Impact of Document Structure on Hierarchical Summarization

Fu Lee Wang and Christopher C. Yang

[1] Department of Computer Science, City University of Hong Kong,
Kowloon Tong, Hong Kong
flwang@cityu.edu.hk
[2] Department of Systems Engineering and Engineering Management,
Chinese University of Hong Kong,
Shatin, Hong Kong
yang@se.cuhk.edu.hk

Abstract. Hierarchical summarization technique summarizes a large document based on the hierarchical structure and salient features of the document. Previous study has shown that hierarchical summarization is a promising technique which can effectively extract the most important information from the source document. Hierarchical summarization has been extended to summarization of multiple documents. Three hierarchical structures were proposed to organize a set of related documents. This paper investigates the impact of document structure on hierarchical summarization. The results show that the hierarchical summarization of multiple documents organized in hierarchical structure outperforms other multi-document summarization systems without using the hierarchical structure. Moreover, the hierarchical summarization by event topics extracts a set of sentences significantly different from hierarchical summarization of other hierarchical structures and performs the best when the summary is highly-compressed.

1 Introduction

Many automatic summarization models have been proposed previously [1, 3, 4]. Traditionally, summarization systems consider a document as a sequence of sentences. The system calculates the significance of sentences to the document. The most significant sentences are then extracted and concatenated as a summary. Research of automatic summarization has been extended to multi-document summarization [6, 10]. Multi-document summarization system provides an overview of a topic based on a set of related documents. It is very useful in digital libraries.

It has been shown that the document structure is important in both automatic summarization [12] and human abstraction [2]. Hierarchical summarization model was proposed based on the hierarchical structure of documents [15]. Experiment results have shown that hierarchical summarization is a promising summarization technique. Nowadays, many digital libraries have begun to provide summarization service. Many documents exhibit a hierarchical structure, such as, books, websites, newsgroups, etc. Hierarchical summarization can effectively extract the most

S. Sugimoto et al. (Eds.): ICADL 2006, LNCS 4312, pp. 459 – 469, 2006.

important information from the documents with hierarchical structures. It provides an important tool for digital libraries.

In most digital library systems, a collection of related documents are returned for a query. However, there is not a trivial way to organize a large collection of documents into a hierarchical tree structure. Three hierarchical structures were proposed to organize a collection of documents into a tree structure [13]. This paper investigates the impact of different hierarchical structures on the summarization technique. Experiments have been conducted to study how the extraction of information is affected by the hierarchical structures.

The results show that the hierarchical summarization of multiple documents outperforms other multi-document summarization without using the hierarchical structure. Moreover, the hierarchical summarization by event topics extracts a set of sentences significantly different from hierarchical summarization of other hierarchical structures and performs the best when the summary is highly-compressed. It is shown that the hierarchical summarization system can extract the critical information effectively among a large collection of documents.

2 Hierarchical Summarization Model

The information overloading problem can be solved by the application of automatic summarization. A number of automatic summarization techniques have been developed [1, 3, 4]. The hierarchical summarization model was proposed to summarize a large document based on the hierarchical structure and salient features of the document [15]. Experimental results have shown that the hierarchical summarization model is a promising summarization technique.

Traditional automatic text summarization is the selection of sentences from the source document based on their significances to the document [1, 4]. The selection of sentences is conducted based on the salient features of the document. The thematic, location, and heading are the most widely used summarization features.

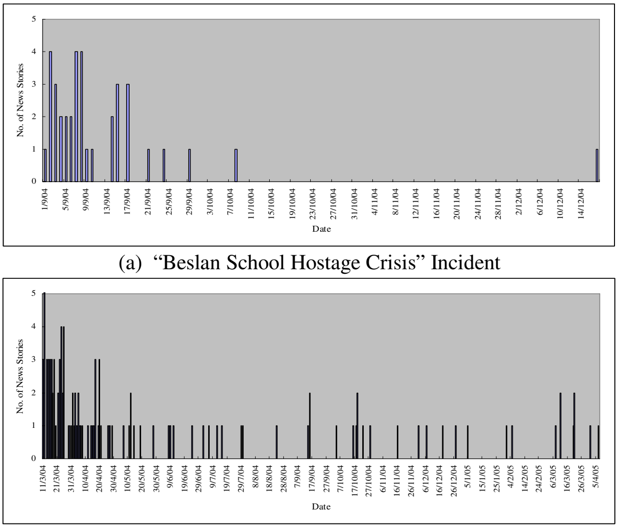
- The thematic feature is first identified by Luhn [4]. Edmundson proposed to assign the thematic weight to keyword based on term frequency, and the sentence thematic score as the sum of thematic weight of constituent keywords [1]. Nowadays, the *tfidf* (Term Frequency, Inverse Document Frequency) method is the most widely used method to calculate the thematic weight of keywords [11].
- It is believed that the topic sentences tend to occur at the beginning or the end of documents or paragraphs [1]. Edmondson proposed to assign positive weights to sentences as location score according to their ordinal position in the document.
- The heading feature is proposed based on the hypothesis that the author conceives the heading as circumscribing the subject matter of the document. When the author partitions the document into major sections, he summarizes them by choosing appropriate headings [1]. A heading glossary is a list of words, consisting of all the words in headings, with weights. The heading score of sentence is calculated by the sum of heading weight of its constituent words.

