doi:10.2498/cit.1002281

# Research on Cluster Analysis Method of E-government Public Hotspot Information Based on Web Log Analysis

Suozhu Wang, Jianlin Zhang, Fuze Yang and Jia Ye

School of Management, Capital Normal University, Beijing, China

To master the hotspot information of E-government public opinion macroscopically, and so that the government can provide more target-oriented information and services for the public. A method for calculating the public concern degree of E-government public opinion is proposed in this paper. In this method, two indices that can reflect the hotspot information of public opinion to a certain degree are calculated through the web log mining, the attention degree of the information through the weighting computation to find the hotspot problems that the public cared about in a certain period of time. A fuzzy cluster analysis method that can find hotspot problems is presented in this paper. And an illustrative example is given to describe the clustering process.

*Keywords:* public hotspot information, web log analysis, cluster analysis, e-government

## 1. Introduction

E-government is the joint product of information technology and new principles of public administration. The primary function of e-government is to support communication between governments and citizens via web-enabled computer technologies [1]. E-government is currently being implemented worldwide. Most government institutions now have their own space on the internet, thereby allowing citizens to find information and, increasingly, to engage in e-government services [2]. E-government uses technology to support a government's interaction with multiple stakeholders including employees, businesses, and other government agencies. It has been a key element of public service reconstruction undertaken by many countries in the 21st century.

A number of studies have been conducted in developed countries, including the U.S., EU countries, Australia, New Zealand, South Korea, and Singapore, as well as Taiwan and Hong Kong, which are regarded as leaders in e-government development. In these developed countries and regions, e-government is characterized by advanced services, such as e-democracy and the integration of services and resources. Government portal websites have become one of the most important channels for public service delivery and for citizen-government interaction. As a result, e-government system, represented by government websites, has a mass of data and rich information resources. How to identify hot issues of public concern quickly and timely from this information, and grasp public interest in information comprehensively is crucial for both the party and the government, to make right decisions and achieve the established goals. Because of this, survey and mining analysis of on-line public interest information are also considered as important content of e-government information mining analysis [3].

When users are visiting web sites, their browsing behaviors are recorded in web logs. These data imply knowledge about users' information browsing interest and so on. They can provide decision-making grounds for site structural improvements and applications such as information sending by using web log data,

The first and abridged version of this paper appeared originally in LISS 2013, Reading, UK, August 21–24, 2013. Published in partnership with ICIR of Beijing Jiaotong University.

clustering technology, web usage mining techniques and other methods to find interesting patterns of user browsing behavior and other aspects. Cluster analysis is a branch in statistical multivariate analysis and unsupervised machine learning, having extensive applications in various domains, including financial fraud, medical diagnosis, image processing, information retrieval and bioinformatics. Web usage mining is an application of data mining algorithm to weblogs to identify trends and regularities in web users' navigation patterns. Among the research work in this area, scholars have done a lot of research and exploration. For example, a clustering algorithm for on-line user behavior analysis proposed in [4,5], considers that the number of times a user clicks on a web page indicates the attitude of the user, and we can conduct clustering analysis on users on the basis of this assumption; the authors conducted a cluster analysis of users based on the time they stay on web pages in [6]. These methods also achieved some results in practical application.

Currently, mining analysis research specifically on e-government public interest information is rare. General collection of public interest information is mainly based on questionnaires, inevitably causing data sources narrowing and data subjectivity. Consequently, results of the mining analysis lack intensity and objectivity. Since public interest information is real-time, dynamic and involving many fields, it is hard to achieve real-time and comprehensive quantitative analysis on public interest information with the data collected by questionnaires alone, it's also difficult to ensure validity of the following public opinion hotspot information classification. So this paper, based on data from web logs, proposes a concept and measurement method of public information concern, which provides clustering analysis method of public interest information. This method implements classification of public interest information from the perspective of users' behavior, which helps the government grasp macro needs of public information, in order to integrate this information as much as possible and provide more targeted services.

The paper is organized as follows. The concept of hotspot information on public opinion and public concern degree is defined in Section 2, while clustering analysis method of hotspot information is presented in Section 3. Finally, Section 4 draws a conclusion of the research and describes the future work.

