International Journal of Database Theory and Application
Vol.9, No.7 (2016), pp.83-94
http://dx.doi.org/10.14257/ijdta.2016.9.7.08

Internet Public Opinion Hot Spot Mining Based on Voting Mechanism and MapReduce Platform

Wei Ling and Gao Chang-Yan

Management Institute, Harbin University of Science and Technology, China
245460112@qq.com

Abstract

In the era of big data, the value of Internet public opinion hot spot extraction is particularly prominent. In order to further develop the application value of Internet public opinion hot extraction, this paper constructs the Internet public opinion hot spot extraction framework, and puts forward an Internet public opinion hot spot extraction method VM-Rep based on voting mechanism and MapReduce framework. A heuristic algorithm based on voting mechanism is proposed for the hot spot of Internet public opinion, and MapReduce is used to improve the ability and efficiency of processing massive data. Experimental results show: VM-Rep's coverage is significantly better than Top-k, K-means and Agglo, and the redundancy of VM-Rep is the least; VM-Rep takes the least time in the four methods, embodies the advantages of VM-Rep method for massive data.

Keywords*: Public opinion hot spot, Internet public opinion, Big data, Voting mechanism, MapReduce platform*

1. Introduction

Public opinion refers to social groups within a certain period and scope of the subjective reflection of the social phenomena and the reality, the public opinion organized and purposeful collection and processing to form public opinion information [1]. Public opinion information is the carrier and reflect the form of public opinion, the obtain of public opinion hot spot can help people understand the social hot spot in time, which also can help the enterprises to timely grasp the latest development in related areas, and it also help the government to quickly understand the social important events and the public opinion direction. With the rapid development of social network and mobile intelligent terminal, the Internet become the focus and spread most open "places" in public opinion. Internet public opinion information mainly distributed in government news sites, large commercial websites, news media websites, representative local websites and BBS, online community, *etc.* [2]. The development of the Internet makes the people express their views and understanding more and more convenient, public opinion information showing explosive growth. According to a CISCO report, which shows that the global Internet newly increased 4.1ZB data volume in 2014, the data volume is expected to 2020 will reach an order of 40ZB magnitude .The mass data has brought a certain amount of distress to the gain of public opinion, then, how to get public opinion from the mass public opinion information has become the focus of the research. Internet public opinion hot spot extraction is an important means to grasp public opinion, and it's also the important part of the Internet public opinion information mining.

The obtain of internet public opinion hot spot is the core of realization of public opinion management work and foundation. Some scholars have applied the clustering method to the internet public opinion hot spot extraction aspect. For example, Shouhua Zhang (2013) has used the hot keywords to hot cluster, and realized the public opinion hot spot monitoring [3]. Chenqing Han (2013) has used the improved similarly calculation

ISSN: 2005-4270 IJDTA

formula and the extraction method of feature words, who used the clustering algorithm based on density is found that public opinion hot spot [4].Wei Han (2012) through the improved Single-Pass clustering algorithm to overcome the defects of the algorithm is sensitive to text input sequence, and applied to the network public opinion hot spot found [5].Based on the clustering method's internet public opinion mining mainly focus on the organization in public opinion information ,which unable to reflect effectively directly the Internet public opinion hot spot. And some scholars have applied the Large Data related technologies to the Internet public opinion hot extraction aspect. For example, Yan Ma (2014) has designed micro-blog public opinion hot spot mining system structure model under the large data environment, and used tools such as ICTCLAS and AntConc to extract hot words [6]. Jianhua Zhou (2014) has put forward a kind method of network public opinion mining and analysis based on Hadoop, which can help solve the distress in mass data processing [7]. Min Huang (2011) has proposed a extraction method based on complex networks in Internet public opinion hot spot, combined with the PageRank and Hits algorithm [8].

Throughout most of existing research results, the Internet public opinion hot spot discovery and analysis is still in the exploratory stage. This article from the nature of public opinion hot spot(that is, the public opinion information can reflect most of the content of the typical information), and proposes based on voting mechanism and MapReduce framework of Internet hot public opinion mining method , VM - Rep method. Voting mechanism has applied more extensive to all kinds of algorithm, to the representative points selection algorithm [9], which is applied to clustering algorithm [10], and which also applied to the recommendation system [11]. This article introduce the voting mechanism to extract the Internet public opinion hot spot's innovation, to improve the coverage of the Internet public opinion hot spot, and reduce the redundancy. The design of MapReduce framework has used the divide-conquer method, it's a simple but powerful parallel and distributed computing architecture [12]. This article has combined it with voting mechanism for large data of Internet public opinion hot extraction, reduced the Internet public opinion hot spot calculation time, and improved improve the ability to deal with large data the method.

