Extraction of Topic and Event Keywords from News Story

Hsi-Cheng Chang

† Department of Information Management, Hwa Hsia Institute of Technology

hcchang@cc.hwh.edu.tw

Abstract

The topic/event related keywords, i.e. key-verbs and key-nouns, identification in news stories always dominates the performance of the news processing. However, little literature has been published on the key-verbs and key-nouns identification in news stories. This paper proposes a topic/event detection method that exploits the characteristics of news writing and the grammar properties of language to identify the topic and event keywords that adequately capture the topical information of the news stories for improving the performance of the automatic news processing, such as news classification, topic detection and tracking, retrieval and summarization, etc. We apply a news clustering system to evaluate the effectiveness of the topic/event related keywords exaction approach. Experimental results show that the proposed method can extract commendably accurate topic and event keywords to represent the news and can efficiently produce news clustering with higher quality compared with the news clustering without topic/event detection processing.

Key words: Event detection, Keyword extraction, News classification.

1 Introduction

News report is always a key way of information dissemination. The well developed of Internet causes that Web news becomes one of the most important channels which people acquire the newly-emerged things in the daily lives. However, great quantity of information on the Internet are reproduced, disseminated and stored. Consequently, the information on the Internet are highly susceptible to redundancy, noise, and inconsistent. How to process the data and improve the quality of the data so as to promote the convenience of the information acquired of people becomes very important. The Topic Detection and Tracking (TDT) [1] study intends to explore techniques for detecting the appearance of new topics and for tracking the reappearance of them, and thus makes the people to handle the development of the news story easily. Three major tasks are involved in the TDT study:

1. The task of news story segmentation: segmenting a continuous stream of broadcast news stories into distinct constituent news stories based on the events they describe.
2. The task of event detection: identifying new events that are the first to occur in the news.
3. The task of event tracking: given a number of sample news articles about an event and finding all the subsequent news articles that discuss the same event.

The event tracking task is close to the document classification problem of information retrieval. Effective Web news classification can remove the redundancy

data and expedite the new information acquirement of people at the time of the explosive growth of the Internet. Many document clustering and classification studies have been presented for browsing documents or organizing the retrieval results for easy viewing [2],[3],[4],[5],[6],[7]. Most of these studies pay closer attention on the development of classification algorithm than the exploitation of admirable feature extraction scheme by assuming that document representations are available as inputs to the classification algorithm. However, there are many studies [8],[9],[10],[11],[12] proved that effective feature exaction will substantially improve the accuracy of document classifying.

Topic and event keyword extraction in news stories is the process of identifying the important keywords, i.e. the key-verbs and key-nouns, in news that bear most of the topical content of news stories. For a long time, topic and event related keywords extraction dominates the performance of news articles processing. Moreover, some problems make the conventional document classification methods frequently with poor performance for Web news classifying. First, Web news articles are always more concise than newspaper articles, suiting the habitual behavior of Web users. Thus, the topic and event related keywords are less apparent repeatedly, and furthermore the reporters may adapt different terms to descript the news events. Using simple feature selection manner for news categorization have difficulty in obtaining the meaningful topic/event keywords for representing the news and consequently most of the conventional document classification methods do not achieve satisfactory classification results. In order to obtain high accuracy news classification results, to identify the topic/event keywords effectively becomes very important.

Every kinds of text writing always comply with a definite writing style. News writing follows a set pattern of what is usually called as “inverted pyramid” that describes the five W’s – Who, What, When, Where and Why of the news story [13]. In addition, some grammar properties of language exist in different language, such as one of the most notable characteristics of Chinese sentence structure is that Chinese can be classified as a “topic-prominent’ language [17]. Based on the grammar properties of language and the characteristics of news writing, a topic/event detection algorithm is developed for detecting the topic/event pairs in news stories, and then extracts the topic/event keywords. The identified topic/event keywords portray the topics and events in the news collection. The experimental results show that using the topic and event keywords to represent the news can significantly enhance the accuracy of the news classification.

