Multi-Document Summarization based on Concept Space

Smooch T. K. Tang, Jerome Yen, and Christopher C. Yang
Department of Systems Engineering and Engineering Management
The Chinese University of Hong Kong
Hong Kong

Abstract—**Capturing relevant information is important in supporting decision making. In this paper, we propose a new summarization method based on cluster analysis, concept space, and statistical approach to extract the essence from a collection of documents. A prototype system has been developed to condense a set of documents into a list of key issues and expands the key issues to form a summary. Cluster analysis and concept space was used as a bridge to connect convergent and divergent processes. Such approach reduces information loss due to vocabulary switching in the summarization process. In the divergent process, it selects the anchored sentences from the original documents to form a summary based on the concept terms generated previously. A user evaluation has been conducted for its usefulness and other performance indices. The results indicate that such approach is promising.**

1. INTRODUCTION

Over the past three decades, information and communication technologies have changed the way how people access and use information [22]. Newspapers, academic journals, reports from information service providers, and magazines used to be the only sources of information [16]. Today, with the advances in Internet and World-wide Web (WWW) technologies, individuals and organizations have begun to enjoy the benefits and power of greater and broader information access. However, as more and more information became accessible, information overload also became a serious problem to challenge the researchers. According to the *Model of Human Processor*, there is a limit to the capacity of human information processing [4]. Such limitation can be found in vision, listening, short-term memory, long-term memory, and association. As Kandt and Yuender have noted, information technologies, especially the Internet and WWW, are the major contributor to information overflow problem:

> *Companies are generating a lot of data but lack the tools to analyze that information for better decision making. Future data acquisition capabilities will be able to collect enormous quantities of information regarding conditions in distribution channels. This information, together with the vast information resources linked to the highly interconnected computer networks, represents a data mining, filtering, and display problem of substantial magnitude [14].*

These remarks summarize two problems: The first is to overcome the information overflow problem and the second is to increase the utilization and the utility of the information that available to the decision makers. In other words, the first problem is at the level of information management and the second problem is at the knowledge management level. Our research is based on the assumption that processing information manually is so inefficient and ineffective that it may become the bottleneck to the management of organizations. Such problem cannot be solved just at the information management level, but at the knowledge management level.

Artificial intelligence techniques to support human information processing have been an important research area for many decades. Text analysis, concept classification, natural language processing, event tracking and tracing, as well as topic or issue identification have long been the research topics that related to utilization of information [9][20]. Examples of using artificial intelligence for knowledge extraction or concept classification include: adding value to financial news, understanding issues in foreign currency options trading, processing money transfer messages, identifying patterns in frequent flyer databases, identifying patterns of spending of credit card holders, and examining corporate intelligence reports [1][8]. However, most of these applications are limited to very small and selected domains, such as, using symbolic pattern analysis to predict the changes and trends in prices of stocks [15] and in foreign exchange rates [1].

In this paper, we propose a new approach based on artificial intelligence and statistical techniques, to condense and expand a set of selected articles or documents. The system is able to extract structures, patterns, heuristics or semantics to develop a set of concept spaces that represent the key issues or concepts of the selected articles. Based on the concept spaces, a summary, which can be one or two pages, is generated for the users. Users normally take only a few minutes to read such summary. If users are interested in knowing in more detailed about any particular issue, then all the related or associated paragraphs or articles will be retrieved.

2. RELATED WORK

The information explosion or information overload proved that text summarization is extremely helpful in saving time and efforts for the users; therefore, during the past twenty years many approaches have been proposed. As time moved on, the requirements became higher and the systems also became more powerful. At the beginning, text summarization only dealt with single document with one specific domain. Some systems that developed during the past five years were able to generate summaries from multiple documents with multiple domains. That is, the trend of text summarization has moved from single document to multiple documents, from domain dependent to domain independent, and from single language to multiple languages.

0-7803-7724-9/30/03/$17.00

There are four approaches to the text summarization [21]: Location Approach, Cue Phrase Approach, Statistical Approach, and Nature Language Process (NLP) Approach.

