Network Topic Detection Model Based on Text Reconstructions

Zhenfang Zhu
School of computer science and technology, Shandong University, 250100, Jinan, China
School of Information Science and Electric Engineering, Shandong Jiaotong University, 250357, Jinan,China
E-mail: zhuzhfyt@163.com

Peipei Wang
Shandong management University, 250357, Jinan, China
E-mail:wpp870213@163.com

Zhiping Jia
School of computer science and technology, Shandong University, 250100, Jinan, China
E-mail:jzp@sdu.edu.cn

Hairong Xiao, Guangyuan Zhang and Hao Liang
School of Information Science and Electric Engineering, Shandong Jiaotong University, 250357, Jinan,China
E-mail:{hairong.xiao, xdzhanggy}@163.com, lianghao3141@126.com

Keywords: topic detection and tracking, single pass algorithm, text reconstruction, network topic detection

Received: January 25, 2013

Single pass clustering algorithm is widely used in topic detection and tracking. It is a key part of network topic detection model. In the process of single pass algorithm, clustering results are not satisfactory, and the similarity matching would be reduced. Focusing on these two defects, this paper physically reconstructs web information into a volume, in which every document contains "theme area" and "details area". To improve single pass clustering algorithm, this paper uses "theme area" to detect topics and apply the whole document to distinguish subtopics, while central vector model is used to denote topics. Experimental results indicate that the model based on text reconstruction performs well in detecting network topics and distinguishing subtopics.

Povzetek: Razvita je nova metoda za zaznavanje teme omrežja na osnovi tekstovne analize.

1 Introduction

Network public opinion inclines to express the public attitudes towards social problems of the world, which are described as hot topics in Internet. The netizens focus on the hot topics by reading, releasing and quoting information of kinds forms, such as web news, BBS posts, blog articles and so on. Therefore, detecting network hot topic quickly and efficiently is the key to grasp the rules of public sentiment changing. Topic detection and tracking (TDT) technology is to provide a core technology to identify a new topic in the web information and group stories on the same topic from huge volume of information that arrives daily. TDT automatically detects hot information of public opinions, is a kind of critical technology in the field of natural language processing and information retrieval.

2 Related work

The TDT technology is intended to explore techniques for detecting the appearance of new topics and tracking the reappearance and evolution of them, and is widely used to detect network hot topics. Researchers at home and abroad have done lots of researches on network topics.

Earlier researchers focus on selecting and combining clustering algorithms. Ron Panka and James Allan [1] use a single pass clustering algorithm and a novel thresholding model that incorporates the properties of events as a major component. Ref. [2] adopts Group Average Clustering (GAC) clustering techniques to detect a novel event. The task of TDT is to automatically detect novel events from a temporal-ordered stream of news stories. In addition, Ron Papka [3] makes comparisons among many clustering algorithms, and tries to solve problems of OTD by putting clustering algorithms together reasonably.

In above researches, all the stories and there related topic are at one level, and one of stories belongs to one topic. However, the whole topic may pivot on multiple points and one story also can cover more than one topics. In order to express these characteristics, hierarchical topic detection (HTD) is put forward in TDT2004. Participants in this task are no longer required to submit flat cluster partitions, but to generate a directed a cyclic graph (DAG). Each graph has a root node, which is an

ancestor of all other clusters. And each node represents a topic at a specific granularity, which can overlap or subsume each other. Hierarchical Agglomerative Clustering (HAC) is an effective method to generate hierarchical structures. Researchers, such as Trieschnigg [6], present a scalable architecture for HTD and compare several alternative choices for agglomerative clustering and DAG optimization in order to minimize the HTD cost metric.

Civil researches pay attention to topic's hierarchy and time sequence, combine natural language processing techniques to detect network topics. Ref. [7] divides all data into groups and clusters in each group to produce micro-clusters, and then groups all micro-clusters to final topics.

This paper proposes the thought of text reconstruction and applies it to improve single pass clustering algorithm. The method both increases processing speed in single pass clustering and considers hierarchical structures of topics.