Typical summarization systems select a combination of features [1, 4], the sentence significance score is calculated as sum of feature scores. The sentences with sentence significance score higher than a threshold value are selected as summary.

A large document has a hierarchical structure with several levels, chapters, sections, subsections, paragraphs, and sentences. Related studies have shown that the document structure is very useful for human abstraction process [2] and automatic summarization [12]. Hierarchical summarization model was proposed to generate summary based on the hierarchical structure and salient features of the document [15]. The original document is partitioned into range blocks according to its document structure. The document is then transformed into a hierarchical tree structure, where each range block is represented by a node. 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 tree as the quota of sentences. The system calculates the significance score of each node by summing up the sentence scores of all sentences under the nodes. The quota of sentences is allocated to child-nodes by propagation, i.e., the quota of parent node is shared by its child-nodes directly proportional to their significance scores. The quota is then iteratively allocated to grandchild-nodes until the quota allocated is less than a threshold value and the node can be transformed to some key sentences by traditional summarization methods.

3 Hierarchical Summarization for Multiple Documents

Multi-document summarization techniques have been developed for flat-structured documents. However, a collection of related documents may exhibit a much more complicated structure. As it was shown that the document structure is important in summarization, three hierarchical structures were proposed to organize a collection of news stories [13].



(a) "Beslan School Hostage Crisis" Incident

(b) "Madrid Train Bombing" Incident

Fig. 1. Distribution of News Stories vs. Time

Multi-document summarization systems have been developed in the past [6, 10]. Typically, the summarization systems consider a collection of documents as a set of individual documents with flat-structure. Given a set of documents, some summarization systems extract concepts and their relationships, and then integrate the extracted information as a summary [10]. Alternatively, some systems segment the documents into some small text units. They compute the similarities among the text units [6]. Then, the text units are extracted based on their similarity measurement to generate summaries. However, a collection of related documents exhibit a more complicated structure. At the initial step, we investigate the summarization of a collection of news stories related to an incident. Each news story is associated with a time stamp. Moreover, the news stories can be classified into event topics [14]. Current summarization system cannot capture the above information. As a result, a multi-document summarization system for structured document is required.

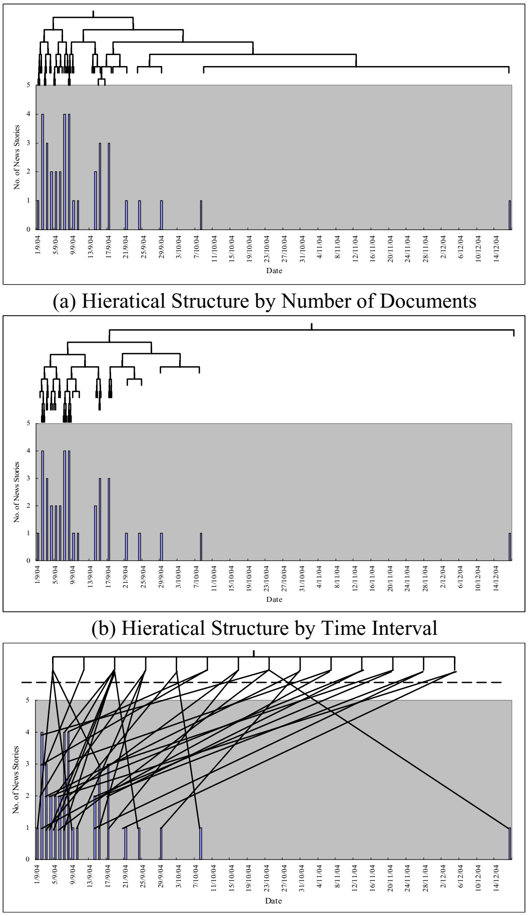
In order to have a better understanding of news stories related to an incident, two incidents have been analyzed. Related news stories have been collected from the CNN.com. The first incident is the “Madrid Train Bombing”. The second incident is the “Beslan School Hostage Crisis”. In the figure of distribution of news stories against time, obvious peaks can be identified at the beginning (Fig 1). The peaks correspond to the burst of the incidents. Then, the number of news stories decreases as time goes by. As shown in the Fig. 1, the “Madrid Train Bombing” has a more long-term impact. Therefore, there are more news stories and last for a longer period.

