Topic Detection with Hypergraph Partition Algorithm

Xinyue Liu[1, 2]
1. School of Computer Science and Technology, Dalian University of Technology, Dalian, China
2. School of Software, Dalian University of Technology, Dalian, China
xyliu@dlut.edu.cn

Fenglong Ma[2], Hongfei Lin[1]
1. School of Computer Science and Technology, Dalian University of Technology, Dalian, China
2. School of Software, Dalian University of Technology, Dalian, China
flma@mail.dlut.edu.cn, hflin@dlut.edu.cn

***Abstract*—An algorithm named SMHP (Similarity Matrix based Hypergraph Partition) algorithm is proposed, which aims at improving the efficiency of Topic Detection. In SMHP, a T-MI-TFIDF model is designed by introducing Mutual Information (MI) and enhancing the weight of terms in the title. Then Vector Space Model (VSM) is constructed according to terms' weight, and the dimension is reduced by combining H-TOPN and Principle Component Analysis (PCA). Then topics are grouped based on SMHP. Experiment results show the proposed methods are more suitable for clustering topics. SMHP with novel approaches can effectively solve the relationship of multiple stories problem and improve the accuracy of cluster results.**

***Index Terms*—topic detection, similarity matrix, hypergraph partition, clustering**

I. INTRODUCTION

Timely and efficiently access to amounts of information is becoming more and more significant. Gaining such access is no longer a problem owing to the widespread availability of Internet. However, it is difficult for users to locate interesting or useful information as the explosion in large volumes of digitized textual content. Evidently, it is impossible for humans to assimilate such vast quantities of information. Topic Detection has risen to challenge this difficulty.

Topic Detection, a subtask of Topic Detection and Tracing (TDT), is becoming a hot research interest in recent years. This subtask attempts to identify “topics” by analyzing and organizing the content of textual materials, thereby helps people separate mixed information into manageable clusters. In the context of news, Topic Detection can be seen as an event detection which groups related stories into clusters, each of which stands for a single topic.

Many clustering algorithms have been presented to group topics [1-3]. However, three crucial issues are still not well solved. (1) *The relationship of multiple stories*. The relationship of multiple stories decides the topic each story might belong to. While existing Topic Detection algorithms use the similarity or distance between two stories only, none of them make use of multiple stories’ relationship. Considering multiple stories’ relationship, the corpora could be analyzed more precisely. (2) *The ability to deal with high-dimensional data*. Due to occurrence of long stories, the curse of dimensionality has heavy impact on the performance of topic detection algorithms. (3) *The influence of noise to the clustering result*. Owing to some stories that distribute inhomogeneous in the data model, the detection system may recognize them as noise and cluster them into the wrong topics.

Taken these problems into consideration, we propose a novel framework for topic detection. In this framework, Top-N and PCA are used for dimension reduction, and then a clustering algorithm is employed. We also propose two clustering algorithms for our concerned problem: the Self-tuning Spectral clustering algorithm (SSC) and the Similarity Matrix based Hypergraph Partition algorithm. The experiments show that both improvements are capable of improving the clustering results.

The remainder of this paper is organized as follows: In Section II, we define key concepts and introduce works that are directly related to the ideas presented in this paper. Section III describes our novel approach for clustering topics via improving TFIDF computing formula and using dimensional reduction strategies. In Section IV, we demonstrate the superiority of our approach with comparative empirical results. Finally, in Section V, we present some conclusions and future research directions.

II. RELATER WORK

A. Topic Definition

A topic is defined to be a seminal event or activity, along with all directly related events and activities [4].

Manuscript received December 28, 2010; revised March 1, 2011; accepted March 28, 2011.
Corresponding author: Xinyue Liu

doi:10.4304/jsw.6.12.2407-2415

Thereby, we can deduce that a topic consists of events and activities, both of which are defined in more detail in [5]. A TDT event is defined as something that happens at a specific time and place, along with all the necessary preconditions and unavoidable consequences [5]. For instance, a TDT event may be an earthquake, a conflagration, or an air accident. A TDT activity is a connected series of events with a common focus or purpose that happens in specific places during a given time period [5]. Such a TDT activity might be a presidential election, a social investigation, or a disaster relief effort.

B. Topic Detection

Traditional Topic Detection system comprises three kinds of methods: (1) Single-Pass clustering [1]. In this kind of methods the first story is viewed as the centroid of the topic cluster. When a new story arrives, it will be compared with all other stories by computing the cosine similarity. If the similarity exceeds a threshold, then the story belongs to an old topic, and put it into the most similar cluster; otherwise it will be considered as a new topic and a new cluster will be created. Two disadvantages of this method arise: one is that Single-Pass clustering algorithm depends evidently on the stories sequence; the other is the unbalanced distribution problem. (2) Hierarchical Clustering [2][3]. This kind of methods combines two stories with the highest similarity each time, until all the stories become a single cluster. The complexity of this kind of methods is usually unacceptable. Moreover, the result of clustering is spherical and average， making it more noise sensitive. (3) K-means Clustering [6], in which the K initial cluster center will be selected. The algorithm terminates when the K clustering centers do not change. The performance of this method mainly depends on the selection of the K cluster center, and cluster number K is required.

In recent years, most work focus on proposing better methods on comparison of stories and document representation. A basic incremental TFIDF model is used in [7-10], and new incremental dynamic topic model named topic based TFIDF (T-TFIDF) is proposed in [11]. At the same time, static TFIDF model is widely used [12-16]. If a term frequently appears in different stories, the term weight may be lower than those rarely emerge in different stories in traditional static TFIDF model.