## 2. Hotspot Information of Public Opinion and Public Concern Degree

The public opinion information refers to all relevant information which reflect wishes, ideas and attitudes of the people, including questions, comments and suggestions the public propose through the network to the government.

The hotspot information of public opinion are questions and information with relatively high degree of public concern, visited rather frequently.

In an e-government environment, two indicators, number of public clicks on some information and the average residence time the public spend on some information, can, to a certain extent, identify hot issues that people are generally concerned about. The value of indicators can be calculated based on web log analysis.

### 2.1. Web Log Data Collection

The original web log data can be obtained from a variety of sources, such as web server, proxy server or client. Here, the original data is mainly obtained from web server log file of the government. The web server log file explicitly records the user access to the site. Each time a page is visited, the web server will add a corresponding record in the log. It is not only a detailed record of the site visitors browsing behavior, but a collection of multiple users' access to the same site behavior. Although a different web server log file format is not exactly the same, it generally follows the W3C standard, including a series of information about the user visiting the site, such as user IP, access date and time. Table 1 is an example of web log data.

### 2.2. Preprocess Web Log

In an e-government environment all users' browsing information is stored in web logs. In the original web log files not all information is useful, only the user's IP address, access time and date. The URL of requested page is valuable,

| IP | TIME | URL |
|---|---|---|
| 192.168.126.196 | 19/Aug/2012:14:47:37 | `http://zhengwu.beijing.gov.cn/gzdt/gggs/t1240141.htm` |
| 124.127.179.64 | 21/Aug/2012:15:23:44 | `http://zhengwu.beijing.gov.cn/gzdt/gggs/t1240141.htm` |
| 61.135.132.86 | 03/sep/2012:08:33:45 | `http://zhengwu.beijing.gov.cn/gzdt/gggs/t1239575.htm` |
| 123.112.44.0 | 05/sep/2012:20:32:55 | `http://zhengwu.beijing.gov.cn/gzdt/gggs/t1239009.htm` |
| 221.220.159.78 | 12/sep/2012:05:11:30 | `http://zhengwu.beijing.gov.cn/gzdt/gggs/t1241296.htm` |
| 117.79.235.94 | 19/sep/2012:09:17:12 | `http://zhengwu.beijing.gov.cn/gzdt/gggs/t1241178.htm` |
| 124.127.179.64 | 23/sep/2012:19:08:36 | `http://zhengwu.beijing.gov.cn/gzdt/gggs/t1240973.htm` |

*Table 1.* Web log data.

other data must be cleaned and conversion pretreated before the data analysis. For example, picture, voice and animation stored in web log files, are irrelevant and should be removed. There are several preprocessing tasks that must be performed prior to mining hotspot information in the data collected from web logs. Traditional preparation includes data cleaning, user identification, session identification and path completion [7]. Here we just need data cleaning and simple session identification. Data cleaning means that all data irrelevant to the method are deleted. Session identification refers to the number of users' accesses to the information and to the duration of their retention in the pages whose calculation is based on a session.

Then, web logs can be expressed as a set of L= ⟨ URL_Referer, URL, Viewing Time⟩ . URL_Referer denotes referrer information page. URL denotes the browsing information page Viewing Time denotes the time spent on the browsing information page.

## 2.3. Degree of Public Concern

Government issues a lot of information related to people's vital interests through e-government websites every day. In general, the more times that information is clicked by public, indicates that more people are interested in this information. Meanwhile, the more visitors are interested in the information content, the longer it will stay.

If there are $m$ kinds of different selections to browse government information on a government website those selections that occur more frequently reveal public intention and preference.

***Definition 1 (Selection Tendency)*** Assuming that ES is a government website, let set $C = \{c_1, c_2, \ldots, c_n\}$ represent all types of information pages in ES, $\forall c_i \in C, c_i = \{c_{i1}, c_{i2}, \ldots, c_{im}\}$ mean that set of all information of web page $c_i$, then selection preference or tendency of website users to information $c_{ik}$ is defined as

$$ST = V_{ik}/((\sum_{j=1}^{m} V_{ij})/m) \qquad (1)$$
$$(i = 1, 2, \ldots, n; \quad k = 1, 2, \ldots, m)$$

where $V_{ik}$ stands for the number of clicks on the information $c_{ik}$ in web page $c_i$ by users within a certain period of time.

