A session identification method of Web user based on K-means algorithm

Xiao Ping

Department of Network Crime Investigation, Criminal Police College, China

Corresponding author's e-mail: xiaopingcpc@163.com

Received 25 December 2013, www.cmnt.lv

Abstract

The session identification is an important work in the early stage of analysis of behaviour, which has a decisive impact to find out the behaviour characteristics. After analyzing these common session identification methods, we put forward a kind of optimization method to identify Web user session based on K-means algorithm. We compared the method proposed by this paper with other two methods including θ equals thirty minutes and the session identification based on time distance in three aspects: the number of session, the value of absolute evaluation function A(h) and the value of relative evaluation function R(h). It shows that the session identification method proposed by this paper can identify the real user session more completely.

Keywords: Web user session identification; K-means algorithm; Data Pre-processing; Web log mining

1 Introduction

User session is a series of activities carried out by the user from entering to leaving during visiting a site. The user may visit the site repeatedly for very long periods, so the session identification divides the pages accessed by the user into different sessions based on the user identification. The efficiency and accurateness of the session identification determines whether the subsequent Web data mining results are meaningful directly. At present, the web session identification methods can be divided into the following three categories:

a. The session identification method based on time thresholds, giving the user an upper boundθfor staying on the entire site, if the page staying time exceeds the threshold θ then the active page is identified as a new session beginning, θ is set to 30 min generally; or giving the user a page staying time threshold Δt. If the time interval between two consecutive pages is not exceed the threshold Δt, the two pages are the same session. Otherwise, these pages belonging two different sessions, Δt is set to ten min generally. The biggest problem of this method is how to set the value of θ and Δt, if the threshold is too short, then some log records in the same session are divided into different session; otherwise, those belonging to different session entries into the same session [1-3].

b. The session identification is based on content. Spiliopoulou M proposed the method of using user access history and reference pages to divide the session, if a request cannot be entered through the linkage of reference pages, it belongs to another session very likely, that is to say the current request reference page not appeared in the pages accessed is a new session start [4]; According to the access habit of the user, website home page is regarded as the start of a new session. These heuristic methods are susceptible to subjective effects of customer, if the viewer is used to store these pages used commonly in favourite and access the page directly, which will become invalid [5].

c. The session identification method is based on clustering algorithm. Ling Haifeng proposed the optimization method based on Web clustering user identification, the clustering distance is computed by time interval of two successive accessed pages [6]. Nirmala Huidrom proposed that the web session can be identified by grouping the TCP linkages through clustering and the hierarchical agglomerative methods [7]. At present, this kind of algorithm is based on the time distance for clustering only, limited by the time threshold to a great extend.

In this paper, we researched how to optimize the web session identification method, put forward a clustering division method using page access time and URL semantic distance. The basic idea is that the cluster process is adjusted by access time distance and between two continuous pages. The access time distance is longer, the possible that they belong to the same session is smaller, then to optimize the each edge of the clustered group through URL semantic distance. The traditional K-means algorithm has the scalable and high efficient advantages, but the time cost to compute large data sets is large. Web user session identification method in this paper is based on the double standards of the page access time and URL semantic distance to cluster. The experimental results shows that the method improves the convergence speed of K-means algorithm, and can recognize web session more effectively compared to the traditional time threshold or the heuristic method based on contents.

2 Web log data Analysis

Web log data records the user's behaviour to access to the website really. The mainstream web server, such as Apache, IIS etc, has a variety of web log format. In this paper, we used data in the Apache log format, including a record as following:

210.47.131.41- - [10/Oct/2013:13:55:36 -0700] "GET /index.htm HTTP/1.1" 200 2326 "http://www.example.com/start.html" "Mozilla/4.08 [en] (Win98; I ; Nav)".

