Performance Evaluation of the Cyberspace Public Opinion Detection and Tracking

Yongping Du
.Institute of Computer Science, Beijing University of Technology. Beijing, China
Email: ypdu@bjut.edu.cn

Changqing Yao
Institute of Scientific and Technical Information of China, Beijing, China
Email: yaocq@istic.ac.cn

Abstract: **The task of Public Opinion Detection and Tracking monitors the hot topics in the forums, BBS and so on. It collects the latest news and views and then the classifying and clustering algorithms are applied. Finally, the monitoring results are presented to the end users. This task has given the significant impact on the large numbers of internet users. The effective features in the story document are selected and we adopt the vector center model to represent the text document. The clustering algorithm merges the story to the corresponding cluster. We adopt three types of performance evaluation metric and that is F-Value, the entropy value and the Square Error Function JC. We analyze the distribution of the monitoring results on different forums. The results indicate that our approach is effective. The future research will focus on the detection of the timing news reports, and extracting the unique features to study the adaptive detection model and strategies.**

***Index Terms*—Public Opinion Detection, Performance Evaluation, Natural Language Processing**

I. INTRODUCTION

The task of Public Opinion Detection and Tracking monitors the hot topics in the forums, BBS and so on. It will collect the latest news and views and then the classifying and clustering algorithms are applied. Finally, the monitoring results are presented to the end users, such as the relevant functional departments.

The main problems of implementing the topic detection and tracking include the representation of information, the similarity measurement between the topic and the report, the selection of clustering strategy and realization.

The transmission mode of network media is different from traditional media. The accuracy of media information and the range of communication can not be effectively controlled. Anyone can publish the opinion in the forum or blog. The cyberspace public opinion has given the significant impact on the large numbers of internet users. This makes the effective public opinion detection and tracking become very important.

The task of topic detection is the relevant research field to the public opinion detection and tracking. The main purpose of the technology is used to monitor information sources in various languages, and give the warning when the new topic appears. It has broad application in the field of information security, financial security and other areas. The international conference of Topic Detection and Tracking, referred to as the TDT, gives the large scale evaluation of this task.

The conference of TDT[1] is sponsored by the United States Defense Advanced Research Development Agency (DARPA) and it is held every year. It attracts many research institutions to participate, including the universities of CMU, UMass, UPenn and also the companies of BBN, Dragon, GE, IBM and so on. Topic detection and tracking gives research to the text information re-organization, and it is also proposed in response to the problem of information overload. Various research institutions discussed it[2,3,4].

The selection of clustering strategy is important. In recent years, the subspace clustering algorithms have been proposed and have the rapid development in the field of data mining and machine learning. Especially, some practical subspace clustering algorithms are emerged to resolve the problem of the high dimensional data. CLIQUE algorithm is widely used[5], specifically for clustering high dimensional data sets.

WaveCluster (Clustering with Wavelets)[6] is a clustering algorithm based on the grid and density. It can effectively handle large data sets, and the computational complexity is low. It is not sensitive to the sequence of input data and any complex shape can be found in the cluster. This algorithm also has its disadvantage that the ability to handle high dimensional data is not strong.

II. FEATURE SELECTION AND THE CLUSTERING ALGORITHM

A. Feature Selection

The method of text representation is important to implement the public opinion detection. Language Model, referred to as the LM, and Vector Space Model, referred to as the VSM, are used mostly. We adopt the VSM to represent the text. For example, m*n matrix denotes that there are n feature items representing m pieces of text documents.

The text is represented as a collection of feature items in VSM. The vector space becomes too large when more

doi:10.4304/jcp.7.5.1284-1288

feature terms are extracted, and this will lead to much more time and space overhead. It is essential to reduce the dimensions of the vector space. The feature selection is important to resolve the dimension reduction within the high- dimensional feature set.

Given the feature set F={f_1, f_2, …,f_s}, select a subset F'={$f_{1'}$, $f_{2'}$, …,$f_{s'}$}. Here, s denotes the size of the original feature set and s' denotes the size of the selected feature set, and s' is less than s. The process of feature selection chooses the important features to from the new low-dimensional vector space. The commonly used feature selection statistics include Term Frequency, Document Frequency, Term Entropy, Mutual Information, Information Gain and so on.

