Multi-Document Summarization for Terrorism Information Extraction

Fu Lee Wang[1], Christopher C. Yang[2], and Xiaodong Shi[2]

[1] Department of Computer Science, City University of Hong Kong, Hong Kong SAR, China
flwang@cityu.edu.hk
[2] Department of Systems Engineering and Engineering Management,
The Chinese University of Hong Kong, Shatin, Hong Kong SAR, China
{yang, xdshi}@se.cuhk.edu.hk

Abstract. Counterterrorism is one of the major challenges to the society. In order to flight again the terrorists, it is very important to have a through understanding of the terrorism incidents. However, it is impossible for a human to read all the information related to a terrorism incident because of the large volume of information. Summarization technique is urgently required for analysis of terrorism incident. In this work, we propose a multi-document summarization system to extract the critical information from terrorism incidents. News stories of a terrorism incident are organized into a hierarchical tree structure. Fractal summarization model is employed to generate a summary for all the news stories. Experimental results show that our system can effectively extract the most important information for each incident.

1 Introduction

After September 11th of 2001, the public realized that terrorist attacks are threatening us anywhere over the world. In order to flight against the terrorists effectively, it is important to build a knowledge base of terrorism. However, there is a large volume of information related to a terrorism incident. It is impossible for a human to digest all the information. The problem of information-overloading can be reduced by automatic summarization. Research in this area is very essential for counterterrorism.

Many summarization models have been proposed [5, 6]. Traditionally, summarization system considers a document as a sequence of sentences. They calculate the significance of sentences. The most significant sentences are then extracted and concatenated as a summary. It was shown that the document structure is important in both automatic summarization [1] and human abstraction [2]. The fractal summarization was proposed based on hierarchical structure of document [10]. Experiment results showed that fractal summarization is a promising summarization technique.

Summarization has been extended to multi-document summarization [7]. Given a set of flat-structured documents, the summarization system identifies the similarities among the documents. The sentences are extracted based on their similarity measurement. However, there is not a trivial way to organize a set of flat-structured documents into a hierarchical tree structure. This paper investigates the impact of different hierarchical structure to the summarization technique. Experiments are conducted on terrorism incidents to determine how the summarization techniques perform in extracting the terrorism information from a set of documents.

2 Fractal Summarization Model

The information overloading problem can be solved by automatic summarization. A lot of summarization models have been developed. Traditional automatic summarization is the selection of sentences from the source based on their significance to the document [6]. The thematic, location and heading are the most widely used features.

A large document can be represented as a tree structure with several levels. At the lower abstraction level of a document, more specific information can be obtained. Studies of human abstraction had shown that the human abstractors extract the topic sentences according to the document structure [1, 2]. Fractal summarization model was proposed to generate summary based on the hierarchical structure of the document and fractal theory [10]. The detail is shown as the following algorithm:

Fractal Summarization Algorithm

1. Calculate the Sentence Quota of the summary.
2. Divide the document into range blocks and transform the document into fractal tree.
3. Set the current node to the root of the fractal tree.
4. **Repeat**
   4.1 **For** each child node under current node,
   Calculate the fractal value of child node.
   4.2 Allocate Quota to child nodes in proportion to fractal values.
   4.3 **For** each child nodes,
   If the quota is less than threshold value
   Select the sentences in the range block by extraction
   Else
   Set the current node to the child node
   Repeat Step 5.1, 5.2, 5.3
5. **Until** all the child nodes under current node are processed

In fractal summarization, the document is partitioned and transformed into a hierarchical tree structure according to its document structure. For each node, the system calculates the Range-block Significance Score by summing up sentence significance scores under the range-block. The fractal value of the root node is 1, and it is prorogated to other nodes directly proportional to its significance score. Then, the system calculates the number of sentences to be extracted according to the compression ratio. The number of sentences is assigned to the root of document tree as the quota of sentences. The quota is allocated to child-nodes by propagation directly proportional to the fractal value of the child-nodes. The quota is then iteratively allocated to child-nodes of child-nodes until the quota allocated is less than a threshold value and the range-block can be transformed to some key sentences.