III. OUR APPROACH

First and foremost, preprocess step is provided for the corpus. For each story we tokenize the text, tag the tokens, remove stop words, and get a candidate set of features for its vector-based model. In this paper, a token and its tag are viewed as a feature. If the tokens are spelled the same way but differ in tags, they will be taken as different features. ICTCLAS [17] is used as the tokenizer and tagger, and the stop word list contains 507 words.

A. Term Weighting

Incremental TF-IDF model is used for term weight calculation in this paper.

TABLE I.
MEANINGS OF SYMBOLS

Symbol	Notation
w_{ij}	Weight of term i in story j.
tf_{ij}	Within-story term frequency (TF).
$\log(\frac{n_j}{N}+0.01)$	Inverse document frequency (IDF).
n_j	The number of training story where term j occurs.
N	Size of training corpus.
n_{ij}	Raw count of term i occurrences in story j.
len_j	Total terms number of story j.
len_{avg}	Average terms number in each story.

$$w_{ij} = tf_{ij} \cdot \log(\frac{N}{n_j} + 0.01) . \quad (1)$$

$$tf_{ij} = \frac{n_{ij}}{n_{ij} + 0.5 + 1.5\frac{len_j}{len_{avg}}} . \quad (2)$$

The meanings of symbols emerged in (1) and (2) are explained in TABLE I.

For the purpose of setting the value of w_{ij} between 0 and 1, (1) is normalized. The normalized TFIDF is given as:

$$w_{ij} = \frac{tf_{ij} \cdot \log(\frac{N}{n_j} + 0.01)}{\sqrt{\sum_{j=1}^{N} (tf_{ij})^2 \log^2(\frac{N}{n_j} + 0.01)}} . \quad (3)$$

The terms that we use are preprocessed, and they are representative. A term emerged frequently in different documents may be the representative word of a topic. In case that the weight of this term is lower than others, the Topic Detection algorithm will not cluster this topic perfectly.

As noted in Section II, the inadequacy of current TFIDF model is due to the fact that existing approach for computing IDF do not consider the importance of term's repetition. We now present solutions to this problem and add title information into TFIDF model.

***Mutual information based TFIDF*.** First of all, the importance of mutual information (MI) is shown by a simple example.

15 stories are selected about the topic of State Council from the corpus. The weights of terms are listed in TABLE II. "State Council" should be more important than "Bank" in this topic. On the contrary, the weight of "Bank" is bigger than the weight of "State Council". Hence, traditional TFIDF is not suitable for computing term's weight. In this paper, mutual information is employed for computing term's weight. And the correctness of TFIDF using MI had been proved in [18].

TABLE II.
TFIDF VALUE OF TERMS

Term	TF	DF	IDF	TFIDF
State Council	0.2	1	0.001	0.0002
⋮	⋮	⋮	⋮	⋮
Bank	0.05	3	1.102	0.0551
⋮	⋮	⋮	⋮	⋮

Therefore, mutual information based TFIDF (MI-TFIDF) is

$$w_{ij} = \frac{tf_{ij} \cdot \log(\frac{n_j}{N} + 0.01)}{\sqrt{\sum_{j=1}^{N} (tf_{ij})^2 \log^2(\frac{n_j}{N} + 0.01)}}. \quad (4)$$

In order to give prominence to the key terms and keep down other terms influence, the constant in (4) is taken as 2. The new MI-TFIDF is

$$w_{ij} = \frac{tf_{ij} \cdot \log(\frac{n_j}{N} + 2)}{\sqrt{\sum_{j=1}^{N} (tf_{ij})^2 \log^2(\frac{n_j}{N} + 2)}}. \quad (5)$$

Title based MI-TFIDF. Intuitively, the most important information for a story should appear in the title [19]. The title always describes who, where and what aspect of an event. If the two stories were on the same topic, they would share a part of title information.

We illustrate the above intuition with examples. Terms in **bold face** are reserved in the story's title. And the examples are extracted from the benchmark TDT5.

Story I: A new topic
Docno: XIN_CMN_20030929.0010
Title: (国际)车臣总理食物中毒但已脱离危险
Pro-Processing: **国际**/n **车臣**/ns **总理**/n **食物中毒**/l 脱离/v **危险**/an

Story II: Story on the same topic
Docno: XIN_CMN_20030930.0101
Title: (国际)俄罗斯车臣总理“食物中毒”治愈出院
Pro-Processing: **国际**/n **俄罗斯**/ns **车臣**/ns **总理**/n **食物中毒**/l 治愈/v **出院**/vn

Story III: Story off the topic
Docno: XIN_CMN_20030916.0223
Title: (国际)俄罗斯发生重大交通事故 6 人死亡
Pro-Processing: **国际**/n **俄罗斯**/ns 发生/v 重大/a **交通**/n **事故**/n 死亡/v

Story I is the seed of topic 55044 about “Popov gets food poisoning”. In the title of Story I, we select $Set1$={国际, 车臣, 总理, 食物中毒, 危险} as the feature set. About story II, $Set2$={国际, 俄罗斯, 车臣, 总理, 食物中毒, 出院} is taken as feature set, and $Set3$={国际, 俄罗斯, 交通, 事故} is viewed as title feature set of story III. Analyzing $Set1$, $Set2$ and $Set3$, stories on the same topic always share most of the features, especially the named entities. On the contrary, stories off the topic barely contain the same features in the title.