We adopt the method of TFIDF(Term Frequency Inverse Document Frequency) to select the feature and calculate the weight of the feature item. Text document is treated as the set of the term *t* and each item has its corresponding weight value *w*. Text document is represented as multiple <*t, w*> pairs. Each text document *d* is mapped to a vector in the VSM, such as $V(d)=(t_1,w_1(d),t_2,w_2(d),\ldots,t_n,w_n(d))$. Here, t_i denotes the feature term and $w_i(d)$ denotes the weight value of t_i in the text document *d*.

The weight value is calculated as following equation 1:

$$w_i(d) = \frac{\left| tf_i \times \log \left| \frac{N}{nt_i} \right| \right|}{\sqrt{\sum_{i=1}^{n} (tf_i)^2 \times \log^2 \left| \frac{N}{nt_i} \right|}} \quad (1)$$

Here, tf_i denotes the frequency of t_i that appears in the document *d*. *N* indicate the total number of text documents for feature extraction. nt_i denotes the frequency of t_i that appear in the document set. Therefore, $(w_1(d), \ldots, w_i(d), \ldots, w_n(d))$ is treated as a vector in n-dimensional vector space. The higher TFIDF value indicates that the feature term has better discrimination.

B. Clustering Algorithm for the Public Opinion Detection

We make use of the largest minimum distance algorithm to implement the clustering process. The algorithm takes the objects as farther away as the initial cluster centers and gives the effort to obtain a better partition on the data set. The dissimilarity between the objects is measured by the methods of SIM and Euclidean Distance. The largest minimum distance algorithm is shown as following.

1. Given the text document set $D = \{d_1, d_2, ..., d_n\}$. Select one document object randomly, labeled as d_i, as the initial cluster center, labeled as C_1.

2. Calculate the distance of C_1 and the other document objects in the collection of *D*, and select the document object with the maximum distance or minimum similarity as the second cluster center C_2. We use two methods, identified as DIS and SIM, to implement the calculating.

SIM：

$$sim(d_i, T_j) = \frac{\sum_{k=1}^{n} (w_k(d_i) \cdot w_k(T_j))}{\sqrt{\sum_{k=1}^{n} w_k^2(d_i) \cdot \sum_{k=1}^{n} w_k^2(T_j)}} \quad (2)$$

Here, *n* denotes the dimension of the vector space model.

DIS:

$$dis(d_i, d_j) = \left[\sum_{k=1}^{n} \left| w_{ki} - w_{kj} \right|^p \right]^{\frac{1}{p}} \quad (3)$$

The higher value of DIS means the lower value of similarity. We give the experiments to verify the effect of *p* and find that the value of 2 achieves the better result.

3. For the document d_i in the collection of D, calculate the distance of d_i and cluster center C_i separately. The minimum value is labeled as $MinDist_i$. Further, find the object which has the largest distance to the cluster centers within all the remaining documents, namely:

$$\delta = \max(\prod_{i=1}^{n} MinDist_i) \quad (4)$$

4. Set the threshold value λ. Set the object dj as the new cluster centers when δ is larger than λ.

5. Repeat step 3 until there is no eligible new cluster centers.

It is unnecessary to give the number of clusters in advance, and it can be determined according to certain intelligent rules. The algorithm avoids selecting the object which is near to an existing cluster center but far away from the other objects as a cluster center. Therefore, it can guarantee that each new cluster center is far away from the existing cluster centers.

III. EVALUATION METRIC

We adopt three types of performance evaluation metric and that is F-Value and the entropy value.

A. F-Value Evaluation

For test data set given, the monitoring result of $P=\{P_1,P_2,\ldots, P_s\}$ is annotated by human. It is assumed that the topic T_j is existed within the system monitoring results for the corresponding topic P_i. It is checked through the system result $T=\{T_1, T_2,\ldots, T_c\}$ in order to find the cluster T_j, and the precision, recall and F-value is calculated[7]. Finally, the cluster with the optimal value is selected and the optimal value determines its quality.

