The Improved Clustering Algorithm for Mining User's Preferred Browsing Paths

Xiaojing Li(✉) and Yanzhen Cheng

Jiyuan Vocational and Technical College,
Jiyuan 459000, Henan, China
Lxj55883@163.com

Abstract. The current mining algorithm only consider the user's access frequency, neglecting the interest of users in their visiting path. Compared to the current algorithms for mining user browsing preferred path, clustering algorithm combines the Jacques ratio coefficient and the longest public path coefficient multiplication. This proposed method can estimate the user similarity of page interest and website access structures matrix more accurately for the element value based on the "three tuple" model. Adopting an improved mining algorithm for preference and interest calculation, the bad impact of mining is removed due to pages idle and links. The experimental results showed that the algorithm had higher efficiency and accuracy in web log mining of big data.

Keywords: Clustering algorithm · Data mining · Web logs · Preferred browsing paths

1 Introduction

Web user browsing preference path mining algorithm is used to analyze of web log records and find the user access rules. This algorithm has been successfully applied to personalized web recommendation, system improvement and business intelligence and so on. At present, the most commonly used algorithms in the acquisition of browsing patterns are the maximum frequent sequence method, the reference length method and the tree topology structure method [1]. But these algorithms in fact belong to an improved association rule algorithm and there exist two issues. Firstly, it simply assumed that the frequency of user's browsing is a representation of the user's interest. Secondly, web log data gradually showed distributed, heterogeneous, dynamic and massive properties with the development of network [2], thus the traditional centralized data mining algorithms can't meet the needs of web log mining process with massive data.

In order to solve the problems mentioned above, this paper combines the clustering algorithm and web user browsing pattern mining algorithm, improves the existing algorithm, together with putting the method of multiplying the Jacobi coefficient and longest common path coefficient into consideration to reflect the similarity between users more accurately. This method uses a three tuple to represent the degree of the page interest, considering the user's access time, the size of the page and the number of visits and so on to construct a data matrix which take the URL address of the reference page as the row, the web page URL address as the column and the degree of access

D.-S. Huang et al. (Eds.): ICIC 2016, Part III, LNAI 9773, pp. 329–337, 2016.
DOI: 10.1007/978-3-319-42297-8_31

interest as the element value, and calculate the preference and interest of the matrix with the improved mining algorithm based on this [3]. When to choose to browse the next page, users can get more accurate preference path since the comprehensive consideration of the number of visits, access time and page size.

2 Improved Clustering Algorithm

2.1 The Basic Definition of Clustering Algorithm

Assume n users access path set $U = \{C_1, C_2, \ldots, C_n\}$, one of the access path for $C_i = \{V_1, V_2, \ldots, V_i\}$, where V_i represents a node to be visited.

Definition 1: The number of modes of user access is equal to the length of the path $|c|$.

Definition 2: The Jacobi coefficient

$$s'_{ij} = \left|\frac{c_i \cap c_j}{c_i \cup c_j}\right| \tag{1}$$

For example, there are two web user access path C1 = {V1, V2, V3}, C2 = {V2, V1, V3}, and the calculated results are both 1 using the Jacobi coefficient, but the two paths were obviously not the same. It is because the transaction data which the Jacobi coefficient described does not have precedence relation, but the web user access path is a sequential event. So, it can't be simply described the similarity of access paths by the Jacobi coefficients.

Definition 3: The similarity coefficient of access path refer to s''_{ij}.

Assume $comm(c_i, c_j)$ represents the longest length of common path, $\max(|c_i|, |c_j|)$ represents the longest length of path from c_i to c_j, and the similarity coefficient for user access is:

$$s''_{ij} = \frac{|comm(c_i, c_j)|}{\max(|c_i|, |c_j|)} \tag{2}$$

If there are three web access paths where $C_1 = \{V_1, V_2, V_3\}$, $C_2 = \{V_2, V_3, V_4\}$, and $C_3 = \{V_3, V_2, V_4\}$, the longest length of common path from C_1 to C_2 is V_2 or V_3, the length is 2, and the similarity factor is 0.5. Also the longest length of common path from C_1 to C_3 is V_2 or V_3, the length is 1, and the similarity factor is 0.25. The nodes of the paths C_2 and C_3 are exactly the same with the same order, but the similarity coefficient is only 0.33 using the calculation method and it is lower than the similarity coefficient from C_1 to C_2. This is obviously unreasonable. Thus we propose to improve it as follows:

Definition 4: The similarity coefficient of path from C_i to C_j refer to S_{ij}

$$S_{ij} = \left(S'_{ij}\right)^{\alpha}\left(S''_{ij}\right)^{\beta}, 0 \le \alpha, \beta \le 1 \tag{3}$$

Where α, β refer to the adjusted metric coefficients. The effect of the Jacobi coefficient increases as the increase of the value of α, and the effect of similarity coefficient increases as the increase of value of β, and the effect of sequential increase accordingly.

