Incorporating Content and Time Features for Chinese Story Subtopic Identification

Zhaoman Zhong
School of Computer Engineering, Huaihai Institute of Technology, Lianyungang, China
Email: zhongzhaoman@163.com

Cunhua Li
School of Computer Engineering, Huaihai Institute of Technology, Lianyungang, China
Email: cli@hhit.edu.cn

Hongwei Dai
Jiangsu Marine Resources Development Research Institute, Lianyungang, China
Email: hongweidai1976@hotmail.com

***Abstract*—Subtopic is the division for a topic again, and it is a novel research direction compared with topic. Subtopic identification is the foundation for analysis of topic evolution relations. In this paper, we propose a novel method of Chinese story subtopic identification incorporating content and time features. We focus on analysis of subtopic content features of the story, and study the computation of subtopic word weights, the dimension reduction of subtopic words and the extraction of subtopic names. Five topics including 20 subtopics are used to implement the experiment. Experimental results show that the performance of the method proposed is better than the existing subtopic identification methods.**

***Index Terms*—subtopic identification, content features, time features, subtopic name extraction**

I. Introduction

With the volume of online stories available today, it is difficult to grasp the growing trend and migration process of topics. Nowadays, many technologies of processing internet information mostly centre on this [1-5], such as information retrieval, classification, monitoring and extraction.

The task of topic tracking is to collect dispersed information together and make it easy for users to get a general understanding. Traditional topic tracking approaches can obtain the relevant stories. However, the relationship between stories occurred during the topic developing process can not be exhibited clearly. Traditional topic tracking considers story topics as a flat collection of stories. It does not consider internal evolution of the topic, and it can not reflect sequence of events in the topic. Using the present topic tracking method we can not easily catch the developing process of the whole topic.

In order to have a better understanding of topic evolution, we should divide stories related to the topic again according to their different attention points, namely subtopics. The analysis of internal subtopics for a topic goes beyond TDT research domain, and it is a novel research direction arousing researchers' attentions. The research and development of subtopics will change defects of current systems such as search engine, information monitoring and classification, which just return the collection of stories related to the topic.

Subtopic identification is the foundation for analysis of topic evolution relations. The starting point of this paper is that we have obtained a lot of stories related to a topic, and we need to identify some subtopics from the topic. We make a number of contributions in this paper. First, subtopic content features are fully investigated. Previous researches do not discuss the differences of content features between the topic and the subtopic. Secondly, we propose a method of identifying subtopics from Chinese story collection incorporating content and time features.

The remainder of this paper is organized as follows. We begin with a brief description of background research and related work in section 2. Section 3 describes the connotation of topic, subtopic and event. Section 4 analyzes the content features and time features of subtopics. We propose a novel method of identifying subtopics from Chinese story collection and some experiments are implemented to validate the proposed methods in section 5. Finally, the conclusion is described.

II. Background and Related Work

A. Topic Representation Models

Nowadays, the commonly used topic representation models are language model and VSM in TDT research area, and subtopic representation models also applied these. The language model is based on Bayesian decision theory [6]. Suppose that the word t in document is

Manuscript received January 1, 2013; revised June 1, 2013; accepted July 1, 2013.
Corresponding author: Cunhua Li, Email:cli@hhit.edu.cn

doi:10.4304/jsw.9.3.668-675

independent, and the relevant probability between the document and the topic is given by (1),

$$P(T\mid d)=\frac{P(T)\bullet P(d\mid T)}{P(d)}\approx P(T)\prod_{n}\frac{P(t\mid T)}{P(t)} \quad . \qquad (1)$$

Where $P(T)$ is the prior probability between the document and topic T , $P(t|T)$ is the generating probability of word t in topic T , and $P(t)$ is the distribution probability of word t in whole corpus. One drawback of language model is that they generally under-estimate the probability of any previously unseen word in the sentence. To combat this problem smoothing techniques are used to assign a non-zero probability to the unseen words and as a result improve the accuracy of overall term probability estimation. A linear interpolation smoothing approach is commonly used to overcome this drawback, and the commonly used method is shown by (2),

$$\lambda P(t\mid T)+(1-\lambda)P(t)\ . \qquad (2)$$

In order to reduce tracking cost, the value of coefficient λ is usually 0.25 [7].