The fractal values of the nodes in the hierarchical fractal structure are computed based on the traditional salient features. To fully utilize the fractal structure of document, the traditional salient features are fractalized as fallowing:

- It is believed that a term carries different weight in different location of a full-length document [3]. In fractal summarization, the *tfidf* of a term in a range block is defined as proportional to the term frequency within a range-block and inversely proportional to the frequency of range-block containing the term.
- Fractal summarization calculates the location weight based on which document-level we are looking at. We calculate the location weight for a range-block, all sentences inside a range-block will receive same position weight.

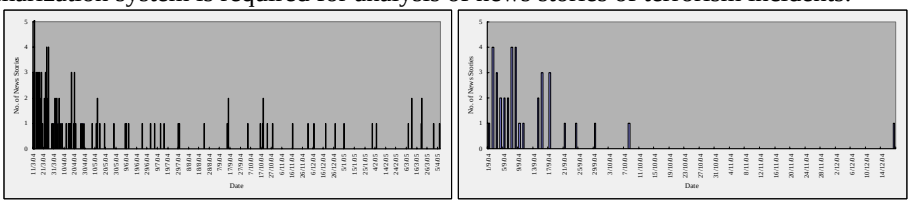
- At different abstraction level, some headings should be hidden and some headings are emphasized. Moreover, the significance of the heading is inversely proportional to its distance from the sentence. Propagation of fractal value [4] is a promising approach to calculate the heading weight for a sentence.

Our experiments showed that the fractal summarization model produce a summary with a wider coverage of information subtopic than traditional (non-hierarchical) summarization model. The precision of fractal summarization model outperforms the traditional summarization significantly at 99% confidence level [10].

3 Hierarchical Structure of News Stories for Terrorism Incident

To prevent a terrorism incident, it is very important to have a through understanding of the incident. However, analysis of the incident is infeasible without help of summarization tool because of a large number of documents. Multi-document summarization systems have been developed for flat-structured documents. Advanced technique is required for analysis of structured documents.

Typically, multi-document summarization systems consider the documents as individual documents in a flat-structure [7]. However, a set of news stories related to a terrorism incident has a more complicated structure. A timestamp is associated with each news story. The distribution of news stories is not uniform along the timeline. Moreover, the news stories can be classified into events [11], and the number of news stories is not uniform in all events. As a result, a more advance multi-document summarization system is required for analysis of news stories of terrorism incidents.



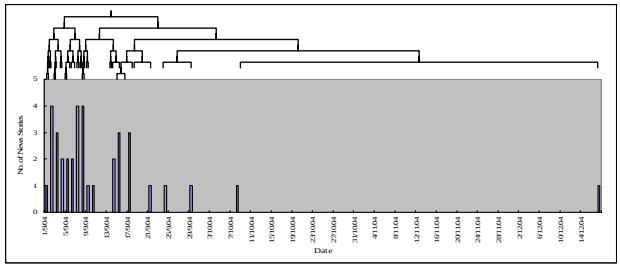
(a) “Madrid Train Bombing” (b) “Beslan School Hostage Crisis”

Fig. 1. Distribution of News Stories of Terrorist Attack Incident vs. Time.

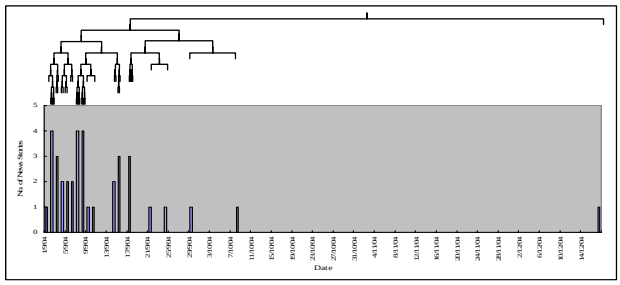
In order to have an insight understanding of terrorism incidents, two terrorist attack incidents have been analyzed. Related news stories have been collected from the CNN.com. In the figures of distribution of news stories against time (Fig. 1), obvious peaks can be identified at the beginning of each incident. The peak is caused by a large number of news stories soon after the burst of the incident. Then, the number of news stories decreases as time goes by. As shown in Fig. 1, the “Madrid train bombing” has a more long-term impact; therefore, 115 news stories continually appear over more than one year. However, the “Beslan school hostage crisis” has only a short-term impact, 36 news stories appeared in the first month, and then it remains silent for about two month until last story in the third month.