2. Problem Description

Internet public opinion hot spot refers to the representative information in Internet public opinion, and it can reflect the most common and views of people's understanding. Internet public opinion hot spot is required for extraction of maximum coverage the original internet public opinion information, at the same time, to ensure the information sets itself redundant small as far as possible. Internet public opinion information set D and Internet public opinion hotspot R have the relationship, such as, the R has covered the original information set D, that is, R has overall maximum similarity degree to D; the information set R has minimum information redundancy, that is, the similarity between each other is small enough in R's information. Extraction process of internet public opinion hot spot can be described as follows:

$$\text{Finding } R$$

$$\text{t.}\begin{cases} R \subseteq D \\ \max(data\_coverage(R, D)) \\ \text{// public opinion information set maximum coverage} \\ \min(data\_redundancy(R)) \\ \text{// Public opinion hot spot set has minimum redundancy information itself} \end{cases} \tag{1}$$

Assuming that the original public opinion information sets $D=\{d_1,d_2,...d_n\}$, from the original public opinion information sets D to find the public opinion hot spot sets $R=\{d_{r1},d_{r2},...,d_{rk}\}$, moreover, the sets R satisfies the following conditions:

$$\min |R|$$
$$\text{s.t.}\begin{cases} D_{r1}^{\lambda} \cup D_{r2}^{\lambda} \cup \ldots D_{rk}^{\lambda} = D \;//\; \text{Coverage constraint} \\ \min\left(\dfrac{\sum_{i=1}^{k-1}\left(\sum_{j=i+1}^{k} sim(d_{ri}, d_{rj})\right)}{|R| \times (|R|-1)}\right) \;//\; \text{Redundant constraint} \end{cases} \quad (2)$$

The λ respects the coefficients in it, sim(d_i, d_j) refers to the similarity between d_i and d_j, and the D_i^{λ} refers to the information d_i which can be instead of the collection of information, $D_i^{\lambda} = \{d_k \mid sim(d_i, d_k) \geq \lambda, d_k \in D\}$.

3. Internet Public Opinion Hot Spot Extraction Framework

According to the characteristics of Internet public opinion information and hot spot to extract demands. This article constructed by public opinion information collection, pretreatment, processing, and public opinion hot spot show VM-Rep of Internet public opinion hot spot extraction framework. As shown in Figure 1.

(1) Internet public opinion information collection

Internet public opinion information collection mainly from micro-log, search results, social networking sites, online comments, and blogs, *etc*. Through the web URL way of collection to collect the public opinion information about government economic work and key work, major social events and incidents, and so on. The process of Internet public opinion URL information acquisition as shown in Figure 2.

