HK.NCCP
International Journal of Intelligent Information and Management Science
ISSN: 2307-0692 *Volume 2, Issue 7, October 2013*

Research on Internet Public Opinion Hotspot Detection Based on Document Cluster

Gensheng Wang
Electronic Business department
Jiangxi University of Finance and Economics
Nanchang, China

Abstract: In order to strengthen management and monitor to Internet, collection and analysis of public opinion information is a realistic problem solved urgently for the present government departments. The paper present a new algorithm for detecting internet public opinion hotspot based on K_means and particle swarm optimization algorithm. First, the limitations of the K_means and particle swarm optimization algorithm are analyzed. Second, the algorithms of K_means and particle swarm optimization algorithm are integrated and corresponding improvements including algorithm principle, the exploration ability of global solution , algorithm calculation process are presented to overcome the limitations of original algorithms. Finally, the experimental results verify that the new algorithm can improve effectiveness and validity of hotspot discovery of internet public opinions when used for internet public opinion hotspot detection practically.

Keywords: Internet public opinion hotspot detection; Document cluster; K_means; Particle swarm optimization

1. Introduction

With the development of science and technology, as a new form of media , Internet has been playing a decisive position in fierce competition of the media. It is inevitable that the Internet public opinion has influenced the society profoundly. In regard to the Internet public opinion, their occurrence range is wide, spreading speed is high, and their eruption spot is difficult to be detected and controlled. Therefore, not only governments but also individuals are necessary to master the hot topic timely, know the direction of popular opinions correctly, and take action to control and guide the trend of hot topic in order to reduce the disadvantageous influence. So research on internet public opinion hotspot detection has become a hotspot for the researchers in the fields related.

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 of language, measure of multidimensional vector space on article theme; [2-3] In the aspect of internet data mining, contents involved are information acquisition of public opinions, automatic classification, automatic clustering, and etc., having obtained certain achievements. For instance, Huang Xiaobin, on the basis of analyzing text mining, put forward the mining and analyzing model of internet public opinions information, and illustrated the application of text mining in the analysis of internet public opinions; Qian Aibing[4] analyzed the basic situation of internet public opinions, and designed an analyzing model of internet public opinions based on themes; Guo Jianyong[5] and etc., combing the advantages of comprehensive partitional clustering and agglomerate clustering, put forward an incremental hierarchical clustering algorithm applied to theme discovery; Yu Manquan and etc.[6], combining natural language processing with information retrieval technology, put forward a very effective single-granularity topic identification method as to the features of events; Liu Xingxing and etc.[7] Designed a hotspot events discovery system which is, geared to the needs to internet news coverage, able to automatically find the hotspot events on the internet within any period; Wang Wei[8], according to the demands on the analysis of internet public opinions, built the discovery and analysis system of internet public opinions hotspots problems based on clustering. As to mass internet public opinions information, how to improve the effect and efficiency of analysis and processing as well as the accuracy and efficiency of the analysis of internet public opinions 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, clustering based on models, and etc. 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. 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 algorithm is one of the best information clustering methods in data mining which can extract and find new knowledge and. But it is found that using K-means algorithm to process the data of isolated points has great limitations[7-9]. An improved particle swarm and K_means hybrid clustering algorithm put forward in the paper firstly makes use of the population fitness variance to determine the operating time of K_means algorithm, so as to realized the organic combination of PSO algorithm and K_means algorithm, as well as accelerate the rate of convergence of the algorithm while enhancing local search ability of the algorithm. In order to increase the rate of convergence of hybrid algorithm in early period, increase the updating mechanism of particle position based on extrapolating direction in the evolutionary process solves the problem that K_means algorithm is slow in rate of convergence. Bring random mutation operation in the evolutionary process of PSO algorithm and only carry out K mean search on the particles participating in the mutation; hence, it will not affect the rate of convergence of the algorithm while enhancing population diversity, further remedying the defect that K_means algorithm falls into local optimization.

3. Research Method

3.1. Particle Swarm Optimization (PSO)

PSO is an evolutionary method based on swarm intelligence. Every potential solution optimizing problems is a particle of the search space. Every particle has its corresponding speed, position and a fitness which is determined by objective function and through which algorithm evaluates the strengths and weaknesses of particles. Algorithm firstly initializes a swarm of random particles, and finds optimal solution through iteration. In each iteration, the particles update themselves through tracking two "extreme values" which are the optimal solution found by the particle itself, i.e. individual extreme value $pBest$ and the optimal solution currently found by the particle swarm, i.e. global extreme value $gBest$. After finding the above two extreme values, the particles update their speed and position according to equation 1 and equation 2.

