ATLANTIS PRESS

Advances in Computer Science Research, (ACSR), volume 52

2016 International Conference on Computer Engineering and Information Systems (CEIS-16)

Research of Clustering Model of Network Public Sentiment based on Suffix Array

Liang Hu

Department of humanities and management, JiangXi Police College, NanChang, China

E-mail: huliang_thu@163.com

***Abstract*-According to the characteristics of the network public sentiment in Chinese colleges and universities, a clustering model of network public sentiment is proposed. This paper extracts the keywords from the network text based on suffix array algorithm, combines with the principle of locality and clustering of point position, and then analysis the text of public sentiment that could facilitate the management to understand the network public sentiment focus in the college and university students.**

Keywords-public sentiment; clustering model; suffix array; network

I. INTRODUCTION

With the development of network information technology, a large number of text with emotional color appeared on the university network, the form of presentation is also diverse, such as BBS and Blog etc.[1][2]. Network provides convenience to promote more and more college students choose to use the Internet to express their views, published in the network of public sentiment, and gradually formed a network of public sentiment. The analysis of network public opinion is of great significance for the maintenance of the stability and development of the school [3].

University teachers and students are not only a group of high network utilization of, but also the main generation of network public sentiment [4]. College network public sentiment formation mainly in three aspects: the first aspect is due to domestic and international hot issues related to trigger, the second aspect is the social events inside and outside of online discussion, and the third aspect is something closely related to the interests of college teachers and students online demands [5].

Network text has relatively high dimension and included noise data, there are a lot of redundancy. Therefore, the text data to and through the method of clustering extract keywords, becomes the key work of user groups public sentiment analysis. This paper proposes a clustering model for network public sentiment, through the suffix array to find the key phrases, and then use the data mining method to analyze the relationship between key phrases, and then the network public sentiment data clustering.

II. SUFFIX ARRAY CONSTRUCTION

Suffix tree and suffix array is a common technique to deal with string. But suffix array is optimized by suffix tree, it is easy to program, in the completion of the suffix number can be completed under the functional conditions of time complexity and space complexity are smaller. Suffix array technology to achieve the longest common sub sequence, you can achieve multiple text strings of the longest common sub sequence, the output in a number of short text in the presence of a number of times greater than half of the longest common sequence[6][7][8].

Definition 1 Substring: string R [m...N], m<n, S, N, from the R position to the end of the M string to the end of this string of a string.

Definition 2 Suffix: string S suffix refers to a location from the m S to the end of the beginning of the end of the end of the string s string, the suffix S Suffix (m).

Definition 3 Frequent pattern discovery: compute the frequency of the sub string in the network text is greater than the specified threshold.

Definition 4 Suffix array: given a consisting of n words of a text T, which corresponds to the suffix array substring is a element values in the 1 to n between the integer array, which represents the position suffix[i] suffix is a sequence of characters in the suffix i.

Definition 5 Longest common prefix array (Longest Common Prefix, LCP): for a text T and its corresponding suffix array substring , substring corresponds to the longest common prefix array LCP is an array of integers, which LCP[0] is always 0. LCP[i] is between suffix[i] corresponding to the position of suffix [i-1] corresponding to the position of the longest common prefix length.

According to the nature of the prefix comparison we can get the following very important properties:

Properties 1: For any constant k=n, Suffix (i) <k*Suffix (J) is equivalent to all the i and j, both Suffix (i) <Suffix (j) was established.

Properties 2: Suffix(i)=2*k*Suffix(j) is equivalent to Suffix(i)=k*Suffix(j), and Suffix(i+k)=k*Suffix(j+k).

Properties 3: Suffix(i)<2*k*Suffix(j) is equivalent to Suffix(i)<k*Suffix(j), or Suffix(i)=k*Suffix(J) and Suffix(i+k)<k*Suffix(j+k).

