One Method for On-Line News Event Detection Based on the News Factors Modeling

Hui Zhang[1] and Guo-hui Li[1,2]

[1] Department of Engineering, School of Information System and Manegement,
National University of Defense Technology, Changsha, 410073, China
[2] Science and Technology Foundation of State Key Laboratory of Information System
Engineering, National University of Defense Technology, Changsha 410073, China
zhanghui_nudt09@yahoo.com.cn, guohli@nudt.edu.cn

Abstract. On-line news event detection is detecting the first news report of a news event from various news sources in real time. Related to on-line news event detection, in this article, the author firstly introduces a news representation method for the news factors modeling based on the time, locations, characters (or organization), contents, and so on, and deducing a method related to the features of different types of news factors to calculate the weight of those news factors. Considering the insufficient of the traditional detection algorithms, then the author presents the algorithm of Micro-clusters-based on-line news event detection with Window-Adding and conducts an experiment based on news data which is collected in reality. The author achieved a satisfied experimental result verifying the validity of the proposed method.

Keywords: News factors modeling, News event detection, Hierarchical clustering algorithm with agglomeration.

1 Introduction

Nowadays, Internet technology is developing rapidly. On-line news is gradually surpassing the traditional mediums such as newspapers and TV and becoming the first-hand access for people to receive news and information. Even many traditional mediums also have established websites to report news and supply video-news and download services. Internet is a vast amount of information database. As one of the most primary information carriers on Internet, on-line news reports thousands of news events happening anywhere in the world day after day, including political, economic, military, culture, religion, conflict, and other aspects.

How to search the latest news events from the vast amount of news data on Internet is the problem that on-line news event detection, ONED [1] going to solve. In this article, according to news factors, the author establishes news reports modeling. Aiming at the insufficient of the traditional detection algorithms, the author introduces the algorithm of Micro-clusters-based on-line news event detection with Window-Adding.

Y. Wang and T. Li (Eds.): Knowledge Engineering and Management, AISC 123, pp. 427–434.
springerlink.com

2 The News Report Modeling Based on News Factors

A complete news report generally includes the following four elements: time (When), locations (Where), characters (organization) (Who), and contents (What), named the four Whs. These elements are the key to understand news event. The traditional detection methods usually combine the four elements to form a simple vector to make the similarity comparison between news reports. However, because jumping the key role to classify events through the time, locations, characters (organization), and other news elements, those traditional methods can not solve some special circumstances efficiently. Such as, the same or similar contents (consistent theme) which belong to different news events.

2.1 News Reports Modeling

In the article, for each news elements, the author assigned a semantic class, which is a set of words having the same type meaning. Location semantic class contains all place names related to a news report; time semantic class contains all time nouns related to a news report; characters (organization) semantic class contains all person's and organization names related to a news report; content semantic class contains other news report elements other than time, locations, characters (organization).

According to news four elements (When, Where, Who, WHAT, 4 Whs), the author creates modeling of one news report responding to different semantic classes. This modeling method is to express each news elements as a sub-vector, in other words, the time sub-vector $T = \{t_1, t_2,... , t_n\}$, to location sub-vector $L = \{l_1, l_2,..., l_n\}$, characters (organization) sub-vector $P = \{p_1, p_2,..., P_n\}$, content sub-vector $C = \{c_1, c_2,... , c_n\}$. Therefore, a news report can be expressed as $S = \{T, L, P, C\}$, which each elements in S is a real sub-vector. The time, locations, characters (organization) sub-vector elements are extracted from news report by named entity relation extraction technology (NEE) [2] . Sometimes, the same time, locations, characters (organization) may appear in the same report for several times and on different places; thus, the degree of importance that the four elements reflects in reports will be various. The following formula [3] is for calculating weight for time, locations, character (organization), and other name entity features.

$$\omega(f,d) = \log\left[(1 + \frac{C_d}{N}) \times \frac{1 + (N - L_d)}{\sum_{j=1}^{N} j} \right] \tag{1}$$

C_d is the frequency how the feature f appears in the report d. N is the sum of times that each entity appears in the name entity which belongs to the feature f . L_d is the first place where the feature f appears in d.

2.2 The Calculation Method of the Featured Weight of Content Sub-vector

To calculate the featured weight of content sub-vector is an important part of the modeling based on news elements. In the article, the traditional $tf * idf$ method to calculate weight is applied [4]. The method works by using a single news report as the basic statistics unit to calculate featured weight and reflects the importance degree that features appear in reports.

