Hindawi Publishing Corporation
Computational Intelligence and Neuroscience
Volume 2013, Article ID 230946, 6 pages
http://dx.doi.org/10.1155/2013/230946

Research Article

Research on Hotspot Discovery in Internet Public Opinions Based on Improved K-Means

Gensheng Wang

Electronic Business Department, Jiangxi University of Finance and Economics, Nanchang 330013, China

Correspondence should be addressed to Gensheng Wang; wanggenshengjc@163.com

Received 31 March 2013; Revised 5 June 2013; Accepted 21 July 2013

Academic Editor: Saeid Sanei

How to discover hotspot in the Internet public opinions effectively is a hot research field for the researchers related which plays a key role for governments and corporations to find useful information from mass data in the Internet. An improved K-means algorithm for hotspot discovery in internet public opinions is presented based on the analysis of existing defects and calculation principle of original K-means algorithm. First, some new methods are designed to preprocess website texts, select and express the characteristics of website texts, and define the similarity between two website texts, respectively. Second, clustering principle and the method of initial classification centers selection are analyzed and improved in order to overcome the limitations of original K-means algorithm. Finally, the experimental results verify that the improved algorithm can improve the clustering stability and classification accuracy of hotspot discovery in internet public opinions when used in practice.

1. Introduction

The rapid development of the Internet exerts a profound impact on the country, society, and individuals and how to effectively master mass data and extract the hotspot information therein have been a problem urgently to be solved in the management of internet public opinions. Solving this problem has an extensive application prospect: first, for individuals, it is an important means to promptly and conveniently obtain the hotspot information in current society; second, for enterprises, it can help enterprises master the most cutting-edge information and hot technology in their fields, increase their competitiveness for enterprises through this method; especially for the country, it can provide important clues for relevant departments of the governments to promptly know about the direction of public opinions in current society, be conductive to the governments to analyze and guide the public opinions, actively guide the healthy development of internet public opinions; meanwhile, help the governments to grasp the problems mostly cared by the people in each period as well as the viewpoints and attitudes on these problems, so as to make scientific and correct decision, keep the society stable, and truly reach the aim that the Internet serves for the society and the people. In the past, public opinions workers rely on manual work to sort the contents on the webpage to discover the hotspot information of the society, not only low efficiency in work, but also easy to be subjectively influenced and make the result deviate from the truth. At present, search engines, to some extent, meet people's demand on rapidly acquiring information needed among massive and messed information; however, its adoption of simple key words matching to find information causes a great deal of redundant and irrelevant contents in search results, results in redundant information overwhelming the information needed, leads to the incomplete analysis on topics of relevant personnel, and makes it difficult to have a comprehensive mastery. The premise for discovering hotspot information by search engines is that analysts know in advance the existence of such topics, so such method is obviously lagging and it is not good for discovering new problems, easy to miss the best timing to solve problems, making the problems spread and difficult to be controlled. Therefore, if the real-time hotspot information in a period is to be obtained and the internet hotspot topics in current society are to be periodically discovered, automatic solutions are becoming a valuable research orientation.

2. Literature Review

At present, the study on hotspot discovery of internet public opinions at home and abroad mainly focuses on such two aspects as internet information processing and data mining. (1) In the aspect of internet information processing, the main research contents of scholars at home and abroad include word segmentation technology, measuring of multidimensional vector space on article theme [1]. (2) In the aspect of internet data mining, contents involved are information acquisition of public opinions, automatic classification, automatic clustering, and so forth, and this kind of methods has obtained certain achievements. For instance, Hamerly and Elkan, on the basis of analyzing the shortages of original K-means and its reasons, put forward a new model to mine and analyze internet public opinions information, and illustrated the application of text mining in the analysis of internet public opinions [2]; Kristina analyzed the basic situation of internet public opinions and designed an analyzing model of internet public opinions based on themes [3]; Andreas combined the advantages of comprehensive partitional clustering and agglomerate clustering and put forward an incremental hierarchical clustering algorithm and applied it to hot topic discovery in internet public opinion [4]; Wagstaff and Rogers combined natural language processing with information retrieval technology and put forward a very effective single-granularity topic identification method as to the event features [5]; Ya designed a hotspot events discovery system which is geared to the needs to internet news coverage and able to automatically find the hotspot events on the internet within any period [6]; Bradley and Managasarian, according to the demands on the analysis of internet public opinions, built the discovery and analysis system of internet public opinion hotspots problems based on clustering [7]. As for mass internet public opinion information, how to improve the effectiveness and efficiency of analysis and processing as well as the accuracy and efficiency of the analysis of internet public opinion hotspots remains a hotspot for current research.

