Applied Mechanics and Materials Vols. 278-280 (2013) pp 2058-2064
Online available since 2013/Jan/11 at www.scientific.net

doi:10.4028/www.scientific.net/AMM.278-280.2058

Hot Topics Detection of Network News

Cheng-Ying CHI [1, a], Hong LI [1,b], Xuegang Zhan [1,c] and Shengnan Jiang [2,d]

[1] School of Software Engineering, University of Science and Technology Liaoning, Anshan

[2]Tourism management, Dongbei University of Finance and Economics, Dalian

[a]1149301853@qq.com, [b]liaoninglihong@163.com, [c] zhanxg@ustl.edu.cn, [d] 645018201@qq.com

Keywords: single-pass cluster, topic identification, hot topic, heat analysis

Abstract. In this paper, through analysis of the structure of web news texts, we have proposed an improvement measure for term weighting in hot topics detection, and a topic weighting scheme for hot topics ranking. Experiment result comparison shows that our method is effective and ranking of hot topics is closer to reality.

Introduction

Topic detection and tracking (TDT) is an area of information retrieval research that focuses on news events. In TDT, a news event is defined as something happening somewhere at some time associated with some specific actions. A further application of TDT techniques is hot topics detection of web news. In this paper, we present our experiment on hot topics detection of Chinese news texts, using the online clustering algorithm for topic detection.

Zeng, et al., proposed a splicing algorithm based on multi-level filtering[1]. Their method effectively replaces the strings that represent hot topics of web news texts, but it is hard to guarantee the efficiency of the Chinese word segmentation and multi-level filtering. Sun et al. proposed a second feature selection and clustering for web text, and gained pretty well results[2]. The restriction of their method is that it can only applied to web documents with apparent features. Web news texts are often semi-structured, not all of them can meet the requirement. Zhou et al. proposed an algorithm for computing the degree of relevance hot topic words in news stream, and extracted hot topic clusters by word density[3]. Their method improves the efficiency of topic detection, but cannot meet the real time requirement of hot topic detection. YE et al. proposed an algorithm based on $TF * PDF$ (Term Frequency * Proportional Document Frequency) topic heat calculation, their result shows better performance than other methods[4], but the ranking of hot topics slightly deviates from reality.

The event-based topics are dynamic, have only few relevant documents, are mostly unpredictable, and have usually a short lifespan. A hot topics detection system needs to distinguish between the same and similar events, for example, different instances of elections, riots, and train wrecks. However, news texts reporting such events are often semi-structured, the title, issue time, page view number and number of comments, etc, contain much potential information.

YE's method shed on us, and by incorporating the above information items into topic weighting process, we hope to improve the precision of hot topics detection system. Our experiment shows that the result is satisfying.

Pre-processing

Chinese Word Segmentation

Chinese text is written without natural delimiters such as white spaces, so word segmentation is often an essential first step in Chinese language processing. Though it seems simple, Chinese word segmentation is actually not a trivial problem, and has been an active research area in computational linguistics for more than 20 years. In our system, we use the ICTCLAS system developed by the Chinese Academy of Sciences. ICTCLAS's functions include Chinese word segmentation and part-of-speech tagging, named entity recognition, new word recognition, and at the same time support user dictionary. Its segmentation accuracy had reached 98.45% [5].

The document vector

News texts are represented using vector space model (VSM). For a news text vector (d) see Equation (1).

$$d_i = (t_{i1}, w_{i1}, t_{i2}, w_{i2}, \cdots, t_{in}, w_{in}) \quad (1)$$

Where, $t_{i1}, t_{i2}, \cdots, t_{in}$ are term frequency of the document, and $w_{i1}, w_{i2}, \cdots, w_{in}$ are corresponding term weight.

Term Weighting

The document vectors are constructed by ICTCLAS segmentation system, and stop words are removed through a stop-list[6]. We use the classical $TF * IDF$ for term weighting. A document is represented in a vector form with $tf * idf$ term weights through Equation (2) and Equation (3)

$$W_{w,d} = tf \cdot idf \quad (2)$$

Where, $W_{w,d}$ is a term in the document in which the Weight;

$$tf = \frac{t}{t + 0.5 + 1.5\dfrac{dl}{dl_{avg}}} \quad (3)$$

Where, tf is term frequency in the document; t is a term's number occurrences in the document; dl is the length of the document; dl_{avg} is the average document length in the document collection;

$$idf = \frac{\log(\dfrac{N + 0.5}{df})}{\log(N + 1)} \quad (4)$$

Where, idf is a term's inverse document frequency; N is total number of documents in the collection; df is the number of documents in which the term appears in the collection.