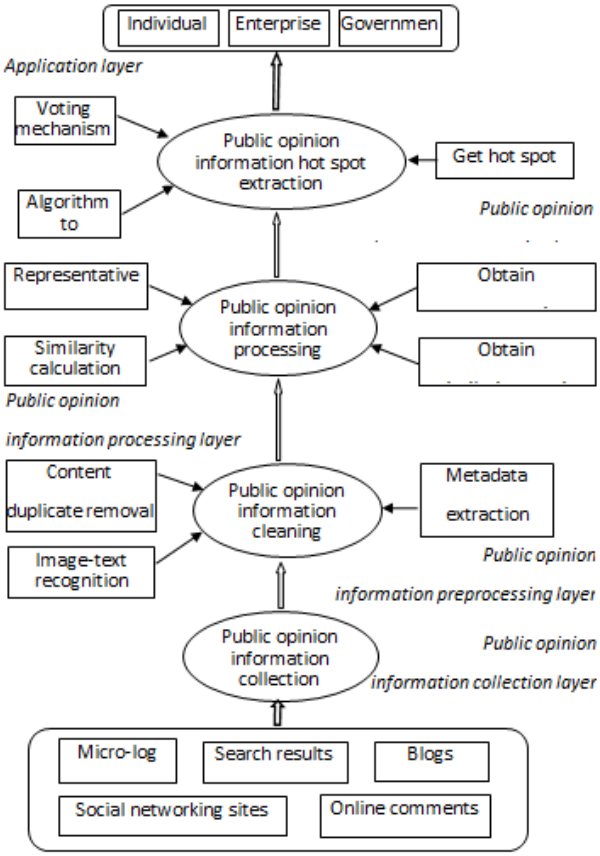

Figure 1. VM-Rep of Internet Public Opinion Hotspot Extraction Framework

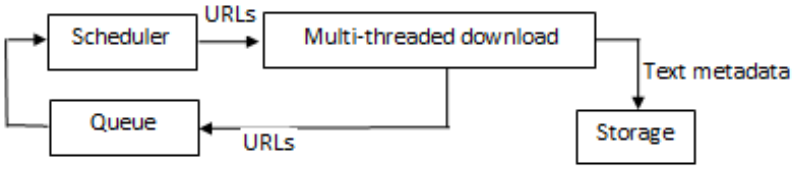

Figure 2. The Process of Internet Public Opinion Information Collection

(2) Internet Public opinion information preprocessing

Internet public opinion information preprocessing mainly to clean the original Web, and extract the metadata. Filter out the public opinion of the content of web, the word's count is less than a certain threshold of public opinion information, then, transform the traditional Chinese characters into simplified Chinese, and let the pictures on the web to directional escape.

(3) Internet Public opinion information processing

To extract metadata information mapping to the multidimensional space, the metadata d mapping as following:

$$V(d) = (feature_1, w_1; feature_2, w_2; ...; feature_n, w_n) \quad (3)$$

Calculating similarity between metadata each other, to get the similarity matrix. Calculate the similarity of metadata d_i and d_j:

$$sim(d_i, d_j) = \cos\theta = \frac{\sum_{k=1}^{n} w_k(d_i) \times w_k(d_j)}{\sqrt{(\sum_{k=1}^{n} w_k^2(d_i))(\sum_{k=1}^{n} w_k^2(d_j))}} \quad (4)$$

According the coefficient λ to nature similarity matrix represents of 0, 1, and gain the representative set of metadata. If $sim(d_i,d_j) \geq \lambda$, the d_i can represent the d_j on the degree of λ, and denotes it as $rep_\lambda(d_i,d_j)=1$. In other words, if $sim(d_i,d_j)<\lambda$, the d_i unable to represent the d_j on the degree of λ, and denotes it as $rep_\lambda(d_i,d_j)=0$.

(4) Internet public opinion information hot spot extraction

According to the representative set, using the voting mechanism to calculate every representative set of the vote, and find with the most votes belonging metadata to join the hot spot sets. Uses the Map-Reduce module to parallel computing, and improve the efficiency of algorithm implementation. The specific process detailed in Section 4.

4. Based on Voting Mechanism of Internet Public Opinion Hot Spot Extraction Algorithm

4.1. Based on Voting Mechanism of Internet Public Opinion Hot Spot Extraction Steps

This article presents a heuristic method which is based on the voting mechanism to seek public opinion information hot spot, each round through Internet public opinion information set D's all the metadata for the data corresponding to represent "vote" set, and the highest votes join the Internet public opinion hot spot in R, by virtue of the voting basic idea. Based on voting mechanism of internet public opinion hot spot extraction specifics steps as following:

Step 1: To calculate representational set. According the original public opinion information set $D=\{d_1,d_2,...,d_n\}$to calculate the similarity matrix M, and according to represent coefficient λ, if the similarity≥λ,then round up to 1, otherwise, round up to 0, and get the representational matrix M. According to the representative matrix M_λ to get all the representative sets λ-D_i^λ.

Step 2: To calculate representational set votes. Define the initial value of Each metadata in the D is 1. If the d_j exists in D_i^{λ}(that is, $rep_{\lambda}(d_i,d_j)$=1), the d_j has voting qualification for D_i^{λ}. Then, records the d_j emerges in the representative sets' numbers as n_j, the d_j has appeared to cast votes for each representative sets is $\mathrm{vote}_j^i = \frac{1}{n_j}$. Finally, to calculate each round of votes for the representative set D_i^{λ}, the voting results is both of all d_j's votes as following:

$$\mathrm{vote}_i = \sum_{d_j \in D} \mathrm{vote}_j^i \times \mathrm{rep}_{\lambda}(d_i, d_j) = \sum_{d_j \in D} \mathrm{rep}_{\lambda}(d_i, d_j) / n_j \quad (5)$$

Step 3: Redundancy optimization. According to each representative sets' votes, to find the votes during $[\alpha \cdot \max(\mathrm{vote}_j^i), \max(\mathrm{vote}_j^i)]$, just as the top several representative sets to select d_j that makes the redundancy of R smaller to join in R. The d_j should satisfy the following conditions:

$$\begin{cases} \mathrm{vote}_j \in [\alpha \cdot \max(\mathrm{vote}_j^i), \max(\mathrm{vote}_j^i)] \\ \min\{\frac{1}{|R|} \times \sum_{d \in R} \mathrm{sim}(d_j, d)\} \end{cases} \quad (6)$$

Step 4: To wipe out d_j and it's representative data in R, get a new D and new representative set D_i^{λ}. Repeat steps 2 and 3 until there is no need adds data to R.