Currently, domestic and overseas studies on the clustering methods of internet public opinions are mainly divided into the following categories: partitional clustering, hierarchical clustering, clustering based on density, artificial neural network clustering, clustering based on internet, and so forth in which clustering is widely applied. According to different objects, application fields, and aims of clustering, there are specific requirements on the quality, efficiency, and result visualization degree of clustering for clustering methods. Hence, proper clustering algorithm shall be selected as required by specific conditions, among which as to text clustering, K-means clustering, due to its features like increment, batch processing, speediness, and efficiency, as well as its advantage in applicable to dynamically process mass data of internet media information, is widely applied in the detection of internet hotspot topics. However, the clustering quality in K-means algorithm relies too much on the initial number of clusters and initial clustering centers, which shall be conquered in actual application.

K-means algorithmis one of the best information clustering methods in data mining which can extract and find new knowledge. But it is found that K-means algorithm used in processing the data of isolated points has great limitations [6–8]. The paper tries to present some improvements to over- come these limitations and takes advantage of powerful classification ability of the algorithm to discovery hotspot in internet public opinions.

3. Text Preprocessing

Hotspot discovery depends on website text clustering which can be described as a given text set $D = \{d_1, d_2, \ldots, d_n\}$, eventually get a cluster's set $C = \{C_1, C_2, \ldots, C_n\}$, $\cup_{i=1}^{K} C_i = D$ derive for all d_i ($d_i \in D$), $\exists C_j$ ($C_j \in C$) and $d_i \in C_j$, and also make the objective function $Q(C)$ reach the minimum or maximum value, of which n is total text number, K is final clustering number, and $C_j \cap C_i \neq \phi$, $j \neq i$.

3.1. Characteristic Selection and Expression of Website Text. Vector space model (VSM) is commonly adopted to express each text. In this model, each text d is considered as a vector in a vector space. $tfidf$ is used as a measure of characteristic vector in this paper, and this measure gives the weight of each word t. See (1) for the calculation of the weight:

$$tfidf\,(d,t) = tf\,(dt) * \log_2 \frac{N}{df\,(t)}. \tag{1}$$

In (1), $tf(d,t)$ is the word frequency of word t in the text d, $df(t)$ is all the text numbers of word t contained in the text set D, and N is total text number. After the characteristic selection, text $d \in D$ is the form of the vector, and the value of each dimension is the corresponding $tfidf(d,t)$ weight value, so the text can be expressed as follows:

$$d = \{(t_i, tfidf\,(d,t_i)) \mid 1 \leq i \leq m\}, \tag{2}$$

of which t_i is the lexical entry and m is the dimension of the characteristic vector. However, after the characteristic selection, m is still very large, thousands of dimensions at least and tens of thousands of dimensions at most while nonzero word frequency of each corresponding text vector is very few, which makes text VSM show the high dimension.