The following section describes the characteristics of news writing and grammar properties of language. Section 3 introductions the topic/event detection and topic/event keyword extraction method. Section 4 illustrates our experimental methodology and results. Section 5 concludes.

2 Topic and Event Definition

This section describes the characteristics of news writing and the grammar properties of language especially in the Chinese language. These properties give some evident notices for identifying the topic and event related keywords in news stories.

SARS is declared to be contained around the world *Hong Kong, July 5*

The World Health Organization declared today that SARS had been contained around the world, with no new cases reported to the agency by any country since June 15. But the agency warned that the disease could still pose a threat. ***[Who What When Where Why]***

The W.H.O. had removed the last place on its list of SARS-affected areas, Taiwan. "No new cases have been found there for 20 days, a span the agency believes to be twice the disease's incubation period. ***[Why]***

"SARS, or severe acute respiratory syndrome, has infected 8,439 people in 30 countries on five continents and has killed 812 people. Nearly 200 people with SARS are still being treated in hospitals around the world under strict isolation procedures to prevent them from infecting health-care workers." ***[What Where]***

Fig. 1. A news article extracted from the New York Times

2.1 Characteristics of News Writing

There are different kinds of writing, including essays, fiction, non-fiction, poetry, etc. In relation to news, there are only two major forms of writing: news writing and feature writing. News report portrays strictly on the occurrence and the course of a news event and so news writing is strictly based on news events. News writing follows a set pattern of what is usually called as "inverted pyramid". News writing must include answers to the five W's -Who, Where, When, What and Why [13]. News about SARS epidemic situation is shown in Fig. 1.

According to the definition of topic and event of TDT study [1], a topic is defined as "seminal activity or event, along with all directly related events and activities." Furthermore, an event is "something that happens at a specific time and place along with all necessary preconditions and unavoidable consequences." Base on the definitions of topic and event, and the properties of news writing. To determine which words are best used to represent the news, that are those words that answer the questions "Who","Where", "What", "When", and "Why", since they characterize the topic and event of the news. Accordingly, we define the following classes of keywords in news:

1. The words that answer the question "Who": for instance, the person name, organization name, legal person name, etc. These are central to the news stories, such as the World Health Organization, an organization name, shown in Fig.1.
2. The words that answer the question "Where": any locations occurring in the news that depict the news story take place, such as the Hong Kong shown in Fig.1. The proper name (includes person name, organization name, and place name, etc.) can be detected by a proper name identification algorithm [14],[15].
3. The words that answer the question "When": the time or date that the event occurs. That may be extracted from the off-the-shelf named-entity tagger of news articles [16].
4. The words that answer the question "What" and "Why": the words that depict the process of the event happened or explain why the event occurs. Generally, those are the key-verbs and key-nouns in news that bear most of the topics or events of news stories.

Except the person name, organization name, location name, time and date, etc., related studies and experiments have shown that accurate extraction of the key-verbs and

key-nouns from news stories can aid in better organization of news stories by their events [16],[19]. But, the studies also found to extract the key-verbs and key-nouns from news is very difficult and on research findings are yet available to date. The primary research questions to be addressed in this study are to develop an approach to identify the key-verbs and key-nouns in news stories to promote the convenience of the news articles processing. To identify the topics of news stories is especially difficult. We solve this problem with the help of grammar properties of language that will be described in following section.

2.2 Grammar Properties of Language

Theoretically, a sentence comprises two major components: the subject and the predicate. The subject is a nominal expression and the predicate comprises a verb and some optional elements. While these basic principles appear to be universally applicable in all languages, some differences still exist among languages. For example, the sentences of both Chinese and English are following the order of **S**ubject + **V**erb + **O**bject (SVO). However, this basic SVO word order can be altered once other factors are considered. For example, languages all have methods of expressing negation, existence and causation, modifying nouns and verbs, asking questions, and combining simple sentences into compound/complex sentences. However, languages also differ from one another many interesting ways. The languages of the world can be classified into three main groups based on the order of the verb and nouns in simple sentences, i.e. the SVO, SOV and VSO word order types. In the great majority of languages the subject comes before the object, such as English that with the typical word order for most sentences is to have the verb in the middle would be SVO type. A language with the typical word order for most sentences is to have the verb at the end would be SOV type, such as Japanese [17].