From the analysis of the example above, it is reasonable to adjust term weighting according to their location. If the term emerges in the title, the weight should be higher than other terms in the context. Hence, the weight functions using title information, namely, title based MI-TFIDF (T-MI-TFIDF) is denoted as followed:

$$w_{ij} = \frac{\alpha \cdot tf_{ij} \cdot \log(\frac{n_j}{N} + 2)}{\sqrt{\sum_{j=1}^{N} (tf_{ij})^2 \log^2(\frac{n_j}{N} + 2)}}. \quad (6)$$

which holds

$$\alpha = \begin{cases} t & \text{if } n_{ij} \in \text{title} \\ 1 & \text{else} \end{cases}$$

where t is an integer greater than 1. In this paper, T-MI-TFIDF is used to construct VSM (Vector Space Model) for the corpus.

B. Top-N and PAC strategies

After preprocessing, the story vector is high-dimensional. Hence, some noise still exists. Two dimensional reduction strategies (i.e. top-N and PCA) are introduced, to remove the noise.

Top-N strategy simply selects the N words with highest term weight for each story from all the terms. However, it is hard to decide the value of N. Therefore, based on story length harmonic average top-N strategy is put forward. Harmonic average is less than arithmetic average when each element is positive number in the set [20]. Besides, harmonic average is to give support to the shorter length of stories, which satisfies the demand of dimensional reduction method. Therefore, harmonic average is selected for dimensional reduction.

The PCA strategy is a usual dimension reduction method [21]. In this paper, we will extract 99% principal components in the experiment.

C. Topic Detection Algorithm

To overcome the drawbacks of existing algorithms, we bring two methods. The one is SSC, and the other is SMHP. We will introduce them in detail in next two sections.

Self-tuning spectral clustering algorithm. Manor et al. proposed a local scale similarity measure $s(i, j) = \exp(-\| x_i - x_j \|^2 / \sigma_i \sigma_j)$ [22] where σ_i is the distance between point x_i and its k-th nearest neighbor. With this local scale similarity measure, we have $\sigma_c > \sigma_b$, and then $\sigma_a \sigma_c > \sigma_a \sigma_b$. That means point a get closer to point c than to point b. The corresponding spectral cluste-

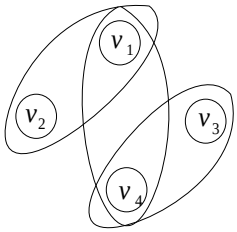

Figure 1. Hypergraph

ring based on this local scale similarity measure is called self-tuning spectral clustering algorithm.

Similarity Matrix based hypergraph partition algorithm. The definition of a hypergraph is given as:

Definition: A hypergraph G can be defined as a pair (V, E), where V is a set of vertices, and E is a set of hyperedges between the vertices. Each hyperedge is a set of vertices: $E \subseteq \left\{\{u, v, \cdots\} \in 2^V\right\}$ [23].

In SMHP algorithm, the most important step is how to transform similarity matrix to hypergraph expression. In this paper, the difference of stories is used and stories are taken as vertices. If the value of difference is smaller than a threshold β, the vertices may be partitioned in the same hyperedge. Following the process of the transformation process is illustrated. For the stories similarity matrix M shown as below,

$$M = \begin{Bmatrix} 1.00 & 0.92 & 0.45 & 0.54 \\ 0.92 & 1.00 & 0.87 & 0.79 \\ 0.45 & 0.87 & 1.00 & 0.36 \\ 0.54 & 0.79 & 0.36 & 1.00 \end{Bmatrix}$$

When the value of β is set to 0.1, the hyperedges are $e_1 = \{v_1, v_2\}$, $e_2 = \{v_3, v_4\}$ and $e_3 = \{v_1, v_4\}$. And the hypergraph is showed in Figure 1.

After transformation, the hypergraph partition tool [24] is called to obtain the clustering result.

Algorithm I shows the exact pseudocode used in this paper.

Algorithm I: SHMP algorithm for Topic Detection

```
For each document d in the corpus do
    Do pre-processing for d
    Compute terms weights
    Flag the length of document d
End for
For each document d do
    Compute harmonic average of all the documents' length
End for
Use H-TOPN and PCA strategies constructing VSM
Compute Similarity Matrix
For each column of the Matrix do
    Descend the column according the value
    Compute the differences diff of two rows
    If diff < β then
        construct a hyperedge
    End if
End for
Use HMETIS toolbox for clustering topics
Output the topic's labels
```

IV. Experimental Results and Analysis

A. Dataset and Evaluation Metrics

In this paper, we used two datasets for our experiment. The one is LDC [19] dataset TDT5, which contains news stories from April to September 2003. TDT5 contains 407,505 stories from source like Xinhua, Associated press, CNN, New York Times, China News Agency, Zaobao News etc. Unlike other TDT corpora, TDT5 does not contain any broadcast news data; all sources are newswire. Only mandarin stories in the collection were considered. In our experiment, we selected 324 stories about 56 topics. The other dataset used in the experiments is gathered from several online news channels. Each news story saved in a separate text file. There are totally 113 stories coming from 21 different topics.

TDT uses a cost function C_{Det}:

$$C_{Det} = C_{Miss} \cdot P_{Miss} \cdot P_{Target} + C_{FA} \cdot P_{FA} \cdot P_{Non-target}. \quad (7)$$

where $P_{Non-target} = 1 - P_{Target}$, and the meanings of symbols in (7) are listed in Table III.

Following a convention in the TDT evaluations, we assign $C_{Miss} = 1.0$, $C_{FA} = 0.1$ and $P_{Target} = 0.2$ .Then C_{Det} is normalized to $(C_{Det})_{Norm}$.

$$(C_{Det})_{Norm} = \frac{C_{Det}}{\min(C_{Miss} \cdot P_{Target}, C_{FA} \cdot P_{Non-target})}. \quad (8)$$

The minimum value of $(C_{Det})_{Norm}$ denoted as $MIN((C_{Det})_{Norm})$, is the optimal value that a system could reach at the best possible threshold [25].