Vector space model is a widely used model of text representation. It is simple and effective in practice. A document is represented as a vector $V(d)=\{t_1,w_1(d);t_2,w_2(d);...;t_n,w_m(d)\}$, where $t_i(i=1,2,...,m)$ is a separate term, and $w_i(d)$ is the weight of t_i in document d . One of the best known schemes is $TF*IDF$ weighting. TF is term frequency of term t_i in document d , and IDF is a measure of the general importance of the term, computed by $\log(N/n_k+0.01)$, in which N is the number of a document collection, and n_k is the number of documents containing the term t_i . The weight of term t_i is computed by (3),

$$w_i(d)=TF_i*Log\ (N/n_k+0.01)\ . \qquad (3)$$

A high weight in $TF*IDF$ is reached by a high term frequency (in the given document) and a low document frequency of the term in the whole collection of document. The weights hence tend to filter out common terms.

For accurately representing documents, some researchers applied more than one VSM to represent a document. For example, Wai Lam et al. [8] regarded a document as two vectors: name entity vector and content vector. And literature [9] represented a document using four vectors: location, time, name and content.

B. Subtopic Identification

Identifying subtopics from a topic is the foundation work for analyzing the evolution relations between subtopics. IBM developed a relatively successful system around topic identification, which uses two layer clustering strategy and computes the similarity between documents by symmetrical Okapi equation. First, the system classified documents into different small clusters, and then sorted out them into final topic clusters after limited time delay.

Event evolution is a new concept developed recently. Makkonen [9] was the first to conduct investigation on event evolution as a subtopic of TDT. He applied ontology including name, place and time to measure the similarity between events. However, he did not define subtopics in detail, and did not provide any experimental results.

Wang [10] divided a topic into subtopics based on results returned by search engines, and proposed two methods of identifying subtopics, which are keywords-based and time-based. For the method of keywords-based partition, keyword weights are computed at first, and then documents are divided in accordance with keywords. But he did not analyze content features of subtopics. For the time-based method, he simply merged documents existing at the same time fragment, and did not consider the situation that more than one subtopic appeared at the same time fragment.

Zhang et al. [11] proposed a method of term-committee-based event identification within topics, which only detected events and tested on English corpus.

Hong et al. [12] proposed a new event detection method based on division comparison of subtopics, which divided each story into different subtopics and identified new topic basing on the proportion and distribution relations of the relevant subtopics. TextTiling algorithm was used to divide subtopics for a document.

C. Existing Problem Overview

The main problems of existing researches related to subtopic identification are as follows:

- There will be more than one topic in a certain time fragment, and the method of time-based subtopic identification can not overcome this problem. For example, after topic ‘地震 (earthquake)’ occurred, some subtopics will be aroused in a short time, such as ‘救援 (rescue)’, ‘人员伤亡 (injury and death)’, ‘交通中断 (traffic interruption)’ and ‘捐款捐物 (contribution)’.
- The existing methods of keywords-based subtopic identification did not analyze content features of subtopics for a topic, and further applied suitable feature selection strategies for subtopics.
- The existing researches related to subtopic identification mainly focus on English stories, and do not pay more attention to Chinese stories.

III. Comparisons of Topic, Subtopic and Event

Topic is one of basic concepts for TDT (Topic Detection and Tracking), and TDT evaluating conference sponsored by DARPA defines an event as a ‘narrowly defined topic for search’ [13]. An event is something that happens at a specific time and place. A topic is a set of events that are strongly interconnected with each other. The current topic is a broad concept, and it includes a core event (or a seed event) and related events or activities. So to speak, a topic consists of some related subtopics.

According to the definition of the topic, a story belongs to the topic so long as it relates to the seed event of the topic. Usually, the seed event is thought to be the name of the topic. For example, the stories about ‘灾后重建 (post-disaster reconstruction)’ and ‘捐款捐物 (contribution)’ are related to the seed event ‘汶川地震 (Wenchuan earthquake)’, and they are all components of topic ‘汶川地震 (Wenchuan earthquake)’.

Subtopics of a topic will change with the passage of time. Taking ‘汶川地震 (Wenchuan earthquake)’ as an example, soon after it occurred, reported stories mainly round the epicenter, the earthquake magnitude and the situation. With the passage of time, they may get to concern about the casualties and disaster relief situation, and then about the epidemic prevention, post-disaster reconstruction. This is the evolution of the topic.

The relations of topic, subtopic and event are shown in Fig. 1.