Chinese is not easy to classify in terms of word order, since the notion of subject is not structurally well defined in the grammar of Chinese. Generally, the words and phrases occur is determined by considerations of meaning rather than of grammatical functions. This means that verbs in Chinese sentences can be at the beginning, in the middle or at the end. Moreover, Subject is difficult to identify based on sentence structure in Chinese. One of the most notable characteristics of Chinese sentence structure, and one that differentiates many other languages is that besides the grammatical relations of “subject” and “direct object”, the sentence structure of Chinese includes the element “topic”. Because of the importance of “topic” in the grammar of Chinese, Chinese can be classified as a topic-prominent language [17].

Basically, the topic of a sentence is what that sentence is about. The topic always comes at the beginning of the sentence and always refers to something about which the speaker assumes the person listening to the utterance has some knowledge. Furthermore, a topic can always optionally be followed by a pause in speech, which serves to set the topic, that which is being talked about, apart from the rest of the sentence. The difference between topic and subject is that the subject must always have a direct semantic relationship with the verb as the one that performs the action or exists in the state named by the verb, but this is not necessary for the topic.

Topic-prominent sentence structure is a notable topological feature of Chinese and is important in comparisons with other languages, such as shown in Fig. 2. This topic-prominent sentence structure frequently appears in the headline of news

articles. According to these grammar properties, we can identify the topics and topic keywords in news stories easily.

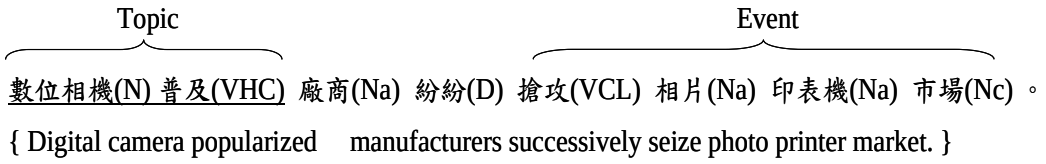

Fig. 2. An example of topic-prominent sentence structure in Chinese.

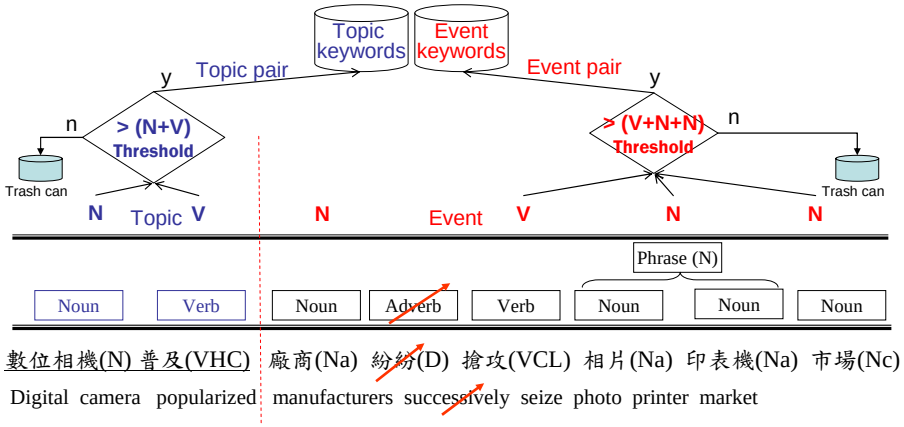

Fig. 3. Topic/Event keywords extraction.

3 Topic and Event Identification

In this section, we discuss the main ideas of the development of the topic/event detection algorithm and the topic and event keywords extraction method.