Every field of the log record is shown in table 1

TABLE 1 Apache log format

Data item	Meaning
210.47.131.41	IP Address
[10/Oct…]	Time and date of request
GET	Request method
/index.htm	Request page
HTTP/1.1	Transmission protocol version
200	Status code returned to client
2326	Number of bytes returned to client
http://www…	Reference page
Mozilla/4…	Browser information of client

3 The Abstract description of Session identification

In order to make the session containing all the information for clustering, we applied a triple to represent the core concept of the session, where URL means the Webpage address, T means the visit time when the server receives the user's request, and B means the number of bytes returned by the server in response to the user's request.

Definition 1 User session is a item sequence of ordered triples: S={ (url_1, t_1, b_1), (url_2, t_2, b_2),... , (url_m, t_m, b_m) }, where m denotes the number of items in a session, (url_i, t_i, b_i) denotes the i^{th} item in a session, urli is the number i *ith* webpage address accessed by users, t_i denotes the time when urli received by server, b_i is the number of bytes returned by the server in response to the user's request urli.

Definition 2 Session identification is to recognize an effective sequence of a user accessing to the Web server. In a session, all pages are related in content, and continuous in access time. We suppose that a user's access logs contain k-sessions, that is $S=\{S_1, S_2,... , S_k\}$; where $S_j = \{(url_1, t_1, b_1), (url_2, t_2, b_2),... , (url_n, t_n, b_n)\}$, $S_j \in S$.

4 An optimization clustering method based on K-means algorithm

The k-means algorithm is a clustering method based on partitioning, the basic idea of which is that randomly choose K initial clustering centroids. According to the similarity of each data item and each clustering centroid, assign the data item to the most similar cluster until the objective function convergence. This algorithm makes the nearest distance of each data item in the same cluster and the fastest distance of each data item in the different cluster. As K-means algorithm has the characteristics of good scalability and high efficiency, etc, it is widely applied to big data set processing.

Ling Haifeng proposed an improved K-means session identification algorithm, the main idea of which is that according to the time distance between log data items to cluster. The process is as following [6]:

(1) Number the data items in a user's web log data set according to time sequence, such $a_1, a_2,\ldots, a_n$, where n is the number of items in data set.

(2) Confirm the time threshold of $\theta=\mu+3\delta$, where μ denotes the user's average time staying in pages.

(3) Confirm the initial clustering centroid and k value. Compute the distance $d(a_i, a_i+1)$ between ai and a_i+1 , which are two adjacent data items, if $d(a_i, a_i+1) \le \theta$, then ai and ai+1 are together into a same cluster, otherwise they are divided into different clusters. Repeat to make the above process until all data items divided into clusters. Compute the number of clusters as k value and the average value of each cluster as the initial clustering centroid itself.

(4) According to the time sequence, renumber the initial clustering centers obtained by the third step, such as $b_1, b_2,\ldots, b_k$, denote the initial clustering center closest to a_1 as b_1, where $b_i \in S_i$ ($i \le k$).

(5) The clustering process is that we compare each data item with the two adjacent cluster centroids in time sequence order. $d(a_i, b_j)$ denotes the euclidean distance from the *ith* data item to the *jth* clustering centroid where $i \le n$, $j \le k$. If $d(a_i, b_j) \ge d(a_i, b_j+1)$ then $a \in S_j+1$, otherwise $a \in S_j$.

(6) Update the clustering centroids. For the new divided clusters, select the average time of each clustering centroid as the new centroid to replace the old one in above step.

(7) Repeat the circle of the fifth and the sixth step until the clustering centroids do not change.

We found that, the algorithm only considers the one-dimensional feature of the time difference between pages, does not take URL semantic feature into account, which can identifies the user session, has many limitations, so this paper analyzes the drawbacks of the original algorithm, proposes an optimized K-means algorithm based on page browsing time and URL semantic distance to identify user sessions.

4.1 THE DEFECTS AND OPTIMIZATION METHOD OF TIME THRESHOLD

The time threshold in the original algorithm is the sum of the average page residence time of each user and the variance, that is $\theta=\mu+3\delta$. The formula of the average page stay time of the user is shown as following:

$$\mu = \frac{\sum_{i=1}^{n} t_{i+1} - t_i}{n-1}, \tag{1}$$

where ti denotes the time the web server received the urli request, ti+1 denotes the time the web server received the urli+1 request, n denotes the number of items in an user's log data set.