In traditional methods, frequency of access to the page is used to measure the user interest degree, which is expressed as $c_k = (\sum_{i=t}^{n} c_i)$. This method is not applied to measuring the preference of users for browsing information, because if there is only single information in certain types of web pages, user preference degree to this web page is on the high side. For example,

$$c_1 = \{c_{11}\}, c_2 = \{c_{21}, c_{22}, \ldots, c_{28}\},$$
$$V_{11} = 10, V_{22} = 10, \sum_{j=1}^{8} V_{2j} = 50,$$

if it's defined by frequency the preference degree for $c_{11}$ is $1(10/10 = 1)$, the preference degree for $c_{22}$ is $0.2(10/50 = 0.2)$, because of one web information in page $c_1$, $c_1$ is always 1, there are eight web information in $c_2$, the preference degree for $c_2$ is impossible greater than 1, thus it's difficult to say that users prefer $c_1$ to $c_2$. The definition 1 in this paper solves the problem. For $(1)$, the preference degree for $c_{11}$

is $1(10/10 = 1)$, the preference degree for $c_{22}$ is $1.6(10/((50/8)) = 1.6)$, which means that users pay more attention to $c_{22}$ than to $c_{11}$.

According to the viewing time of an information page, navigation can be classified into passing, simple viewing, normal viewing, preferred viewing, which presents how interested the user is. The longer viewing time, the more interested and preferred visiting, and vice versa. According to the 3 ways of viewing, by using the discretization technique, the viewing time can be classified as follows

$$T = \begin{cases} 1 & 0 < t \leq T_{\max\_passing} \\ 2 & T_{\max\_passing} < t \leq T_{\max\_normal\_viewing} \\ 3 & T_{\max\_normal\_viewing} < t \end{cases}$$

where the time range $0 \sim T_{\max\_passing}$ represents skipped browsing and is defined as 1, the time range $T_{\max\_passing} \sim T_{\max\_normal\_viewing}$ represents normal browsing and is defined as 2, the time range above $T_{\max\_normal\_viewing}$ represents preferred browsing and is defined as 3.

For example, supposing that $T_{\max\_passing}$ is 60 s and $T_{\max\_normal\_viewing}$ is 300 s, respectively, the time could be partitioned as follows:

$$T = \begin{cases} 1 & 0 < t \leq 60\text{ s} \\ 2 & 60\text{ s} < t \leq 300\text{ s} \\ 3 & 300\text{ s} < t \end{cases}$$

If there is $n$ different e-government information, those information that users view for a long time attract more public attention.

***Definition 2 (Viewing Time Tendency)*** Assuming that ES is a government website, let set $C = \{c_1, c_2, \ldots, c_n\}$ represent all types of information pages in ES, $\forall c_i \in C, c_i = \{c_{i1}, c_{i2}, \ldots, c_{im}\}$ mean that set of all information on web page $c_i$, then the viewing time preference or viewing time tendency to page information $c_{ik}$ is defined as

$$VTT = T_{ik}/((\sum_{j=1}^{m} T_{ij}/m)) \qquad i = 1, 2, \ldots, n; \quad k = 1, 2, \ldots, m) \tag{2}$$

where $T_{ij}$ stands for total viewing time of information $c_{ij}$ in web page $c_i$ by users within a certain limited time.

Although indicator $ST$ and $VTT$ can reflect the concern of people on a variety of information to a certain extent, it's not very accurate to measure it by one indicator only. Therefore, this paper proposes an analysis method which takes both the number of clicks and the average residence time into consideration and measures the concern degree of information by weighted calculation, in order to identify the hot issues of public concern in a certain period of time. For different government websites, these two indicators have different impact in the analysis of public concerns about various information.

***Definition 3 (Public Concern Degree)*** Assuming that $ES$ is a government website, let set $C = \{c_1, c_2, \ldots, c_n\}$ represent all types of information pages in ES, $\forall c_i \in Cc_i = \{c_{i1}, c_{i2}, \ldots, c_{im}\}$, mean that set of all information on web page $c_i$ then the public concern degree for $c_{ik}$ is defined as

$$PCD = a * ST + b * VTT \tag{3}$$

where $a$ and $b$ are weights, $a, b \in [0, 1], a + b = 1$.