3.1 Topic and Event pairs Detection

An event is defined as "something that happens at a specific time and place along with all necessary preconditions and unavoidable consequences [1]." This definition means that an event implies an action or a state of being. According to the definition of grammar the word that can be used to express action or a state of being is the verb. There are two different types of verbs: action verbs and linking verbs that describe the subject and form different sentence pattern. Here are the examples:

Action verbs	**Linking verbs**
Davie *wrote* a paper.	The coffee *smelled* wonderful.

A verb phrase is a phrase headed by a verb that expresses the action or a state of being of the sentence. Most verb phrases consist of a verb head together with that verb's complement. A complement is whatever is required by a particular verb to make a complete sentence. Accordingly, we can extract the event keywords in the news stories through identifying the verb phrases in the sentences. The identified verb phrases are regarded as the candidate topic/event pairs in the news stories. In this paper, some major candidate topic/event pair patterns are considered and illustrated in the following:

- Someone/Country/Organization does,

that is, the “Noun+Verb” pattern. For example: the verb phrase of 加拿大防疫 (Canada epidemic control).

- Someone/Country/Organization do someone/something, that is, the “Noun+Verb+Noun” pattern. For instance: the verb phrase of 中國公佈疫情 (China announces the epidemic situation).
- Do something, that is, the “Verb+Noun” pattern. For instance: the verb phrase of 排除障礙物 (… obviate an obstacle).

Based on the grammar properties of language and the probable variation of forms of topic/event patterns in Chinese, the candidate topic/event pairs are defined as the patterns that consist of verbs and nouns occurring in succession and form the patterns of N+N+V, N+V, N+V+N, V+N+N, V+N, etc. Fig. 3 illustrates the candidate topic/event pair detection. The nouns and verbs contained in the candidate event pairs are probable the key-nouns and key-verbs of news stories. A key-verb in an event pair describes an action occurring in the story, and a key-noun occurring in an event pair that may play a player of an action or the objective of an action. The identified key-verbs and key-nouns portray the topics and events in the news stories that also answer the question “What” and “Why”.

3.2 Topic and Event Keyword Extraction

Fig. 4 gives a sketch of the keyword extraction algorithm. Five major steps are involved in the keyword extraction approach.

Sentence parser: a sentence parser unit is used to partition the paragraphs into sentences and identify terms from sentences.

Word segmentation: Chinese is a character-based language. A sentence in Chinese must be segmented into meaningful single- and multi-character terms. A Chinese word segmentation program is used to process Chinese text and determine the part-of-speech of each term in the news corpus.

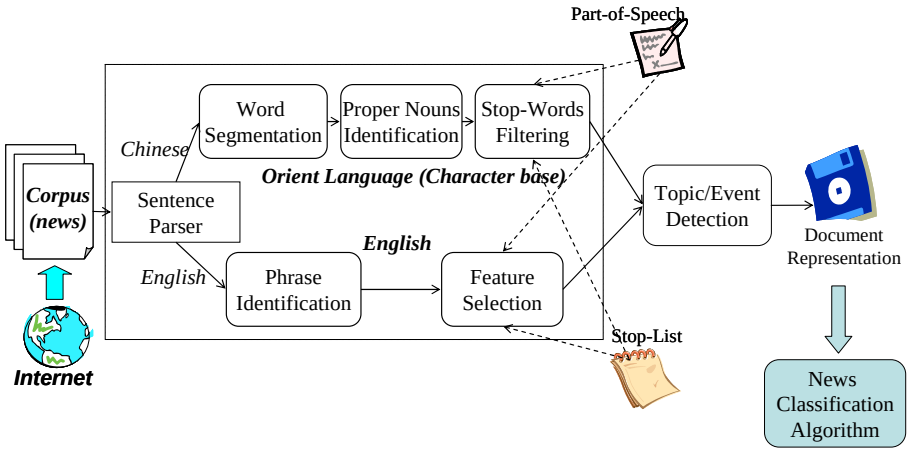

Fig. 4. The process of keyword extraction

Proper nouns identification: a proper name identification was used to identify the person name, organization name, legal person name, etc. [14],[15]. Moreover, meaningful phrases in documents must be identified since they profit for topic representations and improve the possibility of obtaining human-understandable results [3],[4]. A phrase identification program, a statistical technique and a modified DHP algorithm [15],[18], that is suitable for use in English and Chinese are developed and

applied to identifying meaningful phrases in the document collection.