This similarity coefficient considers the advantages of both the Jacobi coefficient and similarity coefficient. We get the data matrix S of similarity coefficient of user access by using the coefficient to calculate the similarity coefficient of all the paths:

$$S = \begin{bmatrix} S_{11} & S_{12} & \cdots & S_{1n} \\ & S_{22} & \cdots & S_{2n} \\ & & \cdots & \vdots \\ & & & S_{nn} \end{bmatrix}$$

2.2 Improved Clustering Algorithm

In the process of combining the access paths, the maximum similarity coefficient plays a decisive role. The most similarity coefficients less than the threshold value are not effective for clustering [4]. In order to solve the "dimension disaster" problem of the traditional clustering algorithm in the high-dimensional web log data clustering, filtering the smaller similarity coefficient could greatly reduce the data size.

Algorithm 1. output the similar user access path clustering
Enter: The web log file, and set a threshold;
Export: Similar path clustering $C = \{C_i\}$.
Description of algorithm:

```
C = {φ} ; // initialization
While (not to the end of the file)
{
 Read records from a data table;
 While (not to the end of the file)
 {
   Read records from a data table;
   Calculate the similarity coefficient of access path
   S = (S')^α (S'')^β ;
   If (S>θ ) //compare of S and threshold θ
    {Keep the current number of path;
    }
    }
    Get temporary clustering C_i ;
    If(C_i is not a subset of set C)
    {
     Put C_i into set C;
    }
 }
```

Calculate the intersection of each category of membership, and eliminate duplicates according to the degree of membership.

Output to get set C.

3 Improved Preferred Path Mining Algorithm

3.1 Browsing Frequency Preference

If the user has m ways to leave a web page, the user is more interested in the choice with a higher selected number [5].

Definition 1: Assume S_i represents the frequency of user selected into the next page through i option. Based on the traditional definition of confidence and formula 1, regardless of the site structure in the confidence limits on the traditional case, assume the threshold of i option as:

$$C_K = \frac{S_K}{\left(\sum_{i=1}^{n} S_i\right)} \tag{4}$$

Definition 2: On a website, assume all the URL as set U, all the subset as W. If $w \subset W$, $\forall x \in w$ (x represents the page browsing sequence composed by $\forall u \in U$, where the i represents the i browse page), the m numbers before the page browsing sequence are the same, but the m + 1 exists in n different pages, and it represents there are n different browse ways from m to m + 1. So, we assume the j(j = 1, 2, …, n) reference of browse way as:

$$P_j = \frac{S_j}{\left(\sum_{i=1}^{n} S_i\right)/n} \tag{5}$$

Thus, when n > 1, the possibility of i approaches to surfing the internet is considered in the preference coefficient of P in n options. Therefore, it could reflect the user's interest degree more accurately compared to the traditional confidence.

3.2 Browsing Interest Preference

The algorithm of formula 5 only consider the frequency of user browsing, and it is not comprehensive [6, 7]. As the interest of users is related to the size, time and frequency of use access, large page results in a long time, and the long browsing time represents the high interest of browsing. At the same time, the interest degree of user browsing also depends on the number of users access.