Actually, users maybe open multiple pages at one time and turn to browse in the access process, then the calculation method of average residence time on pages can deviate the actual situation largely, so this paper put forward the improvement method to calculate the average time on pages, which is based on the response bytes number in web server log.

When a user in the client sends a page request to the server, the server will generate a number of log records. Because of many pictures, video, animation and other information contained in web pages, it will generate many separate log records in the server, which will be cleared in the log preprocessing stage usually, this paper only consider the text reading time [8]. For a given data item (urli, ti, bi) in an user's log record set. According to the sample survey of "Chinese community newspaper" shows that the general adults read 300 words per minute, The formula of the average page stay time of the user is:

$$\mu b = \frac{\sum_{i=1}^{n} b_i}{n*300*2}, \tag{2}$$

where b_i denotes the number of page bytes response to urli, n denotes the number of items in an user's log data set.

4.2 THE DEFINITION OF URL SEMANTIC DISTANCE

We supposed that all sites use web directory structure, those pages of which can be represented in tree for m well. The page file addressed by each URL corresponds to a leaf node in the site tree [9]. Part of the tree structure of a government test website is shown in figure 1.

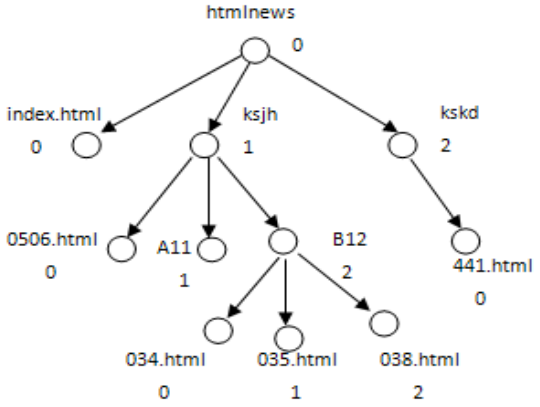

FIGURE 1 Part of the tree structure of a government test website

In order to compare the distances between URL, each node in layers can be marked by different number, so we can use a sequence of digits to represent a URL. Through the comparison of the respective digital sequence, we can find out the same sub data sequence, and then calculate the distance between the two pages according to weight in each layer of data. For example, the page "/htmlnews/ksjh/B12/038.htm" and "/htmlnews/kskd/441.html" can be indicated by "0122" and "020" respectively, the algorithm to calculate the URL distance between the two pages is shown as:

(1) Comparing digital sequence one by one to find the same sub sequence

Repeat to compare each pair of number in two digital sequences one by one until one pair of number is different. The comparison process of two digital sequences is shown in Figure 2; we can see that the two visible digit sequences have one pair of matched digital here.

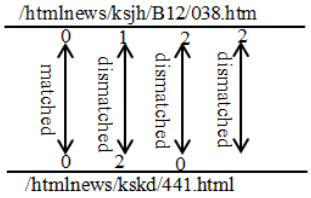

FIGURE 2 Match two digital sequences

(2) To assign weight to nodes, calculate the distance between URL pages

We supposed that the length of the first digital sequence is L_1, the length of the second digital sequence is L_2. Select the longer sequence in L_1 and L_2 firstly, then assign a weight for each layer of the longer digital sequence according the proportion, such as the last layer's weight is 1, the bottom second layer's weights is 2, the bottom third layer's weight is 3, and so on, until assign the given weight L to the first layer. The similarity between two digital sequences can be defined as the result that the sum of those matching digital weights divides the sum of the total weight. Accordingly, the distance between two digital sequences can be defined as the result that one takes away the similarity of the two digital sequences. For the above example, the digit sequence "0122" length is 4, the digit sequence "020" length is 3, so the valve of L is 4, the weight of each layer is shown in figure 3. The URL distance between the two pages is: DURL=1-4/ (4+3+2+1) =0.6

Digit sequence1:	0	1	2	2
Digit sequence1:	0	2	0	
Corresponding weight:	4	3	2	1

FIGURE 3 The weight of each layer

The URL distance DURL is calculated by above method between 0 and 1, if the two pages are same completely, then DURL=0; accordingly if the two page are different completely, then DURL=1.