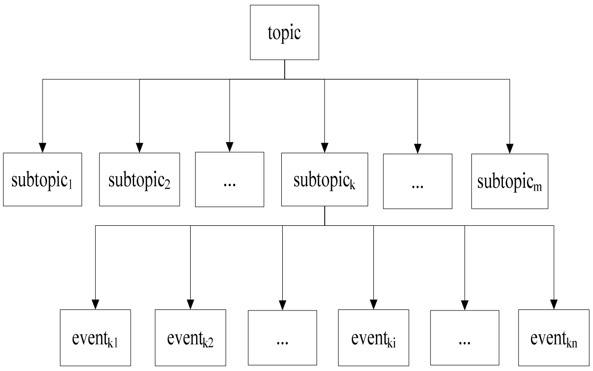

Figure 1. Relations of topic, subtopic and event.

Relations of topic, subtopic and event in Fig. 1 can be formalized as follows: $T = \{ST_1, ST_2, \cdots, ST_m\}$, and $ST_k = \{E_{k1}, E_{k2}, \cdots, E_{kn}\} (1 \le k \le m)$. T is a topic and it consists of m subtopics. ST_k is a subtopic and it contains n events.

The stories reporting a topic usually contain a number of events. If a story d introduces some events of subtopic ST_k , then d belongs to ST_k . And if a story d contains more than one event relating some different subtopics, then d belongs to different subtopics. Namely, the relations of stories and subtopics are many to many.

IV. Subtopic Content and Time Features

A. *Subtopic Content Features*

Subtopic content features are as follows:

- A seed event exists in every subtopic for explaining subtopic source. For example, ‘汶川地震 (Wenchuan earthquake)’ is not only a topic, but a seed event. It exists in every subtopic such as ‘人员伤亡 (injury and death)’, ‘捐款捐物 (contribution)’ and ‘反贪调查 (corruption investigation)’.
- A subtopic is described by some events, and an event usually associates some elements such as event triggers, time, locations and participators. For subtopics, event triggers have more distinguishing ability compared to other elements. For example, subtopic ‘捐款捐物 (contribution)’ pays close attention to some events related to ‘捐款捐物 (contribution)’, and does not consider which organization or person contribute.
- The title and a few start sentences of a story can nearly account for a subtopic because of story writing habits.

If we use all features of a story to identifying subtopics, then noise features will inevitably reduce the subtopic recognition performance. Subtopic content features show that:

- A seed event should be removed in selecting subtopic features, including event triggers, locations, time and participators of the seed event, because it often exists each subtopic. For topic ‘汶川地震 (Wenchuan earthquake)’, ‘5月12日汶川地震 (Wenchuan earthquake of May 12th) is a seed event, and location element ‘汶川 (Wenchuan)’, time element ‘5 月 12 日 (May 12th)’ and trigger element ‘地震 (earthquake)’ should be removed. The seed event is effective for distinguishing different topics, but has no effect for identifying different subtopics of a topic.
- An event trigger usually refers to a class of events, and it represents a specific event under the circumstance of including clear event elements. Event triggers should be given higher weights, and the weights of other elements, such as time, locations and participators, should be lowered. This idea is different from commonly used topic identification technologies, which augment weights of not only event triggers, but people, time and locations [15-17].
- The positions of features are different, and their weights are different. The title and a few start sentences of a story should obtain higher weights.

B. *Subtopic Time Features*

Time can reflect the developing trend of topics. For example, the seed event of a topic occurred at the earliest, and stays stories for a long time. In contrast, new events of a topic tend to occur later. Internet stories have very strong time sequential nature, and take time fragment as the basic narrative unit. The time of stories can be divided into two classes: the published time of the story and the time of event occurrence contained by the story. We mainly use the published time of the story in this research, which is located in the bottom of the title in general.

The published time of the story can be divided into two kinds: the absolute time and the relative time. The absolute time refers to clear time of the published story, such as ‘2012-3-20’, ‘2012/3/20’, ‘2012:3:20’, which is relatively simple. And the relative time is relative to the current time of the system, which is slightly more complicated. Through the statistics, the commonly used formats of the relative time are shown in Table I .