Definition 3: Set the interest degree of user browsing as:

$I(URL.time, URL.size, URL.num)$, if we assume the access number of page $URL_i \rightarrow URL_j$ as num, the access time of URL_j as $URL_{i\rightarrow j}.time$, and the size of page URL_j as $URL_j.size$. Then the interest degree of a user is:

$$I_{ij} = Int(\sum_{i=1}^{num} URL_{i\rightarrow j}.time)/URL_j.size) + 1 \tag{6}$$

3.3 Improved Preferred Path Mining Algorithm

Assume there are n URL pages on a website, then construct a matrix of URL-URL, where the row element is URL-Reference, the column element is URL, and the element value is the interest degree of a user from a reference page links to access page. Set an additional null value in row and column. If the user directly enter the web address or access page links from other sites, null appeared in a row. If the user visits the page on the website or exit from the site links to external sites, null appeared in the column [8, 9]. In addition, all reference web page are not themselves, so the diagonal elements of the matrix is 0 (Table 1).

$$\begin{array}{c} \\ Null \\ URL_1 \\ URL_2 \\ \vdots \\ URL_n \end{array} \begin{array}{c} \begin{array}{ccccc} Null & URL_1 & URL_2 & \cdots & URL_n \end{array} \\ \begin{bmatrix} 0 & I_{01} & I_{02} & \cdots & I_{0n} \\ I_{10} & 0 & I_{12} & \cdots & I_{1n} \\ I_{20} & I_{21} & 0 & \cdots & I_{2n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ I_{n0} & I_{n1} & I_{n2} & \cdots & 0 \end{bmatrix} \end{array}$$

Table 1. The process of mining user preferred path in example 1

	Null	A	B	C	D	E	F	G	H	The sum
Null	0	63	15	2	5	0	0	0	5	90
A	6	0	36	35	0	0	0	0	0	77
B	43	0	0	0	6	40	0	0	8	97
C	3	0	0	0	0	10	30	0	0	43
D	16	0	0	0	0	8	0	35	20	79
E	3	0	0	0	0	0	31	0	0	34
F	5	0	0	23	0	0	0	0	0	28
G	4	45	0	0	0	0	0	0	0	49
H	6	0	0	0	0	0	0	0	0	6

Calculation 2: The improved clustering algorithm for mining Web preference path

Input: Assume Web browsing matrix M[n+1][n+1], Sup represents the threshold of browsing support, Pre represents the threshold of browsing preference.

Output: Web preferred browsing path set as NPS.

```
i=0;
while(i<n+1)
{
m=non-zero number of colum;
coun=0;//coun represents the sum of interest expressed
in the row
j=0;
while(j<n+1)
 { if(Sij>0)
   coun+=Supij;
   if((Sij>=Sup)&&( Sij/(coun/m)>=Pre)
   item2= item2+{i,j};
   j++;
 }
i++;
}// merge the same preferred path
x=2;//x represents a set of matrix data item
Flag=0; // detection if X concentration is a merge op-
eration
while(Flag==1)
{
 i=1;
 while(i< path numbers in itemx-1)
 {
  P1= the i preference sub path in itemx-1;
  comb=0;// judge whether to do the merge operation
  j=i+1;
  while(j<= path numbers in itemx-1)
  {
   P2= the j preference sub path in itemx-1;
   if((the (x-2) of before P2)==(the (x-2) after P1))
   {
    Merge the sub path of P1 and P2 to X item of set
    itemx;
    comb=1;
   }
  j++;
 }
```

```
if(comb==0)
Write the preference sub path P1 to set NPS;
Flag=Flag ∪ comb;
i++;
}
x++;
}
```

4 Analysis of Experimental Results

Assume $|URL|$ represents the number of web page, by using the algorithm one can draw that the time complexity of sub path of browsing preference is $O(2(|URL|+1)^2)$. The time to merge with the same paths is $O((|URL|-2)(|URL|+1)^2)$. Thus the total time is $O((|URL|+1)^3)$.

In the process of experiment, 25930 records and 35 pages of web log were experiment objects. The preferred path mining algorithm proposed in this paper and the MFP algorithm in path mining is used to control the threshold setting. In the scenario where the two kinds of mining method were used to explore the same number of preference sub path and frequency browse sub path, the respective accuracy of the

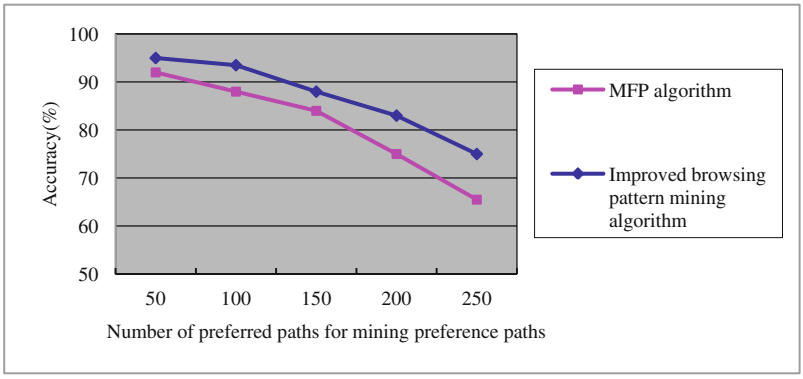

Fig. 1. The accuracy of the algorithm (Color figure online)

proposed algorithm is greater than the known preference path of site access (Fig. 1).