4.3 DETERMINE THE K VALUE AND THE INITIAL CLUSTERING CENTROIDS

We supposed that a data set of $X=\{x_1, x_2, ..., x_n\}$ is required to cluster, it has n objects, each object denotes a log record. In the web session identification application, the attributes of each log record used to calculate the distance are b, t and url, so if $X_i \in X$, then X_i can be represented by an attribute set, namely $X_i=\{x_ib, x_it, x_{iurl}\}$, the distance between X_i and X_i can be calculated by the Manhattan distance formula.

$$d(x_i,x_j) = (1-\alpha)|x_it-x_jt| + \alpha|x_{iur}-x_{jurl}|, \quad (3)$$

where xit is the time attribute of X_i, x_{iurl} is the digital sequence corresponding to the X_i's URL attribute, x_jt is the time attribute of the X_j, x_{jurl} is the digital sequence corresponding to X_j's URL attribute, and α equals 0.4.

In the K-means algorithm, K denotes the number of clustering, which determines the accuracy of clustering algorithm to the great degree, and the initial cluster centers will also influence the final clustering results greatly. Therefore, how to determine the K value and the initial clustering centers becomes an important problem to be solved in K-means algorithm.

In the traditional K-means algorithm, k value is pre-defined according to experiences of the user. Due to the different characteristics between user's personal interest, the method of selecting k value is added many personal sub-jective factors, it is very subjective and mechanical, will cause the algorithm to cluster not exactly. The improved K-means session identification algorithm proposed by Ling Haifeng is comparing the time distance of the adjacent log with the time threshold θ, if it is less than or equal to θ, then the adjacent log are classified into the same cluster, other-wise they are classified into different clusters. This algo-rithm does not consider the impact of URL semantic dis-tance on the session, so in this paper we proposed a new method to calculate the K value and the initial clustering centers based on two dimensional feature distance of time and the semantic URL. The specific process is as following:

(1) Number the data items in a user's web log data set according to time sequence, such $x_1, x_2, \ldots, x_m$, where m is the number of items in data set.

(2) Calculate the average distance between objects in the log sequence by the double standard of the user average residence time on pages and URL semantic, the formula is as follows:

$$d_{avg} = (1-\alpha)\frac{\sum_{i=1}^{n} X_{ib}}{n*300*2} + a\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n}\left|X_{iur} - X_{jurl}\right|, \quad (4)$$

where X_ib is the property of the number of bytes for Xi, xiurl is the digital sequence corresponding to X_i's URL attribute, xjurl is the digital sequence corresponding to X_j's URL attribute, α equals 0.4, n denotes the number of log records.

(3) Calculate the distance of two adjacent log records as $d(x_i, x_i+1)$, if $d(x_i, x_i+1,)$ is less than or equal to davg, then x_i and x_i+1 are classified into a same cluster; otherwise, they are classified into different clusters.

(4) Repeat the third step until all log records are classified into clusters.

(5) Compute the number of clusters as k value.

(6) In each initial cluster, compute the fitness function f (xi) of each log record, the formula is as follows:

$$f(x_i) = \sum_{j=1}^{m} \frac{1}{d(x_i, x_j) + d_{avg}/m}, \quad (5)$$

where m denotes the number of log records in the current cluster, $d(x_i, x_j)$ denotes the distance between the log record x_i and x_j, davg denotes the average distance of objects within a cluster. In a cluster the object possessing the maximum fitness function value, as the feature point of the cluster, is also the initial clustering centroid.