Location approach

The idea of Location Approach simply determines the importance of the sentence or paragraph according to its position in the article. Luhn proposed the location approach [18] for constructing abstract. Intuitively the title and heading represent the topic of the article, and the sentence in first and last paragraph often function as summary of the passage. Thus, it shown those wording can be useful for forming summary.

Cue Phrase Approach

Cue Phrase Approach finds the cue words or phrases to identify the important sentences, which are then used to generate the summary. Cue-phrase, such as 'in summary', 'in conclusion', 'this paper', 'this paper present', 'paper describes' etc., can indicate the following sentences are important. That helps us to identify important sentence. Frequently we added cue-phrase module in the summarization process to increase the overall performance.

Statistical Approach

Statistical Approach applies information retrieval (IR) technique, for which the mathematical or statistical calculation is used to find out the importance of the key words or sentences. By calculating term frequency and document frequency, we assign each key word and key phrases a weight. Straightforward we increase the sentence score for high frequency word. In addition, it applies co-occurrence method to determine the associations of a word to word, paragraph, or document. The summary is generated based on the significance of the terms in sentences.

NLP Approach

NLP approach solves the lexical problem by identifying, the subject, verb, etc. But it is difficult to generate good results if structures of sentences are complicated or writing style varies. Besides, different language has different lexical structure. There are still rooms for improvement, but it is very labor-intensive. Lehnert used lexical chain to summarize narrative articles [28]. The algorithm tried to capture the primitive and complex plotting units in order to identify the connectivity and symmetry of how characters or words interact. This approach is specific designed for generating summaries from narrative articles. More general approaches that based on lexical chain were proposed by Barzulay and Elhadad [2]. The summarization process identifies string lexical chains and extracts significant sentences based on the identified lexical chains.

The system presented in this paper summarizes multi-documents with general domain - domain independent. After computing the term frequency and document frequency to determine the key words or key issues of the documents, cluster analysis algorithms will be used to generate the concept spaces. Our system uses terms in concept spaces to represent the key ideas of related documents. Finally a summary will be generated based on the identified concept spaces and their associated terms.

In the summarization process, the meaning or key ideas of each document needs to be identified. However, a vocabulary switching problem exists in human writing or speaking. Different people use different sets of vocabulary to describe the same idea. Early research has identified that for two people to use the same term to describe one subject or object, the possibility is less than 0.2 [11]. By using concept space approach, it allows the use of more than one word or phrase to represent one idea. It provides greater flexibility than the approaches that based on keywords. And also it is domain independent; no thesaurus is needed to support segmentation.

3. SYSTEM ARCHITECTURE

Our system is able to generate summary from domain independent multiple documents. It involved several steps to finish this task. As a human reader, the summarization process includes [3]:

1. Understanding the contents of the documents.
2. Identifying the most important pieces of information or the key concepts.
3. Writing up a summary that contains such information or key concepts.

There are several definitions of text summarization for computer system. As appeared in earlier research, summarization = topic identification + interpretation + generation [12], or summarization process can be characterized as analysis, refinement, and synthesis [19]. Basically, the idea of summarization is the same among different approaches. In our system, the summarization process includes two sub-processes —Converge and Diverge.

3.1 Convergent Process

During the Convergent Process, the system identifies key ideas from the documents and uses these key ideas to generate a summary. This process condenses the source documents into few words or phrases. During this process, how to minimize the information loss is a critical issue. If too few key words or concept spaces selected, then it is possible that part of the contents in the original documents will be lost. In this sub-process, we use statistical approach to condense the content by generating a list of concept space to represent the major idea of the content.

3.2 Divergent Process

The Divergent Process generates a summary based on the selected key ideas. In this process, the system expands the size of the document - use key words or concept spaces to determine the anchor sentences to generate summary, which can be one or two pages in length. Such process fits quite well to how human writers generate summarization. They extract a set of key phrases from documents in the convergent process to generate the skeleton, then combine or reorganize the sentences to form the summary in divergent process. Based on the summary, the users should be able to retrieve the paragraphs from the original documents that discuss those issues that appeared as sentences in the summary.