The similarity of each of sub-vectors is measure by new events detection Hellinger [5] methods, in other words, the sum of square root of the product of the corresponding featured weight. Hellinger methods express as following:

$$sim(d_1,d_2)=\sum_{1\le i\le|d_1\cap d_2|}\sqrt{\omega(f_i,d_1)\times\omega(f_i,d_2)} \tag{2}$$

3 The Algorithm of On-Line News Event Detection

The traditional on-line news event detection employs the statistics principle-based text representation form. The most commonly used represetation method is vector space model. The similarity between events and reports is accordingly calculated by cosine included angle formula [6] and Hullinger distance calculation formula [5]. Detection algorithms commonly use agglomerative method [1], single-Pass method [7] and increment *k*-mean method [7], etc. Carefully analysing traditional method, it can be seen that the traditional methods have a lot of drawbacks: 1) the biggest flaw of the method is not able to efficiently distinguish different events having the same topic; 2) the noise informations in event space interfere new event detection; 3) it increases the time comsumed for the traditional similarity comparisons strategy in on-line detection; especially, when data flow rate is large.

According to the deficiency of the traditional algorithms, in this article, the author applys the algorithm of Micro-clusters-based on-line news event detection with Window-Adding (MONEDW). In the window, the system can adopt condensedly hierarchical clustering algorithm to conduct micro-clustering, generating micro clusters, and use a earliest report representation of time stamp for micro cluster. By setting a higher similarities threshold value, the similarity of micro cluster internal report is extremely high. After the completion of window-interior micro-clustering, each micro cluster will make similarity comparisons to ready-detected event to determine whether the micro clusters describe a new events.

Using Window-Adding strategy in the algorithm, the size of a window is defined by time. In this article, the time interval for periodical collection news from web is defined to be the size of a window. The purpose is to avoid the case where intervals that cover fixed numbers of news at different time points may vary dramatically. Such as, setting the window size in hours. For example, setting one hour for the window size means adding all collected news reports in an hour to the window.

This on-line news event detection algorithm include two sub-algorithms: Algorithm 1 is micro clustering within window algorithm; Algorithm 2 is the similarity comparisons algorithm between micro clusters and events. Window micro clustering choose condensedly hierarchical agglomerative clustering algorithm. Hierarchical agglomerative algorithms find the clusters by initially assigning each object to its own cluster and then repeatedly merging pairs of clusters until a certain stopping criterion is met.

Algorithm 1: Micro Clustering within Window Algorithm

Input: News reports set $D=\{d_1,d_2, \cdots, d_n\}$ collected from window, threshold θ;

Output: Micro clusters;

Step 1: Start by assigning each item of D to a cluster, so that D have n clusters, each cluster containing just one item. Let the distances (similarities) between the clusters the same as the distances (similarities) between the items they contain;

Step 2: Find the closest (most similar) pair of clusters and merge them into a single cluster as long as the similarity between pair of clusters above the θ;

Step 3: Compute similarities between the new cluster and each of the old clusters;

Step 4: Repeat steps 2 and 3 until all similarities not above θ.

Algorithm 2: The similarity comparisons algorithm between micro clusters and events

Input: micro clusters from algorithm 1, ready-detected event model (featured weight vector) saved in memory;

Output: whether micro clusters describe new event;

Step 1: Select a micro cluster in order (choose the news reports represented the micro cluster);

Step 2: Make comparisons between the sub vectors of time, locations, characters of micro cluster and corresponding sub-vectors of ready-detected events. if each similarity of sub-vectors is greater than the corresponding threshold values, the micro clusters belongs to the event, adding the micro cluster to the event, and updating the event. Go to the first step otherwise going to the third step;

Step 3: If all the similarities of the sub-vectors of locations and characters are greater than the corresponding threshold values, the micro cluster belongs to the event, adding the micro cluster to the event, and updating the event. Go to the first step otherwise going to the fourth step;

Step 4: If all the similarities of the sub-vectors of time and characters are greater than the corresponding threshold values, the micro cluster belongs to the event, adding the micro cluster to the event, and updating the event. Go to the first step otherwise going to the fifth step;

Step 5: Comprehensively calculate the similarities of four sub-vectors, and weighting sum the result. Finally, make comparison to threshold values. If the similarity is greater than threshold value and the similarities of time and location are greater than the corresponding threshold value, the micro cluster belongs to the event, adding the micro cluster to the event, and updating the event. Go to the first step; Otherwise, the cluster is a new event. The news report which is on behalf of the micro cluster is the first report of the event, creating a new event mark and the representation model of the new event, saving to local, going to the step 1.

4 Experimental Results and Evaluation

4.1 Experiment Data

This section presents our experiments performed on news report texts which come from CNTV's XinWenLianBO. All the news report texts are collected from March

2006 to July 2008, including a total of 25776 news stories, each news unit is treated as a text. In our experiments, we select a total of 4214 news stories, which belong to 18 event, each news story associated with a reporting category being a member of event. As the news programs of XinWenLianBo report only a variety of important news, so the number of news reports in each category is uneven distribution, the number of various types of news stories used in this study range from 26 to 828. We randomly select the 30% samples of the total number of each category as being training samples, the remaining 70% samples as being testing samples.