3 Topic detection model

3.1 Topic/Story Model

Every story di in a topic is expressed as di=(item1,wi1,...itemj,wij,...,itemm,wim); where wij is computed by TF-IDF. In this paper, we considers the term's position in the story when we compute term frequency, as formula (1):

$$w_{ij} = \frac{(t_{ij} \times \log(N/n_i + 0.01))}{\sum_{k=1}^{m} [tf_{ik} \times \log(N/n_k + 0.01)]^2} \quad (1)$$

Where $tf_{ij} = 5 \times tf_{ij}(title) + tf_{ij}(text)$, $tf_{ij}(title)$ is the frequency of term termj occurrence in the title of the story, m is the number of terms, N is the number of stories in the topic, ni is the number of stories which contains term ti.

Many related stories make up a topic, this paper adopt a central vector to build topic model. The central vector is described as (item1,w1,...,itemj,wj,...,itemm,wm), the weight of the term is the average of all the stories, computed by formular (2):

$$w_j(t,T) = \sum_{d_i \in ST} w_{ij} \Big/ StoryNum(t,T) \quad (2)$$

Where wj(t,T) is the weight of term itemj in the t statistical time, StoryNum(t,T) is the number of all stories in topic T at the time.

The new coming story S is expressed as (item1,ws1,..., itemj,wsi,..., itemm,wsm) by vector space model, where itemj is the term, wsi is the weight of term si in story S.

The paper adopts classical cosine similarity to compute the similarity between story S and topic T.

3.2 Topic Detection Algorithm

Single pass incremental clustering algorithm is widely used in TDT; it has simple thought and faster processing. The algorithm sequentially process documents using a pre-specified order. The current document is compared to all existing topics, and it is merged with the most similar topics if the similarity exceeds a certain threshold. Single pass clustering is discussed in detail as Table 1.

Table 1: Single pass algorithm description.

Input: the new stories
Output: some clusters
Process:
Step1 Read in a new story S;
Step2 Compute similarity $Sim(S,T_i)$ between S and each cluster existing at its processing time.
Step3 The story S is assigned to the cluster T, when $Sim(S,T) = \arg\max_i Sim(S,T_i)$ and $Sim(S,T) > \theta$ （θ is a threshold）;
Step4 If the story S fails a certain similarity test it becomes a new cluster T';
Step5 If story S is not the last, go to Step1.

4 Text reconstruction–based hierarchy topic detection model

TDT can detect network hot information which reflects public opinion, and it is the basic work of public opinion analysis. This paper focuses on improving the results of single pass clustering algorithm in TDT and introduces the thought of text reconstruction. It separates information collected from internet into "theme area" and "details area". In addition, this paper adopts central vector to represent a topic, in order to increase processing speed when a story matches each topic.

4.1 Topic Structure

Events change continuously, a topic which is defined as an event or activity, along with all directly related events and activities is also in constant change. A topic can be described as a congregation which contains web news, BBS posts, blog articles and so on. They change along with public attentions and the course of events. Take Pakistani airliner crash for example, media and people focus on "air accident", "source of damage", "care-taking arrangement".

4.2 Text Reconstruction

4.2.1 Short Text Reconstruction.

Interactive BBS forum provides a platform for people to express their sentiment and opinions, which is an area with a high incidence of public opinions. The intercommunion in BBS forum by releasing and replying posts, public sentiment and opinions is merged in these posts. Therefore, how to organize posts effectively is the key to detect network hot topic.

Ref. [8] defines "one clue" as title, main post and all responding posts. They consider that main post and responding posts revolve around the title. They introduce the idea of reconstruction to solve the problem of sparseness of the short text and get a better clustering performance.

Borrowing ideas from the above references, this paper introduces "text reconstruction". It gets typical features of a topic together, namely "theme area" and the remainder, namely "details area".

Posts in "one clue" of BBS forum usually contains plentiful information, such as title, author, main post, posts in responding to original and so on. Therefore, this paper puts title and main post together to form "theme area". "Details area" are made up of random selected responding posts. Other short texts, such as instant message, commenting on blogs, and online chat log, can be processed by the similar method.