There is a large collection of news stories related to an incident. It is difficult for a human to view all the information without a structure. When a human professional writes a document about an incident, he partitions the information into chapters and then sections. As human is the best summarizer, a high quality summarization system should work similarly as human [2]. Therefore, the collection of news stories must be organized into a hierarchical structure before applying the summarization techniques. In Fig. 1, a large number of news stories spread out over an interval of time. By intuition, we propose to organize the news stories by number of documents as well as by time interval. It is also believed that a set of news stories may contain several event topics [14], which are very important during information extraction. As a result, three hierarchical structures are proposed to organize a collection of news stories.

- Results of hierarchical summarization of large documents showed that a good summary must have a wide coverage of information and extract information distributively [15]. Moreover, when an author writes a document, he distributes the information into units. Combining these observations together, we propose to organize the news stories into a hierarchical tree by number of documents (Fig. 2a). The news stories are sorted by chronological order and then organized as balanced hierarchical tree, such that each node at the same level contains approximately the same number of news stories. Because the information contents are evenly distributed into the tree structure, hierarchical summarization will extract information distributively. To simplify our discussion, we focus on binary tree in this section. The figures in this paper show the news tree up to news stories level only. Tree structure exists within the news story.
- Temporal text mining discovers temporal pattern inside the text [7]. Similar technique has been used in multi-document summarization [6], summarization of news stories are generated for fixed number of days, then an overall summary is

generated. Therefore, we propose the hierarchical structure by time interval (Fig 2b). The news stories are organized into a hierarchical structure such that each child node represents an equal and non-overlapping interval. Unlike the hierarchical structure by number of documents, the hierarchical structure by time interval is an unbalanced tree structure. Therefore, the information is not evenly distributed into node blocks.

- It is believed that a collection of news stories may contain several event topics, the detection of event topics is very important in information retrieval [14]. Recent research in automatic summarization proposes to classify the documents into document sets before summarization [9]. Therefore, we propose the hierarchical structure by event topics (Fig 2c). Because the accuracy of event topic detection affects the performance of the summarization directly, the news stories are



(a) Hieratical Structure by Number of Documents

(b) Hieratical Structure by Time Interval

(c) Hieratical Structure by Event Topics

Fig. 2. Hierarchical Structure of "Beslan School Hostage Crisis" Incident by Event Topics

clustered into event topics by qualified human professionals in our experiment. Each event topic is represented as a child node under the root node. The news stories under the event topics are then the child nodes of events. The hierarchical structure by event topic is not a balanced tree.

Hierarchical summarization is applied to summarize the news stories with different hierarchical structures. The system generates a summary for each range block, and then the summaries of range blocks are concatenated as an overall summary for the collection of news stories. When the number of news stories inside a range block is too large, iterative partition of range block into sub-range blocks is required and the hierarchical summarization technique will be applied to summarize the range blocks. The hierarchical summarization for multiple documents is very similar to the hierarchical summarization of a large document [14, 15], only some minor modifications are required to demonstrate the characteristic of the news stories.

- Firstly, there is no heading for the internal nodes in the tree. Hence, the heading feature considers only the headings of news stories and the theme of the incident.
- Unlike traditional summarization, the news stories inside a node are considered as equally significant regardless its location inside the node. Therefore, the location feature is not considered during hierarchical summarization of the tree structure. However, if the range block is small enough, for example, selection of sentences within a news story, the location feature will be considered.

4 Impact of Hierarchical Structure on Summarization

A collection of related documents can be organized into hierarchical tree structures by different classification. They have a different distribution of information contents among the nodes inside the tree. It may have a significant impact on the summarization technique. In this section, we will investigate the impact of hierarchical structure on the accuracy of automatic text summarization.