As seen from definition 3, public concern degree of e-government information is a weighted value of ST and VTT. ST reflects the concern degree of information to the public from the choice intention, and VTT reflects the public's attention from the angle of information viewing time. For some web information, the user will stay a lot of time when visiting it because of the structure of the site location, however this formula can weaken this advantage by the selecting tendency. Similarly, some web information in the page contains many links, so users will view it a lot of times, but its access time is not long compared to other information. This method can reduce the impact of views to the preference by the viewing time preference. The higher public concern degree for information, the more interested in the information, and vice versa.

When users visit a government website, the preference of each user to web information is measured in order to find the hotspots of the users' interest, so as to be able to offer users personalized service, making it the most effective services. Improve users' satisfaction with government, and ultimately improve the level of government services.

***Definition 4 (Public Concern Degree Matrix)*** Matrix $A = (a_{ij})_{m \times n}$ $(i = 1, 2, \ldots, m,\ k = 1, 2, \ldots, n)$ is known as Public Concern Degree Matrix, where $a_{ij}$ represents the public concern degree of user $i$ to information $j$, $a_{ij}$ is calculated by the formula (3).

## 2.4. Method of PCD Calculation

Web log data can be transferred into a standardized, actionable data set through the pretreatment described in subsection 2.2, then conduct a variety of statistics in accordance with data required by indicators. Since user's access records are stored in pretreatment web log data, these indicators data can be analyzed according to URL by statistical analysis. Clicks of a certain information is the count of URL visiting records mapping to this message, and viewing time is the time interval between access time of this information and next one. Through web log mining they can be brought into $ST$ and $VTT$ formulas and the values of $ST$ and $VTT$ are calculated.

As the index values obtained above are different in dimension and the difference between them is large, analysis based on them will lack standardization and comparability. Therefore, these two index values need to be normalized, namely mapping numerical value to the $[0,1]$ interval, in order to obtain evaluation score values on the same dimension. Indicators can be standardized as follows.

According to expert experience method, each indicator can be given a finite interval $C = [\min x_i, \max x_i]$ in which $\min x_i$ stands for the worst critical point or the worst score of these indicators and $\max x_i$ stands for the best critical point or the best score. In this case, we assume $r_i(x)$ is the evaluation score of indicators $ST$ and $VTT$, $r_i(x)$ and the measured value $x$ of the indicator in interval $C$ are linearly proportional. The index value less than $\min x_i$ can be assigned a score of 0 and those above $\max x_i$ can be assigned a score of 1. So we conclude a formula below

$$r_i(x) = \begin{cases} 1 & x > \max x_i \\ \frac{x - \min x_i}{\max x_i - \min x_i} & \min x_i < x < \max x_i \\ 0 & x < \min x_i \end{cases}$$

$$i = 1, 2, \ldots, n$$

Therefore, according to their degree of importance, weights were set as $\{a, b\}$, $a, b \in [0,1]$, $a + b = 1$. Through standardization, the indices $ST$, $VTT$ are transferred into $r_1$ and $r_2$ in the same dimensions, and then we obtain the degree of public concern about information $i$ by weighted calculation, and its expression is $PCD = a * r_1 + b * r_2, PCD \in [0,1]$. Calculate the concern degree of each information issued on government website, according to the value of the size, these information will be sorted in descending order by the value of $PCD$, the one listed on the top is the hot information that people are most concerned about.

## 3. Clustering Analysis Method of Hotspot Information

Cluster is defined as a process classifying the set of unsorted things as reason, according to certain rules and certain properties of things. Cluster analysis is based on the principle that the objects have similarities as large as possible in the same class, and the objects of different classes have differences as large as possible [8-10].

### 3.1. Clustering Analysis Method

There are many commonly clustering algorithms in data mining. They are partition method, method based on the hierarchy, method based on density, and so on. However, these methods have some disadvantages, especially in dealing with large-scale, high dimension, fuzzy, dynamic data. The hot issue people are concerned about is analyzed from the perspective of user's behavior in this paper, so we will use fuzzy clustering analysis method through the comparison of different clustering methods.