Stop-word filtering: to maintain the computing cost of the news articles processing small and increase the processing accuracy, removing the words that are not meaningful and discriminative among topics is very important. In this study, an aggressive data clearing approach is used to remove the meaningless terms from the news articles [19].

Topic/Event detection: according to the patterns of topic/event pair defined in section 3, the topic/event pair detection algorithm identifies all the candidate topic/event patterns in news articles. The topic/event detection algorithm calculates the frequency of each candidate topic/event pair and reserves the topic/event pair whose frequency exceeds a predefined threshold such as delineated in Fig. 5. The thresholds are different with the types of topic/event pairs. For example, a predefined threshold NVN_THRESHOLD for "Noun+Verb+ Noun" pattern, a NV_THRESHOLD for "Noun+Verb" pattern, etc. The values of the thresholds must be dynamically adjusted with the news collection. The key-nouns and key-verbs are extracted from the identified topic/event pairs, which are the most discriminative topic/event keywords among the topics and events in the news collection, can significantly enhance the performance of the news processing.

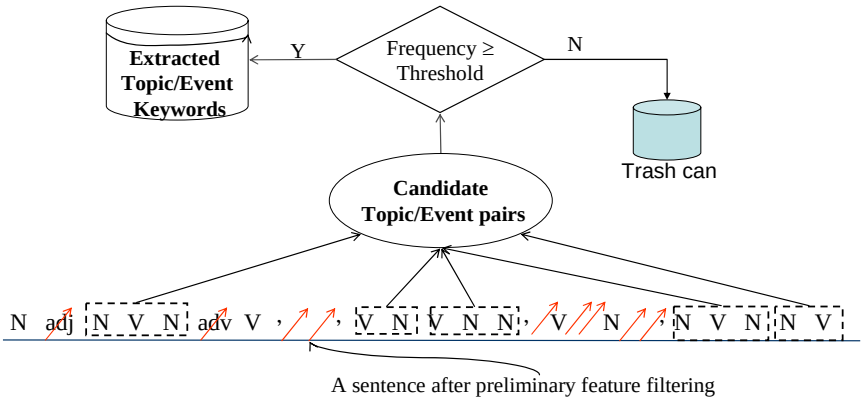

Fig. 5. Topic/Event keyword extraction

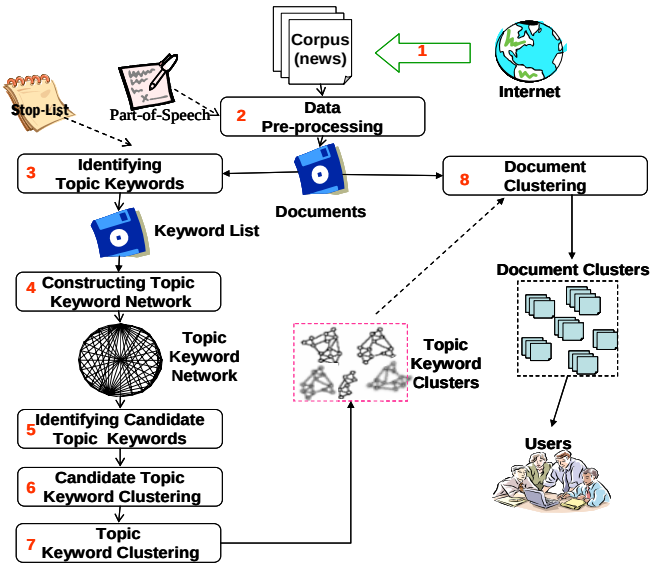

Fig. 6. The document clustering system.