4.2.2 Web news reconstruction.

News title contains plentiful categories information. It is the summary of the web news. The title has simple syntax and structure, and the accuracy is higher when it is used to classify web news. Ref. [9] adopts 2003 text corpus of People's Daily to test, up to 93.7% titles contains category information.

Topic is the support point of title structure. Under the same topic, every report's titles are the same or similar. Title information has significant ability of topic distinguish in TDT. But with the events' development, the distance between follow-up report's title and initial event's title may become farther and farther. The longer the time from the initial event is, the greater possibility of title drifting is. Therefore, the accuracy of utilizing title similarity to identify topic will certainly decline.

The first paragraph of news webpage outlines the basic information which includes time, place, events, characters and so on. And also numerous of category information was included in this paragraph. According to the idea of text reconstruction, combining the news webpage's first paragraph and title information which effectively gathers the typical features of topic.

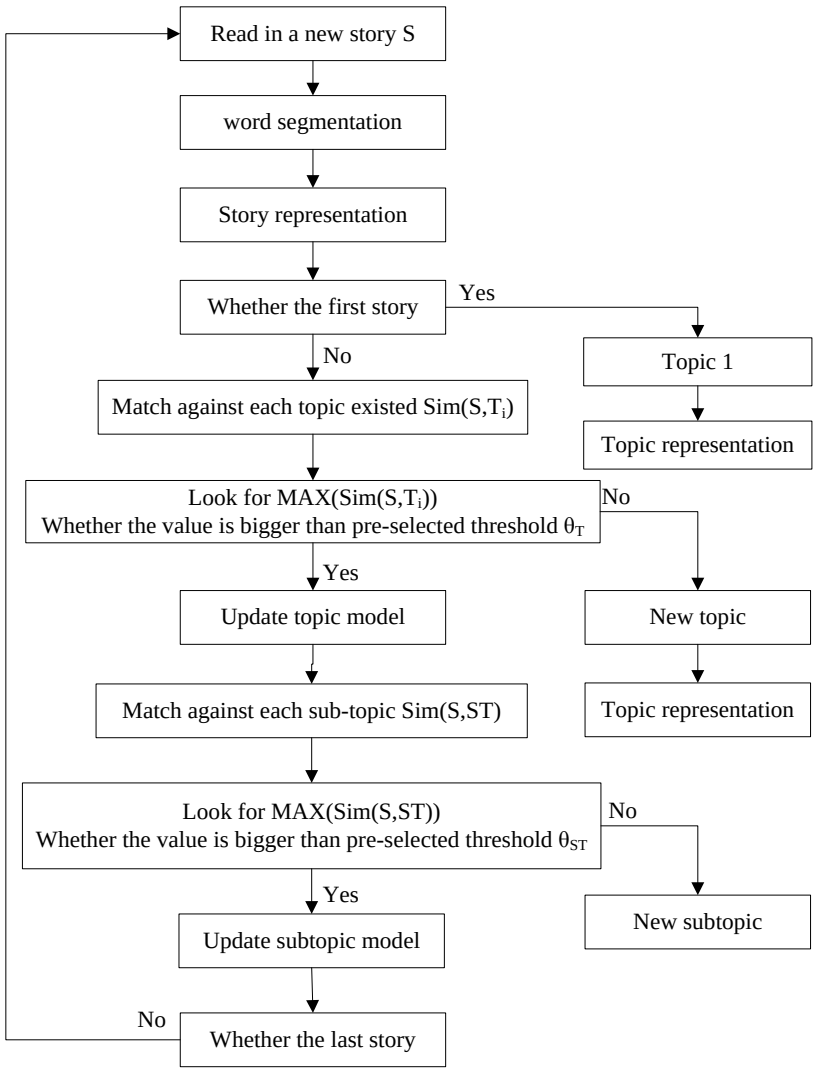

Figure 1: The improved Single pass clustering algorithm.