In a certain period of time, through a web log mining extraction, in order to get the related information on the users who access government websites, and find the hot issue people are concerned about during this period objectively. And then clustering of the hot issue can be conducted by using the fuzzy analysis method. The specific method is as follows [11-12].

(1) Suppose there are $n$ hot spots obtained by using method described in Section 2 in a period of time, using $Y$ represents set of page URL, $Y = \{Y_1, Y_2, \ldots, Y_n\}$; at this time there are $m$ users accessing information, using $U$ represents the set of these users $U = \{U_1, U_2, \ldots, U_m\}$. $Y$ is an object to be classified.

(2)The time of each user $U_j$ $(j = 1, 2, \ldots, m)$ access to each message $Y_i$ $(i = 1, 2, \ldots, n)$ is calculated in this period. If $T(Y_i, U_k)$ stands for the time user $U_k$ visiting $Y_i$, the time of each $Y_i$ $(i = 1, 2, \ldots, n)$ that was accessed by all users, that is $\sum\limits_{k=1}^{m} T(Y_i, U_k)$, and $b_{ij} = \frac{T(Y_i,U_k)}{\sum\limits_{k=1}^{m} T(Y_i,U_k)}$ means degree of association the information $Y_i$ and between the user $U_j$. And $n$ rows and $m$ columns of the original data matrix $B$ are obtained.

$$B = \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1m} \\ b_{21} & b_{22} & \cdots & b_{2m} \\ b_{31} & b_{32} & \cdots & b_{3m} \\ \vdots & \vdots & \ddots & \vdots \\ b_{n1} & b_{n2} & \cdots & b_{nm} \end{bmatrix}$$

(3)Construct fuzzy similar matrix $R$: fuzzy similarity matrix is used for storing similarity points between classifying objects. The elements of the matrix $R$ are set, it represents the similarity between objects $i$ and $j$. The commonly used method of calculating, there are dot product method, angle cosine method, the maximum and minimum method, the arithmetic average method. This paper adopts Max-min method, making the original data matrix $B$ into $n * n$ fuzzy similarity matrix. Expressed as $r_{ij} = \frac{\sum\limits_{k=1}^{m} \min(b_{ik},b_{jk})}{\sum\limits_{k=1}^{m} \max(b_{ik},b_{ik})}$, where $r_{ij}$ is the degree of similarity between the information $i$ and information $j$ $(i,j = 1, \ldots, m)$ and $r_{ij} = 1$, if $(i = j)$.

(4)Fuzzy similar matrix transitive closure is obtained by method of combining fuzzy similarity and graph theory. Specific steps are as follows:

1) Because fuzzy similar matrix $R$ has symmetry, only similarity on one side of diagonal of $R$ needs to be sorted in a descending order, and written down the corresponding number of rows and columns.

2) Generate an undirected graph $G$, each node corresponds to each individual object of classifying objects.

3) Connect the corresponding nodes according to the order generated in step 1), and mark the weight, that is the similarity between two individual subjects.

4) If the graph $G$ appears to loop after an edge is added to graph $G$, the edge is not to be added. The join process continues until the graph $G$ is connected, and the resulting graph $G$ is a tree.

5) Construct fuzzy equivalent matrix $R^F$, where $rF_{ij}$ $(i,j = 1, 2, \ldots, m)$ is the element of the matrix $R^F$. Marking $rF_{ij} = 1$, $(i = j)$. $rF_{ij}$ $(i \neq j)$ is the minimum weight node in the tree $G$.

(5)Select cutting matrix method for fuzzy clustering analysis: Select $\lambda, \lambda \in [0, 1]$, with the largest to smallest value, the classification $Y$ corresponding to fuzzy equivalent matrix $R^F$ constantly changes, so that the classification forms a dynamic clustering figure. The greater the value $\lambda$, the higher the classification precision. And if the value $\lambda$ is smaller, the classification accuracy is more rough. Generally the value is the weight of the tree $G$. After the above steps, the hot information the people are concerned about is obtained.