$$v_i(n+1) = wv_i(n) + c_1 \cdot rand_1() \cdot (pBest - p_i(n)) + c_2 \cdot rand_2() \cdot (gBest - p_i(n)) \quad (1)$$

$$p_i(n+1) = p_i(n) + v_i(n+1) \quad (2)$$

In equation 1 and equation 2, $v_i(n)$ is the current speed of particles, $p_i(n)$ is the current position of particles, $i = 1, 2, 3, ...N$, N is the dimension of current space, $rand_1()$ and $rand_2()$ are random numbers among [0 ,1], and c_1 and c_2 are learning factors, generally dereferencing $c_1 = c_2 = 1$, w is weighting coefficient, generally dereferencing from 0.1 to 0.9; many experiments show that if w is linearly decreased with the algorithm iteration, the convergence performance of the algorithm will be significantly improved. Suppose that $w_{\max}$ is the largest weighting coefficient, $w_{\min}$ is the smallest weighting coefficient, run is the current times of iteration, and $runMax$ is the total times of algorithm iteration, then equation 3.

$$w = w_{\max} - run \frac{(w_{\max} - w_{\min})}{runMax} \quad (3)$$

3.2. Particle Swarm and K_means Hybrid Clustering Algorithm

K_means clustering algorithm is widely used in a lot of fields like data mining, image segmentation, pattern recognition, feature extraction, and etc. due to such advantages as simple in algorithm and fast in rate of convergence. However, traditional K_means algorithm has two inherent shortcomings: sensitive to initial value and easy to fall into local optimization. The occurrence of PSO optimization algorithm provides a new idea for solving such problem. Nevertheless, through the analysis of previous literatures, we find that the key to improve the accurate local solution searching ability of K_means algorithm lies in accelerating the rate of convergence of algorithm while increasing the accuracy of solution.

3.3. Improvement of Algorithm Principle

Main idea for the improvement is to change the calculation of distance in linear space into the calculation of kernel function, i.e. equation 4.

$$k(x_i, x_j) = \varphi(x_i) \cdot \varphi(x_j) \tag{4}$$

In equation 11, $\varphi(\cdot)$ is vector function, the specific form of which is uncertain; $k(\cdot,\cdot)$ is the kernel function meeting Mercer condition. Positional change of original particles is determined by its speed, individual optimal position and swarm optimal position. As $\varphi(\cdot)$ is uncertain, it is unable to determine the position and speed of particles in kernel space. In the original particle swarm algorithm, particles update their position through equation 5 and equation 6, however, in kernel space, it is impossible to find x_{id}, v_{id}, p_{id} or p_{gd}.

$$v_{id}^{k+1} = w \times v_{id}^{k} + c_1 rand(\cdot) \times (p_{id} - x_{id}^{k}) + c_2 rand(\cdot) \times (p_{gd} - x_{id}^{k}) \tag{5}$$

$$x_{id}^{k+1} = x_{id}^{k} + v_{id}^{k+1} \tag{6}$$

In equation 5, the most important two parts are ($p_{id} - x_{id}^{k}$) and ($p_{gd} - x_{id}^{k}$), which represent the individual optimal direction of the particle and swarm optimal direction respectively. As we cannot even determine the dimension of p_{id} and p_{gd} in kernel space, we can only find the reflection of two optimal direction of the kernel space in linear space in the manner of probing.

The above analysis shows that the most critical process in PSO algorithm is the moving of particles. If the process is to be carried out in kernel space, the specific form of $\varphi(\cdot)$ is indispensable. This paper puts forward the idea of using pattern search method, still changing the position of particles in linear space, but the standard of optimal direction comes from kernel space, thus avoiding obtaining the specific form of $\varphi(\cdot)$, and PSO algorithm improved by pattern search method can be used for kernel space calculation. Specific method is that find the number and initial centroid of data clustering through KPSO algorithm, and cluster in the kernel space by making use of K_means algorithm based on kernel function. Thus, not only the clustering ability of K_means algorithm can be greatly improved, but also the appropriate initial centroid can be found, also determining the number of clustering.

3.4. Improvement of the Exploration Ability of Global Solution

As the search ability of global solution of particle swarm and K_means hybrid clustering algorithm totally relies on the exploration result of global solution space in the early phase of particle swarm algorithm, random mutation operation shall be brought in particle swarm to prevent PSO algorithm from falling into local extremum resulting in early convergence. As mutation is not necessary for particles with good fitness, this paper only carries out random mutation operation on part of the particles with poor fitness, and other optimized particles remain the original population structure to carry out local search, thus realizing the balance between improving the rate of convergence and keeping the population diversity, as shown in equation 7.

$$f \quad (r_i \le C_v) \quad then \quad v_{id} = r_2 \times r_3 \times V\max / C_m \tag{7}$$

In equation 7, r_i ($i \in M$) is the random variable uniformly distributed among [0,1], M is part of the poorer particles through fitness sequencing, r_2 is the random variable uniformly distributed among [0,1], and r_3 is random variable, when the random number is less than 0.5, r_3 is 1, when the random number is larger than 0.5, r_3 is -1, controlling the flight direction of particles. Through bringing random mutation operation in particle swarm, early convergence of PSO algorithm is avoided, so as to enhance the exploration ability of global solution of particle swarm and K_means hybrid clustering algorithm. In order to realize the balance between global solution ability in clustering algorithm and rate of convergence, new hybrid clustering algorithm only carries out one accurate K_means search on the particles participating in the mutation, strengthening the exploitation ability of particles in new space participating in random mutation operation; moreover, limited to the number and probability of particles participating in random mutation, there is little impact on the rate of convergence of algorithm.