The comparison of summarization system is very difficult, because different research uses different data sets and different ground-rules. The TIPSTER Text Summarization Evaluation (SUMMAC) is the first large scale, developer-independent evaluation of automatic summarization systems [5]. The SUMMAC has identified two categories of methods for evaluating text summarization. Both intrinsic evaluation and extrinsic evaluation will be conducted on the previous two incidents in our experiment. Moreover, we will analyze the intersection of sentences in the summaries by summarization using different hierarchical structures.

4.1 Intersection of Summaries

In most literatures, the compression ratio for summarization is chosen as 25% because it has been shown that extraction of 20% sentences can be as informative as the full text of the source document [8]. However, it is believed that the highly-compressed abstracting is more useful [12]. Therefore, we have conducted the experiments from 5% to 25% for each interval of 5%. The intersections of summaries by summarization of different hierarchical structures are analyzed.

Table 1. Intersection Percentage of Summaries (Compression Ratio = 5%)

		By Event Topic	By No. of Document			By Time Interval		
			Deg. 2	Deg. 3	Deg. 4	Deg. 2	Deg. 3	Deg. 4
By Event Topic		-	44.3%	44.3%	41.6%	47.0%	43.8%	47.6%
By No. of Document	Deg. 2		-	84.1%	85.6%	67.8%	77.5%	79.1%
	Deg. 3			-	82.9%	66.1%	73.2%	76.4%
	Deg. 4				-	64.5%	76.6%	78.2%
By Time Interval	Deg. 2					-	69.3%	69.9%
	Deg. 3						-	86.6%
	Deg. 4							-

In our previous discussion, the number of children (degree) of a tree is limited to two for hierarchical tree by number of documents and by time interval. However, there may be a large number of children in the hierarchical tree by event topics. The number of children nodes will significantly affect the distribution of information. In order to have a fair comparison, we have conducted the experiment to summarize hierarchical tree with different degrees for these two hierarchical structures. For a fixed compression ratio, the summaries have an equal number of sentences. We calculate the intersection of two summaries as the number of sentences which appear in both summaries. The intersection of summaries with 5% compression ratio is reported in Table 1. As shown in the table, the intersection for summarization by event topics to another two hierarchical structures is not high. The intersection for summarization by number of document and summarization by time interval is higher. Moreover, the summarization of hierarchical structure of same classification with different degree has a high level of intersection.

Table 2. Average Intersection Percentage of Summaries

Compression Ratio	5%	10%	15%	20%	25%
Intersection Percentage	65.0%	59.5%	72.2%	74.7%	79.5%

Table 2 shows the impact of compression ratio on intersection of summaries. As the compression ratio increases, the intersection of summaries for summarization with different hierarchical structures increases. Because extraction of 20% sentences can be as informative as the full text of the source document [8], when the compression ratio is large, summarization of different hierarchical structures can extract the common set of essential information from source documents. Therefore, intersection percentage is high. However, the intersection of summaries cannot show the performance of summarization. Therefore, intrinsic evaluation and extrinsic evaluation will be conducted in next two subsections.

4.2 Intrinsic Evaluation of Summarization

Intrinsic evaluation is the most straight forward method to measure the quality of system summaries. It judges the quality of summaries by direct analyses in terms of some set of norms. One of the most common approaches is to match a system summary against an ideal summary.

Because highly-compressed abstracting is more useful [12] and there are a huge number of sentences within a collection of related news stories, user evaluation are conducted only at 5% compression ratio to reduce the workload of human abstractors. The collection of news stories is presented to human professionals, and they are asked to compose a covering summary for the incident. In order to have a fair comparison between the system summaries and the human abstract, the human professionals are asked to select specific number of most important sentences among the news stories as indicated by the compression ratio.

Table 3. Precision of Summaries with Different Degrees by Gold Standard

	By Even Topic	By No. of Documents			By Time Interval		
		Deg. 2	Deg. 3	Deg. 4	Deg. 2	Deg. 3	Deg. 4
Precision	77.1%	60.4%	57.7%	62.1%	57.7%	61.0%	57.3%
		60.1% (Mean)			58.7% (Mean)		

The system summaries are compared with human abstracts to measure the quality of summaries by gold standard [3]. The precision are shown in Table 3. The ANOVA shows that there is no significant difference among the precisions of summaries of one hierarchical structure with different degrees. Therefore, the mean of precision of one hierarchical structure with different degrees is taken as the precision of the hierarchical structure. One-way ANOVA reveals a significant difference between different document structures ($p < 0.002$). The t-test shows that the hierarchical summarization by event topics outperforms the hierarchical summarization by number of documents and the hierarchical summarization by time interval at 88% and 92% significance levels respectively. There is no significant difference identified between the hierarchical summarization by number of documents and the hierarchical summarization by time interval.