It has been shown that the document structure is important in summarization [1, 2]. Therefore, the news stories are organized into a hierarchical tree. Taking consideration of the temporal and semantic information of news stories, we have investigated three alternatives:

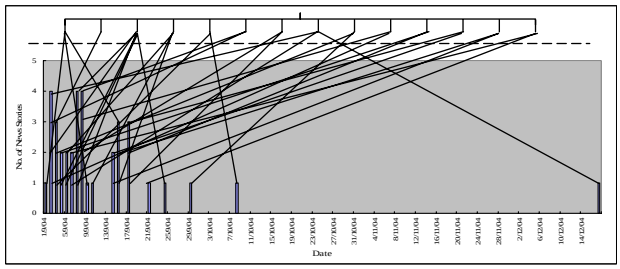
1. The news stories are organized by the number of stories (Fig. 2a). Our previous result showed that a good summary must extract information distributively [10]. Moreover, when an author writes a document he distributes information evenly into sections. Therefore, we consider all stories equally significant and they are evenly distributed into a tree structure. We propose to organize the tree such that the nodes at same depth contain same number of news articles.
2. The news stories are organized by the time interval (Fig. 2b). Temporal text mining technique has been applied to multi-document summarization [**Error! Reference source not found.**]. Summarization of news stories are generated for each period of fixed number of days, then an overall summary is generated. We propose to organize news stories in a tree such that each child represents an equal and non-overlapping interval.
3. The news stories are organized by the event topics (Fig. 2c). The research of information retrieval focus on detecting of event topics of news stories [11]. A good summary must extract information from each event topics [9]. Recent research in automatic summarization proposes to classify the documents into document set before summarization [8]. Therefore, we propose to organize by event topics. As the accuracy of event topic detection affects the performance of the summarization directly, the stories are clustered into events topics by human professional [11].



(a.)by No of News Stories



(b) by Time Interval



(c) by Event Topics

Fig. 2. Hieratical Structure of “Beslan School Hostage Crisis” Incident

4 Impact of Hierarchical Structure to Summarization

News stories of a terrorism incident can be organized into different hierarchical structures. They are great different from each other (Figure 2). To investigate the impact of hierarchical structure to summarization, the machine generated summaries of news stories using different structures are compared with abstracts composed by human professional to measure the quality of summaries.

To investigate the impact of hierarchical structure to the performance of automatic summarization, experiments have been conduced on previous two terrorism incidents. The news stories are organized as three alternatives in previous section. As there are relatively more children in the hierarchical tree by event topics, in order to have a fair comparison, we have considered hierarchical tree with different number of children for first two structures in addition to binary tree. Fractal summarization model is then applied to summarize these two incidents.

The fractal summarization for news stories is very similar to the fractal summarization of large text document, only some minor modifications are required to demonstrate the characteristic of the news stories.

- First, there is no heading for the internal node in the tree. As a result, the heading feature will only be considered at the news headings of individual news stories and the theme of the incident.
- The location feature in traditional summarization assumes that the text unit in the beginning or ending is more important. The news stories inside a node of news tree are considered as equally significant. Therefore, the location feature is not considered during summarization of news stories.

As the high-compression ratio abstracting is more useful, the news stories in our experiment are summarized with 5% compression ratio. To measure the precisions, the summaries generated by machine are compared with abstracts composed by human professional by gold standard [5]. The precision of a summary is calculated as the percentage of sentences selected by human professional (Table 1).