TABLE I.
RELATIVE PUBLISHED TIME FORMATS OF STORIES

No.	Time formats	Examples
1	秒前 (seconds ago)	20 秒前 (20 seconds ago)
2	分钟前 (minutes ago)	5 分钟前 (5 minutes ago)
3	半小时前 (half an hour before)	半小时前 (half an hour ago)
4	小时前 (hours before)	2 小时前 (2 hours ago)
5	昨天 (yesterday)	昨天 (yesterday)
6	前天 (the day before yesterday)	前天 (the day before yesterday)
7	今天 (today)	今天 (today)
8	天前 (days ago)	2 天前 (2 days ago)
9	秒之前(seconds ago)	10 秒之前 (10 seconds ago)
10	分钟之前 (minutes ago)	30 分钟之前 (30 minutes ago)
11	小时之前 (hours ago)	1 小时之前 (1 hour ago)

V. INCORPORATING CONTENT AND TIME FEATURES FOR CHINESE STORY SUBTOPIC IDENTIFICATION

A. Method Flow

The task of subtopic identification is to divide a topic again, and extract subtopic names to present. Suppose the set of stories related to topic T is $D = \{d_1, d_2, \cdots, d_n\}$, and the result of subtopic identification is to build a subtopic set $T = \{ST_1, ST_2, \cdots, ST_m\}$. Each subtopic is not empty, and $\forall d_i, \exists ST_j \in T, d_i \in ST_j (1 \le i \le n, 1 \le j \le m)$.

The method flow of incorporating content and time features for Chinese story subtopic identification is shown in Fig. 2.

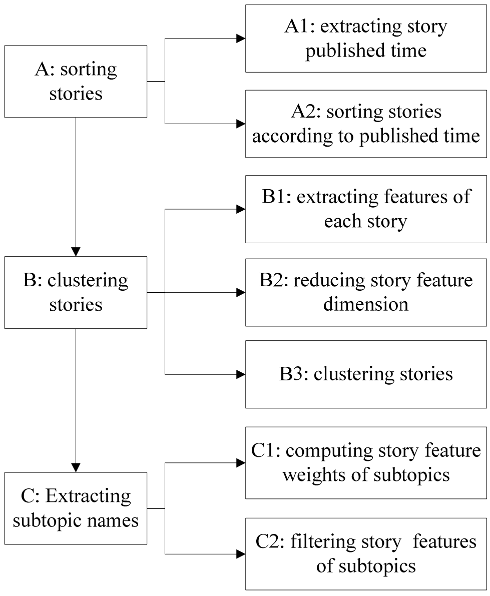
A: sorting stories
A1: extracting story published time
A2: sorting stories according to published time
B: clustering stories
B1: extracting features of each story
B2: reducing story feature dimension
B3: clustering stories
C: Extracting subtopic names
C1: computing story feature weights of subtopics
C2: filtering story features of subtopics

Figure 2. Flow of Chinese story subtopic identification integrating content and time features.

Fig. 2 mainly includes three steps: sorting stories (A), clustering stories (B) and extracting subtopic names (C).

Step A includes two sub-steps: extracting story published time (A1) and sorting stories according to published time (A2). Sub-step A1 extracts the absolute time or relative time of the story, translates the relative time into the absolute time, and formalize the absolute time as $Time = \{年，月，日\}$ ($Time = \{year，month，day\}$). Sub-step A2 takes '天 (day)' as the basic time unit to sort stories. Suppose the set of stories is $D = \{d_1, d_2, \cdots, d_n\}$, we get sorting result is $D = \{D_1, D_2, \cdots, D_h\}$ after sorting stories, where h is the number of days lasted by stories. $D_i = \{d_{i1}, d_{i2}, \cdots, d_{ij}\} (1 \le i \le h)$, and j is the number of stories reported in ith days.

Step B includes three sub-steps: extracting features of each story (B1), reducing story feature dimension and clustering stories (B3). For sub-step B1, VSM is used to represent the document. On the one side, the weight of word t depends on the frequency and position of t in the story, denoted by w_t^{TF}, and on the other side, it depends on its inverse document frequency, denoted by w_t^{IDF}. The words existing in the title of the story is considered to be the first level. The first sentence of the main body is the second level, and so on. Three punctuations, '。', '！' and '？', are used to segment the sentence. The method of computing w_t^{TF} is given by (4),

$$w_t^{TF} = \sum_{k=1}^{m} \frac{1}{2^{\ln t_k}} \quad . \tag{4}$$

where m is the number of word t in the story, and t_k is the level of word t.

The method of computing w_t^{IDF} is given by (5),

$$w_t^{IDF} = Log\ (N / n_t + 0.01) \ . \tag{5}$$

where N is the number of stories, and n_t is the number of stories containing word t.