TABLE III.
MEANINGS OF SYMBOLS

Symbol	Notation
C_{Det}	Cost of missing a new story and a false alarm.
C_{Miss}	Cost of missing a story.
P_{Miss}	Probability of missing a story.
P_{Target}	Probability of seeing a new story.
C_{FA}	Cost of a false alarm.
P_{FA}	Probability of a false alarm.
$P_{Non-target}$	Probability of seeing an old story.

TABLE IV.
SUBSYSTEMS OF SYSTEM-1

Name	Strategy
system-1-1	HCA+TFIDF
system-1-2	HCA+T-MI-TFIDF
system-1-3	HCA+T-MI-TFIDF+A-TOPN
system-1-4	HCA+T-MI-TFIDF+H-TOPN
system-1-5	HCA+T-MI-TFIDF+PCA
system-1-6	HCA+T-MI-TFIDF+A-TOPN+PCA
system-1-7	HCA+T-MI-TFIDF+H-TOPN+PCA

TABLE V.
SUBSYSTEMS OF SYSTEM-2

Name	Strategy
system-2-1	KCA+TFIDF
system-2-2	KCA+T-MI-TFIDF
system-2-3	KCA+T-MI-TFIDF+A-TOPN
system-2-4	KCA+T-MI-TFIDF+H-TOPN
system-2-5	KCA+T-MI-TFIDF+PCA
system-2-6	KCA+T-MI-TFIDF+A-TOPN+PCA
system-2-7	KCA+T-MI-TFIDF+H-TOPN+PCA

TABLE VI.
SUBSYSTEMS OF SYSTEM-3

Name	Strategy
system-3-1	SSC+TFIDF
system-3-2	SSC+T-MI-TFIDF
system-3-3	SSC+T-MI-TFIDF+A-TOPN
system-3-4	SSC+T-MI-TFIDF+H-TOPN
system-3-5	SSC+T-MI-TFIDF+PCA
system-3-6	SSC+T-MI-TFIDF+A-TOPN+PCA
system-3-7	SSC+T-MI-TFIDF+H-TOPN+PCA

TABLE VII.
SUBSYSTEMS OF SYSTEM-4

Name	Strategy
system-4-1	SMHP+TFIDF
system-4-2	SMHP+T-MI-TFIDF
system-4-3	SMHP+T-MI-TFIDF+A-TOPN
system-4-4	SMHP+T-MI-TFIDF+H-TOPN
system-4-5	SMHP+T-MI-TFIDF+PCA
system-4-6	SMHP+T-MI-TFIDF+A-TOPN+PCA
system-4-7	SMHP+T-MI-TFIDF+H-TOPN+PCA

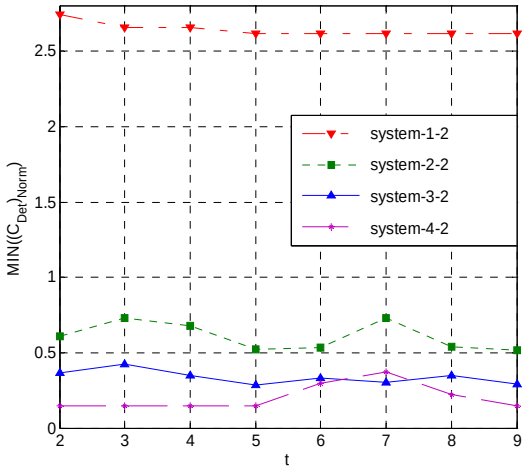

Figure 2. MIN((C_{Det})$_{Norm}$) with different subsystems and *t*

TABLE VIII.
VERIFY T-MI-TFIDF MODEL'S VALIDITY

System	MIN(C_{Det})$_{Norm}$	System	MIN(C_{Det})$_{Norm}$
system-1-1	2.7454	system-1-2	**2.6608**
system-2-1	0.9515	system-2-2	**0.5233**
system-3-1	0.3285	system-3-2	**0.2858**
system-4-1	0.2227	system-4-2	**0.1485**

B. Systems in Experiments

To validate the approaches proposed in the model, we implemented and tested 4 systems including 28 subsystems:

System-1 uses HCA (hierarchical clustering algorithm) and its subsystems listed in TABLE IV. KCA (K-means clustering algorithm) is used in **System-2** and its subsystems listed in TABLE V. **System-3** uses SSC algorithm and its subsystems listed in TABLE VI. SMHP algorithm is used in **System-4** and its subsystems in TABLE VII.

In this paper, the single-pass clustering algorithm is not selected because this method requires the document in order. The corpus collected is out-of-order, so the single-pass clustering algorithm is not tested.

C. Result and Discussion On collected Dataset

Verifying T-MI-TFIDF model's validity. Figure 2 shows the $MIN((C_{Det})_{Norm})$ in subsystems with different *t* values. The lower the $MIN((C_{Det})_{Norm})$ is, the better the system's performance is. From Figure 2, when $t = 5$, all of the four systems can achieve best performance. Thus we will choose $t = 5$ in the next experiments.