In TDT system, “theme area” is composed by news webpage’s first paragraph and title information. Considering the effect of news commentary in public opinion analysis, this paper uses the remaining paragraphs’ descriptive information and news comments to compose “details area”.

4.3 Hierarchy topic detection algorithm

Single pass clustering algorithm can detect network hot topics, which satisfies public opinion analysis system basically, but here lists several shortcomings:

(1) The effective of the algorithm is dependent on the order in which documents are processed. This is not a problem when the documents are temporally ordered, because the order is fixed.

(2) The common strategies for combining similarity values are known as single-link, complete-link, and average-link clustering. All comparison strategies need to compare with all documents in each existed topics. If the number of documents in a topic is on a large scale, processing speed reduces.

(3) Single pass clustering algorithm can group documents with similar content. It collects stories which belong to one topic, but it ignores topical hierarchical structure.

In order to overcome the above shortcomings, this paper uses the “theme area” to gather similar stories which belong to one topic, and adopts “details area” to divide subtopics. It also adopts central vector model to denote topics to increase processing speed, then the improved single pass clustering algorithm is described as Fig. 1.

5 Experiments and Analysis

5.1 Experiments corpus

The paper takes web news to validate effectiveness of text reconstruction–based network topic detection model. We collect thematic information and much-talked-about topic web news from sina.com.cn, 163.com.cn, sohu.com, ifeng.com, people.com.cn, xinhuanet.com. We select and clean up eight topic information, such as Tang Jun, "fake door" (TJFD), Dalian oil pipeline explosion (DOPE), Hubei officials wife incident (HOWI), Qian Wei-chang's death (QWCD), 10-year goal of developing the western region (GDWR), Luanchuan Bridge collapse incident (LBOI), Nanjing plant Explosion (NJPE), Zijin Mining Pollution (ZJMP), Table 2 gives description in detail.

To verify the effectiveness of text reconstruction in TDT, we construct two data sets: data set one contains the above eight topics, and each story is original documents; data set two still contains eight topics, but each story is reconstructed as “theme area” and “details area”.

5.2 Evaluation indexes

In the TDT setting, we chose the miss rate Pmiss and false alarm rate Pfa to measure the effectiveness of topic detection model based on text reconstruction. Pmiss is the probability that a model produces a miss, and Pfa is the probability that a model produces a false alarm. The method for calculating the measures are summarized below using the following table3:

Where the retrieved texts in the table are those that have been classified by the system as positive instances of a topic, and the relevant texts are those that have been manually judged relevant to a topic. The measures used in this paper can be computed from the table as follows:

$$P_{miss} = C/(A+C) \tag{3}$$

$$p_{fa} = B/(B+D) \tag{4}$$

A cost function (Cdet) is usually used to analyze detection effectiveness. The general form of the TDT cost function is as formular (5):

$$C_{\det} = C_{miss} * P_{miss} * P_{t\arg et} + C_{fa} * P_{fa} * P_{non-t\arg et} \tag{5}$$

Where Cdet is a cost function, Cmiss is lost cost, Cfa is false alarm cost, Ptarget is the prior probability that a document is relevant to a topic. In our expriment, we

Table 2: The experiment corpus.

Number	1	2	3	4	5	6	7	8
Topic	TJFD	DOPE	HOWI	QWCD	GDWR	LBOI	NJPE	ZJMP
Collected stories	135	119	14	42	30	139	54	94
Selected stories	100	100	14	42	30	100	54	94
Sub-topics	5	4	3	2	2	4	3	4

Table 3: The related parameter.

