News Topic Detection based on Hierarchical Clustering and Named Entity

Sheng HUANG, Xueping PENG, Zhendong NIU*, Kunshan WANG
The School of Computer Science and Technology
Beijing Institute of Technology
Beijing, China
{20812036, pengxp, zniu, wks}@bit.edu.cn

***Abstract*—News topic detection is the process of organizing news story collections and real-time news/broadcast streams into news topics. While unlike the traditional text analysis, it is a process of incremental clustering, and generally divided into retrospective topic detection and online topic detection. This paper considers the feature changes of modern news data experienced from the past, and presents a new topic detection strategy based on hierarchical clustering and named entities. Topic detection process is also divided into retrospective and online steps, and named entities in the news stories are employed in the topic clustering algorithm. For the online step's efficiency and precision, this paper first clusters news stories in each time window into micro-clusters, and then extracts three representation vectors for each micro-cluster to calculate the similarity to existing topics. The experimental results show remarkable improvement compared with recently most applied topic detection method.**

Keywords-news topic detection; agglomerative hierarchical clustering; vector space model; named entity

I. INTRODUCTION

With the highly developed modern Internet media, the web capacity doubles approximately every two years. News stories have sprung up in abundance via news websites, discussion forums and blogs every day. The fast growth of information has posed new challenges on how to efficiently and accurately manage and retrieve the textual collections. Most existing systems lack the ability to cope with the challenges. Topic Detection and Tracking (TDT) is one of the research projects that aim at developing effective techniques for solving these challenges. Topic detection is one of the core tasks in TDT, aims at discovering news topics from news story collections and online story/broadcast streams.

A topic is usually defined as "a set of news stories that are strongly related by some seminal real-world event", and an event is defined as "something (non-trivial) happening in a certain place at a certain time" [1]. For instance, Japan Earthquake happened in 11.3.2011, is the seminal event that triggers Japan's earthquake and nuclear pollution topic. Any stories that discuss the earthquake, or the rescue attempts for survivors, the caused tsunami, rebuilding plan and so on, are all parts of the topic. Topic detection systems are often used in domains like financial markets, news analysis, and intelligence gathering etc.

In this paper, under deep analysis of modern news stories data, we propose a novel topic detection strategy based on hierarchical clustering and named entities. We divide topic detection process into retrospective and online topic detection steps. For the sake of system efficiency and precision, agglomerative hierarchical clustering and single-pass clustering algorithms are used in retrospective detection and online detection respectively. Named entities have been proved to be useful in information extraction field [6, 7, 9]. They are also topic-indicative features in topic detection process. Two different news stories with the same named entities are more likely belonging to the same topic. Experiment results have showed remarkable improvement in section 6. Although we only conduct experiments on news stories data, our proposed approach is also applicable for latest news media data, such as discussion forums, blogs, micro blogs and social networks data.

The rest of the paper is organized as follows. In section 2, there is a brief review on related works of topic detection. Section 3 describes the basic model for news topic detection that we share with most current systems. The agglomerative clustering strategy based on named entities is described in section 4. Section 5 is the introduction of the dataset and evaluation metrics used in our experiment. Section 6 shows comparisons of experiment results between our method and recent most applied method. At last, we summarize the whole paper and predict future works.

II. RELATED WORKS

Topic detection has always been a challenging problem in text analysis area [3, 4, 8, 10, 14, 15, 16]. Most of the state-of-art topic detection systems are based on clustering techniques and Vector Space Model (VSM) model. Under the frameworks, researchers have proposed several practical methods for topic detection and tracking, like agglomerative hierarchical clustering, single-pass clustering and incremental K-means clustering etc.

By using density function to initialize the cluster centers, [3] proposed an improved incremental K-means clustering method for news event detection. Ref. [10] presented a scalable architecture for hierarchical topic detection. Ref. [11] proposed

*Corresponding author, email:zniu@bit.edu.cn

978-1-61284-729-0/11/$26.00 ©2011 IEEE

a new incremental hierarchical clustering algorithm which combined both partition and agglomerative clustering approaches. A method of comparing two stories by finding three cosine similarities based on names, topics and the full text was presented in [4] to finish the new event detection task. It treated the new event detection as a binary classification problem with the comparison scores serving as features. And [8] proposed an improved agglomerative hierarchical clustering to accomplish topic detection and an improved single pass clustering algorithm to finish topic tracking. Above mentioned methods mostly only focused on improving the clustering algorithms and system efficiency, and mostly based on traditional VSM model. But they totally ignored that some kinds of feature terms are more topic-indicative than others.