TABLE VIII shows the best result of using T-MI-TFIDF model. The best performance of system-1-2 decreases by 8.46% comparing to system-1-1. For system-2-2, the performance decreases by 42.82% compared with system-2-1. And system-3-2's performance is 4.27% lower than system3-1. In similar manner, system-4-2's performance is lower than system-4-1, and the decrement is 7.42%. This phenomenon explains it reasonable to use mutual information and

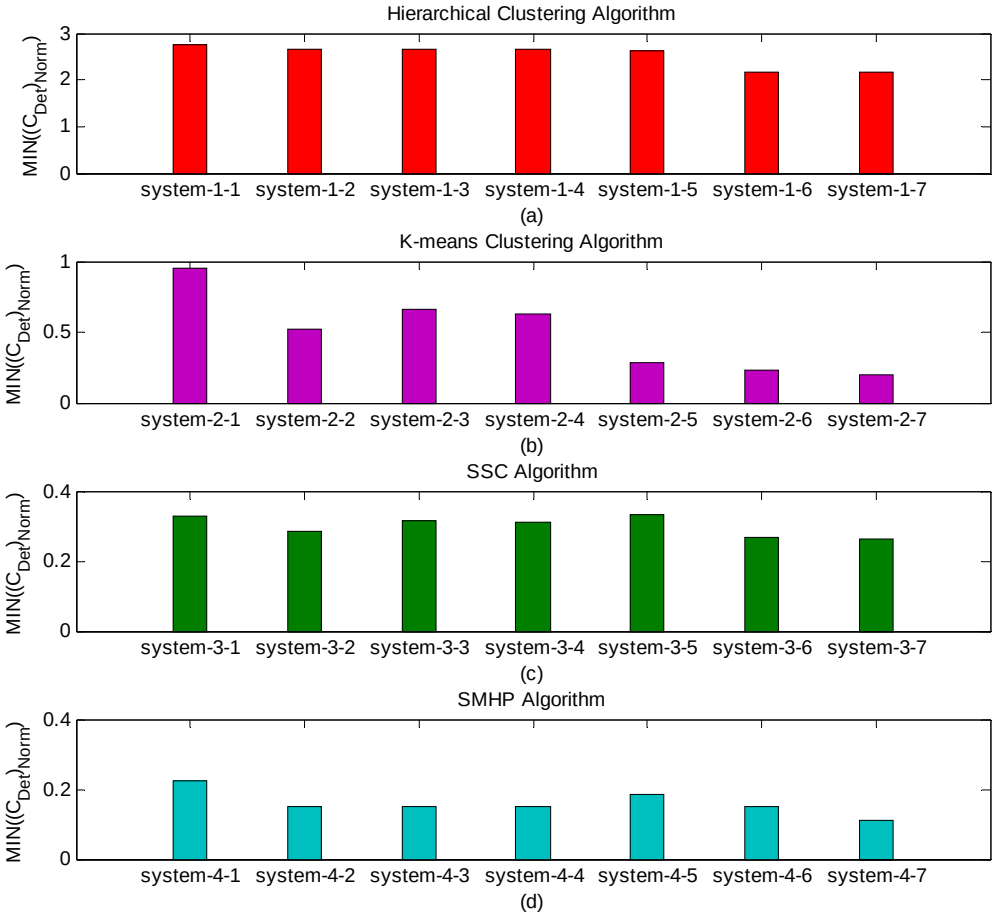

Figure 3. Results of using top-N and PCA strategies

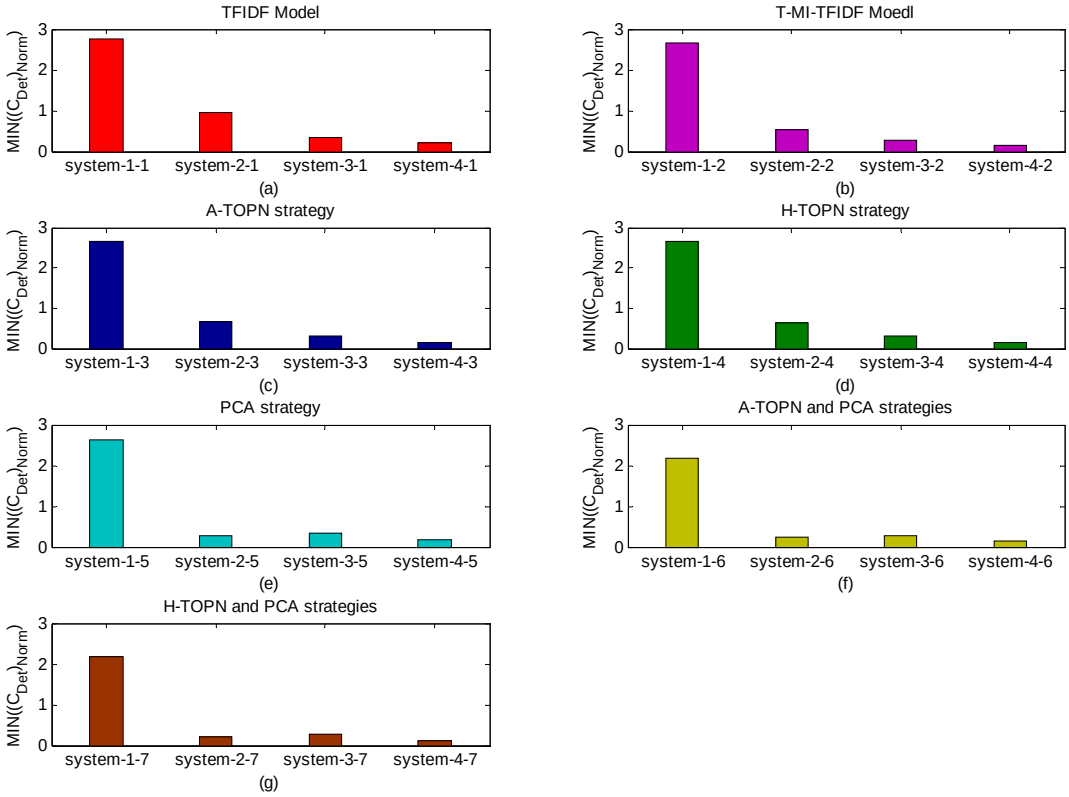

Figure 4. Results of using SMHP algorithm

the title information. Using the mutual information, the weight of features appeared frequently in the stories had been improved; thereby the ability of distinguishing different topics is enhanced. The title information heightened the system's performance as well. The title is the concentration of a story. Hence, T- MI-TFIDF model is more easily to take apart different topics through improving the feature's weight contained in the title.