4 Experiments and Discussion

To clarify the relative contribution of this study, we apply a document clustering system [19] to evaluate the performance of the topic/event keywords extraction method. Fig. 6 illustrates the sketch of the document clustering system. The document clustering method adopts a multi-stage process. First, an aggressive data cleaning approach is employed to reduce the noise in the free text and further identify the topic keywords in the documents. All extracted keywords are mapped into an undirected weighted graph where a vertex denotes a keyword and an edge denotes an association between two keywords. The keywords with high component weights are selected as *candidate* topic keywords. Then, the *k*-nearest neighbor approach is employed to find the candidate topic keyword clusters and further to form the topic keyword clusters. Finally, the generated topic keyword clusters are used to find documents related to the topics. We omit the detail due to the limitation of the paper length.

4.1 Testing Corpus

The testing data is accumulated from some well-known news Web sites, including "*Udn news*" (http://udn.com/NEWS/mainpage.shtml), "*Yahoo!*" (http://tw.yahoo.com), "*Chinatimes*" (http://news.chinatimes.com/), etc. Some keywords, covering 18 topics listed in Table 1, are entered as queries, and 100 documents about each topic are gathered. The kind of the news corpus is focused on news event reported and the average length of the news articles is about 226 terms.

Table 1: The topics of the testing corpus.

Topics: (18)	A_ Football game	G_ Notebook	M_ Computer virus
	B_ Cellular phone	H_ Environment protection	N_ Plasma TV
	C_ Fixed Network	I_ SARS	O_ Baseball game
	D_ Typhoon	J_ Credit card	P_ Air casualty
	E_ Finance	K_ Educational reform	Q_ Highjack.
	F_ Weight reduce	L_ The presidential election	R_ Music

In this study, some tests are proceeded to evaluate the effectiveness of the topic/event detection method upon the document clustering results by reporting the precision rate of the clustering results. The precision measure is a well-known metrics in IR community. The clustering precision rate is defined as follows:

$$\text{Precision} = \frac{\text{The number of documents found by a clustering method and belonging to the correct cluster}}{\text{The number of documents in the cluster}} \quad (1)$$

4.2 Experiment results

The experiment attempts to evaluate the effectiveness of the topic/event keywords extraction for news clustering. Some experiments are processed and one of the experimental results is shown in Fig. 7, where Y-coordinate represents the precision rate of the clustering results and X-coordinate depicts the number of keywords selected from each news articles in the news corpus. The experiment result shows the clustering accuracy promoted average 9.8% compared with the news clustering without using topic/event detection processing.

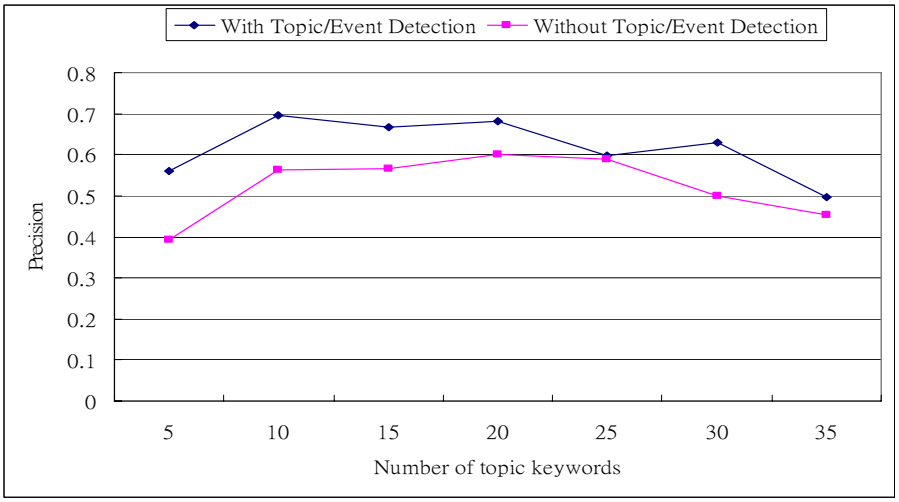

Fig. 7. The comparison of the clustering results

5 Conclusion

Studies proved that better feature selection of data can substantially improve the performance of the data processing. In the studies of event tracking or news classification, keyword extraction strongly influences the accuracy of the classification. Simple feature filtering metrics or statistical models can remove numbers meaningless terms from documents but have difficulty in obtaining the appropriate topic and event related keywords to represent the news stories. In this study, we consider and make use of the knowledge of the characteristics of news writing and the grammar properties of language to develop a topic/event detection algorithm for extracting the key-nouns and key-verbs that are usually linked to the name-enteritis and place name typically to express actions in news story. The experimental results shown the effectiveness of the topic/event keyword extraction method evidently promoted the accuracy of the news clustering.