Recent researchers have also realized the important role of topic-indicative terms. Ref. [8] considered that the feature terms appeared in story title will contribute more during the similarity calculation between two stories, and should often be given more weights. But most story titles are too brief to represent whole story content features. Ref. [14] represented each story with 3-dimension vector model and give different weight value to each dimension feature. But it has poor time efficiency in online topic detection process, and cannot be applied to new media data, like micro blog data.

In this paper, we find that modern news data have experienced many major changes from the past. They can happen suddenly and randomly, and have more wide scope and influence. The stories about the same topic become more emergent and intensive in a short period of time, and may span a longer time interval. We can compare the topic evolution process to a life cycle, and its evolution process can be divided into trigger event, evolution stage and withering stage etc. In this situation, since two stories of the same topic usually have little features in common, named entities become more topic-indicative features in the story collection. Therefore named entities are identified, and given more weights. The two-steps clustering strategy is used to handle massive data.

III. BASIC MODEL

A. Preprocessing and Feature Representation

The stories data are collected from different web sources, so the data cleaning is first needed. We remove duplicated stories in the data collection, and use ICACLAS 3.0 Chinese Words Segmentation System for word segmentation and part-of-speech tagging. Common stop words are removed, some story domain-specific stop words are also removed.

For the sake of simplicity and effectiveness, each story is represented by a vector of weighted terms (features). In the related research of topic detection and tracking, TF-IDF and incremental TF-IDF are the two most popular models of calculating feature weight. We used TF-IDF model for retrospective topic detection, and the incremental TF-IDF model for incremental clustering. Each cluster was represented by a vector of frequently occurring features' weights in our strategy.

B. Feature selection and Weighting

Modern news stories often have been normalized by web portals, and usually have the properties of anchor, title, keywords, description and content etc. We treated story anchor, keywords and description the same as the title.

In retrospective topic detection, we used traditional TF-IDF model to compute the feature weight. Then we normalized the gained feature vector. The TF-IDF weight is calculated as follows:

$$Weight(t,d) = Weight_t * \frac{d_t}{\| d \|} * \log(\frac{N+0.5}{N_t+1}) \quad (1)$$

Where d_t means the frequency of term t in story d, $\| d \|$ represents the terms length in d. N is the stories collection size, N_t and represents the number of stories which consist of the term t.

$Weght_t$ is the weighting factor for terms in the title part or the named entities in the story. We set $Weght_t = 1$ when term t is not the title words or named entities. In the experiment section, we compare our proposed reweighting method with the title words based reweighting method proposed in [8] using a large dataset. We set the $Weght_t$ value when the system achieves the best performance.

In online topic detection, we use incremental TF-IDF model to compute the feature weights [5, 8]. Then we normalize the obtained feature vector.

$$Weight(t,d) = Weight_t * \frac{d_t}{\| d \|} * \log(\frac{N_c+0.5}{N(t,c)+1}) \quad (2)$$

Where c represents the current time, N_c denotes the total number of stories at the current time c, $N(t,c)$ means the total number of stories which contain term t at the current time c [8].

C. Similarity Calculation

We use cosine similarity to calculate the similarity between two stories. It is a good reflection of vectors' similarity and the variations of vector's elements. For stories d_1 and d_2, the similarity between them is calculated as follows:

$$Similarity(d_1,d_2) = \frac{\sum_{t \in \{d_1 \cap d_2\}} Weight(t,d_1) * Weight(t,d_2)}{\sqrt{\sum_{t \in d_1} Weight(t,d_1)^2} * \sqrt{\sum_{t \in d_2} Weight(t,d_2)^2}} \quad (3)$$

IV. AGGLOMERATIVE HIERARCHICAL CLUSTERING

Different from traditional textual collection clustering, we consider news topic detection as an incremental clustering process, and divide it into two steps: the retrospective topic detection step and the online incremental detection step.

