and CSS components. We can track user behavior by JavaScript log and database saved in web pages.

A. Data Structure Design

Web data matrix: Let \( U = \{X_1, X_2, \ldots, X_n\} \) be a collection of Web objects, each Web object and its properties expressed by the \( m \) indicators:

\[ x_i = \{x_{i1}, x_{i2}, \ldots, x_{im}\} \quad (i = 1, 2, \ldots, n) \]

So get the original Web data matrix:

\[
\begin{bmatrix}
  x_{11} & x_{12} & \cdots & x_{1m} \\
  x_{21} & x_{22} & \cdots & x_{2m} \\
  \vdots & \vdots & \ddots & \vdots \\
  x_{n1} & x_{n2} & \cdots & x_{nm}
\end{bmatrix}
\]

(1)

![Figure 3. Detail structure of web pages requesting and rendering](image)

Figure 3. Detail structure of web pages requesting and rendering

![Figure 4. WFCM model](image)

Figure 4. WFCM model

In practical problems, different data generally have different dimensions. The amount can be compared in order to use different dimensions represented by the proper transform, usually the data. Web fuzzy clustering operation object Web fuzzy matrix, thus requiring Web data matrix data must be in the interval [0,1], that is, according to the requirements of the Web fuzzy matrix, data standardization, Web data the matrix data compression to the interval [0,1].

Normally do transform (Equation 1) and very poor conversion (Equation 2) two standard deviation.

\[
x_i' = \frac{x_i - \overline{x}}{s_k} \quad (i = 1, 2, \ldots, n; k = 1, 2, \ldots, m)
\]

\[
\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i, s_k = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2
\]

\[
x_i' = \frac{x_i - \min_{i \in \text{cluster}} x_i}{\max_{i \in \text{cluster}} x_i} \quad k = 1, 2, \ldots, m
\]  

(2)

Let \( U = \{X_1, X_2, \ldots, X_n\} \) For clustering Web objects collection, where \( x_i = \{x_{i1}, x_{i2}, \ldots, x_{im}\} \), the main task of the calibration step is to create a Web fuzzy similarity matrix. The key task is to determine Web objects \( x_i \) and \( x_s \) similar degree of similarity coefficient \( r_{is} = R(x_i, x_s) \) main method in accordance with the traditional fuzzy clustering methods:

Direct Hamming distance method

\[
r_{is} = 1 - c \sum_{k=1}^{m} |x_{ik} - x_{is}| \quad (3)
\]

where \( c \) is properly chosen parameters, it makes \( r_{is} \in [0,1] \).

Number of plot method

\[
r_{is} = \begin{cases} 
1 & i = h \\
\frac{1}{M} \sum_{k \neq i} x_{ik} \cdot x_{is} & i \neq h 
\end{cases}
\]  

(4)

© 2014 ACADEMY PUBLISHER
where $M$ is suitably parameters to the $r_{ik} \in [0,1]$.

Central angle cosine method

$$r_{ik} = \left(\frac{\sum_{h} x_{ih} x_{ikh}}{\sqrt{\left(\sum_{h} x_{ih}^{2}\right)\left(\sum_{h} x_{ikh}^{2}\right)}}\right)$$

(5)

Maximum and minimum method

$$\sum_{h} (x_{ih} \land x_{ikh}) / \sum_{h} (x_{ih} \lor x_{ikh})$$

(6)

Definition, if $U$ is a finite non-doctrinaire domain, $X \subseteq U$, $R \subseteq U \times U$ is arbitrary binary relations to $U$.

$$c(R_i(x), X) = \begin{cases} 1 - \frac{|R_i(x) \cap X|}{|R_i(x)|}, & R_i > 0 \\ 0, & R_i = 0 \end{cases}$$

In which $|R_i(x)|$ represents the cardinality of set $R_i(x)$, $c(R_i(x), X)$ is called the relative error rate about set $R_i(x)$ on $X$. Obviously, $0 \leq c(R_i(x), X) \leq 1$.

Definition, $0 \leq \beta \leq 0.5$, most contain general relations relationship is defined as

$$X \supseteq R_i(X) \leftrightarrow c(R_i(x), X) \leq \beta$$

The amount of data in the Web log is massive, directly mining based on the Web log will undoubtedly algorithm inefficient. In addition, the Web log a lot of noise data and invalid data, noise data and invalid data is likely to affect the accuracy of clustering results. Therefore, we need cleaned before the Web log analysis. It can be confirmed from the statistical analysis of the Web log, that some extensions URL (such as suffixed by. ico, gif, jpg, css, wmv, swf, etc) has nothing to do with mining Web log. Also transfer status as "404", "301" and "500" are unrelated contents. These clear these unrelated record, no doubt will greatly reducing the amount of data, and improve the accuracy and quality of the mining.