Topic detection and tracking(TDT) is often used Recall Rate(R), the Precision Rate (P), the Miss Rate(P_M), the False_alarm Rate(P_F), Normalized System Cost(C_N) to evaluation the performance of news event detection [7]. This study refer to the performance evaluation criterion of TDT, using the same evaluation system.

4.2 The Result of Experiment

First of all, for the sake of obtaining the performance difference between our method and traditional method, we compare this news event detection method with the traditional method which use single vector of features splited from news story for new event detecton. There are two tables which display the result of comparison, table 1 shows the performance of news event detection based on multi-semantics class, table 2 shows the performance of news event detection based on single vector of features.

Table 1. The Performance of News Event Detection based on Multi- semantics Class

ID	Event Name	R(%)	P(%)	P_M(%)	P_F(%)	C_N
1	Fire	97.50	98.73	2.50	0.02	0.0262
2	Flood	96.72	96.72	3.28	0.05	0.0351
3	Earthquake	87.08	83.61	12.92	2.98	0.2754
4	Storm	97.17	90.35	2.83	0.27	0.0414
5	Olympics	89.13	88.60	10.87	2.81	0.2462
6	Avian Influenza	100.00	96.3	0.00	0.02	0.0012
7	Taiwan Problem	96.46	97.32	3.54	0.07	0.0390
8	Korea Nuclear	97.37	97.37	2.63	0.02	0.0275
9	United Nations	99.44	98.90	0.56	0.05	0.0080
10	America	89.81	89.32	10.19	1.01	0.1516
11	Russia	88.80	86.47	11.20	0.91	0.1566
12	Japan	89.86	86.71	10.14	0.96	0.1487
13	Iran	97.54	94.82	2.46	0.33	0.0406
14	Iraq	95.77	93.15	4.23	0.25	0.0543
15	Terrorists Attacks	96.47	95.35	3.53	0.10	0.0400
16	Countryside	88.96	86.27	11.04	2.55	0.2353
17	Oil Price	99.12	99.12	0.88	0.02	0.0100
18	Football	96.55	100.00	3.45	0.00	0.0345
Total		94.65	93.28	5.35	0.69	0.0873

Table 2. The Performance of News Event Detection based on Single Vector of Features

ID	EventName	R(%)	P(%)	P_M(%)	P_F(%)	C_N
1	Fire	85.00	91.89	15.00	0.15	0.1571
2	Flood	80.33	87.50	19.67	0.17	0.2050
3	Earthquake	86.12	82.44	13.88	3.21	0.2959
4	Storm	87.74	85.32	12.26	0.39	0.1417
5	Olympics	88.65	88.22	11.35	2.89	0.2553
6	Avian Influenza	88.46	82.14	11.54	0.12	0.1212
7	Taiwan Problem	84.07	87.16	15.93	0.34	0.1760
8	Korea Nuclear	84.21	82.05	15.79	0.17	0.1661
9	United Nations	88.33	89.83	11.67	0.45	0.1385
10	America	85.95	89.40	14.05	0.96	0.1876
11	Russia	87.64	85.34	12.36	0.99	0.1719
12	Japan	88.77	85.37	11.23	1.07	0.1646
13	Iran	89.34	88.26	10.66	0.73	0.1424
14	Iraq	88.73	89.36	11.27	0.37	0.1307
15	Terrorists Attacks	84.71	86.75	15.29	0.27	0.1660
16	Countryside	87.71	85.45	12.29	2.69	0.2546
17	Oil Price	87.72	90.09	12.28	0.27	0.1360
18	Football	89.66	83.87	10.34	0.12	0.1093
Total		86.84	86.69	13.16	0.85	0.1733

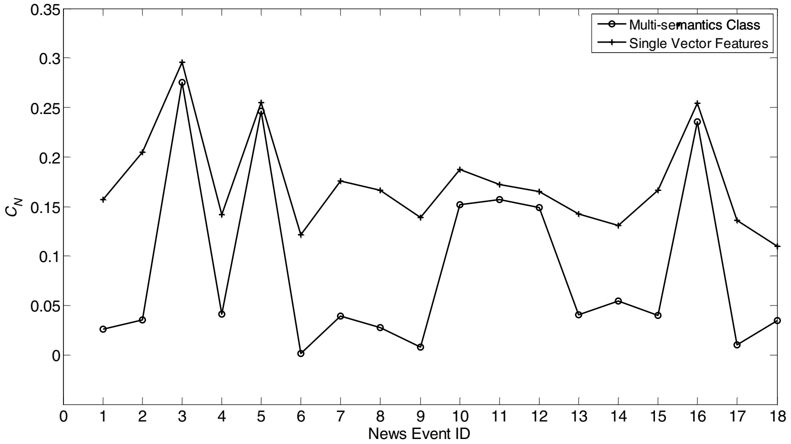

Fig. 1. The Normalized System Cost(C_N) Comparison with Different Text Representation Model