4.2. Based on Voting Mechanism of Internet Public Opinion Hot Spot Extraction Pseudo Code

Input: Data set $D = \{d_1, d_2, \cdots d_n\}$; representative coefficient λ; parameter $\alpha \in [0,1]$
Output: Representative data set R
$Setup()$
$R = \phi$;//Declare the representative data set
$M[\,][\,] = Compute\_Similarity(D)$; // Calculate similarity between the metadata
// to obtain the matrix M
$M_{\lambda}[\,][\,] = \mathrm{Compute\_Represent}(M)$; // Treat the matrix M in 0 and 1
While $D \neq \phi$ do{
$avgsim = 1$;//Define the initial redundancy as 1
$D^{\lambda}[\,] = \mathrm{Compute\_Represent\_sets}(M_{\lambda})$;//Calculate the representative set
$\mathrm{Vote}[\,] = \mathrm{Compute\_vote\_value}(M_{\lambda}, D_{\lambda})$;//Calculate each representative sets votes
$Max\_vote\_value = Find\_\max(vote)$;// Get the maximum votes' number
For d_i in D {
if($vote[i] \in [\alpha \times Max\_vote\_value, Max\_vote\_value]$){
if($Compute\_avg\_sim(d_i, R) < avgsim$){ //Redundancy optimization
$avgsim = Compute\_avg\_sim(d_i, R)$;
$d_r = d_i$;}}}
$R = R + d_r$; //Add the qualified data to R
$D = D - D^{\lambda}[d_r]$; //To wipe out d_j and it's representative data in R
$Output(R)$;
End;

4.3. Algorithm Accelerate to Strategy

(1)MapReduce Framework

MapReduce is a programming model proposed by Google, whose main juche idea is" Divide and rule". Just as first part, after the whole [13]. The MapReduce module's essence can be summarized as the system divides mass data into small chunks, the Map function parallel processing each small data sets, and gives to the Reduce function to statute. The Map function structures the input line data into a key/value pair <key, value>, and order by the key. The Reduce function processing phase is merges the key value of the same key values, and input the key-value into the Hadoop to parallel processing. Thus, to further increase the speed of algorithm [14]. The MapReduce execute process shown as Figure 3.

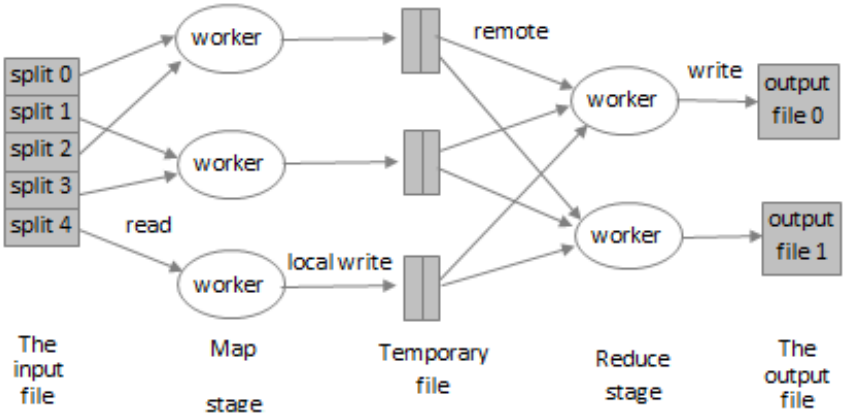

Figure 3. MapReduce Execute Processing

(2) Introduce the framework of MapReduce

This article introduce the framework of MapReduce at which is based on voting mechanism of internet public opinion hot spot extraction step 2, and greatly improve the processing ability and efficiency of huge amounts of data. It has Gotten a set of data similarity matrix as following:

$$\mathbf{M}_{\lambda} = \begin{array}{c} \\ d_1 \\ d_2 \\ d_3 \\ d_4 \\ d_5 \\ d_6 \\ d_7 \\ d_8 \end{array} \begin{array}{c} \begin{array}{cccccccc} d_1 & d_2 & d_3 & d_4 & d_5 & d_6 & d_7 & d_8 \end{array} \\ \begin{bmatrix} 1.00 & 1.00 & 0.00 & 1.00 & 0.00 & 0.00 & 1.00 & 0.00 \\ 1.00 & 1.00 & 0.00 & 1.00 & 1.00 & 0.00 & 1.00 & 0.00 \\ 0.00 & 0.00 & 1.00 & 0.00 & 1.00 & 1.00 & 0.00 & 0.00 \\ 1.00 & 1.00 & 0.00 & 1.00 & 0.00 & 0.00 & 1.00 & 0.00 \\ 0.00 & 1.00 & 1.00 & 0.00 & 1.00 & 1.00 & 0.00 & 1.00 \\ 0.00 & 0.00 & 1.00 & 0.00 & 1.00 & 1.00 & 0.00 & 1.00 \\ 1.00 & 1.00 & 0.00 & 1.00 & 0.00 & 0.00 & 1.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 1.00 & 1.00 & 0.00 & 1.00 \end{bmatrix} \end{array}$$