Pointed out that there may be such visitors access records too scarce and cannot constitute the main body of the log file records; constitute logging request visitors access records also requested page there is too little of the total number of clicks. These data is the apparent lack of a sufficient amount of information and should not be involved in the mining process. Therefore, insufficient amount of information filtering log data can be further reduced to the size of data to reduce the dimension of the space of the clustering algorithm to improve the efficiency and quality of clustering.

B. Fuzzy Clustering Algorithm

1) Transitive Closure Method

The fuzzy matrix calibration from the Web just a Web fuzzy similarity matrix, not necessarily transitive, that $R$ is not necessarily Web fuzzy equivalent matrix. Classification transformed into the $R$ Web fuzzy equivalent matrix. Fuzzy similar matrix $R$ from the Web,

the square method and the transitive closure of $R$, $R$’s Web fuzzy equivalent matrix, i.e. $R' = \tau(R)$. Then, the obtained Web fuzzy different equivalent matrix $R' r_{ik}$ are arranged in descending order in the number set $l = (1 = \lambda_1 > \lambda_2 > \cdots > \lambda_n)$, so that the input parameter $\lambda$ iterate number set $l$ , obtained by $\tau(R)_{\lambda}$ -series classification that Web fuzzy classification.

2) Direct Cluster Analysis

Direct clustering method to create a Web fuzzy similar matrix, do not seek to pass closures, but directly from the Web departure fuzzy similar matrix obtained Web fuzzy classification [12]. Theory has proven; direct clustering method and transitive closure method is equivalent to the result. The concrete steps are as follows:

(1) Take $a = 1$ (maximum), each for similar class, about to meet, and placed in a class, a similar class. Similar classes and equivalence classes’ difference lies in the different similarity class may comprise common elements, which may be the case of the formula (7). Similar class containing common elements combined can be obtained in the case of equivalence classification.

$$\left[ x_{jr} \right] = \left[ x_{jr} , x_{jr} \right]$$

(2) Is taken as the second largest value, find out directly from the degree of similarity as elements, i.e., corresponding to the classification corresponding to the equivalence class where the class where combined, all of these cases were combined after corresponding at the equivalent classification.

C. Session and Usage Identification

The main goal of user identity is to distinguish and identify the user. One of the best methods of user identification is to implement user registration visit, but this is definitely limiting site visitors access to some extent reduce the site visits.

The session process of web requesting is showed as below. Browser stores data of web pages to local storage and local storage connected to numerous web servers all over the internet by http request of other requesting, and these different web servers are connected to a session sever. In some particular cases, browser also could request web data from web server directly.

![Figure 5. Session model of web request](image-url)
In addition, a better approach is to analyze Web logging visitors IP and visitors’ status of these two fields. If two records visitor IP and visitors to the state field values are not identical, that this is two different user access records; otherwise, according to the topology of the site to determine whether the two records requested URL link exists, if not The presence of a link can be considered to exist on the same machine by two different users, otherwise the two records correspond to the same user.

Session the user a continuous effective access to the site, its manifestations is the user’s access path sequence. Different users access to the site should be considered a different session. If the same user has two pages requested interval within a predetermined time (i.e. the session validity), it can be considered both pages belonging to the same session; otherwise, these two pages belonging to different sessions of the same user. The session is valid for different sites can be set according to the actual situation.

User behavior model during surfing on internet is showed as below figure, it was consisted of seven modules: habit hierarchies, social factors, affect, perceived consequence, intentions, facilitating conditions and behavior is related to habit hierarchies, intentions and facilitation conditions. Therefore, only these three modules are considered in our usage model identifications.

Web user browsing behavior directly reflects the user's interest, because the users to access the Web site are generally carried out in accordance with the interest. The process of analyzing user access to the site, it is not difficult to find Web user browsing behavior generally involves three the behavior factors [4]: time on page, the number of page views, and the order of access paths.

Because of the correlation matrix of the structure is only concerned about the user access to a single page and cannot examine the order of the access path to the size of the degree of interest, so we through the introduction of multiple linear regression model to describe the page residence time and the number of page views on the degree of interest. The linear regression equations can be expressed as: \( P_\mu = aX_\mu + bY_\mu + c \), wherein \( P_\mu \) the degree of interest of the \( jth \) user on the \( ith \) site; \( Y_\mu \) expressed in a time repeated clicks of the \( jth \) user to the number of the \( ith \) site; \( a, b, c \) are empirical value of the site.