Reference

1. Allan,J., Carbonell, J., Doddington, G.., Yamron, J., Yang, Y.: Topic Detection and Tracking Pilot Study Final Report. Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop, (1998) 194-218
2. Aggarwal, C. C., Gates, S. C., Yu, P. S.: On the merits of building categorization systems by supervised clustering. in proceedings of the fifth ACM SIGKDD international conference on knowledge discovery and data mining, (1999) 352-356
3. Lai, Y. S., and Wu, C. H.: Meaningful term extraction and discriminative term selection in text categorization via unknown-word methodology. ACM transactions on Asian language information processing, vol. 1, no. 1, March (2002) 34-64
4. Jain, A. K., Murty, M. N., Flynn, P. J.: Data Clustering: A Review. ACM computing surveys, vol. 31, no. 3, September (1999) 264-323
5. Karypis, G.., Han, E. H., Kumar, V.: CHAMELEON: a hierarchical clustering algorithm using dynamic modeling. IEEE computer (1999) 68-75
6. Lin, S. H., Chen, M. C., et al.: ACIRD: Intelligent internet document organization and retrieval. IEEE transactions on knowledge and data engineering, vol. 14, no. 3, May/June (2002) 599-614

7. Clifton, C., Cooley, R., Rennie, J.: TopCat: data mining for topic identification in a text corpus. IEEE transactions on knowledge and data engineering (2003) 2-17
8. Koller, D., Sahami, M.: Hierarchically classifying documents using very few words. in proceedings of the 14th International conference on machine learning, Nashville, Tennessee, July (1997) 170-178
9. Yang, Y., Pedersen, J. O.: “A comparative study on feature selection in text categorization. in proceedings of ICML97, (1997)
10. Aggarwal, C. C., Wolf, J. L., Yu, P. S., Procopiuc, C., Park, J. S.: Fast algorithms for projected clustering. in proceedings of ACM SIGMOD, vol. 28, no. 2, June (1999) 61-72
11. Yang, Y.: Noise reduction in a statistical approach to text categorization. in proceedings of ACM SIGIR, (1995) 256-263
12. Yang, Y., Wilbur, J.: Using corpus statistics to remove redundant words in text categorization. in proceedings of JASIS, (1996)
13. Chiang, O.: The Narrative Structure of a News Report. Time for Students, vol.47, September (2003) 11-12
14. Chen, H. H., Lee, J. C.: Identification and Classification of Proper Nouns in Chinese Texts. Proceedings of 16th International Conference on Computational Linguistics, (1996) 222-229
15. Tsai, K. H.: On the Chinese Document Clustering Based on Dynamical Term Clustering. a dissertation of master, National Taiwan University of Science and Technology, (2003)
16. Nallapati, R., Allan, J., Mahadevan, S.: Extraction of Key Words from News Stories. CIIR Technical report #IR-345, Center for Intelligent Information Retrieval, Department of Computer Science, University of Massachusetts, (2004)
17. Charles, N. L. Sandra, A. T.: Mandarin Chinese-a functional reference grammar. The Crane Publishing Co., (1982)
18. Tseng ,C. M, Tsai ,K. H., Hsu, C. C., Chang, H. C.: On the Chinese Document Clustering Based on Dynamical Term Clustering. Lecture Notes in Computer Science, (2006)
19. Chang, H.C., Hsu, C. C.: Using Topic Keyword Clusters for Automatic Document Clustering. IEICE Transactions on Information and Systems, vol.E88-D, August (2005) 1852-1860