According to above that it also gets the representative sets shown as Table 1, and to explain the procedure of calculating public opinion hot spot on MapReduce framework.

Table 1. λ Representative Sets

λ Representative sets(D_i^{λ})	The data object
D_1^{λ}	$\{d_1, d_2, d_4, d_7\}$
D_2^{λ}	$\{d_1, d_2, d_4, d_5, d_7\}$
D_3^{λ}	$\{d_3, d_5, d_6\}$
D_4^{λ}	$\{d_1, d_2, d_4, d_7\}$
D_5^{λ}	$\{d_2, d_3, d_5, d_6, d_8\}$
D_6^{λ}	$\{d_3, d_5, d_6, d_8\}$
D_7^{λ}	$\{d_1, d_2, d_4, d_7\}$
D_8^{λ}	$\{d_5, d_6, d_8\}$

1)Preprocessing According to the characteristics of the MapReduce, which needs to construct the key/value pair <key, value>. On the basis of the representative sets shown in the Table 1, and sets the representative data d_i of D_i^{λ} as Key value. However, the Value is made up of by representative data. For example, the d_i could represent data such as d_1, d_2,d_4 and d_7 in representative sets D_1^{λ}, the Key to store d_i and the Value stores (d_1, d_2, d_4, d_7). It will get the following Table 2 by MapReduce preprocesses the Table 1:

Table 2. The Preprocessing Stage Output Format

Input	Key value	Value value
Map1	d_1	(d_1, d_2, d_4, d_7)
Map2	d_2	$(d_1, d_2, d_4, d_5, d_7)$
Map3	d_3	(d_3, d_5, d_6)
Map4	d_4	(d_1, d_2, d_4, d_7)
Map5	d_5	$(d_2, d_3, d_5, d_6, d_8)$
Map6	d_6	(d_3, d_5, d_6, d_8)
Map7	d_7	(d_1, d_2, d_4, d_7)
Map8	d_8	(d_5, d_6, d_8)

2) Map Procedure As mentioned in the sample table needs eight Map to complete follow-up processing. The map procedure needs to load drivers and initialized, leading to consume a large amount of resources, so, decrease the number of calls for the Map procedure can reduce the consumption of system resources. Here, this article put a large amount of complicated data calculation into Map, the Reduce just to do simple statistical work. Then, it will improve system's calculation efficiency. According to this thought, separate dispose each Key - Value pairs on Map, calculate the number of each representative sets as n, and the data node voting Value is 1 / n. The disposed results as shown in Table 3. For instance, the $d_i \rightarrow d_1,d_2,d_4,d_7$ through the Map procedure gets several new Key-Value pairs $d_1 \rightarrow 1/4$、$d_2 \rightarrow 1/4$、$d_4 \rightarrow 1/4$ and $d_7 \rightarrow 1/4$.

Table 3. The Map Output

Input	Output
Map1	$d_1 \rightarrow 1/4$; $d_2 \rightarrow 1/4$; $d_4 \rightarrow 1/4$; $d_7 \rightarrow 1/4$
Map2	$d_1 \rightarrow 1/5$; $d_2 \rightarrow 1/5$; $d_4 \rightarrow 1/5$; $d_5 \rightarrow 1/5$; $d_7 \rightarrow 1/5$
Map3	$d_3 \rightarrow 1/3$; $d_5 \rightarrow 1/3$; $d_6 \rightarrow 1/3$
Map4	$d_1 \rightarrow 1/4$; $d_2 \rightarrow 1/4$; $d_4 \rightarrow 1/4$; $d_7 \rightarrow 1/4$
Map5	$d_2 \rightarrow 1/5$; $d_3 \rightarrow 1/5$; $d_5 \rightarrow 1/5$; $d_6 \rightarrow 1/5$; $d_8 \rightarrow 1/5$
Map6	$d_3 \rightarrow 1/4$; $d_5 \rightarrow 1/4$; $d_6 \rightarrow 1/4$; $d_8 \rightarrow 1/4$
Map7	$d_1 \rightarrow 1/4$; $d_2 \rightarrow 1/4$; $d_4 \rightarrow 1/4$; $d_7 \rightarrow 1/4$
Map8	$d_5 \rightarrow 1/3$; $d_6 \rightarrow 1/3$; $d_8 \rightarrow 1/3$