Finally, the weight of word t is computed by (6),

$$w(t) = w_t^{TF} * w_t^{IDF} \ . \tag{6}$$

The method of identifying event triggers introduced in Literature [18] is used for sub-step B2, which divides stop-verbs into two kinds. The first kind of verbs is not considered as event triggers such as existential verbs, volitive auxiliary, jugging verbs, causative verbs and others expressing feeling and guess. The second is not event triggers, but others together with it can be used as event triggers such as verb, denominal verb and noun. For phrases '发生火灾 (have fire)' and '开始演讲 (start presentation)', '发生 (have)' and '开始 (start)' are not event triggers, but '火灾 (fire)' and '演讲 (presentation)' can be considered as event triggers.

We should augment event trigger weights after identifying them from the document, and the method is given by (7),

$$w_i = w_i * \log_2{}^{n} \,. \tag{7}$$

Where w_i is the weight of event trigger e_i, and n is the number of e_i in the document.

The k-means algorithm is used to cluster stories for Sub-step B3. We select stories to cluster according to their time sequence. In this paper, the k-means algorithm is omitted.

Some keywords should be extracted as the subtopic name for presenting. Step C includes two sub-steps: computing story feature weights of subtopics (C1) and filtering story features of subtopics (C2). The weight of keyword t is computed by (8),

$$w(t) = w_t^{TF} * w_t^{DF} \tag{8}$$

Where w_t^{TF} refers to (4), and w_t^{DF} is the number of stories containing word t .

The words used as the subtopic name are usually nouns, verbs or denominal verbs, and the word length is greater than one Chinese character. According to this characteristic, we can filter subtopic features. Suppose the set of words in descending order according to their weights is $Term = \{t_1, t_2, \cdots, t_s\}$, s is the number of words for a subtopic, and $y(1 \le y \le s)$ is the number of words selected as the subtopic name. We can get the subtopic name is $t_1, t_2, \cdots, t_y$. Generally, the subtopic name should not be too long, and contains 3 to 5 words.

B. Experimental Results and Analysis

In order to verify the effectiveness of the method proposed in this paper, we select real topics and subtopics to conduct the experiments. The selecting topics and subtopics are shown in tableⅡ, and numbers in parenthesis represent the number of collecting stories.

TABLE II.
TOPICS AND SUBTOPICS FOR THE EXPERIMENT

Topics	Subtopics

汶川地震(48 篇)
Wenchuan earthquake (48)
	①人员伤亡(4 篇), injury and death (4)
	②捐款捐物(14 篇), contribution (14)
	③灾区救援(13 篇) , rescue (13)
	④灾后重建(9 篇) , rebuilding (9)
	⑤反贪调查(8 篇), corruption investigation (8)

温州动车事故(28 篇)
Wenzhou high-speed train accident
	①人员伤亡(10 篇) , injury and death (4)
	②原因调查(5 篇) , reason investigation (5)
	③责任追究(5 篇) , accountability (5)
	④事故赔偿(8 篇), compensating for loss (8)

中菲黄岩岛事件(30 篇)
China-Philippines Huangyan island incident
	①对峙(5 篇) , confronting each other (5)
	②黄岩岛改名(9 篇) , Huangyan island renaming (9)
	③美国态度(10 篇) , American attitude (10)
	④中菲影响(6 篇), influence (6)

三鹿牛奶事件(56 篇)
Sanlu milk powder incident
	①官员复出(15 篇) , official return (15)
	②赔偿(14 篇) , compensating for loss (14)
	③问责(16 篇) , accountability (16)
	④影响(11 篇), influence (11)

中日钓鱼岛撞船(18 篇)
China-Japan Diaoyu allision
	①撞船(5 篇) , allision (5)
	②起诉(7 篇) , charge (7)
	③索赔(6 篇), claim (6)

For tableⅡ, the experimental corpus is manually collected by undergraduate students. The partitions of subtopics and their stories have been serious discussions.

For the subtopics of topic ‘汶川地震(Wenchuan earthquake)’, $TF * IDF$ is used to compute word weights, only filtering stop-words, and no reducing feature dimension. TableⅢ lists top 10 words in descending order according to their weights for each subtopic.