	Relevant	Non-relevant
Retrieved	A	B
Not Retrieved	C	D

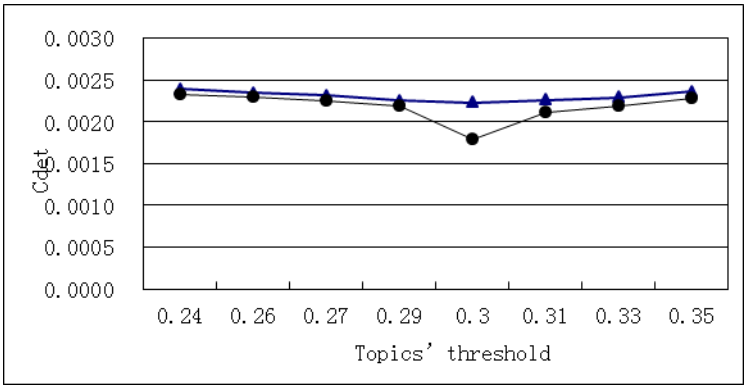

Figure2: The comparison in C_{det}.

preset Cmiss= Cfa=1.0, Ptarget=0.2.

For one topic, we use formular (5) to compute Cdet, but for all topics, we adopt (Cdet)norm, which is defined as the weighted Cdet average value of all topics, the weigh is that the number of all stories in a topic divided by all stories in all topics [10].

5.3 Experiments and results analysis

5.3.1 Topic detection and interpretation of results.The experiment is to check the ability of improved single pass clustering algorithm in detecting topics. A reasonable threshold θ_T is the key to identify topic correctly. Stories that have similarity exceeding the threshold are classified into one topic. If the θ_T is too big, the granularity of a topic is too large, in contrast, if the θ_T is too small, there will be too many topics. So it is more difficult to determine a good score that can be used as a threshold.

We select 10 stories randomly from each topic, reconstruct each story. We select cosine similarity between every story and the topic in which it belongs to. According to the values of similarities, we can conclude that:

(1)In data set one, the similarities between story and its topic are above 0.30, and the similarities with the other topics are under the 0.24.

(2) In data set two, the similarities between story and its topic are above 0.35, and the similarities with the other topics are under the 0.28.

Therefore, the topics' similarity threshold θ_T should range from 0.24 to 0.35.

We use the above corpus to compare the original topic detection model which use single pass clustering algorithm and the text reconstruction–based hierarchy topic detection model which improve the single pass clustering algorithm. We adopt 10-fork method and adopt average Cdet of all experiments to evaluate performance. The Cdet changes along with the different threshold. Fig.2 gives the detail description.

Fig.2 tells that, given the topics' similarity threshold θ_T, compared with original topic detection model, the topic detection model based on text reconstruction has smaller Cdet. It shows that the improved topic detection model performs better in identifying and tracking topics. The average cost function Cdet fluctuates along with the different similarity threshold θ_T. θ_T ranges from 0.24 to 0.30, Cdet keep in decreasing, but Cdet is in upswing when θ_T is more than 0.30. So θ_T =0.3 is reasonable in experiments.

5.3.2 Topic Structure Identification and Results Analysis.The purpose of this experiment is to check the ability of improved single pass clustering algorithm in detecting topics. We determine subtopic threshold θ_{ST} in a similar way as θ_T, it ranges from 0.4 to 0.6, and θ_{ST} =0.48 is the best in the experiment.

Take topic "DOPE" for an example, "DOPE" covers few subtopics, such as "Event overview (EO)", "Accident cause and responsibility (ACR)", "Deal with pollution (DP)", "Accident impact and compensation (AIC)" and so on. We collect five subtopics and its stories of "DOPE" artificially.

In the experiment, we present $\theta_T = 0.3$, $\theta_{ST} \in [0.4,0.6]$, and adopt text reconstruction–based hierarchy topic detection model to test. The model identifies the topic "DOPE" correctly, and separate subtopics at a certain extent. It detects five categories, which is consistent with the results of artificial markers in principle. We count numbers of stories in each subtopic collected both by hand and by the model, Fig.3 gives the details.

Fig.3 shows that the results of topic detection model based on text reconstruction are similar with the results of artificial markers. It indicates that the improved model is able to identify topic structure to some extent.