Table 1. Precision of Summaries with Different Hierchical Stucture

Incident	By No. of Stories			By Time Interval			By Event Topics
	Degree-2	Degree-3	Degree-4	Degree-2	Degree-3	Degree-4	
Madrid Train Bombing	62.1%	58.9%	63.2%	58.9%	58.9%	55.8%	71.6%
Beslan School Hostage Crisis	58.7%	56.5%	60.9%	56.5%	63.0%	58.7%	82.6%

When an author writes a large document with a lot of information, he groups similar information into same sections [2]. Therefore, classification of news stories into event topics simulates the process of an author writing a large document. It gives a more natural classification of the news stories. However, the classifications of news stories by number of news stories and by time interval partition the news stories by brute force. The themes among stories are not preserved. Therefore, the precision of summaries of news stories with hierarchical structure by event topics is significantly higher than the other two (Table 1).

News stories in hierarchical structure by number of stories and by time interval are both organized by their order along the timeline. Therefore, there is not much difference among two structures, and their precisions are similar (Table 1). On the other

hand, there is no significant difference among the precisions for different degrees (Table 1). Changing the degree will not change the intra-stories relationship; they are still organized by time ordering. It will only change the amount of information inside each node. The fractal summarization model calculates the significance of each node by the amount of information inside the node, and the quotas are assigned accordingly. However, organizing news stories into hierarchical structure by event topics makes a fundamental change in the organization. As observed in the experiment, there is a substantial improvement in the precision.

Moreover, it is believed that a good summary must cover as many topics as possible and the redundant information within a topic must be eliminated [9]. If the news stories are organized into hierarchical tree by event topics, the fractal summarization extracts sentences distributively among the event topics. It also ensures that the most significant node will not dominate the summary. A more balanced quota among event topics can eliminate the possibility of redundant information as well. The fractal summarization model summaries a large document in the similar way as a human abstractor. Therefore, it is promising technique to summarize multiple documents.

5 Conclusion

Automatic summarization of multiple news stories is very useful to extract terrorism information from a large volume of information. Three hierarchical structures of news stories have been investigated in this paper. Experimental results show that the summarizations of news stories with hierarchical structure classified by event topics outperform the other two structures. The fractal summarization model together with hierarchical structure classified by event topics becomes a promising multi-document summarization system for multiple news stories. This novel approach provides an essential analytical tool for terrorism incident.

References

1. Endres-Niggemeyer B. et al, How to Implement a Naturalistic Model of Abstracting: Four Core Working Steps of an Expert Abstractor. Info. Proc. & Manag., 31(5):631-674, 1995.
2. Glaser B.G. et al., The Discovery of Grounded Theory. Aldine de Gruyter, NY, 1967.
3. Hearst M. Subtopic Structuring for Full-Length Document Access. SIGIR'93, 56-68, 1993.
4. Koike, H. Fractal Views: A Fractal-Based Method for Controlling Information Display. ACM Tran. on Information Systems, 13(3), 305-323, 1995.
5. Kupiec J. et al. A Trainable Document Summarizer. SIGIR'95, 68-73, 1995.
6. Luhn H.P. The Automatic Creation of Literature Abstracts. IBM J R&D, 159-165, 1958.
7. McKeown K.R., et al., Columbia multidocument summarization: Approach and evaluation. Proc. the Document Understanding Conference (DUC01), 2001.
8. Nobata et al., A Summarization System With Categorization of Document Sets, Proc. Third NTCIR Workshop, 2003.
9. Nomoto T. et al., A New Approach to Unsupervised Text Summarization, SIGIR'01, 2001.
10. Yang C.C. et al., Fractal Summarization: Summarization Based on Fractal Theory, SIGIR 2003, Toronto, 2003.

11. Yang Y. et al., Learning Approaches for Detecting and Tracking News Events. Intelligent Information Retrieval, July 1999, 32-43, 1999.