3)Reduce Procedure After the Reduce port access every Map output key/value pair<key, value>, accumulate the Value of the same Key. And calculate the final votes. For example, calculate the representative set $D5^{\lambda}$ and accumulate Key is d5' s key/value, {d5→1/5, d5→1/3, d5→1/5, d5→1/4, d5→1/3}. So the d5' s votes as shown following:

$$vote_5 = \frac{1}{5} + \frac{1}{3} + \frac{1}{5} + \frac{1}{4} + \frac{1}{3} = 1.32 \quad (7)$$

Calculate the representative set D_1^{λ} to D_8^{λ} votes respectively shown as :0.95, 1.15, 0.78, 0.95, 1.32, 1.12, 0.95 and 0.78. Take the largest vote representative D_5^{λ} 's d_5 to join into R, the representative data of first join into R don't need to consider the redundancy. Then, remove d_5 which joined into R and its representative data d_2, d_3, d_6 and d_8, and get the new data set D={d_1,d_4,d_7}. Repeat the steps 2 and 3 until $D = \phi$.

5. Experimental Analysis

5.1. Experimental Environment

Set up a Hadoop cluster under the Linux environment, the Hadoop platform version is 2.2. This experiment run a total of 15 servers of cloud computing environment, the one as the main nodes, and the others as a slave node. Each node of the processor is Intel (Intel) 22 nm Core i5 quad-core processors, the memory is 4 GB. Operating systems are used Ubuntu12.04, JDK version is Sun JDK 1.7. This article choose sohu, tencent, netease, sina, *etc.* of these unofficial and free contains the netizen comments and views of the 30 sites as the network public opinion information source. Because these web have both HTML and XML format, and unified conform the semi-structured HTML document into XML text form of description to feature extraction. After that processing features items deposited in the document library, and respectively marked as U1,U2,...,U15. This article sets based on voting mechanism of the Internet public opinion hot spot extraction algorithm in the data parameter experiment, effects, as well as the efficiency experiment was carried out. For the convenience of test coverage and redundancy measure, the extraction results coverage calculation formula is $\mathrm{cov}\,erage = \frac{|R|}{|D|}$, the extraction results of redundancy formula as following:

$$redundancy = \sum_{d \in R} (1 - 1/\sum_{d' \in R} sim(d, d')) / |R|.$$

5.2. Experimental Results and Analysis

This article has introduced the experiment of different λ value, and inspected the representative coefficient of λ to the hot spot extraction scale, the results coverage and the impact of redundancy. The results shown as Figure 4 and 5.