A news report generally consists of a title and content. Since a title can unveil or imply the main content of the news report, terms that appear in the title are given more importance in our system. Such terms are weighted by multiplying a weighting factor, as shown in Equation (5)

$$f(term) = \alpha f_{tittle}(term) + f_{\text{content}}(term) \quad (5)$$

α through the experimental data discussed（$\alpha = 3$）.

Hot Topic detection and analysis

The procedure of hot topics detection is shown in figure (1).

New report texts are extracted from the Internet and preprocessed. Daily news texts are clustered using the method described in section 2.3. These clusters form a topics list. Using the heat calculation formula, we obtain a hot topics list. Then we calculate the topic index to get topic development curve, so as to predict the future trend of the topic.

Topic model usually have two kinds of representation, it is the center vector and cancroids vector. Select Center vector is incorrect, resulting in incremental clustering there are a lot of errors in the future. Cancroids vectors while sensitive to noise and isolated point，but there will not be many errors. In this paper topic model using cancroids vector representation, topic vector is the topic of the average of the news text vector. See Equation (6).

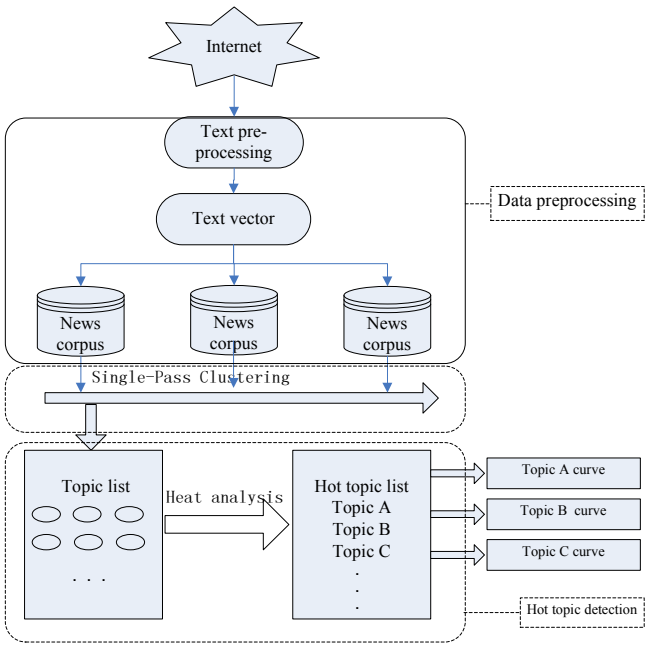

Fig. 1.Hot Topic automatically detection

$$T = \frac{1}{N} \cdot \sum_{k=1}^{N} (d_k) \qquad (6)$$

Where, T is a topic cluster center of vector; N is topic clusters containing the number of news reports ; d_k is topic clusters news Reports in vector.

Similarity between Document and Topics

We use the traditional Euclidean cosine coefficient to measure the similarity between a news report and topic representation. Suppose D is a document vector and T is a topic representation vector [7]. The cosine similarity is defined as (7)

$$Sim(D,T) = \frac{\sum_{i \in H} q_i d_i}{\sqrt{(\sum_{i \in H} q_i^2)(\sum_{i \in H} d_i^2)}} \qquad (7)$$

where, q_i and d_i reflect the weights assigned to term i in the document and topic vectors, respectively.

Online clustering algorithm for topic detection

Topic detection is new in the news will be placed in different clusters of topics, and establish a new topic cluster in time of need. In essence, the process is equal to unsupervised clustering [8]. In our system, we use an incremental clustering algorithm to detect new events. See Fig. 2.