Testing Top-N and PCA strategies. Figure 3 demonstrates the results of using top-N and PCA strategies for HAC, KCA, SSC and SMHP algorithms.

Through analyzing the results of Figure 3(a) with using HCA, the performance of system-1-5 (using PCA strategy only) is the best comparing to system-1-3 (using A-TOPN strategy only) and system-1-4 (using H-TOPN strategy only). The reason is that some noise had been removed with the PCA strategy. But some of useless information exists still when using H-TOPN strategy, which influences the performance of HCA.

For Figure 3(b) used KCA, system-2-5 shows the best performance compared with system-2-3 (using A-TOPN strategy only) and system-2-4 (using H-TOPN strategy only). The reason is the same as Figure 3(b)'s.

System-3-4 gets the best performance when using SCC comparing to system-3-3 (using A-TOPN strategy only) and system-3-4 (using H-TOPN strategy only) in Figure 3(c), which is different from Figure 3(a) and Figure 3(b). Because self-tuning spectral clustering algorithm is robust to noises. But the dimension decreases obviously with H-TOPN strategy, most of useful information had been restored. Thus the performance of system-3-4 is better than others.

The system-4-4's performance is the best compared with system-4-3 (using A-TOPN strategy only) and system-4-4 (using H-TOPN strategy only) in Figure 3(d). The reason is the same as Figure 3(c). In figure 3, the best performance for each sub graph is the last system (using H-TOPN and PCA strategies). The reason is maintaining the H-TOPN and PCA strategies' advantages and counteracting each other's disadvantages. Thus, the whole system's performance enhanced evidently.

Verifying the validity of SMHP algorithm. Figure 4 shows the results of SMHP algorithm with different strategies.

For each sub graph in Figure 4, the last second system's performance is better than the first two systems. Because SSC algorithm is good at clustering the data distributed inhomogeneous and this method resists the influence of noise strongly. However, other two methods except SMHP algorithm are adequate for clustering low noisy and low-dimensional data.

The best performance for each sub graph is the last system. Through the last paragraph's analysis, a part of reasons had listed. The most important reason is that this method considered the relationship among multiple stories. Just using the relationship, the system recognized the topics easily and exactly.

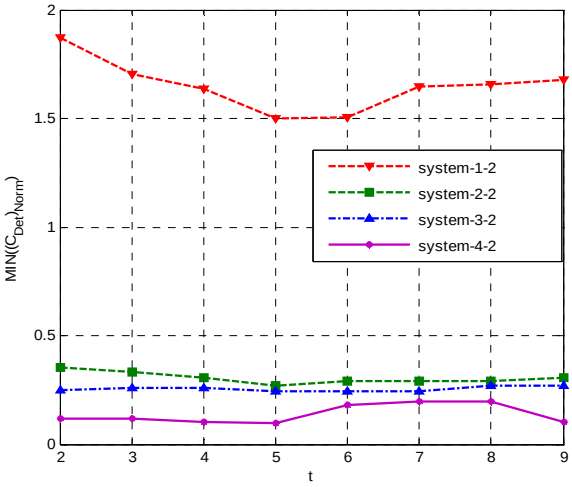

Figure 5. MIN((C_{Det})$_{Norm}$) with different subsystems and t

TABLE IX.
VERIFY T-MI-TFIDF MODEL'S VALIDITY

System	MIN(C_{Det})$_{Norm}$	System	MIN(C_{Det})$_{Norm}$
system-1-1	1.9234	system-1-2	**1.5023**
system-2-1	0.3607	system-2-2	**0.2720**
system-3-1	0.2734	system-3-2	**0.2420**
system-4-1	0.1534	system-4-2	**0.0983**

D. Result and Discussion On TDT5 Dataset

In order to testify the effectiveness of SMHP algorithm, additional experiments are taken on the TDT5 dataset. In Figure 5, when t = 5, all of the four systems can achieve best performance. And using t = 5, T-MI-TFIDF model is better than traditional TFIDF model in TABLE IX. Figure 6 shows the systems can enhance performance using H-TOPN and PCA strategies. From Figure 7, the last system in each sub graph achieves the best performance. It indicates the importance of relationship among multiple stories again.