The algorithm is the main idea of the use of the method of doubling the length of each character to start the length of the 2K to sort the sub string, ranking, that is, rank value. K starting from 0, each plus 1, when the 2K is greater than N, the length of each character starting with the length of the 2K sub string will be equivalent to all suffixes. And these sub strings must have been compared to the size, that is, rank value does not have the same value, then the rank value is the final result. Every sort is the last string of length 2k-1 of the rank value, then the string of length 2K can use two string of length 2k-1 ranked as the key, and then sort, I have come to the rank 2K the length of the string value. Algorithm is as follows:

```
int wa[maxn],wb[maxn],wv[maxn],ws[maxn];
  int cmp(int *r,int a,int b,int l)
  {return r[a]==r[b]&&r[a+l]==r[b+l];}
  void da(int *r,int *sa,int n,int m)
  {
    int i,j,p,*x=wa,*y=wb,*t;
    for(i=0;i<m;i++) ws[i]=0;
    for(i=0;i<n;i++) ws[x[i]=r[i]]++;
    for(i=1;i<m;i++) ws[i]+=ws[i-1];
    for(i=n-1;i>=0;i--) sa[--ws[x[i]]]=i;
    for(j=1,p=1;p<n;j*=2,m=p)
    {
      for(p=0,i=n-j;i<n;i++) y[p++]=i;
      for(i=0;i<n;i++) if(sa[i]>=j) y[p++]=sa[i]-j;
      for(i=0;i<n;i++) wv[i]=x[y[i]];
      for(i=0;i<m;i++) ws[i]=0;
      for(i=0;i<n;i++) ws[wv[i]]++;
      for(i=1;i<m;i++) ws[i]+=ws[i-1];
      for(i=n-1;i>=0;i--) sa[--ws[wv[i]]]=y[i];
      for(t=x,x=y,y=t,p=1,x[sa[0]]=0,i=1;i<n;i++)
        x[sa[i]]=cmp(y,sa[i-1],sa[i],j)?p-1:p++;
    }
    return;
  }
```

At the end of the text to add a character tag, with a depth first traversal of the suffix tree group, get the sort of suffix array in accordance with the dictionary order. The suffix array is constructed as shown in Table 1.

III. KEYWORDS VALUE INDEX

The value of the string is related to the frequency of the frequency, the greater the frequency of the string is the greater the possibility of keywords. If the string n appears in the text S times, each of which appears as P1, P2, P3, ... , Pn, the algorithm of character string is completed by the following procedure:

Step 1. Compute the reference distance. This paper use a string in the text of the distance between the adjacent positions as a reference point, then the string S reference distance is computed as the formula:

$$RD = \frac{\sum (P_i - P_{i-1})}{n-1} \qquad (1)$$

Step 2. Cluster keywords. The current clustering C={P1}, for 1 <i<n+1, cycle compute Pi-Pi-1, if Pi-Pi-1>RD, will C join R, empty C, otherwise the Pi to join C.

Step 3. Compute keywords value index. R={S1, S2, S3, ... , Sn} is obtained by using the frequent pattern algorithm of suffix array, then compute the keyword value index by the following formula:

$$KVI = k \times \frac{\sum (P_i - \frac{\sum P_i}{n})^2}{N-1} \qquad (2)$$

Where k is the index influencing factor, which is set up in order to influence the value of KVI, which can be used to analyze the influence of KVI on the, find the best value, and improve the accuracy of clustering analysis.

IV. PERFORMANCE TEST AND ANALYSIS

In the test, this paper compute the frequency of the string that is greater than 100, and the value of k is varied from 0 to 1, which the step length is 0.05, computing the string at KVI and sorting, were selected for ranking of the 200, 400, 800 string. The test results as shown in Fig.1.