4.4 THE CLUSTERING PROCESS

The optimized K-means clustering algorithm is shown as following:

(1) Number the data items in a user's web log data set according to time sequence, such as $x_1, x_2, \ldots, x_m$, where m is the number of items in data set. Select the K feature points as the initial cluster centroids according to the step of section 3.3, which will be numbered according to the time sequence, such as $f_1, f_2, \ldots, f_k$. The initial clustering centroid closest to the x_1 is denoted as f_1, and $f_i \in S_i (i \le k)$.

(2) According to the time order of log record, compute the distance of x_i to it's adjacent cluster centroids and to it's own cluser centroid one by one, namely $d(x_i, f_j-1)$, $d(x_i, f_j)$, $d(x_i, f_j+1)$, indicate the distance between the *ith* log record and the *j-1th* clustering centroid, the *jth* clustering centroid and the *j+1th* clustering centroid respectively. If $d(x_i, f_j-1) \le d(x_i,f_j) \cap d(x_i,f_j-1) \le d(x_i, f_j+1)$ then $x_i \in S_j-1$. If $d(x_i, f_j+1) \le d(x_i,f_j) \cap d(x_i,f_j+1) \le d(x_i,f_j-1)$ then $x_i \in S_j+1$ else $x_i \in S_j$.

(3) For the new divided clusters, select the object of each cluster possessing the maximal fitness function value as the new centroid to replace the old one in above step.

(4) Repeat the second step and the third step until every cluster's centroid do not change, that is the objects in clusters not updating.

5 Performance Analysis

5.1 EVALUATION CRITERION

We supposed that the file L is composed of log records sorted by access time, where L[i] is the i^{th} member including url, timestamp, the number of bytes response to the request, host marking information and so on. R denotes the set of real user sessions in L, and C_h denotes the set of user sessions obtained by the session identification method h. Ideally, L can be divided into a set of real session through a session identification method, that is C_h=R. Actually, the evaluation of a session identification method h is to compare C_h and R. The current commonly used evaluation methods mainly have two kinds: one kind is the absolute type according to the number of true session was completely identified to evaluate the estimation method, definition evaluation function A (H) the number of members of a number of / be completely identified real session contained in $=C_h$ real session set R; there are two kind of evaluation method mainly, one is the absolute type according to the number of real session identified completely, to definition evaluation function as following [10]:

$$A(h)=N_i/N_r, \quad (6)$$

where N_i denotes the number of real sessions identified by method h completely, and N_r denotes the number of real sessions in set R.

Another kind is the asymptotic estimation method according the dividing overlap properly, the overlap function is defined as following:

$$deg(R,C)=N_{rc}/N_r, \quad (7)$$

ıhere R is the real session set having n members, C is the divided session set having m members; N_{rc} denotes the number of the same members to R and C, N_r denotes the number of real sessions in set R. If all real sessions are identified completely, then the overlap value equals 1. For evaluating to the method h, select the average of overlap values of R and all session members in C_h, the evaluation function is defined as:

$$R(h)=avg\{deg(r,c)|r \in R, c \in C_h\}. \quad (8)$$

5.2 EXPERIMENTAL RESULTS

All log data in the experiment produced by the laboratory site server during the period from 5 to 15th in 2013.4. In order to reflect the browsing behavior of the user correctly, the site is banned to use cache. Before the experiment these web log data are pre-processed as following:

(1) Data cleaning: remove the request log records including extension of jpg, gif, jpeg, rm, css and the error log records that the value of Status is less than 200 or greater than 299.

(2) Using the IP address to identify each user to access the site, if the IP addresses of these logs are same, but the browser or operating system information of which change, these are belong to different users. After the pre-processing, we got 572 users, took the access logs of IP address 118.248.160.124 randomly to identify 298 real sessions, which includes 1511 valid pages.