V. CONCLUSION

In this paper, we have proposed a method for detecting topics in news articles. The algorithm performs well in practice compared to the baseline model. The complexity of SMHP algorithm is $O(|E|)$, which is lower than other algorithms'. It can be seen from the results that mutual information and title information help improving effectiveness of topic detection. Top-N and PCA strategies also improve the performance of clustering result. Our work makes three novel and important contributions:

1. T-MI-TFIDF model for computing the features weight.

2. H-TOPN and PCA strategies' combination for dimensional reduction.

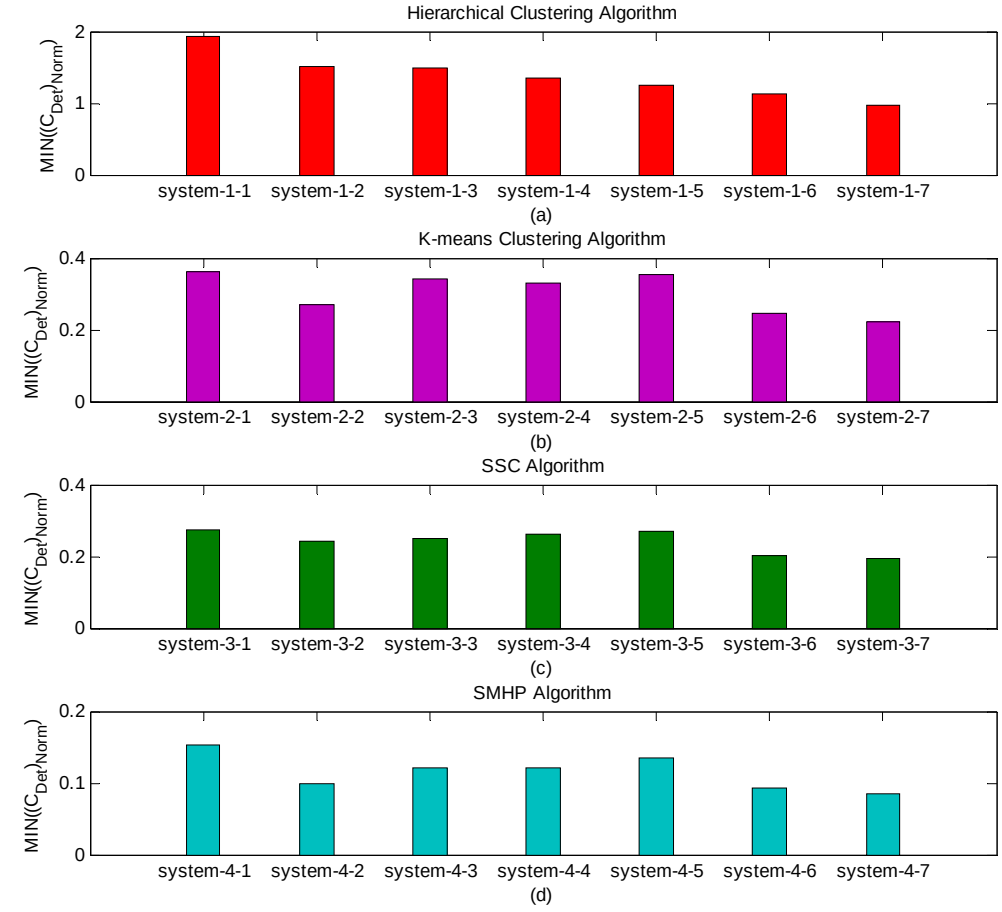

Figure 6. Results of using top-N and PCA strategies

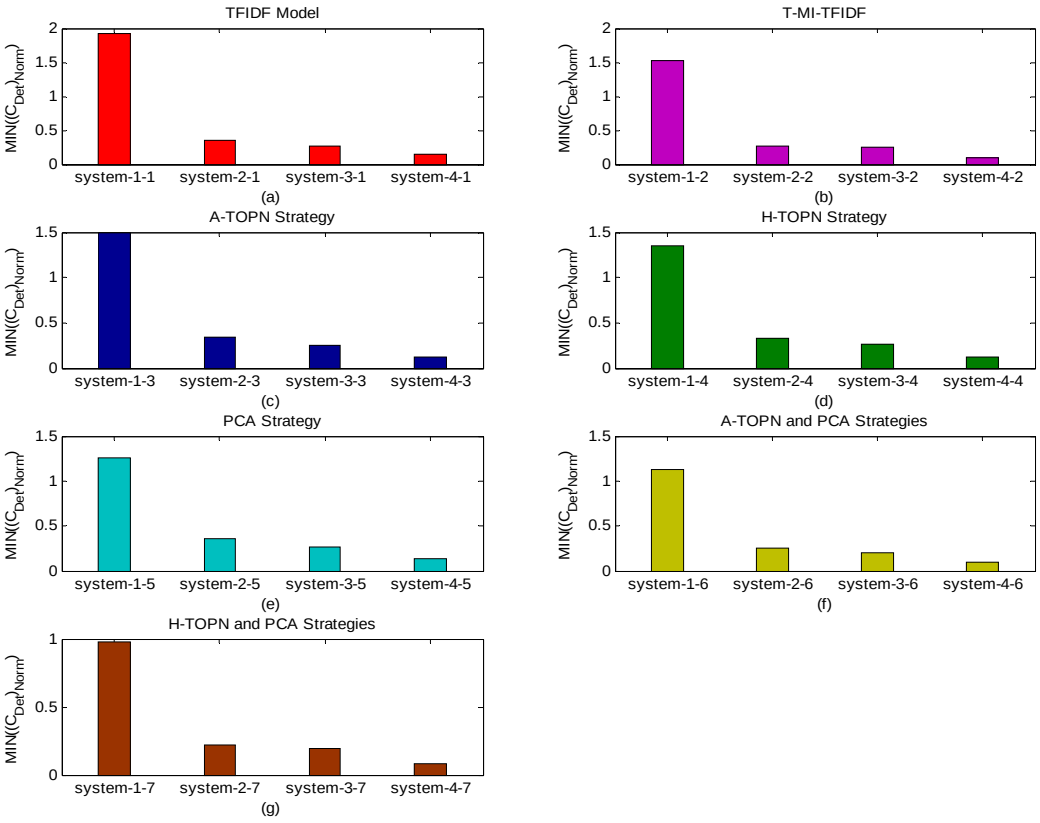

Figure 7. Results of using SMHP algorithm

3. SSC and SMHP algorithms for detecting topics in news stream.

For the Future work, we want to make use of named entities and time information. We also want take advantage of semantic information for distributing different topics.