The summarization process of the system is shown in Figure 1.

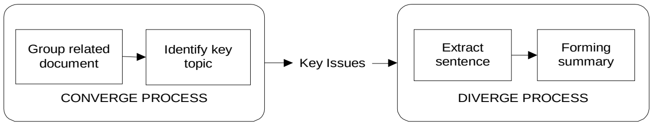

Figure 1: System Flow

4. CONCEPT SPACE APPROACH

In this research, the system uses concept space to support the Convergent Process. The system generates a set of terms or key words that represent the key ideas or key concepts of the processed documents [5]. It first merges and indexes the documents to develop a formatted file, then groups the key terms or ideas that always mentioned in the similar places by cluster analysis to generate concept spaces by cluster analysis algorithms, for example, Hopfield net classification [6][7], Machine Learning.

1. Document Processing
   Text pre-processing merges articles, labels paragraphs and sentences. All the documents are grouped together and each document is partitioned into many paragraph-size units, which called processing units. Size of the processing units is determined by the size of the original documents. In this work, we select paragraph as a processing unit. It is because a paragraph contains sufficient contents, in terms of key words, to support the analysis that based on the term frequency and document frequency. Each sentence will be marked in the paragraph by its position in the paragraph. Each paragraph is also marked and indexed by its position in the set of the documents. The output of the document preprocessing is a single file that contains all the indexed documents.

2. Automatic Indexing
   As evident in the sample letter, many ideas or issues were expressed in specific terms or phrases e.g., “rebounding US economy”, “overseas”, “trading”, “capital markets”, “financial advisory services”. Words in a “stop word” list which consisted of about 1,000 common function (non-semantic bearing) words, such as {*on, in, at, this, there*}, etc. and “pure” verbs (words which are verbs only), e.g., {*calculate, articulate, teach, listen,*} etc. were removed from the document. Due to the unique terminology used in financial documents, we also included an additional set of common financial words, which appeared to be too general to carry any specific meanings, e.g., “business”, “company”, etc. However, in running our agent-based system, any user would be able to modify the stop-word list to help sharpen the system's suggested issues (an iterative process). Adjacent words are grouped to form phrases. After examining similar financial documents, we decided to form phrases that contained up to three words because most subject descriptors are less than four words. Our system generated 1-word, 2-word, and 3-word phrases from adjacent words, e.g., “Wall”, “Street”, “firm”, “Wall Street”, “Street firm”, and “Wall Street firm” from the three adjacent words “Wall Street firm”.

3. Cluster Analysis
   Base on the vector model, we use *document frequency* and *term frequency* to determine that relationships among terms and based on such relationships to identify the key terms. The term frequency, tf_{ij}, is number of a term j that appeared in a document i. The document frequency, df_j, is number of document i that contained this term j.

   The combined weight of term j in document i, d_{ij}, based on the product of *term frequency* and *document frequency* is

$$d_{ij} = tf_{ij} \times \log df_j$$

   computed as follows:

   Based on the asymmetric *Cluster Function* shown below, we generated two concept space matrices of terms and their weighted relationships.

$$ClusterWeight(T_j,T_k) = \frac{\sum_{i=1}^{n} d_{ijk}}{\sum_{i=1}^{n} d_{ij}}$$

$$ClusterWeight(T_k,T_j) = \frac{\sum_{i=1}^{n} d_{ijk}}{\sum_{i=1}^{n} d_{ik}}$$

$$d_{ijk} = tf_{ijk} \times \log df_{jk}$$

   where tf_{ijk} represents the number of occurrences of both term j and term k in document i and df_{jk} represents the number of documents in a collection of n documents in which both term j and term k occur.

