A Concept-Space Based Multi-Document Text Summarizer

By

TANG Ting Kap

A Thesis Submitted in Partial Fulfillment
of the Requirements for the Degree of
Master of Philosophy

In

Department of Systems Engineering and Engineering Management

July 2001

Abstract

People made decision based on the information they collected and analyzed. If the volume or completeness of such information increased, the quality of their decisions also improved. However, information overload and low utilization are two major challenges. In this thesis, we propose a new summarization generation method that based on concept space and statistical approach to solve these problems. Such approach mimics the process of how humans generate summary.

The automated summarization process will divide it into two stages – converge and diverge. Based on statistical analysis and concept space method, our system condenses a set of documents into a list of key issues. Then, based on location and frequency of terms, system selects a set of sentences to form a one-or two-page summary. According to the experiment conducted, summaries that developed based on concept spaces are more representative and flexible than those that based on keywords. This approach significantly reduces the information loss in the retrieval process. A user evaluation has also been conducted for its usefulness and other performance indices. The results indicated that such approach is promising in helping users to reduce time and efforts spent in reading documents.

摘要

我們會透過收集和分析資訊去作出決定，而收集到的資訊數量越多，決策的準成就越高。可是現今社會出現資訊爆炸的情況，使我們往往未能有效地運用現有的資訊，這對於決策者來說是一個很大的難題。為了解決這問題，一些自動摘要系統(Automated Summarization System)因此出現，而在本論文中我們就此作了進一步研究，將人工智能系統應用於這範疇上。在新的系統中，柔合了概念空間方法(Concept Space Approach)和統計學的技巧，從而改善現有的系統。

本系統的自動摘要過程包括了簡化和擴展兩部份，首先我們運用統計學分析和概念空間方法，將一批文章簡化為重點模式，然後基於重點在文章中出現的次數和位置，而從文章抽取不同的句子組成一遍摘要，但字數不多於兩頁紙。根據實驗數據，運用了概念空間方法的系統所得出來的摘要，比只是運用了單字計算的舊系統更能有效地表達原文的內容，明顯地減低在摘要過程中資訊的流失。我們更透過用者測試去評估新系統的實則可用性，結果証明了新系統能能幫助用者減少閱讀和分析所需要的時間。

Acknowledgments

I would like to give thanks to all people who have offered their help to me over the last two years. In this studying period, my friends, colleagues, brothers and sisters always support me.

I particularly would like to thank my supervisor, Professor Jerome Yen, for his kind and patient guidance for my research. Without his invaluable advice and endless encouragement, this thesis may never be completed.

I would like to thank my colleagues in Room 224. I am grateful to Chu Sir and Ivan Lau for giving me much advice and teaching me a lot of skills and telling me the right attitude in writing this thesis. They also teach me the techniques in problem solving. I especially thank Kylie Tang for helping me a lot in editing this thesis, it has consumed her so much valuable time for doing the grammar checking and teaching me how to set up type for the thesis. She is really kind and willing to offer me help at any time.

Finally, I would like to thank Professor Wai Lam and Professor Jeffrey Yu for giving me valuable advice and idea to improve the quality of the research and to enrich the content of the thesis as well.

Contents

List of Figures vi

List of Tables vii

1. INTRODUCTION **1**

1.1 INFORMATION OVERLOADING AND LOW UTILIZATION 2
1.2 PROBLEM NEEDS TO SOLVE 3
1.3 RESEARCH CONTRIBUTIONS 4
1.3.1 Using Concept Space in Summarization 5
1.3.2 New Extraction Method 5
1.3.3 Experiments on New System 6
1.4 ORGANIZATION OF THIS THESIS 7

2. LITERATURE REVIEW **8**

2.1 CLASSICAL APPROACH 8
2.1.1 Luhn's Algorithm 9
2.1.2 Edumundson's Algorithm 11
2.2 STATISTICAL APPROACH 15
2.3 NATURAL LANGUAGE PROCESSING APPROACH 15

3. PROPOSED SUMMARIZATION APPROACH **18**

3.1 DIRECTION OF SUMMARIZATION 19
3.2 OVERVIEW OF SUMMARIZATION ALGORITHM 20
3.2.1 Document Pre-processing 21
3.2.2 Vector Space Model 23
3.2.3 Sentence Extraction 24
3.3 EVALUATION METHOD 25
3.3.1 Recall, Precision and F-measure 25
3.4 ADVANTAGE OF CONCEPT SPACE APPROACH 26

4. SYSTEM ARCHITECTURE **27**

4.1 CONVERGE PROCESS 28
4.2 DIVERGE PROCESS 30
4.3 BACKWARD SEARCH 31

5. CONVERGE PROCESS **32**

5.1 DOCUMENT MERGING 32
5.2 WORD PHRASE EXTRACTION 34
5.3 AUTOMATIC INDEXING 34
5.4 CLUSTER ANALYSIS 35
5.5 HOPFIELD NET CLASSIFICATION 37