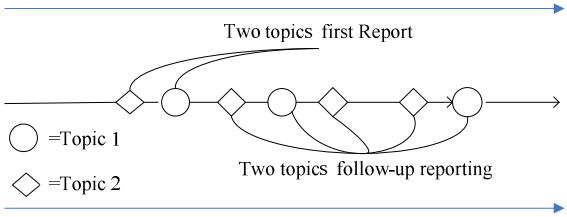

Fig. 2. Incremental clustering topic detection

Algorithm Description is as follows.
Input: The news collection.
Output: More than one topic clusters.
Algorithm process:

(1) Read a news report D_k, and using vector space model
(2) If it is the first report, then jump to (3), otherwise, jump to (4).
(3) Create a topic, and generate topic centroid vector.
(4) Find the most similar topic with D_k.

$$D^* = \arg\max_{D_{j<k}} sim(\vec{D}_k, \vec{T}_j)$$

(5) If $sim(\vec{D}_k, D^*) \geq \theta$, it will be included in the topic T_i, then updated topic centroid vector, otherwise jump to (3).
(6) If there are new reports, then jump to (1), otherwise output results.

Topic heat analysis

The algorithm proposed by YE et al. was based on $TF * PDF$ topic heat calculation. Their algorithm emphasized the number of documents a topic contained, but neglected the weight of the topic itself. Since a topic's weight is proportional to the topic's hotness (hereafter temperature), we introduce the notion of topic weight to reflect the temperature distribution and help us find hot topics more accurately. It can express the topic of heat, better find hot topics. Based on this formula for calculating heat need to be improved, as follows (8).

Topic Temperature

$$THD(j_t) = (\sum_{s=1}^{s=k} \left|D_{js_{(t)}}\right| \times \exp(\frac{D_{js_{(t)}}}{N_{s(t)}})) \times \ln(TW_j)$$
$$\left|D_{js(t)}\right| = \frac{D_{js(t)}}{\sqrt{\sum_{c=1}^{c=C} D_{cs(t)}^2}} \qquad (8)$$

where, $THD(j_t)$ is the Topic Temperature within time interval t. t can be a day, a week or a month and so on. $D_{js(t)}$ is the document number in topic j within t on website s. K is the number of websites. C is the number of topics on website s. $|D_{js(t)}|$ is the normalized topic frequency of a topic on a website. The topic frequency needs to be normalized because when different website has different size of archive, the topic from a website with more documents will have a proportionally higher probability to appear more frequently. However, we would like to give equal importance to the same topic from different website, so normalization is necessary.

TW_j is the weight of topic j. It is the average of the weight of terms in the topic vector. The calculation formula is shown in formula (9).

$$TW_j = \sum_{i=1}^{n} w_i \qquad (9)$$

where, w_i is a weight of term in topic centroid vector.

$\exp(\frac{D_{js_{(t)}}}{N_{s(t)}})$ is the *PDF* (proportional document frequency) of a topic on a website. It is the exponential function of the number of documents on topic *j* to the total number of documents on website *s*. We believe that topics that include many documents are more valuable than those that include fewer documents, hence the documents about a topic appear more frequently in documents archive on a website should have higher possibility to be a hot topic on this website. *PDF* thus should grow exponentially in respect to the number of documents about a topic, instead of linearly, so that we can give a more significant weight to the topic that is being discussed with many documents compared with those that have only a few documents. Mathematically, for a specific topic, the more number of documents on a web site, the larger *PDF* it has. There are a series of values *PDF* can take, from 1 (e^0) to 2.718 (e^1).

Topic index and topic development curve

Not only do we want to discover hot topics, but also to study its development trends [9]. Just like stock index in stock exchange, we introduce the notion of *topic index* to reflect the development trend of hot topics. Topic index is calculated as follows.

$$t_x = \frac{THD(j_x)}{HTD_{base}} \times t_{base} \qquad (10)$$

where, $THD(j_x)$ is the x^{th} day's (month's) hot topic temperature. HTD_{base} is the first day's (month's) hot topic temperature. Both of these values in formula (9) can be known. t_{base} is the first day's topic index ,when $t = 1$. The value of t_{base} is 100 in our experiment. Based on the topic of the day, we an form a hot topic development curve. By analysis of the curve, we can forecasts of the trend of the hot topics in the future.