The precision of hierarchical summarization by event topics is significantly higher than the other two structures (Table 3). It can be explained by that news stories organized by event topics gives a more natural segmentation. When an author writes a large document with a lot of information, he groups similar information into same sections. Therefore, classification of news stories into event topics simulates the process of an author writing a large document. It is the most human-like classification of news stories. The other two structures partition the news stories by brute force, therefore, the themes among stories are not preserved. In conclusion, the hierarchical structure by event topics is the most natural partitioning of news stories. The hierarchical summarization is developed based on the hierarchical structure of document, and it does summarization in the similar way as a human abstractor. Therefore, it is perfectly matched with the document tree by event topics.

The intrinsic evaluation is based on human abstraction. However, it is very time-consuming for a human professional to compose an abstract. Therefore, it is extremely difficult to conduct the intrinsic evaluation with different parameter of settings. A more comprehensive experiment of extrinsic evaluation will be conducted in the next subsection.

4.3 Extrinsic Evaluation of Summarization

The extrinsic evaluation judges the quality of the summarization based on how it affects the completion of some other tasks. Among the extrinsic evaluations, the question-answering task is to find the "informativeness" of a summary, namely, the degree to which it contains answers found in the source document to a set of topic-related questions [5]. The question-answering task has been proved as a promising method for automated evaluation of summarization [5]. The quality of summaries will be measured by question-answering task in our study.

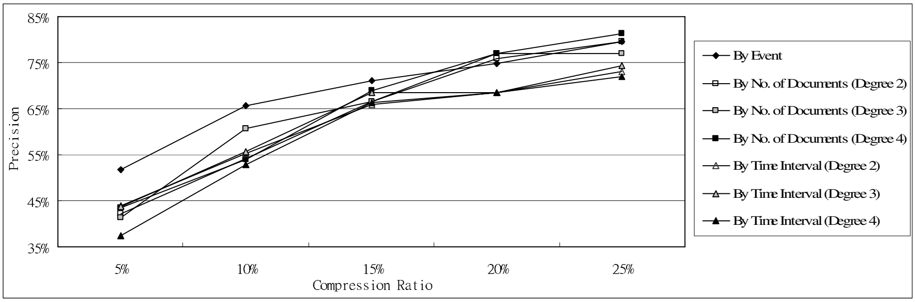


Fig. 3. Recall of Summaries in the Q&A Task for All Hierarchical Structures

Given a collection of news stories, human professionals are requested to prepare a set of topic-related questions and the answer keys using a common set of guidelines. These questions cover some essential information that is provided in any of the news stories. We have conducted experiments on the previous two incidents. The recall of the summarization is defined as the percentage of answers that can be found in the system summaries [5]. In the question-answering task, the set of questions and their answer keys can be used for evaluation at different compression ratios. Therefore, it is feasible to conduct experiments with different settings without increase in the workload on the human professionals. We have conducted experiments from 5% to 25% for each interval of 5% (Fig. 3).

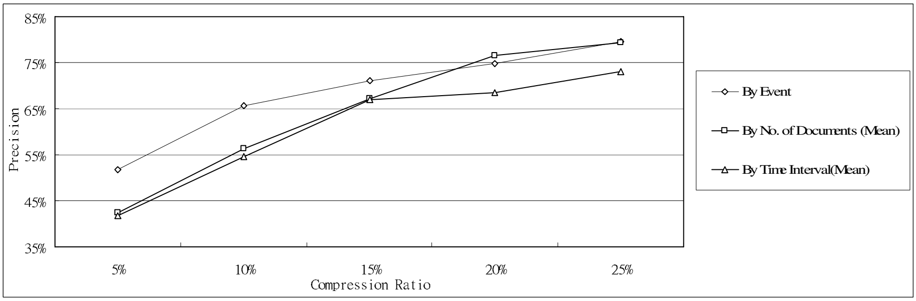


Fig. 4. Average Recall of Summaries in Q&A Task for Three Hierarchical Structures

In the intrinsic evaluation, no significant difference is identified among the precisions for the hierarchical trees with different degrees. For the extrinsic evaluation, we have also compared the recall of summarization of hierarchical trees with different degrees by ANOVA. It further confirms that there is no significant difference between different degrees. As a result, we take the mean of recalls of one hierarchical structure with different degrees as the overall recall of the hierarchical structure (Fig 4). The results in intrinsic and extrinsic evaluation have shown that the degree of a hierarchical tree will not affect the accuracy of hierarchical summarization. Similar observation is identified in the intersection analyses. It could be explained by the fact that the hierarchical summarization calculates the significance score of a node by measuring the amount of information contents inside the node, and the quotas are assigned to the nodes directly proportional to their significance score. Therefore, the summarization process is not affected by the degree of a hierarchical tree.