Table Ⅲ shows that the event trigger ‘地震(earthquake)’ of the seed event has relatively higher ranking in 5 subtopics. The location element ‘汶川(Wenchuan)’ of the seed event has relatively higher ranking in 4 subtopics: ‘人员伤亡 (injury and death)’, ‘灾区救援 (rescue)’, ‘捐款捐物 (contribution)’ and ‘反贪调查 (corruption investigation)’. And two subtopics, ‘人员伤亡 (injury and death)’ and ‘灾区救援 (rescue)’ all contains location element ‘成都 (Chengdu)’ of the seed event. The time element ‘5 月 12 日 (May 12th)’ of the seed event has relatively higher ranking in 2 subtopics: ‘人员伤亡 (injury and death)’ and ‘灾区救援 (rescue)’.

TABLE III.
TOP 10 WORDS FOR 5 SUBTOPICS OF TOPIC ‘WENCHUAN EARTHQUAKE’

Subtopics	人员伤亡 (injury and death)	灾区救援 (rescue)	捐款捐物 (contribution)	灾后重建 (rebuilding)	反贪调查 (corruption investigation)
Top 10 words	地震 (earthquake) 汶川 (Wenchuan) 人员 (person) 伤亡 (injury and death) 级 (level) 报告 (report) 震感 (earthquake feel) 大 (large) 成都 (Chengdu) 5 月 12 日 (May 12th)	救援 (rescue) 灾区 (disaster area) 地震 (earthquake) 人 (person) 汶川 (Wenchuan) 队 (team) 人员 (person) 救灾 (relief) 5 月 12 日 (May 12th) 成都 (Chengdu)	灾区(d isaster area) 捐款 (contributing money) 元 (Yuan) 企业 (enterprise) 地震 (earthquake) 中国 (China) 公司 (Company) 集团 (group) 汶川 (Wenchuan) 捐物 (contributing goods)	重建 (rebuilding) 地震 (earthquake) 后 (after) 恢复 (recovery) 灾害 (disaster) 规划 (planning) 建设 (constructing) 人 (person) 大 (large) 设施 (facilities)	后 (after) 元 (yuan) 钱 (money) 重建 (rebuilding) 办公室 (office) 工作 (work) 捐款 (contributing money) 证据 (evidence) 地震 (earthquake) 汶川 (earthquake)

Subtopic is elaborated by detailed events, but events in 5 subtopics are not much in table 3. Subtopic ‘人员伤亡 (injury and death)’ contains three events: ‘地震 (earthquake)’, ‘伤亡 (injury and death)’ and ‘报告 (report)’; subtopic ‘灾区救援 (rescue)’ contains three events: ‘救援 (rescue)’, ‘地震(earthquake)’ and ‘救灾 (relief)’; subtopic ‘捐款捐物 (contribution)’ contains three events: ‘捐款 (contribute money)’, ‘地震 (earthquake)’ and ‘捐物 (contribute goods)’; subtopic ‘灾后重建 (rebuilding)’ contains six evens: ‘重建 (rebuilding)’, ‘地震 (earthquake)’, ‘恢复 (recovery)’, ‘灾害 (disaster)’, ‘规划 (planning)’ and ‘建设 (construct)’; and subtopic ‘反贪调查 (corruption investigation)’ contains three events: ‘重建 (rebuilding)’, ‘捐款 (contribute money)’ and ‘地震 (earthquake)’. The proportion of event triggers in all words for table 3 is $18/50=36\%$.

For all subtopics of 5 topics, we use $TF*IDF$ to compute word weights, and select top 10 words in descending order to form set W. The proportion of event triggers in set W is $92/250=36.8\%$. The method proposed in this paper is used to analyze the proportion of event triggers. After filtering the seed event elements and augmenting the weights of event triggers, Eq. (6) is used to compute word weights. Likewise, top 10 words of each subtopic are selected to form a set. The proportion of event triggers in the set is $141/250=58.4\%$. The proportion increases by 21.6% compared with the method of simply using $TF*IDF$ to compute word weights.

We mix all subtopics of five topics, use the flow in Fig. 2 to cluster stories, and obtain each subtopic. Three methods are used to implement the experiment: Method M1, which does not augment word weights; Method M2, which augments the weights of event triggers, locations, time and participators; and Method M3, which removes four elements of the seed event, and only augments event trigger weights.

F-measure is used as the assessment index, and its computing method is given by (9),

$$F=\frac{P*R*2}{P+R}\ . \tag{9}$$

Where P is precision, and R is recall. The methods of computing P and R are given by (10) and (11) respectively,

$$P=\frac{U}{V}\,. \tag{10}$$

$$\mathrm{R}=\frac{U}{W}\ . \tag{11}$$

Where U is the number of correct clustering stories, V is the number of real clustering stories, and W is the number of standard clustering stories.