TABLE 2 Performance Analysis Result

Algorithm	Number of sessions	A(h)	R(h)
θ=30 min	256	0.52	0.75
Clustering session identification based on time distance	273	0.67	0.82
Clustering session identification based on the two dimensional distance of time and semantics	284	0.79	0.93

This paper selected three algorithms including the traditional identification method based on threshold θ=30 minutes, the web user identification method proposed by Ling Haifeng, and the optimization identification algorithm proposed in this paper based on the dual cluster standard of URL semantics and time threshold. The experiment compared these algorithms from two aspects of absolute and relative, and the result is showed as table 1, from which we can see that the session identification based on cluster is prior to the traditional identification method based on threshold θ=30 minutes in absolute and relative evaluation. The proposed session clustering identification method in this paper is better than the session identification algorithm based on time distance only in both of absolute and asymptotic estimation due to the semantic identification distance added.

6 Conclusion

The optimization method based on K-means algorithm is proposed in this paper, which computes the average time staying on through response bytes in web server log, the URL semantic distance between pages in a tree structure framework, and the distance between objects in log sequence based on the two dimensional feature of time and semantic, can identify the session more accurately. There are some limitations, which could be directions for future working. First of all, the algorithm is applied to the well-structured website, so this work can be extended in order to deal with other types of website which is not well structured. In addition, how to reduce the frequency of clustering, improve work efficiency, is also the direction of the future research for session identification.

Reference

[1] Ishikawah, Ohtam, Yokoyamas 2003 On the effectiveness of Web usage mining for page recommendation and restructuring *Lecture Notes in Computer Science* **2593** 253-67

[17] Dai Zhi Li, Wang Xin Yu 2010 A session identification method based on dynamic threshold of time *Computer applications and software* **27**(2) 244-6

[18] YuanGeng Fang, XueGang Hu 2009 An improved method for session identification in Web log preprocessing *Computer Engineering* **35**(7) 49-51

[19] Spiliopoulou M, Mobasher B, Berendt B, et a1 2003 Framework for the Evaluation of Session Reconstruction Heuristics in Web Usage Analysis *Informs Journal of Computing* **15**(2) 171-9

[20] Chao Li, KunWu Xie, LiMing Wen, Jun Xiang 2011 A combination of session identification and segmentation method for Web theme *Computer applications and software* **28**(6) 167-9

[21] HaiFeng Ling, Gen Yu 2012 An optimized session identification method of Web user based on Clustering *Application Research of computers* **29**(8) 2862-4

[22] Nirmala Huidrom, Neha Bagoria 2013 Clustering Techniques for the Identification of Web User Session *International Journal of Scientific and Research Publications* **3**(1) 1-4

[23] Dong Zhi-feng, Chen Jun-jie, Fu Yu-feng 2008 Research on method for session identification in Web log mining *Computer Engineering and Applications* **44**(8) 179-82

[24] FengChao Li, YanSheng Lu 2007 Similarity Measurement of Web Page Access Based on URL Structure and Access Time *Computer science* **34**(4) 207-9

[25] Shuai Zhang, XingShu Chen, Hao Tong, XiaoJing Cui 2014 Methodology for session identification based on combination of referenced heuristic and URL-semantic *Application Research of computers* **31**(1) 102-5

[26] ZiJun Chen, XinYu Wang, Wei Li 2007 Method of Web Log Sessions Reconstruction *Computer Engineering* **33**(1) 96-7

[27] ZanFu Chen, Qing Liu, MinQiang Li, JiSong Kou 2012 A novel web usage mining method based on web session clustering *Journal of systems engineering* **27**(1) 131-34

[28] Tao Huang, ShengHui Liu, YanNa Tan 2011 Research of Clustering Algorithm Based on K-means *Computer technology and development* **21**(7) 54-7

Authors

Xiao Ping, 1978.05, Jixi City, Heilongjiang Province

Current position, grades: teacher of Computer Crime Investigation Department of China Criminal Police College.
University studies: Computer Application master's degree from North Eastern University School of Information on March 2006.
Scientific interests: information security and Public Sentiment Monitoring.
Publications: papers in the journal of Information Network Security including "Website Construction and Analysis based LAMP platform", "ASP.NET website rebuild and investigation", and so on.