4. Hopfield Net Classification
   The Hopfield network [6] is used to identify clusters of relevant terms in documents. Each term in the co-occurrence analysis results was treated as a neuron and the asymmetric weight between any two terms was taken as the unidirectional, weighted connection between neurons. Hopfield algorithm activated its neighbors, combined weights from all associated neighbors, and repeated this process until the output pattern converged. The clusters with strongly related terms form the concept spaces. Therefore the result list of concept spaces is the key idea of the documents.

5. EXPANSION - IDENTIFICATION OF ANCHOR SENTENCES

After the Convergent Process, the system identified a set of key issues from the documents. The next step is to expand these key issues into a summary which covers more detailed information that more readable to the users. This expansion simulates the “copy and paste” process of humans generating summaries. Each identified key topic is expressed by a few anchor sentences extracted from the original documents. The idea is to rank the sentences based on the number of appearances of the key terms and selected sentences that with the highest scores. First, it calculates the score of each sentence that contains any of the concept terms. There are two scoring schemes. (1) If the concept term is at the top position, it receives higher score. (2) If the concept term is at the top and the bottom positions, the score it receives higher. The first scheme is more suitable for news articles. But the second scheme is more general, especially for document that have certain styles.

$$SentenceScore_{ij} = \frac{1}{\sqrt{s}} \times \frac{1}{p}$$

Or

$$SentenceScore_{ij} = \begin{cases} \frac{1}{\sqrt{s}} \times \frac{1}{n-p+1} & \text{otherwise} \\ \frac{1}{\sqrt{s}} \times \frac{1}{p} & if\ p \leq \frac{n}{2} \end{cases}$$

s is position no. of the term in a sentence *j*
p is paragraph no. in document *i*
n is total number of paragraph in document *i*

The document score reflects the degree of relevance of a document with that particular issue and is computed as follow.

$$DocumentScroe = \sqrt{\sum SentenceScore_{ij}}$$

After ranking the documents based on the document score for each issue, the system extracts the most important sentences from the most relevant document to form summary. It utilizes the highest of *document score*$_i$ times *sentence score*$_{ij}$ as anchor sentence to represent this particular issue or concet space.

In this system, readers can adjust the number of maximum anchor extracted from the document set. If the reader would like to have a more detailed summary, then this number can be increased. In contrary, such number can be decreased to obtain a more concise summary. However, there is a trade-off between accuracy or conciseness and recall. If length of summary increased, recall is increased, but precision may be sacrificed [13].

To allow the users to better utilize the contents of the original documents, it is important to provide the content recall function. Each anchor sentence in the summary is a "button". When a "button" clicked, the system processes a "backward search" to search and retrieve the paragraphs or documents that contain the concept terms in this anchor sentence. Therefore, after browsing through the summary, users can make query about the key topics that deemed important.

6. EXPERIMENT AND RESEARCH FINDINGS

Summarization provides a summary of contents from a set of documents, which helps readers to grasp the key issues or important insights in short time. In addition, the system should be able to assist the users to determine which subset of documents is more relevant or more important. In general, detailed information about those issues or insights is still buried in the original documents. After reading the summary, users may need to select documents to read in order to understand the background or to associate with the other available information or documents. This was one of the major objectives of building the system. In the following experiment, we would like to investigate the following three factors.

1. Condensation rate - which affects the reading time.
2. Recall and information loss – which determines how much information are lost after such converging process and no longer be able to be retrieved by the users.
3. Retrieve detailed background information from the summary.

On-line news articles were collected from the websites of *South China Morning Post* and *Hong Kong Standard*, two major English on-line news websites in Hong Kong, as the corpus in the experiment. These news articles covered financial and investment in Hong Kong SAR, USA, and Mainland China. These news articles were selected within one week before the experiment and some articles reported the same events. We prepared two sets of articles for our experiments, one contains 20 articles and the another contains 30.

Ten subjects were invited to participate in the experiment. They were graduate students who have in-depth understanding of the financial investment. They were also users of Internet and constantly used Internet search engines. Subjects are assigned with two tasks.