The average F-measure for five topics by three methods is shown in tableⅤ.

From tableⅣand tableⅤ, the average F-measure using M3 is improved obviously compared with M2 and M1. The average F-measure using M3 is 81.05%. It is increased by 11.01% compared with M1, and 7.35% compared with M2. The average F-measure using M2 is 73.71%, and it is only increased by 3.67% compared with M1.

TABLE V.
F-MEASURE OF TOPIC 'EARTHQUAKE' USING THREE METHODS

Clustering methods Subtopics	W	M1			M2			M3		
		V	U	F-measure	V	U	F-measure	V	U	F-measure
人员伤亡(injury and death)	4	5	3	66.67%	5	4	88.89%	5	4	88.89%
捐款捐物(contribution)	14	12	9	69.23%	13	9	66.67%	13	11	81.48%
灾区救援(rescue)	13	14	10	74.07%	14	10	74.07%	14	11	81.48%
灾后重建(rebuilding)	9	7	6	75.00%	8	6	70.59%	8	6	70.59%
反贪调查(corruption investigation)	8	6	5	71.43%	6	5	71.43%	7	7	93.33%
Average F-measure		71.28%			74.33%			83.15%		

TABLE IV.
F-MEASURE OF FIVE TOPICS USING THREE METHODS

Topics	M1	M2	M3
汶川地震 (Wenchuan earthquake)	71.28%	74.33%	83.15%
温州动车事故 (Wenzhou high-speed train accident)	68.90%	73.40%	79.80%
中菲黄岩岛事件 (China-Philippines Huangyan island incident)	69.90%	75.00%	83.10%
三鹿奶粉事件 (Sanlu milk powder incident)	71.10%	73.60%	79.70%
中日钓鱼岛撞船(China-Japan Diaoyu allision)	69.00%	72.20%	79.50%
Average F-measure	70.04%	73.71%	81.05%

Two methods of extracting subtopic names are used to conduct the experiment. One computes word weights by $TF*IDF$, denoted by M1. And the other uses (8) to compute word weights, remove the seed event, augment event trigger weights, and selects nouns, verbs denominal or verbs which length is longer than one Chinese character, denoted by M2.

The subtopic name extraction result by two methods for five subtopics of topic '汶川地震 (Wenchuan earthquake)' is show in tableⅥ, and the number of selecting words is three.

TABLE VI.
NAME EXTRACTION OF FIVE TOPICS USING TWO METHODS

Subtopics	M1	M2
人员伤亡(injury and death)	地震,汶川,人员 (earthquake, Wenchuan, person)	人员,伤亡,报告 (person, injury and death, report)
捐款捐物(contribution)	救援,灾区,地震 (rescue, disaster area, earthquake)	救援,灾区,救灾 (rescue, disaster area, relief)
灾区救援(rescue)	灾区,捐款,元 (disaster area, contributing money, yuan)	灾区,捐款,企业 (disaster area, contributing money, enterprise)
灾后重建(rebuilding)	重建,地震,后 (rebuilding, earthquake, after)	重建,恢复,灾害 (rebuilding, recovery, disaster)
反贪调查 (corruption investigation)	后,元,钱 (after, yuan, money)	重建,捐款,调查 (rebuilding, contributing money, investigation)

For tableⅥ, the name of subtopic '人员伤亡 (injury and death)' is ' (地震,汶川,人员 earthquake, Wenchuan, person)' and the name of subtopic '反贪调查 (corruption investigation)' is '后,元,钱 (after, yuan, money)' by using M1, which can not reflect themes of subtopic '人员伤亡 (injury and death)' and '反贪调查 (corruption investigation)'. However, subtopic names obtained by M2 can basically reflect the theme of each subtopic.

VI. CONCLUSIONS AND FUTURE DIRECTIONS

We focus on Chinese story identification in the paper, especially content feature comparison of subtopics and topics. After analyzing the features of Chinese story content and time, we proposed a novel method of identifying Chinese story subtopics. In order to verify the effectiveness of the method proposed, we select real topics and subtopics to conduct the experiments. Experimental results show that the performance of the method proposed is better than the existing subtopic identification methods.

As part of our future work, in order to achieve better effect we intend to focus on event element identification, time formalization, and subtopic feature selection and reduction, etc.