3.2. Definition of Similarity. In this paper, cosine distance is used to measure the similarity between the website texts and defines the similarity of two texts d_1 and d_2 as follows:

$$\mathrm{Sim}\,(d_1,d_2) = \cos\,(d_1,d_2) = \frac{(d_1 * d_2)}{(\mathrm{norm}\,(d_1) * \mathrm{norm}\,(d_2))}. \tag{3}$$

In order to reduce the impact of different length of the texts on calculating the text similarity, each text vector has been integrated to the unit length. See (2):

$$d = \frac{d}{\|d\|} = \frac{\{tfidf\,(d,t_1), fidf\,(d,t_2), \ldots, fidf\,(d,t_m)\}}{\sqrt{\{tfidf(d,t_1)^2, fidf(d,t_2)^2, \ldots, fidf(d,t_m)^2\}}}. \tag{4}$$

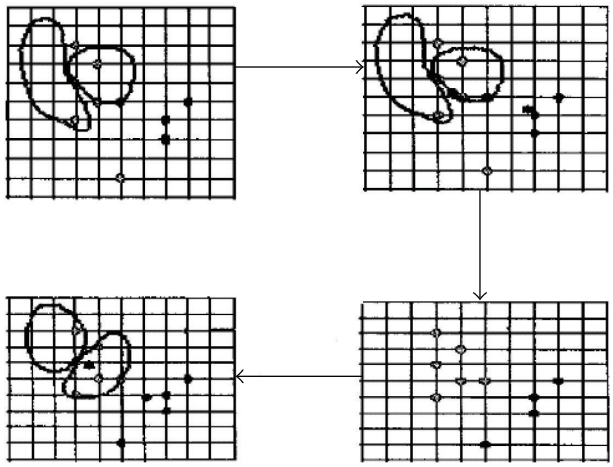


Figure 1: The procedures of K-means algorithm.

Thus, $\|d\| = 1$ and the similarity of the cosine is the dot product of two text vectors; that is, $\text{Sim}(d_1, d_2) = d_1 \cdot d_2$.

4. Derivation of Hotspot Discovery Algorithm

4.1. K-Means Algorithm Principle. Steps for K-means clustering algorithm are as follows [8] (see Figure 1):

(1) Select n objects as the initial cluster seeds on principle;

(2) Reassign each object to the most similar cluster in terms of the value of the cluster seeds;

(3) Update the cluster seeds; that is, recompute the mean value of the object in each cluster, and take the mean value points of the objects as new cluster seeds.

(4) Repeat (2) and (3) until no change in each cluster.

4.2. Limitation of K-Means Algorithm. When K-means algorithm is used to cluster data, the stability of the clustering results is still not good enough; sometimes, the clustering effect is very good (when the data distribution is convex-shaped or spherical), while sometimes the clustering results have obvious deviation and errors, which lies in the data analysis. It is unavoidable for the clustered data to have isolated points, referring to the situation that a few data deviate from the high-dense data in intensive zone. The clustering mean point (geometrical central point of all data in the category) is used as a new clustering seed for the K-means clustering calculation to carry out the next turn of clustering calculation, while under such a situation, the new clustering seed might deviate from the true data intensive zone and further cause the deviation of the clustering results [9]. Therefore, it is found that using K-means algorithm to process the data of isolated points has a great limitation.

4.3. Improving of K-Means Algorithm Principle. The original K-means algorithm selects K points as initial cluster centers, and then the iterative operation begins. Different selection of initial point can achieve different clustering result. For the reduction of the clustering result's dependence on the initial value and the improvement of the clustering stability, better initial cluster centers can be achieved by the search algorithm of the cluster center [9, 10].

In the search process, the sampled data tries to be undistorted and is able to reflect the original data distribution through the random data sampling, as shown in Figure 2, among which, (a) original data distribution, (b) sampled data distribution.