Experiment

Experimental data

Our system took the input data from three popular news websites www.people.com.cn, www.163.com and www.xinhuanet.com from July 23 to August 16. These include the "Wenzhou cars event" and "9th Typhoon plum flower" and other related reports.

Experimental results and analysis

Since there is not a uniformed evaluation or testing standard for hot topic detection and ranking. There are two methods for evaluation of our experiment presented in this paper. First, by comparison with the official website list of hot topics, as shown in Table 1. Second, comparison between the results with improved topic temperature calculation and that of the unimproved, as shown in Table 2.

From Table 1, we can see the hottest topic is the "9th Typhoon plum flower" during the time period from July 23 to August 16. It was followed by "Wenzhou cars event", "Guo Meimei events", "Commending martyrs regulations", "Chromium pollution in Yunnan Qujing events". Because "Dalian PX project events" by "9th Typhoon" caused，So "9th typhoon plum flower" event contains "Dalian PX project event". By comparison with the "Military Dog" public opinion tracking system and found through heat calculation formula to be hot topics of most hot events that contain "Military Dog" public opinion tracking system, so that the hot topic is calculated to be a very good topic for the heat.

Table 1 Comparison Of Hot Topics

No.	Hot Topic（this paper）	Military dogs public opinion system
1	9th typhoon plum flower	Dalian PX project event
2	Wenzhou cars event	Railway wenzhou moving car event dynamic
3	Guo Meimei events	After working GuoMeiMei events of the Red Cross
4	Commending martyrs regulations	False 2011 document no. 47
5	Chromium pollution in Yunnan Qujing events	Chromium pollution in Yunnan Qujing events

Table 2 Result Comparison Before And After Topic Formula Improvement

Hot topic (original formula)			Hot topic (improved formulas)		
No.	Hot topic	Tot topic degree	No.	Hot topic	Tot topic degree
1	Wenzhou cars event	2.974222	1	9th typhoon plum flower	4.274174
2	9th typhoon plum flower	2.790449	2	Wenzhou cars event	4.001353
3	Commending martyrs regulations	0.632163	3	Guo Meimei events	0.634769
4	Guo Meimei events	0.575092	4	Commending martyrs regulations	0.465718
5	The new marriage law	0.08165	5	Chromium pollution in Yunnan Qujing events	0.042054

From Table 2, we can see the ranking of hot topics changed in the improved formula. For example, the “9th typhoon plum flower” rose from 2nd to 1st place. The “Wenzhou cars event” drops to the 2nd place. This is because the further away from the July 23 (Wenzhou cars event’s start time). Major website would be more concerned about other events. The temperature of "Chromium pollution in Yunnan Qujing events" started to grow. Therefore, the topics ranking is closer to reallity.

Topic development curve

Table 2 shows only the sorted ranking of hot topics, and does not reflect the development trends of theses hot topics. The event-based topics are dynamic. Here, we introduce the notion of development curve. It will be very clear about the display of dynamic change. Table 3 is the index of topics by formula (8), (9), and (10). In order to explain the significance of topic index and the development curve, we analyze the hot topic development curve of the “Wenzhou cars event” and the “9th Typhoon plum flower”, as shown in Fig.3 and Fig.4

Table 3 Topic index

Tot topic	7.24	7.26	7.29	8.01	8.04	8.07	8.10	8.13	8.16
9th typhoon plum flower	0	0	0	0	0	100	173.53	169.84	201.11
Wenzhou cars event	100	143.90	197.99	211.27	190.59	200.63	174.46	143.42	130.30
Guo Meimei events	100	112.18	123.78	192.70	216.49	275.79	313.51	93.16	76.79
Commending martyrs regulations	100	120.37	87.83	150.02	145.18	190.25	210.83	140.15	106.15
Chromium pollution in Yunnan Qujing events	100	102.24	138.64	164.73	231.89	258.02	294.14	185.85	135.46

From Fig.3, we can see "Wenzhou cars event" increased significantly from July 24 through August 1. It shows that this piece of news caused widespread concern. Each of the website reported the news with wide coverage. On August 1, the "Wenzhou cars event" curve was in the highest point; it shows that the topic has reached the highest temperature. From August 4 to August 16, the “Wenzhou cars event” curve displays a steady decline, it shows that the topic then gets cooling down. From Fig.4 we can see the topic development curve of the “9th Typhoon plum flower” substantially increased during this period, each website shows greater concern about the “9th Typhoon plum flower” than “Wenzhou cars event”.