**D. Structure Similar Matrix**

User ID is the URL for columns and user interest in the page for the value of the structure User ID-URL associated matrix. After filtration insufficient information data “After this step, the associated dimension of the matrix is greatly reduced. Can know, by the characteristics of the associated matrix row vector similar analysis can be interested in a similar customer base. In order to construct the user interest similarity matrix, we define: \( S_{jk} \) of = \( 1 - \text{Cd} (R_j, R_k) \), where \( S_{jk} \) is the \( ith \) user and the \( jth \) user's interest similarity; \( c \) format rows similarity distance to \([0, 1]\) of the appropriate coefficients; \( D (R_j, R_k) \) represents the \( jth \) row vector of the correlation matrix and the \( jth \) row vector of the similarity distance. The similarity distance can use the Hamming distance to be expressed as: \( D(R_j, R_k) = \sum_{m=1}^{n} | P_{jk} - P_{mm} | \)

Where in \( m \) is the URL of the associated matrix dimension; \( P_{jk} \) said \( ith \) user to the \( kth \) URL of the degree of interest. User ID-User ID row vector similar distance structure similar to the matrix with a reflexive, symmetrical features but do not have to pass. Web users clustering algorithm based on similarity matrix by seeking the transitive closure of the similarity matrix for \( \lambda \) cut clustering when the similarity matrix dimension, find a large amount of computing transitive closure may result in the low efficiency of the algorithm. The theory has proven: direct clustering method and transitive closure method with equivalence [5]. Therefore, you can directly proceed from the similarity matrix by setting set \( \lambda \) cut-off on the similarity matrix. There may contain between \( \lambda \) cut-off after class and class with inclusion relations exist, so you need to contain relationship classes merge. In addition, due to the algorithm itself, different classes may be present the same items. In order to solve the problem of the membership of the same item, we define the entry and connection strength.

\[
J(U_i, c_j) = \sum_{i=1}^{m} \frac{\text{sim}(U_i, U_j)}{m}
\]

connect in strength, \( \text{sim}(U_i, U_j) \) represents the similarity of the options \( U_i \) and \( c_j \) class \( U_j \), \( m \) is \( c_j \) size of the class. Seen from the definition, the connection strength of the item with the class embodies the class cohesion.

The clustering algorithm is described as follows:

**Input:** set the value of \( \lambda \)

**Output:** Web user clustering \( C = \{c_j\} \)

**Step 1:** Initialization; \( C = \Phi \)

**Step 2:** the FOR 1 \( i = 1 \) to \( L \) (\( L \) is the similarity matrix dimension) do

\[
\}

© 2014 ACADEMY PUBLISHER
a) Initialization $c_i = \{U_i\}$

b) For $i < j \leq 1$

\[
/ \text{Sim}\ (U_i, U_j)\text{ is the value of } j\text{-th column of the } i\text{-th row of the similarity matrix.}
\]

If $\text{Sim}(U_i, U_j) \geq \lambda$, then

\[
c_i = c_i \cup \{U_j\}
\]

c) If $C$ in the respective elements with the $c_i$ is not exist that include or contain Off Department, then

$C = C \cup \{c_i\}$

Else $C$ contains only retain relations

Step 3: Calculate the various connections of the same items with the class strength will

The largest item under the corresponding class of the connection strength and eliminate recurring items.

Step 4: output clustering results.

IV. CASE STUDIES

Departure from the original Web log, User-ID can establish the following steps in accordance with the structure of the similarity matrix, similarity matrix.

\[
M = \begin{pmatrix}
1 & 0.749 & 0.020 & 0.259 & 0.755 & 0.999 & 0.996 \\
1 & 0.259 & 0.510 & 0.994 & 0.750 & 0.753 \\
1 & 0.748 & 0.265 & 0.020 & 0.017 \\
1 & 0.504 & 0.260 & 0.263 \\
1 & 0.755 & 0.752 \\
1 & 1 \\
1 & 1
\end{pmatrix}
\]