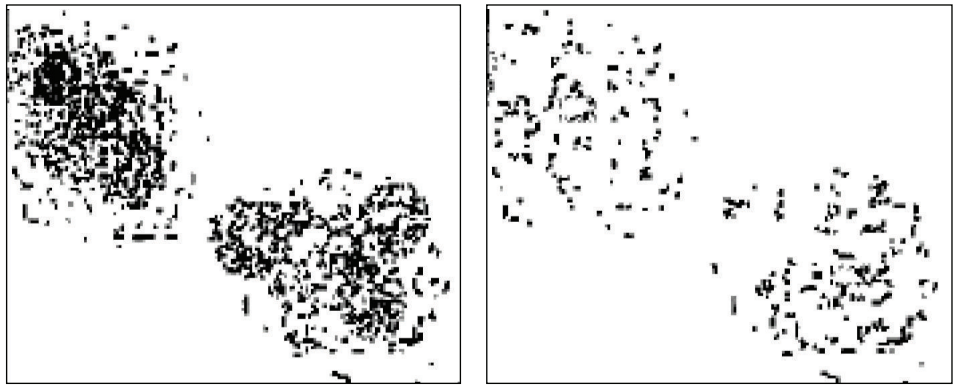
The sampled data and the original data are clustered by K-means algorithm, respectively, and little change of final cluster centers is found. Therefore, the sampling method is suitable for the selection of the initial cluster centers. In order to minimize the sampling effects on the selection of the initial cluster centers, the sample set extracted each time should be able to be loaded into the memory and do best to make the sum of the sample sets extracted J times equivalent to the original data set. Each extracted sample data is clustered by K-means algorithm and one group of cluster center is produced, respectively; the samplings J times produce J groups of the cluster centers in all, and then the comparison of clustering criterion function values is conducted for J groups of cluster centers, and one group of minimum cluster center in J_c value is given as the optimal initial cluster center.

For the protection against segmenting large clusters into small clusters by the criterion function, the algorithm takes the initial cluster as $K'$and $K' \succ K$. According to the quality requirements and the time, K' value does the compromise selection. Larger K' value is able to expand the solution search scope, and the phenomenon of no initial value near certain extremal vertexes is diminished. The utilization of the searched initial cluster center clusters the original data by another K-means algorithm and outputs K' cluster centers, and then the reduction of each cluster quantity to the specified K value is studied.

4.4. Improving the Selection of Initial Classification Centers. The basic idea of new selection method of initial cluster centers is based on the assumption that the distribution of the website text sets has been known; a good initial cluster center should satisfy the following rules in the paper.

(1) The selected initial centers belong to different clusters, respectively; that is, any two initial centers cannot be the same cluster;

(2) The selected initial cluster centers should represent this cluster, that is, be as close as possible to the cluster centers. To select K texts as initial cluster centers and at the same time ensure that K texts just belong to different clusters, such strict constraints are difficult to be achieved through random sampling as much as possible, so it is thought that in order to minimize the sampling's effect on initial cluster centers, m times of samplings are taken and the sample size is n/m, of which, n is the number of the text in the text sets, the value of m is that each sample size should be put into the main storage and as far as possible satisfies the fact that the sum of the samples taken for m times is equivalent to the original text set. Each sample text



(a) Original data distribution (b) Sampled data distribution

Figure 2: Comparison of data distribution before and after sampling.

taken is clustered by K-means algorithm to produce a group of text clusters with K cluster centers, respectively; m times of sampling operation produce $m \times K$ cluster centers in all, and then agglomerative hierarchical clustering algorithm single-link algorithm is used to do the clustering to obtain K clusters, of which, the average value is the final K initial cluster centers. Different from the division strategies taken by K-means algorithm, the agglomerative hierarchical clustering algorithm does not exist in the selection of the initial cluster centers. It regards each text as a cluster at first; the text is the centre of this cluster, and each step of clustering combines the two most similar clusters into a cluster until all the texts are integrated into a cluster or only K clusters. With clustering, the similar text is integrated into a cluster gradually and the hierarchical clustering is able to automatically generate different hierarchical clustering model.