ACKNOWLEDGMENT

This work is supported by the National Natural Science Foundation of China under Grant No. 60873180 and the Fundamental Research Funds for the Central Universities.

REFERENCES

[1] J. Makkonen, H. Ahonen-Myka, and M. Salmenkivi, "Topic Detection and Tracking with Spatio-Temporal Evidence", *Proceedings of the 25th European Conference on IR Research (ECIR 2003)*, Citeseer, 2003, pp. 251-265.

[2] D. Trieschnigg, W. Kraaij. "TNO Hierarchical Topic Detection Report at TDT 2004". *Topic Detection and Tracking Workshop Report*, 2004.

[3] C. Wartena, R.Brussee, "Topic Detection by Clustering Keywords", *Proceedings of the 2008 19th International Conference on Database and Expert Systems Application*, IEEE Computer Society, 2008, pp. 54-58.

[4] The 2004 Topic Detection and Tracing (TDT'04) Task Definition and Evaluation Plan, http://www.nist.gov/speech/tests/tdt/, 2004.

[5] TDT 2004: Annotation Manual Version 1.2, http://www.nist.gov/speech/tests/tdt/, Aug. 2004.

[6] C. Flynn, J. Dunnion, "Domain-informed Topic Detection", *Proceedings of the 6th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2004)*, Springer, 2004, pp. 617-626.

[7] Y. Yang, T. Pierce, J. Carbonell, "A Study of Retrospective and On-line Event Detection", *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, 1998, pp. 28-36.

[8] T. Brants, F. Chen, A. Farahat, "A System for New Event Detection", *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, 2003, pp. 330-337.

[9] M. Zhu, W. Hu, and O. Wu, "Topic Detection and Tracking for Threaded Discussion Communities", *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, IEEE Press, 2008, pp. 77-83.

[10] K. Zhang, J. Zi, and L.G. Wu, "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*, ACM, 2007, pp. 215-222.

[11] X. Zhang, T. Wang, "Topic Tracking with Dynamic Topic Model and Topic-based Weighting Method", *Journal of Software*, Vol. 5, No. 5, 2010, pp. 482-489.

[12] J. Allan, H. Jin, M. Rajman, C. Wayne, D. Gildea, and V. Lavrenko, "Topic-based novelty detection", *1999 Summer Workshop at CLSP Final Report*, http://www.clsp.jhu.edu/ws99/tdt, 1999.

[13] L.S. Larkey, F. Feng, M. Connell, and V. Lavrenko, "Language-specific models in multilingual topic tracking", *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, 2004, pp. 402-409.

[14] 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*, ACM, 2004, pp. 297-304.

[15] B. Li, W. Li, and Q. LU, "Topic tracking with time granularity reasoning", *ACM Transactions on Asian Language Information Processing (TALIP)*, ACM, 2006, pp. 388-412.

[16] J. Zeng, S. Zhang, "Incorporating topic transition in topic detection and tracking algorithms", *Expert Systems with Applications*, Elsevier, Vol. 36, No. 1, 2009, pp. 227-232.

[17] Introduction to ICTCLAS, http://sewm.pku.edu.cn/QA/reference/ICTCLAS/FreeICTCLAS/.

[18] S.S. Kou, and Z.J. Wei, "Improved Weighting Formula in Auto Text Classification", Computer Engineering and Design, Vol. 26, No. 6, 2005, pp. 1616-1618.

[19] J. Qiu, L. Liao, and P. Li, "News Recommender System Based on Topic Detection and Tracking", *Rough Sets and Knowledge Technology*, Springer, 2009, pp. 690-697.

[20] W.S. Jevons, *Investigations in Currency and Finance*, Macmillan & Co., London, 1884.

[21] A Tutorial on Principal Components Analysis, http://kybele.psych.cornell.edu/;edelman/Psych-465-Spring-2003/PCA-tutorial.pdf, 2002.

[22] L. Zelnik-Manor , and P. Perona, "Self-tuning spectral clustering", *Advances in neural information processing systems*, Citeseer, Vol. 17, 2004, pp. 1601-1608.

[23] Hypergraph, http://www.itl.nist.gov/div897/sqg/dads/HTML/hypergraph.html.

[24] HMETIS - Hypergraph & Circuit Partitioning, http://glaros.dtc.umn.edu/gkhome/fetch/sw/hmetis/hmetis-1.5.3-WIN32.zip.

[25] W. Zheng, Y. Zhang, Y. Hong, J.L. Fan, and T. Liu, "Topic Tracking Based on Keywords Dependency Profile", Information Retrieval Technology, Springer, 2008, pp. 129-140.

Xinyue Liu received the M.S degree in Computer Science and technology from Northeast Normal University, China, in 2006. She is currently working toward the Ph.D. degree in the School of Computer Science and Technology, Dalian University of Technology, Dalian, China. Her research interests include multimedia information retrieval, web mining and machine learning.

Fenglong Ma was born in Heilongjiang of P.R.China in 1986. He received his B.S. degree in Software Engineering from Dalian University of Technology in 2010, and currently is a master degree candidate in the same place. His major interests lie in topic detection and tracing.

Hongfei Lin received the Ph.D degree from Northeastern University, China. He is a professor in the School of Computer Science and Technology, Dalian University of Technology, Dalian, China. His professional interests lie in the broad area of information retrieval, web mining and machine learning, affective computing.