For the topic P_i annotated by human and the cluster T_j detected by the system, calculate the precision, recall and F value as following.

$$\text{Precision}(P_i, T_j) = \frac{\left| P_i \cap T_j \right|}{\left| T_j \right|} \quad (5)$$

$$\text{Recall}(P_i, T_j) = \frac{\left| P_i \cap T_j \right|}{\left| P_j \right|} \quad (6)$$

$$F(P_i, T_j) = \frac{2P(P_i, T_j) * R(P_i, T_j)}{P(P_i, T_j) + R(P_i, T_j)} \quad (7)$$

The F-value of topic P_i is got as formula 8.

$$F(P_i) = \max_{j=1,\ldots,c}\{F(P_i,T_j)\} \quad (8)$$

The final F-value of the system is calculated as formula 9.

$$F = \frac{\sum_{i=1}^{s} (|P_i| * F(P_i))}{\sum_{i=1}^{s} |P_i|} \quad (9)$$

The F-value of the system gives the global evaluation of the overall topics. The evaluation metric emphasizes that the clustering results should approximate to the results by the human annotation. The evaluation metric has the strong overall discrimination between the clustering results and it is a valid indicator.

B. Entropy Evaluation

For the given data set $D = \{d_1, d_2, ..., d_n\}$, it is clustered into the set of topic cluster $T = \{T_i \mid T_i \subset D, i = 1...c, \bigcup_{i=1...c} T_i = D\}$. Each topic cluster T_i is evaluated by the following formula based on the entropy. Here, the set of $P=\{P_1,P_2,\ldots, P_s\}$ is the annotated result by human.

$$Entropy(T_i) = \frac{1}{\log(s)} \sum_{j=1}^{s} \frac{|T_i \cap P_j|}{|T_i|} \log\left(\frac{|T_i|}{|T_i \cap P_j|}\right) \quad (10)$$

The entropy value[8] above is normalized between 0 and 1. The value of 0 indicates that the cluster T_i is contained within the a topic P_j completely. The value of 1 indicates that the cluster T_i evenly cover all the topics and this is a poor result. The evaluation metric is a good indicator for the content distribution within the cluster. The more balanced sources, the higher the degree of disorder and the clustering result is bad. The overall system entropy value is obtained by averaging the entropy values of all the topic clusters.

C. Square Error Function J_C

The clustering algorithm is carried out on the document set D and gets the cluster set T.

$$T = \{T_i \mid T_i \subset D, i = 1...c, \bigcup_{i=1...c} T_i = D\}$$

Different cluster contains the different number of sample, such as $n_1, n_2, \ldots, n_c$. The function of square error is defines as following.

$$J_c = \sum_{j=1}^{c} \sum_{k=1}^{n_j} \left\| d_k^{(j)} - m_j \right\|^2 \quad (11)$$

Here, m_j denotes the different cluster center.

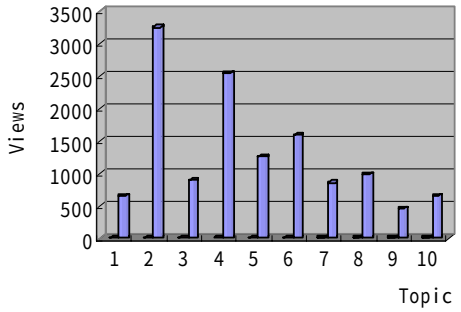
$$m_j = \frac{1}{n_j} \sum_{j=1}^{n_j} d_j \text{ , } j=1,2,\ldots,c \quad (12)$$

The larger of the J_C value and the larger of the error. We hope to get the cluster results with the smaller J_C value.

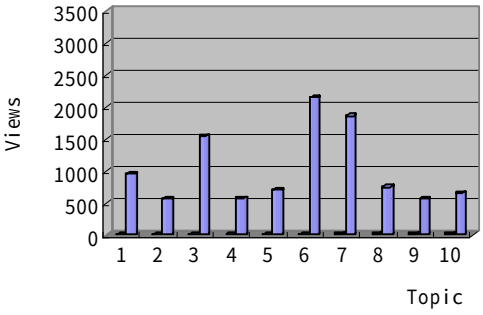
IV. Experimental and Evaluation Result

The experimental data is collected on the website of Tencent (www.qq.com). It contains the information on 3 major forums, and that is the Finance forum, the Sports forum and the News forum. The collection time ranges from March to July in 2010.