6 Conclusion

The basic work of analysing public opinion is to detect hot topics on internet and find out what people concerns, what people meet with, and what people dissatisfy. TDT groups stories to a topic automatically from huge volume of updating information in technology. This paper proposes a network topic detection model based on text reconstruction and improves the usual detection algorithm of the model. Text reconstruction makes every document into two parts: "theme area" and "details area".

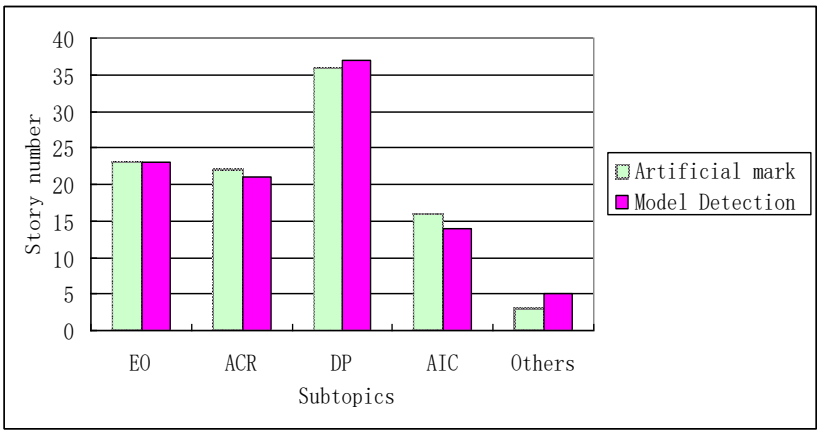

Figure 3: Subtopic Detection.

We use theme area to detect topics and apply the whole document to distinguish subtopics. Experimental results indicate that the model performs well in detecting network topics and distinguishing subtopics.

Text reconstruction–based network topic detection model shows the hierarchical structure of a topic to a certain extent, but increases the complexity of computing. In the next study, we will reform similarity calculation to improve the computational efficiency.

7 Acknowledgement

This work is supported by Shangdong Province Young and Middle-Aged Scientists Research Awards Fund (BS2013DX033), National Nature Science Foundation of China (61373148), Nature Science Foundation of Shandong Province (ZR2012FM038).

References

[1] Ron Papka and James Allan (1998). On-Line New Event Detection using Single Pass Clustering. UMASS Computer Science Technical Report UM-CS-1998-021, Amherst.

[2] Y Yang, T Pierce, J Carbonell (1998). A study on Retrospective and On-Line Event detection. Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval. CMU, USA, ACM, pp.28-36.

[3] Papka R (1999). On-line New Event Detection, Clustering and Tracking. Amherst: Department of Computer Science, UMASS.

[4] The 2004 Topic Detection and Tracking (2004) Task Definition and Evaluation Plan. version 1.2, http:/ /www.nist.gov.

[5] HONG Yu, ZHANG Yu, LIU Ting etc (2007). Topic Detection and Tracking Review. Journal of Chinese in Information Processing, 21(6), pp. 71-87.

[6] D Trieschnigg and W Kraaij (2004). TNO Hierarchical topic detection report at TDT 2004. The 7th Topic Detection and Tracking Conference.

[7] LUO Wei-hua, YUMan-quan, XU Hong-bo etc (2006). The Study of Topic Detection Based on Algorithm of Division and Multi-level Clustering with Multi-strategy Optimization. Journal of Chinese in Information Processing, 20 (1), pp. 29-36.

[8] SUN Cheng-jie, ZHU Wen-huan, LIN Lei, etc (2009). Research on BBS Short Text Clustering. CCIR 2009, pp.470-479.

[9] MIAO Jian-ming, ZHANG Quan, ZHAO Jin-fang (2008). Chinese Automatic Text Categorization Based on Article Title Information. Computer Engineering, 34(20), pp.13-14.

[10] ZHANG Xiao-yan, WANG Ting, CHEN Huowang (2007). Story Link Detection Based On Multi-Vector Model with Support Vector Machine. Chinese Computing Technologies and Related Linguistic Issues---Proceedings of the 7th International Conference on Chinese Computing, pp.390-95.