In the first task, a set of original news articles and a questionnaire were provided to the subjects. They categorized the new articles into different topics or areas and generated a list of topics that cover those news articles. After subjects generated the list, we compared the list of issues that generated by the system against the list generated by the subjects.

In the second task, subjects were asked to assign a grade for the summary that generated by the system for each performance attribute: *Relevance*, *Usefulness*, and *Representative*. A Likert scale of 5 was used for the subjects to choose. *Relevance* asks whether a particular sentence tells something about a key issue. *Usefulness* is how useful the information is that contained in the sentence in helping the user to understand a key issue. *Representative* is the degree of the sentence that represents all the important ideas or information relate to a particular issue.

From earlier research and their results, they indicated that the summary length does not depend on the length or size of the original documents [10]. The compression ratio decreases when the size of documents increases. Therefore, our system was designed to generate one-page summary for both sets of testing samples. The compression ratios are shown in Table 1.

Measurement of precision and recall is utilized to determine how comprehensive the list covers the topics that buried in the news articles. The result is shown in Table 2. Both the precision and recall were way over 70 percent. Compare with other approaches, such as, topic identification integration module in SUMMARIST system [12], which scored 58.07% (Recall and Precision), our results indicated that the performance of the system, based on the two sets of testing documents, were comparable.

	Compression Rate (%)
20 News Articles	4.65
30 News Articles	3.90

Table 1: Compression Ratio of Summary

Recall	72.96 %
Precision	79.36 %

Table 2: Recall and Precision

For the evaluation of divergent process, we studied the performance of the system by three criteria: relevance, usefulness, and representative of the summary. The grading has five levels. For example, for relevance the choices were: Most Relevant (1), Moderate Relevant (2), Little Relevance (3), Too General (4), and Not Relevant (5).

The number in the bracket is the score of each choice. The results for these three categories are very close as shown in Table 3. Numerically, the results are promising; the averages are around Moderate Relevant. The summary generated by the system is not better than that generated by human. But the results showed that the list generated by the system covered broader information.

Relevance	**Usefulness**	**Representative**
2.36	2.39	2.30

Table 3: Average score of Relevance, Usefulness, and Representative

7. CONCLUSIONS

In this paper, we used cluster analysis and concept space as the core to develop system to generate summary from multiple documents. The reason that concept space was selected was due to its flexibility. The results of the evaluations indicated that such approach has good performance in capturing meanings and insights of documents. The system has successfully identified key issues and terms that associate to each issue. Then it clusters the related documents together and uses concept terms to represent those issues. Finally, a set of sentences were selected from the source documents, based on the scores that calculated by term frequency and positions, to generate a one- or two-page summary for the users. Users are also able to retrieve related contents of each sentence from the summary.

We had conducted an experiment to evaluate relevance, usefulness, and coverage of the summary that generated by the system. The results indicated such approach extracted essential information for users, which significantly saved their time and efforts in understanding contents of documents they intended to read. However, there is still limitation on the summarization process and several research issues can be further explored.

- Since the divergent process uses the sentence extraction that based on term frequency and position of the sentence, the extracted sentences are not be modified. In the human summarization process, summarizers usually modify or restructure those sentences. Therefore, the summary that generated by the system may not be as flexible, smooth, concise, or coherent as that generated by human.
- Information loss is still unavoidable even the concept space technique was used in the summarization process. Yet, the convergent and divergent processes can be further fine-tuned to minimize the information loss.
- Our work has successfully used the concept space technique on the summarization process for multi-document as well as multi-domain. The next step is to overcome the multi-language problem. The oriental language has become more important in the globe and in fact, many documents of the first hand information are released in such language. Handling multi-language will be a must for all information providers. The summarization system should be able to assimilate and then combine the information in different languages into one single-language document. Much further research could be worked on such issue.

8. REFERENCE

[1] E. R. Addison. “Using news understanding and neural networks in foreign currency options trading”. *In Proceedings of the First International Conference on Artificial Intelligence Applications on Wall Street*, New York, Oct. 9-11, 1991, pp. 319-323, 1991.