When the cut set $\lambda = 0.36562$

a) $\lambda$ cut clustering after seven sub-categories:

i. $\{U_1, U_2, U_3, U_6, U_7\}$

ii. $\{U_2, U_4, U_5, U_6, U_7\}$

iii. $\{U_3, U_4\}$

iv. $\{U_5, U_7\}$

v. $\{U_6, U_7\}$

vi. $\{U_5\}$

vii. $\{U_1\}$

b) The combined class contains the relations between the three subclasses:

i. $\{U_1, U_2, U_3, U_6, U_7\}$

ii. $\{U_2, U_4, U_5, U_6, U_7\}$

iii. $\{U_3, U_4\}$

c) Processing the same items in each of the sub-class U2 and i like the connection strength:

\[
J(U_2, i) = \frac{0.749+1.0.994+0.750+0.753}{5}
\]

U2 and ii connection strength:

\[
J(U_2, ii) = \frac{1+0.504+0.260+0.263+1.0.755+0.752}{5}
\]

Since $J(U_2, i) > J(U_2, ii)$, U2 only belong to the class $i$.

Other the same items can be classified in accordance with similar practices.

d) Output clustering results

i. $\{U_1, U_2, U_3\}$

ii. $\{U_6, U_7\}$

iii. $\{U_5, U_7\}$

When the cut set $\lambda = 0$, the clustering results $\{U_1, U_2, U_3, U_5, U_6, U_7\}$

When the cut set $\lambda = 1$, the clustering result $\{U_1\}, \{U_2\}, \{U_3\}, \{U_4\}, \{U_5\}, \{U_6\}, \{U_7\}$

When cut set $\lambda = 0.62234$, clustering results $\{U_1, U_2, U_3, U_6, U_7\}, \{U_5, U_7\}$

Various kinds of tests, the experimental data on the Windows Vista platform using the Java language to achieve a clustering algorithm is divided into two kinds of synthetic and real data. The test machine’s hardware configuration: Processor AMD Dual-Core 1.9G, memory 2G. The first test is the of the $\lambda$ cut sets the size of the number of clusters. Site (http://211.66.184.35) on January 10, 2010 00:00:05 to 00:32:47 Web log as the experimental data, the original log recorded a total of 13,509, after data cleaning steps such logging 1005, effective records about 7.44% of the total number of records, identified a total of 16 visitors and 17 main access path. The impact of the different values of $\lambda$ for the number of clusters is shown in Figure 1. You can see when $\lambda$ values greater the higher the classification accuracy, a separate clusters bigger chance; conversely, multiple items together into a kind of bigger chance.

Next step is to test the accuracy of the algorithm. Taken into account when a small number of users, number of clusters is less likely to lead to the extreme cases of the high accuracy of the clustering algorithm, we constructed five test case contains more users. The number of users of these five test cases is among 8 to 15. First of all users in the five test cases manually clustering, then $\lambda$ values as close as possible to the number of clusters and number of clusters manually adjust the clustering algorithm. Define the accuracy of the evaluation standards:

\[
p = \frac{\sum_{i=1}^{n} P_i}{n}
\]

Where in $n$ is the number of clusters, representing the accuracy of the various types of overlapping items and such size ratio, i.e., the corresponding class of the various manual clustering. Finally to obtain the accuracy of the five test cases shown in Figure 2, the accuracy of the algorithm can be obtained from FIG probably in about 80.616%.

The final test is the performance of the algorithm. More meaningful to the experimental results, we selected five large amounts of Web log data as a test case, their size 1203KB, 1340KB, 1466KB, 1668KB and 1856KB.
Adjust the value of $\lambda$ to obtain a more reasonable number of clusters obtained from the experiment the algorithm shown in Figure 3 consumes CPU time and the relationship of the amount of log data. As seen from the Figure 3, when the number of clusters is more reasonable, as the amount of data increases, the moderate increase of the CPU time consumed by the algorithm, a good scalability.

![Figure 8. The accuracy of the test case](image)

![Figure 9. The performance of the algorithm](image)

V. CONCLUSIONS

With the rapid development of the Web, e-commerce has become an integral part of the new business model. The presence of large numbers of the commercial value of the information to the e-commerce site, how to tap these valuable information critical to the development of enterprises and businesses. Identify user interests and cluster users from Web log, is conducive to the realization of the recommendations and personalized service. In this paper, a multiple linear regression analysis of user browsing behavior and adjust clustering results to items with connection strength Web user clustering method. First, to establish a multiple linear regression model associated with a good page user interest and user browsing behavior, user access matrix row vector similarity analysis into user interest similarity matrix, and then direct the similarity matrix $\lambda$ cut poly class, and finally through the merger subclass and calculated item connection strength to adjust clustering results. The experiments show that the high accuracy of the algorithm, and better scalability.

REFERENCES