ACKNOWLEDGMENT

This research (paper) is supported by the National Natural Science Foundation of China (No.60975033), and the Science and Technology Foundation of Lianyungang (CG1121).

REFERENCES

[1] A. Brahmi, A. Ech-Cherif, and A. Benyettou, "Arabic texts analysis for topic modeling evaluation," *Information Retrieval,* Vol.15, pp. 33-53, 2012.

[2] W. H. Dai, X. Q. Wan and X. Y. Liu, "Emergency event: internet spread, psychological impacts and emergency management," *Journal of Computers,* Vol. 6, pp. 1748-1755, 2011.

[3] L. D. Wang, B. G. Wei, and J. Yuan, "Topic discovery based on LDA_col model and topic significance re-ranking," *Journal of Computers,* Vol. 6, pp. 1639-1647, 2011.

[4] M. Xie, C. L. Wu, and Y. L. Zhang, "A new intelligent topic extraction model on Web," *Journal of Computers,* Vol. 6, pp. 466-473, 2011.

[5] C. Y. Yang, X. D. Shi, and C. P. Wei, "Discovering event evolution graphs from news corpora," *IEEE Transactions on Systems, Man, and Cybernetics –Part A: Systems and Humans,* Vol. 39, pp. 850-863, 2009.

[6] R. T. Fernández, D. E. Losada and L. A. Azzopardi, "Extending the language modeling framework for sentence retrieval to include local context," *Information Retrieval,* Vol. 14, pp. 355-389, 2011.

[7] J. Ponte and W. Croft. "A language modeling approach to information retrieval," *Special Interest Group on Information Retrieval (SIGIR) 1998,* Berkeley, pp. 246-253, 1998.

[8] W. Lam, H. M. Meng and K Hui, "Multilingual Topic Detection Using a Parallel Corpus," *In TDT Workshop,* Dec, 2000.

[9] J. Makkonen, "Investigations on Event Evolution in TDT". *Proceedings of Student Workshop of Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics,* Edmonton, Canada, pp. 43-48, 2003.

[10] W. Wang, "Analysis of network topic evolution based on keywords and time", *Dissertation of Master,* Fudan University, Shanghai, 2009.

[11] K. Zhang, J. Z. Li, and G. Wu, "Term-committee-based event identification within topics," *Journal of Computer Research and Development,* Vol. 46, pp. 245-252, 2009.

[12] Y. Hong, Y. Zhang, and J. L. Fan, "New event detection based on division comparison of subtopic," *Chinese Journal of Computers,* Vol. 31, pp. 687-695, 2008.

[13] J. G. Fiscus and G. R. Doddington, "Topic detection and tracking evaluation overview," *LNCS, The Information Retrieval Series,* PP. 17-31, 2002.

[14] Z. T. Liu, M. L. Huang, and W. Zhou, "Research on event-oriented ontology model," *Computer Science,* Vol. 36, pp.191-195, 2009.

[15] J. Makkonen, H. Ahonen-Myka, and M. Salmenkivi, "Applying semantic classes in event detection and tracking," *International Conference on Natural Language Processing (ICON) 2002*, Mumbai, India, pp.175-183, 2002.

[16] Y. Y. Zhao, B. Qin, and W. X. Chen, "Research on Chinese event extraction," *IRCS,* Suzhou, pp. 55-62, 2007.

[17] J. Makkonen, H. Ahonen-Myka, and M. Salmenkivi, "Topic detection and tracking with spatio temporal evidence," T*he 25th European Conference on Information Retrieval Research,* Heidelberg, pp.251-265, 2003.

[18] Z. M. Zhong, P. Zhu, and C. H. Li , "Research on event-oriented query expansion based on Local Analysis," *Journal of the China Society for Scientific and Technical Information,* Vol. 31, pp. 151-159, 2012.

Zhaoman Zhong was born in 1977. He received the Ph.D. degree in computer applications from Shanghai university in 2011. He is an associate professor at Huaihai Institute of Technology. His research interests include information retrieval, text information mining, event ontology, information extraction, etc.

Cunhua Li was born in 1963. He received the Ph.D. degree in computer applications from Southeast university in 2004. He is a professor at Huaihai Institute of Technology. His research interests include data mining, artificial intelligence, image processing, etc.

Hongwei Dai was born in 1975. He received the Ph.D. degree in system science from the university of Toyama in 2007. He is an associate professor at Huaihai Institute of Technology. His research interests include artificial intelligence, data mining, etc.