A. Retrospective topic detection

Retrospective topic detection is the process of clustering existing stories collection into news topic models. We apply the

agglomerative hierarchical clustering to finish the retrospective topic detection task. We choose the last maximum similarity as similarity threshold when the retrospective topic detection achieves the best performance. A similarity matrix is used to represent the similarity between every two topics.

The procedure of agglomerative hierarchical clustering is described as follows:

1) Preprocess all stories, and represent each of them by a feature vector. We view each story as an independent topic at the beginning. Set elements of the similarity matrix as zero.

2) Recalculate the similarity between every two topics. If a topic has been merged into another topic, we set the corresponding row and column elements in the similarity matrix as -1. We use the average-link method to compute the similarity between topics. Then renew the similarity matrix. And reduce the topic number by one.

3) Find the maximum similarity from the similarity matrix. If the topic number is greater than the predefined topic number, go to step 2. Otherwise go to step 4.

4) Set the similarity threshold as the last maximum similarity. We extract topic representation for each generated cluster. Then Exit.

B. Online incremental detection

Online incremental detection is the process of merging real-time news streams into prebuilt cluster models. We use the single pass clustering to finish this subtask. We set a time window w as 24 hours. Our crawler collects the latest stories from several web portals in w. For the sake of clustering efficiency, we first clustered the stories collected in w into micro-clusters [8]. Because of modern news stories of the same topic may span a long time interval, we compare the similarity between the micro-cluster with each existing topic.

The procedure of online topic detection is described as follows:

1) Preprocess each story, and cluster the story collection in w into several micro-clusters using the agglomerative hierarchical method. The feature weight is calculated by incremental TF-IDF.

2) For each micro-cluster, we extract three representation vectors: time vector, subject vector and content keywords vector. Time vector contains the most related time terms in the micro-cluster. Subject vector contains the most appeared subject terms in the micro-cluster. The subject terms here consist of people name, organization, and location. Content keywords vector consists of keywords extracted from the micro-cluster.

3) Get a micro-cluster, calculate the similarity between it with each existing topic. We merge the micro-cluster into the most similar topic if the maximum similarity is greater than the similarity threshold, otherwise split it as a new topic. We calculate the similarity between the micro-cluster and existing topic by:

$$Similarity_{incremental} = \alpha * Similarity_{Timel'ector} + \beta * Similarity_{Subject'ector} + \gamma * Similarity_{Keywords'ector} \quad (4)$$

We set α, β as 0.4, and γ as 0.2. The similarity threshold is set as 0.054 in our experiment.

4) If there is any micro-cluster which has not been processed, go to step 3. Otherwise exit.

V. DATASET AND EVALUATION METRICS

A. Dataset

In our experiment, we use a Chinese news stories dataset, which was constructed from web pages crawled from several Chinese web portals. We collected approximately 15,000 news stories and extracted the stories contents from November to December 2010.

We have annotated 735 news stories manually which are published earlier in November 2010, to perform retrospective topic detection. It contains 29 topics. The maximum topic has 111 news stories, while the minimum one has only 5 news stories. The rest is used to perform the online incremental clustering. In our system, we try to label the rest stories automatically. Table I shows the statistics of the labeled training data.

B. Evaluation metrics

In this paper, we also employ the widely used traditional Information Retrieval evaluation metrics: Precision, Recall and F-Measure to evaluate our experiments, which were described in detail in [8, 13]. We call each stories cluster generated by our system a cluster, while each manually annotated stories topic are regarded as a topic. More specifically, for topic i and cluster j [13]:

$$\Pr ecision(i,j) = \frac{n_{ij}}{nj} \quad (5)$$

$$\mathrm{Re}call(i,j) = \frac{n_{ij}}{n_i} \quad (6)$$

Where n_i is the stories number in i, n_j is the stories number in j, and n_{ij} denotes the number of members of i in j. Then, the F-Measure of i and j is calculated by [13]:

$$F-Measure(i,j) = \frac{2*\mathrm{Re}call(i,j)*\Pr ecison(i,j)}{\mathrm{Re}call(i,j)+\Pr ecision(i,j)} \quad (7)$$