3.5. Improvement of Algorithm Calculation Process

The calculation process of the improved algorithm can be listed as follows.

①Initialize particles in data space, making the range of influence of particles almost cover all the data; and determine the appropriate range of influence of particles r as well as the initial step size of particle perturbation δ_1.

② Each particle calculates its data density within its range of influence. Through traversing data, calculate the distance between particles and data in kernel space; for those less than range of influence, data density of the particles shall be added.

③Search the optimal particle around oneself. Calculate the distance between a particle and all the other particles, the particle with the largest data density within the range of influence is the optimal particle.

④ Determine two approximate optimal directions $p_i^{'} - x_i$ and $g_i^{'} - x_i$.

⑤Particles update positions according to equation 5.

⑥Change step size and return to Step ②, until meeting conditions or reaching the largest cycle times.

⑦Take the optimal particle finally obtained by KPSO algorithm as the initial centroid, then carry out K_means clustering based on kernel function. If there is few data in certain category finally obtained, the data of the category can be deemed as isolated point.

4. Results and Analysis

4.1. 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 etc.; 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 paper adopts the Chinese Lexical Analysis System of Institute o f Computing Technology (ICTCLAS); ③ Text features abstraction, the aim of selecting features is to further filter works with not 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 paper 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.

4.2. Experimental Results

Experimental data come from database of 8919 pieces of news among the politics news on May 1, 2013 to May 20, 2013 as the test samples obtained by features words of webpage cluster, randomly chosen by the simulation experiment team. As webpage comes from real website, webpage data have certain complexity and randomness. After the news was chosen the data preprocessing as section 4.1, here only take the event of Diaoyu Islands and Syria Crisis for example. The feature vector word frequency of Diaoyu Islands and Syria Crisis are 12156 and 10563 respectively. The number of pages of Diaoyu Islands and Syria Crisis are 1876 and 1643 respectively. The feature words for Diaoyu Islands are sovereignty, Shinzo Abe, island purchase, escort, military, fighter, American, China, Japan, and the feature words for Syria Crisis are the opposition, muslim, Shiite, Sunnite, Ba-ShaEr, anti-terrorism, Iran, Russia, American, The Arab League.

As for the performance of the presented algorithm, this paper also realizes the application of the ordinary K-means algorithm[9] and SVM algorithm which are popular used in internet public opinion hotspot detection [6], Cluster performance of different algorithms is shown in Table 1. And the calculation platform as follows: hardware is Dell Poweredge R710, in which processor is E5506, memory 2G, hard disk 160G; software platform is Windows XP operating system, C programming language environment.

Table 1. The Application Performance of Different Algorithms

Algorithm	Algorithm in This Paper	Ordinary K-means algorithm	SVM algorithm
Accuracy Rate	95.85 %	81.33%	71.82%
Time Consuming(S)	17	451	27

Conclusion

Since the particle swarm optimization algorithm could not find the appropriate original centroid and the accurate cluster number under linear inseparable circumstance, an improved K-means algorithm based on particle swarm optimization was given in the paper. Particle swarm optimization was used to seek out the category that should be clustered and many other measures are taken to map the data to higher-dimensional space and to overcome the shortcomings of original K_means algorithm. Then the data was clustered by the cluster algorithm. The experiment demonstrates that the new method runs well and has satisfied webpage cluster when it is used in internet public opinion hotspot detection.

Acknowledgements

This work is supported by the National Social Science Foundation of China under the grant No. 12CTQ042.

References

[1] Liu Hong, Xu Jinhua. Research of Internet Public Opinion Hotspot Detection, Bulletin of Science and Technology, 2011,Vol 27, No.3, pp.421-425.

[2] Nan Li, Desheng Dash Wu. Using text mining and sentiment analysis for online forums hotspot detection and forecast. Decision Support Systems. 2010,Vol 48, No.3, pp.354-368.

[3] Ma Ruixin, Zhu Ming, Modeling and Simulating of User Clustering on Network Based on Particle Swarm Optimization, Computer Science, 2012,Vol 39, No.12, pp.220-223.

[4] Wang Wei, Xu Xin, Online Public Opinion Hotspot Detection and Analysis Based on Document Clustering, Computer Simulation, 2011, Vol 5, No.3, pp.74-79.

[5] Qu Xiaoning, Application of K-means Based on Commercial Bank Customer Subdivision, Computer Simulation, 2011, Vol 28, No.6, pp.357-360.

[6] Ya Been,Research on Public Opinion Hotspot Detection based on SVM, Science and Technology Management Research, 2009, Vol 25, No.2, pp.64-69.

[7] Bradley P S, Managasarian L. k-plane Clustering. Journal of Global Optimization, 2000, 16(1)23-32

[8] Tang Yong, Rong Qiusheng. An Implementation of Clustering Algorithm Based on K-means. Journal of Hubei Institute for Nationalities , 2004, Vol.22 No.1, pp.69-71

[9] Zhang Y.F., Mao J. L., An improved K-means Algorithm, Computer Application, vol.23. no.8, pp. 31-33,2003.