In the combination of agglomerative hierarchical clustering algorithm and K-means algorithm, a hierarchical clustering algorithm based on K-means is addressed to select the initial cluster centers; that is, the cluster centers produced by K-means method restrain the agglomerative space of the agglomerative hierarchical clustering algorithm. The selection method of the initial cluster centers is generally described as follows.

(1) m times of sampling are taken for the text sets, which are divided into m sample sets $\{S_1, S_2, \ldots, S_m\}$.

(2) Each sample set performs K-means algorithm, respectively, to produce m groups of K cluster centers.

(3) Another clustering is done for $m \times K$ cluster centers by the agglomerative hierarchical clustering algorithm (single-link algorithm is used here) until only having K clusters, and the average value of each cluster is taken as the initial cluster centers of next step of K-means algorithm.

From the previous algorithm, it is seen that the text set of the sample taken is smaller than the original text set, so the search process amount of the initial cluster centre is less, the iterative number is less, and the speed is faster; at the same time, it is also ensured that the final cluster centers belong to different clusters and have adequate representation.

The specific algorithm flow used in the paper can refer to the reference [8].

5. Experimental Verification

5.1. Data Acquisition and Preprocessing. Verification data acquisition and preprocessing in this paper mainly include the following steps. (1) Public opinions data acquisition adopts web search technology, traversing the entire Web space within designated scope to collect all kinds of public opinions information, establishing indexes of acquired information through indexer, and save in the index database. Objects of data acquisition are mainly each major web portals, BBS, blogs, and so forth. (2) Word segmentation processing of website text: public opinions information acquired are unstructured data, which shall be preprocessed. Word segmentation study of Chinese language has been mature. This thesis adopts the Chinese Lexical Analysis System of Institute of Computing Technology (ICTCLAS). (3) Text features abstraction: the aim of selecting features is to further filter works with no much amount of information and less influence on the discovery of public opinions hotspots, reaching the effect of dimension reduction of website feature vector, so as to improve the processing efficiency and reduce the complexity of calculation. Form of dimension reduction adopted in this thesis to build evaluation function of webpage theme through statistical methods, evaluating each feature vector and choosing words meeting the preset threshold as the feature item of webpage; (4) Feature representation: this paper adopts vector space model (VSM) to indicate public opinions information; here omit the specific forms.

5.2. Experimental Results. Considering that news is paid high attention in Internet information and it is easy to collect information, this paper takes internet news as verification data. First, randomly choose 8919 pieces of news among the politics news on December 1, 2012 to December 15, 2012 as the test

Table 1: Parts statistics of feature vector word frequency.

Diaoyu Islands	American	China	Syria	Russia	Military	Japan	Shinzo Abe	Obama	Hugo Chavez
12156	8973	9987	4612	3416	1256	3421	1281	2521	1452

Table 2: Parts statistics of news hotspots themes.

News themes	The number of pages	Feature words
Diaoyu Islands	1524	Sovereignty, Shinzo Abe, island Purchase, Escort, Military, Fighter, American, China, Japan
Syria Crisis	642	The opposition, Muslim, Shiite, Sunnite, BaShaEr, Antiterrorism, Iran, Russia, American, the Arab league

Table 3: Algorithm performance F_1 comparison of different algorithms.

F_1	F_1 typical value	F_1 of original K-means algorithm falling into the experimental frequency of this interval	F_1 of improved K-means algorithm falling into the experimental frequency of this interval
[0.15, 0.25]	0.20	1	0
[0.25, 0.35]	0.30	2	0
[0.35, 0.45]	0.40	2	0
[0.45, 0.55]	0.50	4	0
[0.55, 0.65]	0.60	5	0
[0.65, 0.75]	0.70	7	9
[0.75, 0.85]	0.80	2	11
[0.85, 0.95]	0.90	1	8
[0.95, 1.00]	1.0	0	0

samples obtained by features words of webpage cluster. As webpage comes from real website, webpage data have certain complexity and randomness. After word segmentation processing, there are 68213 words in total; 52173 words are obtained after stop words processing to carry out information for subsequent calculation; take top 10% words, that is, 6512 words, as the feature vector of webpage text. Test results are as shown in Tables 1, 2, and 3. Table 1 is the statistical table of vocabulary and word frequency with large information gain value; Table 2 is the statistical result of news hotspots themes; Table 3 is the clustering performance comparison of the algorithm in this paper and ordinary K-means [8].