In the intrinsic evaluation, hierarchical summarization by event topics outperforms hierarchical summarization by number of documents and by time interval when the compression ratio is 5%. We have compared the recalls of summarization using different hierarchical structures at different compression ratios. By t-test analysis, we find that there is no major difference between the hierarchical summarization by number of documents and by time interval. However, we find that hierarchical summarization by event topics outperforms hierarchical summarization by number of documents and by time interval at 90% significance level, when the document is highly compressed, i.e., 5% and 10% compression ratio. However, as compression ratio increases, the recall increases and the difference diminishes. When the compression ratio is 15%, hierarchical summarization by event topics outperforms hierarchical summarization by number of documents, but there is no difference between hierarchical summarization by event topics and hierarchical summarization by time interval. When the compression ratio further increases, there is no significant difference identified among three hierarchical structures.

Because extraction of 20% sentences can be as informative as the full text of the source document [8], when the compression ratio is higher than 20%, most of the summarization systems can produce a summary as informative as the full text. Therefore, there is no significant advantage for hierarchical summarization by event topics over the other two. However, highly-compressed summarization is much more useful [12]. Hierarchical summarization by event topics outperforms the other two structures, when the summary is highly compressed. Therefore, it provides a useful information extraction tool. In this study, the documents are clustered into event topics by human professionals. Further study will be conducted to investigate how the summarization is affected by clustering techniques in the future.

Finally, in the question-answering task of the SUMMAC, it is found that the summarization systems achieve the peak value of recall when the compression ratio is 35% to 40% [5]. Most of the system recorded a recall about 60% [5]. Our system achieves a recall of 60% when the compression ratio is 10%, and a recall of 70% when the compression ratio is 20%. Hierarchical summarization of news stories organized in tree structure outperforms the participants in the SUMMAC. The results show that our system is a promising system for multi-document summarization.

5 Conclusion

Multi-document summarization is very useful to extract information from a large collection of news stories. Three hierarchical structures have been proposed. Experimental results show that the hierarchical summarization of multiple documents organized in a hierarchical structure outperforms significantly the multi-document summarization without using hierarchical structure. It also showed that hierarchical summarizations by event topics outperform the other two hierarchical structures when the summary is highly-compressed. As there is a large volume of information related to an incident, a highly-compressed summarization is more desired. This novel technique extracts essential information from a large number of documents effectively.

References

1. Edmundson H. New methods in automatic extraction. J. ACM, 16(2) 264-285, 1968.
2. 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.
3. Kupiec J. et al., A trainable document summarizer. SIGIR'95, 68-73, 1995.
4. Luhn H.P. The automatic creation of literature abstracts. IBM J. R&D, 159-165, 1958.
5. Mani I. et al., The tipster SUMMAC text summarization evaluation. 9th conference on European chapter of the Association for Computation Linguistics, 1999.
6. McKeown K. et al., Tracking and summarizing news on a daily basis with columbia's newsblaster. Human Language Technology Conference, 2002.
7. Mei Q. et al., Discovering evolutionary theme patterns from text: an exploration of temporal text mining. ACM SIGKDD, 198-207, 2005.
8. Morris G. et al., The effect and limitation of automated text condensing on reading comprehension performance. Information System Research, 17-35, 1992.
9. Nobata C. et al., A summarization system with categorization of document sets, Third NTCIR Workshop, 2003.
10. Ou S. et al., Development and evaluation of a multi-document summarization method focusing on research concepts and their research relationships, ICADL, 2005, 283-292.
11. Salton G. et al., Term-weighting approaches in automatic text retrieval. Info. Proc. & Manag., 24, 513-523, 1988.
12. Teufel S. et al. Sentence extraction and rhetorical classification for flexible abstracts, AAAI'98 Spring Sym., Stanford, 1998.
13. Wang F. et al., Multi-document summarization for terrorism information extraction, IEEE ISI-2006, 2006.
14. Yang Y. et al., Learning approaches for detecting and tracking news events. Intelligent Information Retrieval, 32-43, 1999.
15. Yang C. et al., Fractal summarization: summarization based on fractal theory, SIGIR, 2003.