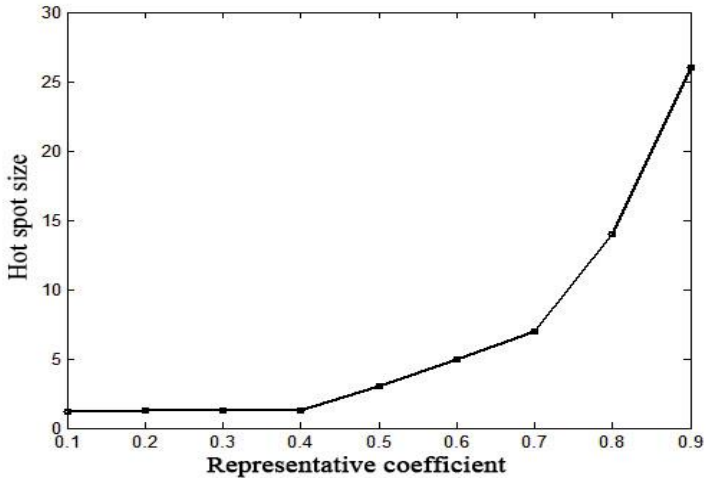

Figure 4. The Scale of Hot Spot Extraction under the Different λ Values

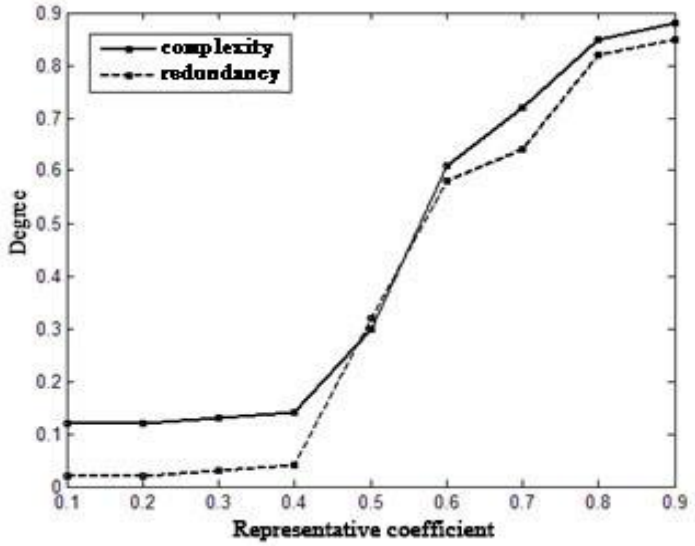

Figure 5. Under Different Value of Λ Extraction Hot Spot Coverage and Redundancy

It can be seen from the diagram when λ blow the 0.5 has little obvious impact. On the contrary, with the λ increases, the size, coverage and redundancy of extraction results has been increased. During the hot spot extraction procedure, the parameter λ values is very important. When the values is big enough, the hot spot scale will big, and lose the extraction's significance. However, if the values is very small, the representative will be more weaker between the two metadata. From the experiment results we can see that the λ value between 0.5 and 0.7 is ideal.

This article has carried on the contrast experiment and observed the effect of the VM - Rep method in extracting the Internet public opinion hot spot. Under the condition of the same data set and parameter λ, separately using the VM-Rep, K-means and Agglo methods to extract hot spot. The extraction results as shown in Figure 6 and Figure 7, the VM-Rep's average coverage is largest and better than the other two methods that means the VM-Rep can cover original content very well. In the meantime, VM-Rep has the minimum average redundancy and better than Agglo method, the mean value less than the K-means, and it means that the VM-Rep method has the minimum hot spot extraction redundancy in itself to suit the intention of the hot spot extraction.

In order to verify under the condition of the same number nodes that the VM - Rep's

ability of dealing with large data, the three different data scale method execution time as shown in Figure 8. Based on the voting mechanism and the VM-Rep algorithm of MapReduce framework have little running time than the K-means and Agglo method, which demonstrates the superiority of the improved algorithm. With the increase of the data scale, the growth rates of running time is increasing, just only investigates the VM-Rep different running time of data scales. That means, the amount of data is gradually increasing, the system performance will eventually reach the bottleneck.