## 3.2. Illustrative Example

The e-government public hot issue clustering analysis method presented in this paper will be showed by a local government website example analysis. Assume the government released a total of 5 messages within a certain period, and that they have the same information entropy. After the pretreatment of web log data of the government web site, the numbers of the public clicks and the public average viewing time of 5 messages can be gotten. Part of the data in web log data is shown in Table 2.

By the expert experience method, if a public message is clicked 40000 or more times, it can be regarded as the best, 0 as the worst. Similarly, the public average viewing time 300s (5min) or more to a message is considered excellent, 0 for the poorest. According to the above formula, $ST$, $VTT$ indicators' values of five messages can be normalized to evaluation scores as shown in Table 3.

| User Id | Information Page | Clicked Numbers/times | Viewing time/s |
|---|---|---|---|
| U$_1$ | M$_1$ | 6100 | 16 |
| U$_1$ | M$_2$ | 2700 | 10 |
| U$_1$ | M$_3$ | 790 | 9 |
| U$_1$ | M$_4$ | 3000 | 6 |
| U$_1$ | M$_5$ | 2010 | 13 |
| U$_2$ | M$_1$ | 5040 | 12 |
| U$_2$ | M$_2$ | 3400 | 13 |
| U$_2$ | M$_3$ | 3750 | 18 |
| U$_2$ | M$_4$ | 6000 | 13 |
| U$_2$ | M$_5$ | 800 | 9 |
| U$_3$ | M$_1$ | 2850 | 10 |
| U$_3$ | M$_2$ | 2080 | 8 |
| U$_3$ | M$_3$ | 2500 | 13 |
| U$_3$ | M$_4$ | 2300 | 3 |
| U$_3$ | M$_5$ | 1350 | 11 |
| U$_4$ | M$_1$ | 1810 | 8 |
| U$_4$ | M$_2$ | 1420 | 4 |
| U$_4$ | M$_3$ | 1760 | 11 |
| U$_4$ | M$_4$ | 7300 | 18 |
| U$_4$ | M$_5$ | 140 | 5 |

*Table 2.* Part data on public clicks and the public average viewing time.

| | M$_1$ | M$_2$ | M$_3$ | M$_4$ | M$_5$ |
|---|---|---|---|---|---|
| r$_1$ | 0.395 | 0.24 | 0.22 | 0.465 | 0.108 |
| r$_2$ | 0.167 | 0.177 | 0.17 | 0.133 | 0.127 |

*Table 3.* Standards evaluation scores.

Suppose that experts set $ST$ weight of 0.5, $VTT$ weights for 0.5, then according to the formula $PCD = a * r_1 + b * r_2$ the public attention of each information is calculated as following:

$$PCD_1 = 0.395 * 0.5 + 0.167 * 0.5 = 0.281$$
$$PCD_2 = 0.24 * 0.5 + 0.117 * 0.5 = 0.179$$
$$PCD_3 = 0.22 * 0.5 + 0.17 * 0.5 = 0.195$$
$$PCD_4 = 0.465 * 0.5 + 0.133 * 0.5 = 0.299$$
$$PCD_5 = 0.108 * 0.5 + 0.127 * 0.5 = 0.1$$

In accordance with the attention from high to small, the hot issue information of public concern is M$_4$, M$_1$, M$_3$, M$_2$, M$_5$.

These five hot issue information the public is concerned about is denoted as $Y = \{Y_1, Y_2, Y_3, Y_4, Y_5\}$, assuming there are four users to access them in this time period, denoted as $U = \{U_1, U_2, U_3, U_4\}$, and each user access time from the web log data preprocessed by statistics information will be stored in a matrix $M$, as follows:

$$\begin{array}{c} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \end{array} \begin{bmatrix} 5 & 20 & 65 & 21 \\ 2 & 0 & 0 & 20 \\ 0 & 20 & 65 & 0 \\ 0 & 0 & 0 & 18 \\ 0 & 5 & 6 & 0 \end{bmatrix}$$

According to the formula $b_{ij}$, construct the degree of association between information and the users, get the original matrix $B$, as follows:

$$B = \begin{bmatrix} 0.04 & 0.14 & 0.46 & 0.36 \\ 0.09 & 0 & 0 & 0.91 \\ 0 & 0.24 & 0.76 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0.45 & 0.55 & 0 \end{bmatrix}$$