From the experimental results can be found that the news event detection algorithm based on news elements of this study obtains the average Recall Rate of 94.65%, the average Precision Rate of 93.28%, it is 7.81% and 6.59% separately higher than the traditional algorithm which is based on single vector of features. Two kinds of detection algorithms for events, there are some performance criterion including the Miss Rate, the False_alarm Rate and the Normalized System Cost being relatively high for the Earthquake, the Olympics, the Taiwan Problem, the United States, the Russia, the Japan, the Countryside, it is because that these types of events cover a wide range and topics are more dispersed. Figure 1 shows that the Normalized System Cost(C_N) Comparison with Different Text Representation Model.

To further verify the validity of the proposed method, we made a set of comparative experiments, the methods used for comparison are reported by literature [8], including C3 method, CMU method and Dragon method, their performance criteria are relatively better. The results of comparison shown in Figure 2. Figure 2 shows that the Normalized System Cost is the lowest for the proposed news event detection algorithm in this study, the comprehensive performance is the best in the four methods. The performance of Dragon method is the worst in the four methods, the Normalized System Cost reach about 28%, the CMU method in regard to the C3 method and the Dragon method performed better. The algorithm presented in this paper significantly improve the system detection performance.

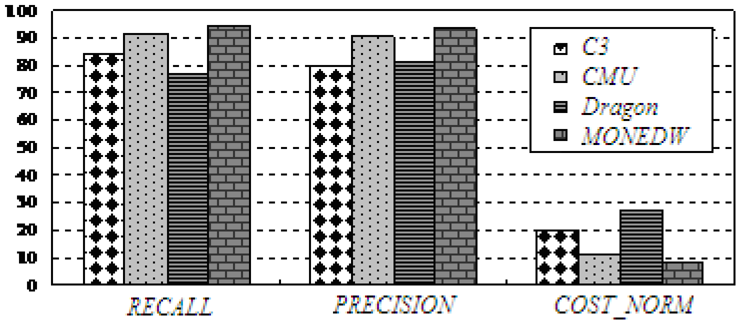

Fig. 2. The Performance Comparison with Different Methods for News Event Detection

5 Conclusion

This paper, being against the lack of traditional event detection algorithm for news and combining with news features, proposed a new algorithm, which is called micro-clusters-based on-line news event detection with window-adding(MONEDW) and make use of news elements for modeling news document. From the experimental results, the algorithm avoied the mutual interference between different events, which subject to the same topic, and promoted the efficiency and accuracy of online news event detection, is a feasible way for news event detection.

Acknowledgments. Thanks to Lian Lin, Chen Jun for discussions on News Event Detection, to Liu Fengzeng for his spell check, and to anonymous reviewers for their comments.

References

1. Allan, J., Papka, R., Lavrenko, V.: On-line new event detection and tracking. In: Proceedings of the SIGIR 1998: 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 37–45. ACM Press, New York (1998)
2. Zhang, H.-P., Liu, Q., Yu, H.-K., Cheng, X.-Q., Bai, S.: Chinese Name Entity Recognition Using Role Mode. Special Issue Word Formation and Chinese Language Processing of the International Journal of Computational Linguistics and Chinese Language Processing 8(2), 29–602 (2003)
3. Fu, Y., Zhou, M.-Q., Wang, X.-S., Luan, H.: On-line Event Detection from Web News Stream. ICPCA 2010 5th International Conference on Pervasive Computing and Applications, 105–110 (2010)
4. Sun, J.: The Topic Tracking Research with Document Title. Beijing City College, Beijing (2006)
5. Makkonen, J., Ahonen-Myka, H., Salmenkivi, M.: Simple Semmantics in Topic Detection and Tracking. Information Retrieval 7(3-4), 347–368 (2004)
6. Chen, F., Farahat, A., Brants, T.: Multiple Similarity Measures and Source-pair Information in Story Link Detection. In: Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, pp. 313–320. Association for Computational Linguistics, Boston (2004)
7. Papka, R.: On-line new event detection, clustering, and tracking. University of Massachusetts Amherst (1999)
8. Seo, Y.W., Sycara, K.: Text Clustering for Topic Detection. Tech. Report CMU-RI-TR-04-03, Robotics Institue, Carnegie Mellon University (January 2004)