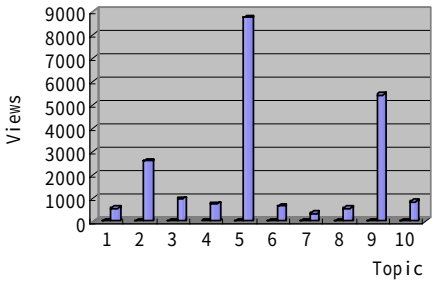
We analyze the distribution of the monitoring results on different forums. The statistic views of the top 10 topics, sorted by the number of the documents within the cluster, are shown in Fig 1. They represent the data distribution on the forum of Finance, Sports and News respectively in (a), (b) and (c).



(a)



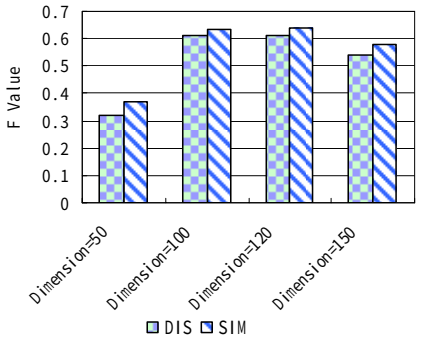
(b)



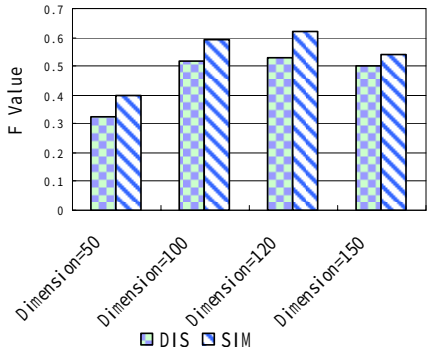
(c)

Figure 1. Views of the Top 10 Topics on Different Forums

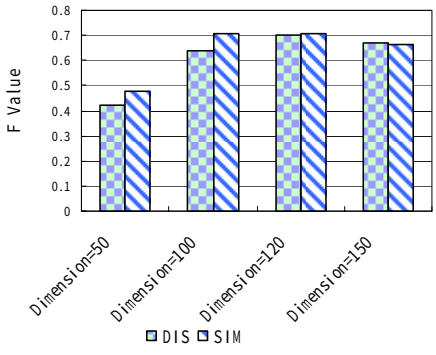
The system performance varies as the vector space dimension increasing and the results on the forum of Finance, News and Sports are shown in Fig 2 (a), (b) and (c) respectively.



(a)



(b)



(c)

Figure 2. The System Performance by the F-Value

The methods of DIS and SIM are applied for the distance computing during the process of clustering. The method of SIM method achieved better performance on the test set of all the 3 forums. Among the test data set, the performance of the forum of News is the lowest, and the difference in performance of the method of DIS and SIM gradually narrowed with the vector space dimension increasing. The dispersion of the news corpus makes it more difficult to implement the clustering. The content of each document are less and there are little effective information. Furthermore, the theme of the forum is relatively dispersive and the coherence between the documents is weak.

The system performance gradually improved and then it leveled off with the vector space dimension increasing. Finally, it even decreased. This indicates that it brings some weak distinguishing features as the dimension increasing, and this not only increase the system overhead of running time and space , but also bring the negative impact on system performance. We also give the t-test results which are shown in table 1. The system gets the significantly improvement when the dimension increased from 50 to 100.

TABLE I. .
TESTS OF STATISTICAL SIGNIFICANCE IN THE INCREMENTAL IMPROVEMENTS

Pairwise comparison	Dimension=50 && Dimension=100	Dimension=100 && Dimension=120	Dimension=120 && Dimension=150
t-test Results	t(5)=6.5007 p=0.0001	t(5)=0.4820 p=0.6402	t(5)=1.3520 p=0.2062

We also give the result by the evaluation metric of entropy and J_c. The entropy and J_c value varies as the dimension of the vector space increasing. The results on the 3 forums are shown in table 2 and table 3.