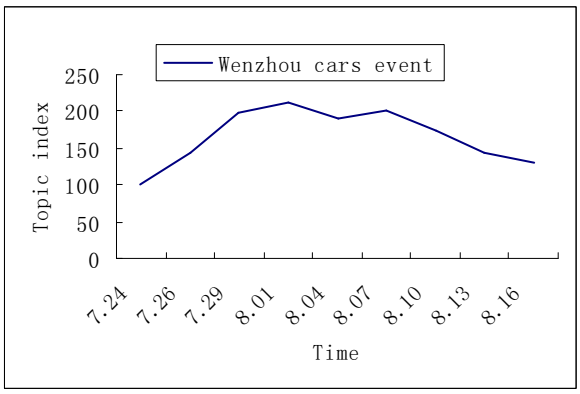

Fig.3 Wenzhou cars event development curve

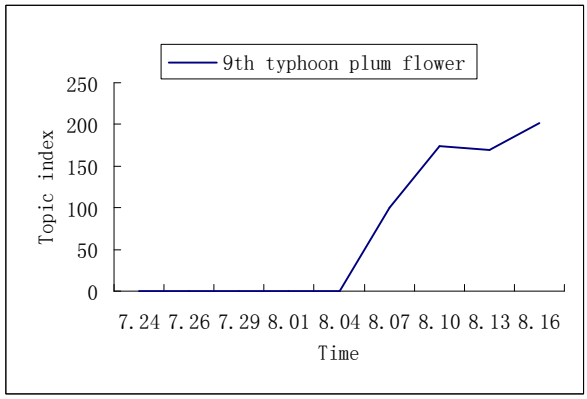

Fig.4 9th typhoon plum flower development curve

Conclusion

In this paper, by using the characteristics that news report generally consists of a title and content, we have introduced the notion of topic weight, and emphasized the term weight of words appearing in the titles. These measures greatly improve the accuracy of the algorithm proposed by YE, et al. Our experiment shows that the result is satisfying.

References

[1] ZENG Yi-lin, XU Hong-bo.Research on Internet hotspot information detection[J]. Journal on ComlTlunications, 2007. 28（12）:141-146

[2] SUN Xue-gang, CHEN Qun-xiu, MA Liang. Study on Topic-Based Web Clustering[J]. Journal of Chinese Information Processing, 2003, (3) : 12-16

[3] Zhou Yadong , Sun Qindong, Guan Xiaohong. Internet Popular Topics Extraction of Traffic Content Words Correlation [J]. Journal of Xian Jiaotong University, 2007, 41(10): 1142-1145, 1150

[4] YE Hui-min, CHENG Wei, DAI Guan-zhang. Design and Implementation of On-Line Hot Topic Discovery Model. Wuhan University Journal of Natural Sciences(WUJNS). 2006. 21-26

[5] http://baike.baidu.com/view/1215398.htm

[6] Zhu Hengmin, ZhuWeiwei. Study on Web Topic Online Clustering Approach Base on Single-Pass Algorithm[J]. XianDai TuShu QingBao JiShu, 2011, 213(12).

[7] Gerard Salton, Christopher Buckley. Term-Weighting Approaches in Automatic Text Retrieval. Information Processing and Management. 1988.5(24):513-523

[8] Li Baoli ,YU Shiwen. Research on Topic Detection and Tracking[J]. Computer Engineering and Applications. 2003, 39(17): 7-10.

[9] LuoYaping, WangCong, Zhou Yanquan. Study on Hot Topic Discovery Model Based on Attention Degree [A]. In: Proc. Of the 7th International Conferences on Chinese Information Processing [C]. WuHan:Chinese Information Processing Society of China, 2007:402-408

Advances in Mechatronics and Control Engineering
10.4028/www.scientific.net/AMM.278-280

Hot Topics Detection of Network News
10.4028/www.scientific.net/AMM.278-280.2058