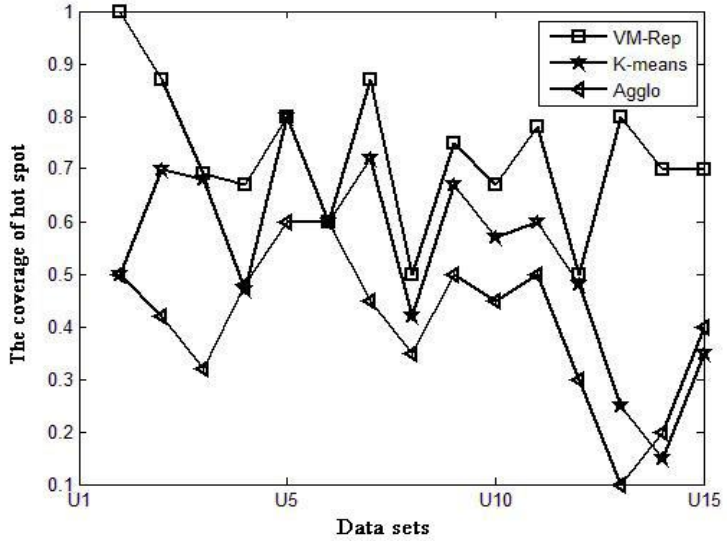

Figure 6. Three Methods to Extract Hot Spot Coverage

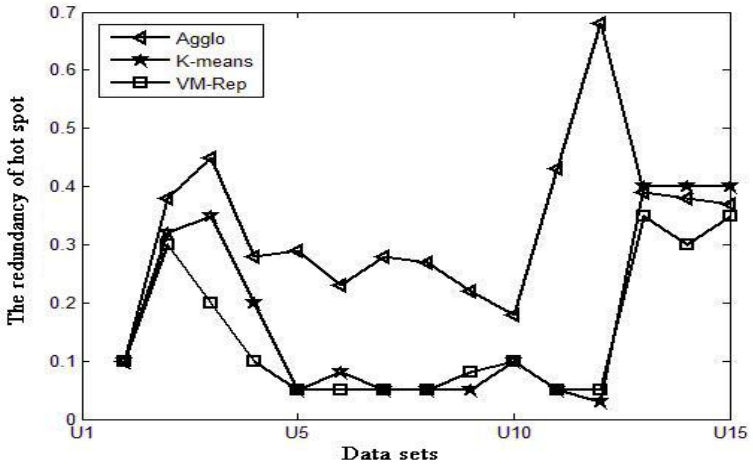

Figure 7. Three Methods to Extract Hot Spot Redundancy

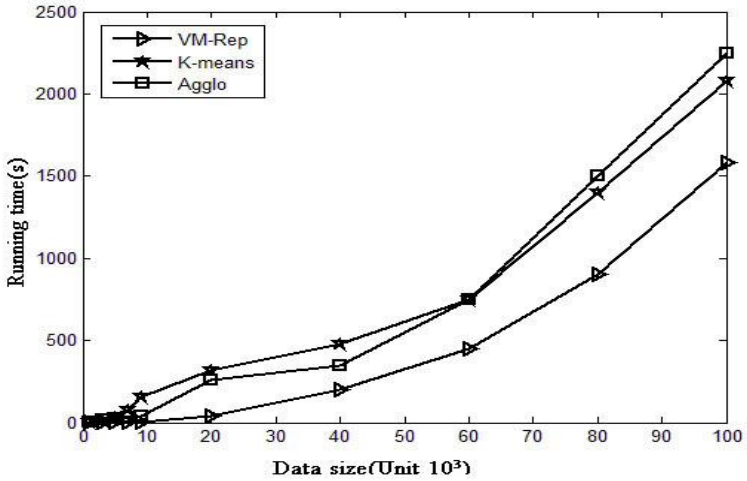

Figure 8. Three Methods in Different Scale Data on the Running Time

Acknowledgements

The Research supported by the National Natural Science Foundation of China (No.71272191).

References

[1] X. Xu and Z. C. Zhang, "Research on Application and Analysis of Internet-Mediated Public Sentiment", Information science, vol. 8, no. 26, **(2008)**, pp. 1194-1200.
[2] A. B. Qian, "A Model for Analyzing Public Opinion Under the Web and Its Implementation", New Technology of Library and Information Service, vol. 4, **(2008)**, pp. 49-55.
[3] S. H. Zhang and Z. P. Liu, "Study on Clustering Method for Internet Public Opinion Hot spot Topic", Journal of Chinese Computer Systems, vol. 3, no. 3, **(2013)**, pp. 471-474.
[4] C. J. Han, "The Hot-Topic Discovery Based on Density Clustering of Feature Words and Similarity Calculation", Chengdu, **(2013)**.
[5] W. Han, "Research on Hot Topic Detection and Tracking in Internet Public opinion", Harbin, **(2012)**.
[6] Y. Ma, "Study on the Method of Micro-blogging Public Opinion Hot spots Mining in Big Data", Modern Information, vol. 11, no. 34, **(2014)**, pp. 29-32.
[7] J. H. Zhou, "Method of Internet Public Opinion Hot Topic Mining under Hadoop", Journal of Hebei North University (Natural Science Edition),vol. 6, no. 30, **(2014)**, pp. 19-21.
[8] M. Huang and X. G. Hu, "Internet Public Opinion Hot Spot Mining Base on Complex Network Theory", Computer Simulation, vol. 9, no. 28, **(2011)**, pp. 114-116.
[9] G. Q. Cui and P. L. Qiao, "The Model SWMS Research and Implementation on Web Mining", Journal of harbin university of science and technology, vol. 11, no. 5, **(2006)**, pp. 15-18.
[10] S. Y. Jiang, "Custer Fusion Algorithm Based on Majority Voting Mechanism", Journal of Chinese Computer Systems, vol. 2, **(2007)**, pp. 306-308.
[11] R. Mukherjee, N. Sajja and S. Sen, "A Movie recommendation system-an application of voting theory in user modeling", User Modeling and User-Adapted Interaction, vol. 13, no. 2, **(2003)**, pp. 5-33.
[12] J. Dai and Z. M. Ding, "MapReduce Based Fast kNN Join", Chinese Journal of Computers, vol. 1, no. 38, **(2015)**, pp. 100-102.
[13] J. Dean and S. Ghemawat, "MapReduce: Simplified data processing on large clusters", Communications of the ACM, vol. 1, no. 51, **(2008)**, pp. 107-113.
[14] R. Karim, A. Hossain and M. Rashid, "A MapReduce Framework for Mining Maximal Contiguous Frequent Patterns in Large DNA Sequence Data sets", IETE Technical Review, vol. 2, no. 29, **(2012)**, pp. 162-168.