According to the formula $r_{ij}$, the original matrix $B$ changes into a $5 * 5$ fuzzy similar matrix $R$, as follows

$$R = \begin{bmatrix} 1 & 0.23 & 0.45 & 0.22 & 0.45 \\ & 1 & 0 & 0.83 & 0 \\ & & 1 & 0 & 0.65 \\ & & & 1 & 0 \\ & & & & 1 \end{bmatrix}$$

According to the fuzzy similarity matrix, transitive closure algorithms obtain the following sequence of steps 1)

$$r = \{0.83, 0.65, 0.43, 0.43, 0.25, 0.22\}$$
$$i = \{2, 3, 1, 1, 1, 1\}$$
$$j = \{4, 5, 3, 5, 2, 4\}$$

According to steps 2), 3), diagram $G$ is obtained as shown in Figure 1.

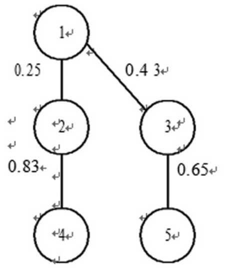

*Figure 1.* Diagram $G$.

According to step 4), construct fuzzy equivalent matrix $R^F$, as follows:

$$R^F = \begin{bmatrix} 1 & 0.25 & 0.43 & 0.25 & 0.43 \\ & 1 & 0.25 & 0.83 & 0.25 \\ & & 1 & 0.25 & 0.65 \\ & & & 1 & 0.25 \\ & & & & 1 \end{bmatrix}$$

Adopt $\lambda$ cutting matrix method for fuzzy clustering analysis select $\lambda=\{0.83, 0.65, 0.43, 0.25\}$, draw dynamic clustering diagram shown in Figure 2.

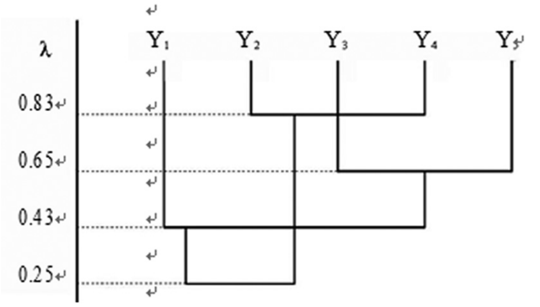

*Figure 2.* Dynamic clustering.

The dynamic clustering figure shows that when $\lambda = 0.83$, the hot information people are concerned about can be divided into four categories $\{1\}, \{2,4\}, \{3\}, \{5\}$.

When $\lambda = 0.65$, the hot information people are concerned about can be divided into three classes: $\{1\}, \{2,4\}, \{3,5\}$.

When $\lambda = 0.43$, the hot information people are concerned about can be divided into two categories $\{2,4\}, \{1,3,5\}$. Government could accord to different needs, the hot information people are concerned about can be classified at different levels of detail.

## 4. Conclusion and Future Work

At present, it is an urgent problem for the applications of e-government how to develop the public preferred patterns accurately, which fertilizes to optimizes government website and improves government services. In this paper, we propose a method for calculating the public concern degree of e-government public opinion, which reveals public intentions accurately, by selection and viewing time situation. Clustering analysis method of e-government information public hot issues based on web log mining is given in this paper. Through cluster analysis of classified hot issues information of public concern is helpful for the government's macro grasp of public services and it provides guidance for the government decision-making, such as canceling service and information that people do not want or are not interested in, providing more targeted services and information timely. Given my limited capacity, this thesis research is still relatively shallow. In fact, public interest information on mining analysis is not studied in this paper only. With the development of the web log mining technology and other network information mining technologies, e-government managers will provide a lot of valuable participants practical reference data. Our future work will add more factors to the public concern degree to calculate user intentions or tendencies more accurately. And, besides, we will develop more efficient clustering algorithms related to public hotspot information.

## 5. Acknowledgments

This research was supported by the Beijing Natural Science Fund Project (Grant No. 9122018).

## References

[1] D. Evans, D. C. Yen, E-government: Evolving relationship of citizens and government, domestic, and international development. *Government Information Quarterly*, **23**(2) (2006), 207–235.