TABLE II.
THE SYSTEM PERFORMANCE EVALUATED BY THE ENTROPY VALUE

Vector Dimension	**Sports**		**Finance**		**News**	
	DIS	**SIM**	**DIS**	**SIM**	**DIS**	**SIM**
50	0.634	0.622	0.608	0.591	0.761	0.704
100	0.421	0.402	0.426	0.408	0.561	0.544
200	0.584	0.542	0.663	0.532	0.782	0.737

TABLE III.
THE SYSTEM PERFORMANCE EVALUATED BY THE J_C VALUE

Vector Dimension	**Sports**		**Finance**		**News**	
	DIS	**SIM**	**DIS**	**SIM**	**DIS**	**SIM**
50	78.74	68.92	73.33	70.40	66.25	56.82
100	42.92	41.84	43.25	40.36	41.31	39.88
200	59.88	52.56	62. 21	58.96	57.32	46.33

We find that there is a common phenomenon on the different test data set. The system gets the minimum entropy value and J_c value as the dimension of the vector space is 100. The clustering algorithm achieves the best performance. However, it reduced significantly as the dimension increased to 200. Some noise data is brought and this affects the clustering effect.

Ⅴ. CONCLUSION

The task of the public opinion detection is monitoring the information sources in various languages, and gives the warning when the new topic appears. It has broad applications, and it will provide the efficient guidance for judging the hot spots on the web.

The vector space model is adopted to represent the text information. The methods of DIS and SIM are applied for the distance computing during the process of clustering. The system performance is evaluated by the F-value, entropy value and J_c value. The results indicate that our approach is effective. We also give the t-test results

which show that the system gets the significantly improvement when the dimension increased from 50 to 100.

The task of topic detection is to detect an unknown topic in advance, and it is faced with the real-time data. The system is lack of the prior knowledge and therefore it increases the difficulty. The future research will focus on the detection of the timing news reports.

ACKNOWLEDGMENT

This work is supported by National Natural Science Foundation of China under grant No. 60803086 and Science and Technology Program of Beijing Municipal Education Commission (KM200910005009).

REFERENCES

[1] James Allan, Jaime Carbonell, George Doddington, Jonathan Yamron, and Yiming Yang. Topic Detection and Tracking Pilot Study: Final Report. In Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop:194-218, San Francisco, CA, 1998, Morgan Kaufmann Publishers, Inc.

[2] Zhang Kuo , Li J uan Zi , Wu Gang. New Event Detection Based on Indexing tree and Named Entity . Proceedings of the 30th Annual International ACM SIGIR Conference, Amsterdam , 2007.

[3] Juha Makkonen, Helena Ahonen-Myka, Marko Salmenkivi. Simple Semantics in Topic Detection and Tracking. Information Retrieval,2004,7(3-4):347～368.

[4] T.Brants, F.R.Chen, A.O.Farahat. A System for New Event Detection. Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press,2003:330～337.

[5] R.Agrawal, J.Gehrke, D.Gunopulos and P.Raghavan, Automatic Subspace Clustering of High Dimensional Data, Data Mining and Knowledge Discovery,2005 Springer Science Business Media, Inc.Manufactured in The Netherlands. 2005,pp.5–33.

[6] Sheikholeslami G., Chatterjee S.and Zhang A. WaveCluster: A wavelet-based clustering approach for multidimensional data in very large databases. The VLDB Journal.8(4),February, 2000,pp.289-304.

[7] Bjornar Larsen and Chinatsu Aone. Fast and Effective Text Mining Using Linear-time Document Clustering. In Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining: 16-22,1999.

[8] Dhillon I., Fan J.,and GuanY. Efficient Clustering of Very Large Document Collections.In Data Mining for Scientific and Engineering Applications, Kluwer Academic Publishers,2001.

Yongping Du, born in 1977. Got the Ph.D. in 2005 and graduated from FuDan University in China. Main research interests include information extraction, information retrieval and natural language processing.

Changqing Yao, born in 1974. Got the Ph.D. in 2006 and graduated from Beijing Normal University in China. Main research interests include the development of the S&T information management system; the digital rights management based on DOI.