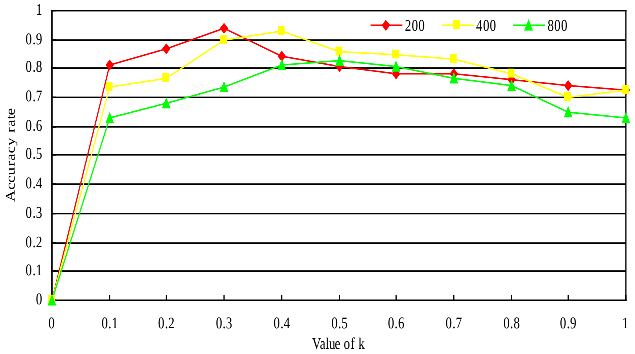

Figure 1. Relationship between index k and accuracy rate.

The best values of k increase with the increase of the number of strings. When a number of the string is 200, the best value of k is 0.3. When the number of the string is 400, the best value of k is 0.4. When the number of the string is 800, the best value of k is 0.5.

After determining the corresponding K value of the different number of strings, the network text was clustered. The test results as shown in Table 2.

V. DISCUSSION AND FUTURE WORK

This paper proposed a network public sentiment information clustering algorithm based on suffix array, which is simple, high efficiency and improve the accuracy of public sentiment information analysis, but the algorithm may be need to consider the number of the corresponding threshold, in order to improve the implementation efficiency of the algorithm.

ACKNOWLEDGMENT

This author’s work is supported by JiangXi Research on teaching reform of higher education (JXJG-15-19-3), JiangXi Science and technology research project of Education Department (GJJ151193), JiangXi Social Science Planning Projects during the 12th Five-Year Plan (14TQ05) and JiangXi Police College Scientific Research Project (2014QN001).

REFERENCES

[1] Zhou Lizhu. An overview of the research on emotional analysis [J]. Computer Application Research, 2008.28 (11): 2726 -2727. (in Chinese)

[2] Zeng R. Internet public opinion information resource sharing research [J]. Intelligence Journal, 2009(8): 187-191. (in Chinese)

[3] Li Junjun. Research on network public sentiment in Chinese universities [J]. Journal of Guangxi Normal University for Nationalities, 2014, 1:126-129. (in Chinese)

[4] Wang L. Introduction to public sentiment research [M]. Tianjin: Tianjin College of Social Sciences Press, 2003: 5-8. (in Chinese)

[5] Li Wenjing. Design and analysis of network public opinion index system [J]. Information Science, 2009 (7): 986-991. (in Chinese)

[6] Pan Daqing. Design of public opinion monitoring system based on data mining [J]. Public Science and Technology, 2014,16 (11): 1-2. (in Chinese)

[7] Zhang Wei. Research and implementation of network public sentiment system [D]. Tianjin: Tianjin University, 2011. (in Chinese)

[8] Tang Tao. Study on monitoring and research of network public opinion based on Information Science [J]. Information Science, 2014, (1): 45-47. (in Chinese).

TABLE I. TABLE CLUSTERING OF NETWORK PUBLIC SENTIMENT

Keyword	**Frequency**	**k**	**Accuracy rate**
university	1925	0.5	0.823
employment	1500	0.4	0.805
CET4	1432	0.4	0.882
dormitory	1325	0.3	0.834
graduation	1215	0.3	0.903
score	1035	0.3	0.915

TABLE II. TABLE SUFFIX ARRAY CONSTRUCTION

Element of suffix array	**Value**
Suffix[2]	axueshengbiyerenshu
Suffix[11]	biyerenshu
Suffix[1]	daxueshengbiyerenshu
Suffix[8]	engbiyerenshu
Suffix[16]	enshu
Suffix[14]	erenshu
Suffix[5]	eshengbiyerenshu
Suffix[10]	gbiyerenshu
Suffix[7]	hengbiyerenshu
Suffix[19]	hu
Suffix[12]	iyerenshu
Suffix[9]	ngbiyerenshu
Suffix[17]	nshu
Suffix[15]	renshu
Suffix[6]	shengbiyerenshu
Suffix[18]	shu
Suffix[20]	u
Suffix[4]	ueshengbiyerenshu
Suffix[3]	xueshengbiyerenshu
Suffix[13]	yerenshu