[2] W. Pieterson, W. Ebbers, J. van Dijk, Personalization in the public sector: An inventory of organizational and user obstacles towards personalization of electronic services in the public sector. *Government Information Quarterly*, **24**(1) (2007), 148–164.

[3] X. Huang, Network information mining. *Beijing: Electronic Industry Press*, (2005), 117–133.

[4] X. Chen, Z. Cui, The Design and Implementation of User Clustering Based on Chameleon Algorithm. *Microcomputer Development*, **4**(15) (2005), 48–50.

[5] B. Song, L. Wang, Research on Anonymous Users Clustering Based on Web Log. *Journal of Nanjing University of Science and echnology*, **5**(30) (2006), 583–586.

[6] G. LI, J. LI, Web Log Mining Based on the Fuzzy Clustering. *Journal of Computer Science*, **31**(12) (2004), 130–131.

[7] D. XING, J. SHEN, Efficient data mining for web navigation patterns. *Information and Software Technology*, **46** (2004), 55–63.

[8] M. LIU, Y. HE, Web fuzzy clustering method and its application. *Journal of Computer Science*, **32**(1) (2005), 155–158.

[9] Y. WANG, S. WANG, E-government Websites Evaluation Method Based on Web Log Analysis. *Journal of Intelligence Science*, **29**(10) (2007), 1495–1499.

[10] H. CHUAN, Z. RONGYIN, Measurement about the electronic government affairs information analysis and decision making government affairs under the network environment. *Journal of Theory and Practice of Intelligence*, **26**(5) (2003), 409–411.

[11] G. LIU, Fuzzy cluster analysis in the application of text classification. *Computer Engineering and Application*, **33**(9) (2003), 110–111.

[12] C. DU, G. L. JI, Fuzzy cluster analysis in the application of Chinese text classification research. *Computer Engineering and Application*, **8** (2006), 170–172.

[13] Y. WANG, Y. Y. GUAN, A new judging model of fuzzy cluster optimal dividing. *Fuzzy Systems and Mathematics*, **20**(4) (2006), 79–85.

[14] G. JUNHUA, Clustering analysis in data mining research. *Wuhan University of Technology*, (2003), 17–28.

*Received:* November, 2013
*Revised:* April, 2014
*Accepted:* July, 2014

*Contact addresses:*

Suozhu Wang
School of Management
Capital Normal University
No. 83, Xisanhuan North Road
Haidian District, Beijing
China
e-mail: wsz_wsz@163.com

Jianlin Zhang
School of Management
Capital Normal University
No. 83, Xisanhuan North Road
Haidian District, Beijing
China
e-mail: jlzhang66@263.net

Fuze Yang
School of Management
Capital Normal University
No. 83, Xisanhuan North Road
Haidian District, Beijing
China
e-mail: y1f2z3@163.com

Jia Ye
School of Management
Capital Normal University
No. 83, Xisanhuan North Road
Haidian District, Beijing
China
e-mail: yejia789@163.com

SUOZHU WANG is a professor of information systems at the School of Management, Capital Normal University, Beijing, China. He obtained his BS (Mathematics) from Shanxi Normal University, MS (Applied Mathematics) from the Dalian University of Technology, and Ph.D. (Management Science and Engineering) from the Xi'an Jiaotong University. His research interests are in the areas of data mining, e-government, and e-commerce. He has published about 50 papers in journals and in the proceedings of several international conferences.

JIANLIN ZHANG is an associate professor of information management at the School of Management, Capital Normal University, Beijing, China. He obtained his BS (Mathematics) from Beijing Normal University, MS (Computer Science and Technology) from the University of Science and Technology Beijing. His research interests are in the areas of data mining, mobile services, and churn management. They also include information management, e-Government, and e-Commerce About 30 of his research papers have appeared in journals and in proceedings of several international conferences.

FUZE YANG is a $2^{nd}$ year post-graduate student at the School of Management, Capital Normal University, Beijing, China. She majored in public management theory and information technology. Her research interests include information management and e-Government.

JIA YE is a $2^{nd}$ year post-graduate student at the School of Management, Capital Normal University, Beijing, China. She majored in public management theory and information technology. Her research interests include information management and e-Government.