For the entire hierarchical clustering, we can get the global Precision, Recall and F-Measure by calculating the weighted mean value for the corresponding metric as follows [8, 11]:

$$\Pr ecision = \sum_i \frac{n_i}{n} \max\{\Pr ecision(i,j)\} \quad (8)$$

$$\mathrm{Re}call = \sum_i \frac{n_i}{n} \max\{\mathrm{Re}call(i,j)\} \quad (9)$$

$$F-Measure = \sum_i \frac{n_i}{n} \max\{F-Measure(i,j)\} \quad (10)$$

TABLE I. STATISTICS OF THE LABELED TRAINING DATA

Total topics	Max topic	Min topic	Total stories
29	111	5	735

VI. EXPERIMENTS AND RESULTS

Since giving more weights to words appeared in title is the most used strategy to improve news topic clustering performance [8], we choose to compare our proposed named entities based reweighting method with this title words based reweighting method. In our experiment, we choose people, time, location, organization, quantity and specific noun as the named entities. The result comparisons of three evaluation metrics are showed in Figs. 1-3 as follows.

Figs. 1-3 describe the comparisons of Precision, Recall and F-Measure performance of the retrospective topic detection process between our proposed method and the title words based reweighting method. From Fig. 1, we find that the precision of our method is almost always better than the title words based method: the best precision of title words based method is 71.2%, while when the $Weight_t$ ranges from 1.3 to 1.4, our strategy get the highest precision 92.8%. Also from Figs. 2, our method almost always performs better than title words based reweighting method: the best recall of title words based method is 93.9% when $Weight_t$ ranges from 1.1 to 1.4, while our strategy achieves the best recall performance 95.8% when $Weight_t$ ranges from 1.2 to 1.4 and from 1.2 to 2. From Fig. 3, we get the best F-Measure performance 71.6% when using title words based reweighting method, and our strategy achieves the best F-Measure 88.6% when $Weight_t$ ranges from 1.3 to 1.4.

With the total consideration of Precision, Recall and F-Measure performance, we set the $Weight_t$ coefficient as 1.4 in our system.

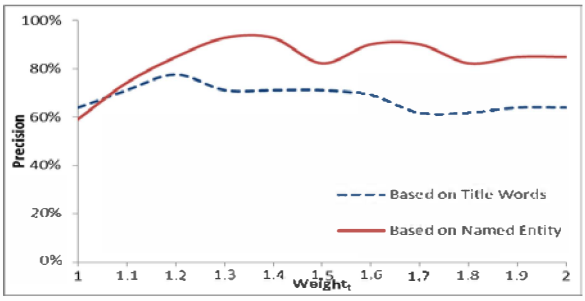

Figure 1. Precision Comparison between our method and title words based method

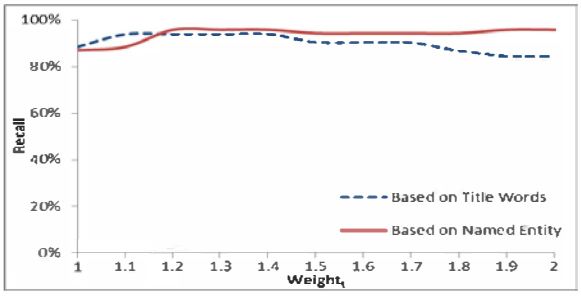

Figure 2. Recall Comparison between our method and title words based method

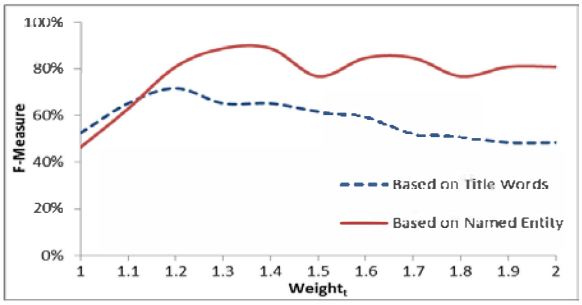

Figure 3. F-Measure comparison between our method and title words based method

VII. CONCLUSION AND FUTURE WORKS

The experiment results described in this paper lead us to believe that named entities are more topic-indicative than other feature terms, and the two-steps clustering strategy can process massive data efficiently and effectively. Compared with traditional methods, we have showed that our proposed method achieves a great improvement.