[2] Regina Barzilay and Michael Elhadad, “Using Lexical Chains for Text Summarization “. *In Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization*, pp. 10-17, 1997.

[3] R. Brandow and K. Mitze and L. Rau, “Automatic condensation of electronic publications by sentence selection”, *Information Processing and Management*, vol. 31, ch. 5, pp. 675-685, 1995.

[4] S. K. Card, T. P. Moran, and A. Newell. *The Psychology of Human Computer Interaction.* Lawrence Erlbaum Associates, Hillsdale, NJ, 1983.

[5] H.Chen and A. Houston and J.F. Nunamaker and J. Yen, “Toward Intelligent Meeting Agents”, *IEEE Computer*, vol. 29, no. 8, pp. 62-72, 1996.

[6] H.Chen, “Machine learning for information retrieval: Neural networks, symbolic learning, and genetic algorithms”, *Journal of the American Society for Information Science*, vol. 3, no. 46, pp. 194-216, Apr 1995.

[7] H.Chen and P.Hsu and R. Orwing and L.Hoopes and J.F. Nunamaker, “Automatic concept classification of text from electronic meetings”, *Communications of the ACM*, vol. 37, no.10, pp. 56-73, Oct 1994.

[8] D. N. Chorafas and H. Steinman. Expert Systems in Banking: A Guide for Senior Managers. *New York University Press*, New York, 1990.

[9] B. Everitt, Cluster Analysis. *Heinemann Education Books*, London, England, ed. Second, 1980.

[10] Jade Goldstein and Mark Kantrowitz and Vibhu Mittal and Jaime Carbonell, “Summarizing Text Documents: Sentence Selection and Evaluation Metrics”, 1999.

[11] Furnas, G. W. and Landauer, T. K. and Dumais, S. T., “The vocabulary problem in human-system communication”, *Communication of the ACM*, 1987.

[12] Eduard Hovy and Chin-Yew Lin, “Automated Text Summarization in SUMMARIST

[13] Hongyan Jing and Kathleen Mckeown and Regina Barzilay and Micheal Elhadad, “Summarization Evaluation Methods: Experiment and Analysis”, *AAAI'98 Symposium*, Stanford Univeristy CA , Apr, 1998.

[14] R. Kalakota and A. B. Whinston. *Frontier of Electronic Commerce.* Addison Wesley, Reading, MA, 1996.

[15] Kandt. K. and P.Yuender. A financial investment assistant. *In Decision Support and Expert Systems*, Trippi R. R. and Turban E. (Eds.), Boyd and Fraser Publishing Company MA., 1990.

[16] Kolb, R. W., "Investment", Foresman and Company, Glenview, Illinoise , 1989.

[17] W. G. Lehnert, “Plot Units: A Narrative Summarization Strategy”, *Advances in Automatic Text*, Massachusetts Institute of Technology.

[18] H.P. Luhn, “The Automatic Creation of Literature Abstracts”, *IBM Journal of Research and Development*, pp.159-165, 1959.

[19] Inderjeet Mani and Eric Bloedorn, “Summarizing Similarities and Differences Among Related Documents”, *Kluwer Academic Publishers*, Paul B.Kantor and Stephen E. Robertson, vol. 1, ch. 2, pp.35-67, 1999.

[20] G. Salton. Generation and search of clustered files. *ACM Transactions on Database Systems*, 3(4):321-346, December 1978.

[21] C. C. Yang, and F. L Wang, “Fractal Summarization: Summarization Based on Fractal Theory,” *Proceedings of the 26th Annual International ACM SIGIR Conference: Research and Development in Information Retrieval*, Toronto, Canada, July 28 to August 1, 2003.

[22] C. C. Yang, and F. L. Wang, “Automatic Summarization for Financial News Delivery on Mobile Devices,” *Proceedings of the International World Wide Web Conference*, Budapest, Hungary, May 20-24, 2003.