As it can be seen, the improved algorithm mentioned in this paper was more accurate than the MFP algorithm. At the same time, the accuracy of algorithm reduced with the increase of the mining path. It is because that the threshold of mining interest reduced with the increase of the number of paths, which in turn lead to the decrease of credibility of the preferred path. In order to detect the mining time performance of the

two methods, we divided the experimental subject into 5,000 records, 15,000 records, 20,000 records and 25,000 records in the experiment. And Fig. 2 showed the comparison of the execution time. We could find that the improved user access pattern mining algorithm had less execution time increase amplitude and better expansibility

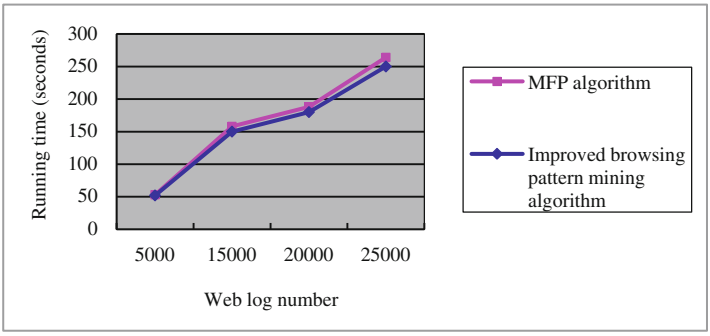

Fig. 2. Time performance comparison of the algorithm (Color figure online)

than the traditional MFP algorithm.

5 Conclusion

This paper puts forward an improved mining method based on clustering web log user preferred browsing paths. First step involves the improvement of the clustering algorithm, duplicates elimination, and the intersection of items to accurately reflect the web user access path similarity. Then, on the basis of a trial model, we explore the preferred browsing paths of multiple pages of a similar user group. Finally, through comparison with other algorithm, the algorithm proposed in this paper has advantages both in accuracy and time performance. Furthermore, it is more comprehensive and accurate data mining algorithm, and has better scalability based on the analysis of different user groups of web browsing preferred path.

References

1. Guan, Y.J., Wang, Y., He, D.N.: A data stream mining approach based on function iterative operation. J. Guangxi Univ. Nationalities **18**(1), 45–49 (2012)
2. Pierrakos, D., Paliouras, G., Papatheodorou, C., Spyropoulos, C.D.: Web usage mining as a tool for personalization: a survey. User Model. User-Adap. Inter. **13**(4), 311–372 (2003). Kluwer Academic Publishers
3. Myra, S., Lukaa, F.: A Data Miner Analyzing the Navigational Behaviour of Web Users, 28 July 2001. http://www.wiwi.hu-berlin.de/~myra/w_acai99.ps.gz

4. Peng, H.Y., Chai, X.G., Chen, X.J.: The clustering algorithm of level iterated theory. J. Tangshan Coll. **24**(3), 86–91 (2011)
5. Caramel, E., Crawford, S., Chen, H.: Browsing in hypertext: a cognitive study. IEEE Trans. Syst. Man Cybern. **22**(5), 865–883 (1992)
6. Salwani, A.: An exponential Monte-Carlo algorithm for feature selection problems. Comput. Ind. Eng. **67**(1), 160–167 (2014)
7. Agrawal, R., Srikant, R.: Mining sequential patterns. In: Proceedings of 11th International Conference Data Engineering, vol. 5, pp. 3–14 (1995)
8. Miao, Y., Song, B.: Research on mining typical anonymous users browsing paths based on web logs. J. Comput. Appl. **29**(10), 2774–2777 (2009)
9. David, D.: Analysis of feature selection stability on high dimension and small sample data. Comput. Stat. Data Anal. **71**, 681–693 (2014)