Although our proposed method achieves significant improvement compared to traditional methods, there still exist unsolved problems, such as stories data sparseness, key phrase extraction and integration of different data sources etc.

In this paper, we have fully analyzed characteristics of modern news stories data. In next step of work, we can apply and improve our method to latest news media data, like discussion forums, blogs, and micro blogs data etc.

ACKNOWLEDGMENT

This work was supported by Program for New Century Excellent Talents in University, China (grant no. NCET-06-0161, 1110012040112), Graduate Student Scientific and Technological Innovation Project of Beijing Institute of Technology (grant no.3070012240901), Key Foundation Research Projects of Beijing Institute of Technology (grant no.3070012231001), Graduate Students Research and Application Programs, Beijing Municipal Commission of Education (grant no.1320037010601) and the 111 Project of Beijing Institute of Technology. We also would like to thank greatly to the authors of ICTCLAS for providing ICTCLAS freely.

REFERENCES

[1] J. Allan, Topic Detection and Tracking: Event-based Information Organization, Kluwer Academic Publishers, 2002.

[2] National Institute of Standards and Technology of America Website, http://www.itl.nist.gov/iad/mig/tests/tdt/tasks/detech.html.

[3] L. Zhen, W. L. da, L. Lei and H. Y. yan, Incremental K-means method based on initialization of cluster centers and its application in news event detection, Journal of the China Society for Scientific Technical Information, July 2006.

[4] G. Kumaran, J. Allan, Using names and topics for new event detection, Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, pp. 121-128, 2005.

[5] Y. M. Yang, J. Pierce and J. Carbonell, A study on retrospective and on-line event detection, Proceedings of the 21st Annual International ACM SIGIR Conference, Melbourne, Australia, pp. 28-36, 1998.

[6] Y. M. Yang, J. Zhang, J. Carbonell and C. Jin, Topic-conditioned novelty detection, Proceedings of the eighth ACM SIGKDD International Conference on Knowledge discovery and data mining, pp. 688-693, 2002.

[7] S. Sekine, Named entity: History and future, Technical report, Proteus Project Report, 2004.

[8] X. Y. Dai, Q. C. Chen, X. L. Wang and J. Xu, Online topic detection and tracking of financial news based on hierarchical clustering, Proceedings of the Ninth International Conference on Machine Learning and Cybernetics, pp. 11-14, July 2010.

[9] Z. Kuo, L. J. Zi, W. Gang, New event detection based on indexing-tree and named entity, Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 215-222, 2007.

[10] D. Trieschnigg, W. kraaij, Hierarchical topic detection in large digital news archives Exploring a sample based approach, Journal of Digital Information Management, Vol. 3, 2005.

[11] A. P. Porrata, R. B. Llavori, J. R. Shulcloper, Topic discovery based on text mining techniques, Information Processing and Management, pp. 752-768, 2007.

[12] G. Kumaran, J. Allan, Text classification and named entities for new event detection, Proceedings of the 27th annual international ACM SIGIR conference on Research and development in Information retrieval, pp. 297-304, 2004.

[13] M. Steinbach, G. Karypis, V. Kumar, A comparison of document clustering techniques.KDD-2000 Workshop on Text Mining, Boston, MA, USA, pp.1-20, August 2000.

[14] Z. Hui, Z. Jingmin, W. Liang, Z. Liping, An adaptive topic tracking model based on 3-Dimension document vector, Journal of chinese information processing, Vol. 24, No. 5, Sep 2010.

[15] H. Wang, J. Zhu, D. Ji, N. Ye and B. Zhang, Time adaptive boosting model for topic tracking, In Proceedings of 2005 IEEE International Conference on Natural Language Processing and Knowledge Engineering, pp.488-492, 2005.

[16] X. Liu, F. Ren, C. Yuan, Use relative weight to improve the kNN for unbalanced text category, In Proceeding of 2010 IEEE International Conference on Natural Language Processing and Knowledge Engineering, pp.1-5, 2010.