In Table 3, F_1 means F-measure value, and F_1 distribution is wildly used to illustrate the performance of different algorithms [3–6, 11]. Using the data introduced previously and specific calculation items can be seen in [8]. It can be seen from Table 3 that there is poor stability in the clustering results obtained by ordinary K-means algorithm and scattered F-measure value, but the improved clustering algorithm has better stability of the clustering results, more concentrated F-measure value, and higher F-measure average value. The experiment shows that the improved clustering algorithm improves its accuracy and stability greatly. In the use of ordinary K-means algorithm, F-value of the clustering results scatters from 0.60 to 0.75; in the use of the improved algorithm, the stability of its value is from 0.75 to 0.85.

6. Conclusion

Nowadays, internet is becoming the main channel for people to obtain and release information, the guiding role of internet public opinions information is larger and larger; it has aroused wide attention in the industry how to carry out public opinions gathering and hotspots discovery on the basis of information acquisition of Internet public opinions as well as track and analyze the hotspots to guarantee the information security. Under such background, this paper, based on analyzing the advantages and disadvantages of all kinds of clustering algorithms, chooses K-means clustering as the website text clustering model and puts forward a new discovery algorithm of internet public opinions hotspots through improving its shortcoming of sensitivity to initial number of clusters and initial clustering centers. The test illustrates the applicability and reliability of method in this paper. The next study shall be focused on clustering of features of internet information text, for the sake of final realization of clustering algorithm applicable to all the languages.

References

[1] H. Liu and J. H. Xu, "Research of internet public opinion hotspot detection," *Bulletin of Science and Technology*, vol. 27, no. 3, pp. 421–425, 2011.

[2] G. Hamerly and C. Elkan, "A new algorithm based on K-means and its application in internet public opinion hotspot detection," *Pattern Recognition*, vol. 32, no. 6, pp. 521–534, 2012.

[3] L. M. Kristina, "Document clustering in reduced dimension vector space," *Journal of Computer Application*, vol. 27, no. 10, pp. 37–49, 2011.

[4] H. J. Andreas, "Research on text document clustering," *Computer Simulation*, vol. 24, no. 7, pp. 84–99, 2010.

[5] C. D. Wagstaff and S. S. Rogers, "Constrained K-means clustering with background knowledge," *Journal of Computer Engineering and Application*, vol. 21, no. 5, pp. 467–479, 2011.

[6] B. T. Ya, "Research on public opinion hotspot detection based on SVM," *Science and Technology Management Research*, vol. 25, no. 2, pp. 64–69, 2009.

[7] P. S. Bradley and L. S. Managasarian, "K-plane clustering," *Journal of Global Optimization*, vol. 16, pp. 23–32, 2010.

[8] Y. Tang and Q. S. Rong, "An implementation of clustering algorithm based on K-means," *Journal of Hubei Institute For Nationalities*, vol. 22, no. 1, pp. 69–71, 2011.

[9] Z. H. Yang and Y. T. Yang, "Document clustering method based on hybrid of SOM and K-means," *Computer Application*, vol. 27, no. 5, pp. 73–75, 2012.

[10] Y. F. Zhang and J. L. Mao, "An improved K-means algorithm," *Computer Application*, vol. 23, no. 8, pp. 31–33, 2009.

[11] N. Li and D. D. Wu, "Using text mining and sentiment analysis for online forums hotspot detection and forecast," *Decision Support Systems*, vol. 48, no. 2, pp. 354–368, 2010.