6. DIVERGE PROCESS **42**

6.1 CONCEPT TERMS REFINEMENT 42
6.2 SENTENCE SELECTION 43

6.3 BACKWARD SEARCHING 46

7. EXPERIMENT AND RESEARCH FINDINGS 48

7.1 SYSTEM-GENERATED SUMMARY V.S. SOURCE DOCUMENTS 52
7.1.1 Compression Ratio 52
7.1.2 Information Loss 54
7.2 SYSTEM-GENERATED SUMMARY V.S. HUMAN-GENERATED SUMMARY 58
7.2.1 Background of EXTRACTOR 59
7.2.2 Evaluation Method 61
7.3 EVALUATION OF DIFFERENT SYSTEM-GENERATED SUMMARIES BY HUMAN EXPERTS 63

8. CONCLUSIONS AND FUTURE RESEARCH 68

8.1 CONCLUSIONS 68
8.2 FUTURE WORK 69

A. EXTRACTOR SYSTEM FLOW AND TEN-STEP PROCEDURE 71

B. SUMMARY GENERATED BY MS WORD2000 75

C. SUMMARY GENERATED BY EXTRACTOR SOFTWARE 76

D. SUMMARY GENERATED BY OUR SYSTEM 77

E. SYSTEM-GENERATED WORD PHRASES FROM TEST SAMPLE 78

F. WORD PHRASES IDENTIFIED BY SUBJECTS 79

G. SAMPLE OF QUESTIONNAIRE 84

H. RESULT OF QUESTIONNAIRE 85

I. EVALUATION FOR DIVERGE PROCESS 86

BIBLIOGRAPHY 88

List of Figures

FIGURE 1: STOP WORD LIST 22
FIGURE 2: SYSTEM FLOW 31
FIGURE 3: PARAGRAPH UNIT FORMAT 33
FIGURE 4: INDEXED WORD-PHRASE FILE 35
FIGURE 5: CLUSTER WEIGHTING 38
FIGURE 6: WEIGHTING BETWEEN DIFFERENT CONCEPT TERMS 40
FIGURE 7: SYSTEM-GENERATED KEY ISSUE 41
FIGURE 9: SYSTEM GUI 51
FIGURE 10: COMPRESSION RATIO 53
FIGURE 11: THE NUMBER OF CONCEPT SPACE USED IN EACH SUMMARY 54
FIGURE 12: THE NUMBER OF CONCEPT TERM USED IN EACH SUMMARY 54
FIGURE 13: THE NUMBER OF CONCEPT TERM USED IN EACH CONCEPT SPACE 55
FIGURE 14: THE INFORMATION LOSS AGAINST NUMBER OF ARTICLE 56
FIGURE 15: THE INFORMATION LOSS AGAINST CONCEPT SPACE 57
FIGURE 16: THE INFORMATION LOSS AGAINST CONCEPT TERM 57
FIGURE 17: THE INFORMATION LOSS OF SAMPLE SIZE 100 58

List of Tables

TABLE 1: OPTIMAL POSITION LIST 14
TABLE 2: THE TWELVE PARAMETERS OF EXTRACTOR 60
TABLE 3: STATISTIC RESULT 62
TABLE 4: AVERAGE GRADING OF THE QUESTIONNAIRE 64
TABLE 5: WORD COUNT OF THE THREE SUMMARIES 64
TABLE 6: AVERAGE SCORE OF RELEVANCE, USEFULNESS, AND REPRESENTATIVE 66
TABLE 7: RELEVANCE AND USEFULNESS SCORES FOR RANKED SENTENCES 67

CHAPTER ONE

1. INTRODUCTION

Over the past three decades, information and telecommunication technologies have changed the way how people accessed and used information. Newspapers, academic journals, reports from financial service providers, and magazines used to be the only sources (Kolb, 1989; Parsaye et. Al., 1990). Today, with the advances in Internet and World-wide Web (WWW) technologies, individuals and organizations have begun to enjoy the benefits from the greater and broader access of information. However, as more information became accessible, information overload and low utilization also became serious problems to challenge the researchers. According to *Model of Human Processor*, there is a limit to the capacity of human information processing (Card et al, 1983). Such limitation can be found in vision, listening, short-term memory, long-term memory, and association. As Kandt and Yuender noted, information technologies, especially the Internet and WWW, are the major contributors to information overflow:

> *Companies are generating a lot of data but lack the tools to analyse 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 (Kalakota and Whinston, 1996).

1.1 Information Overloading and Low Utilization

Information overload has been a common problem in the digital world. In the new digital era, the living style of human has been changed and the ways of accessing and receiving information have also been changed. Especially, once information been digitized, they can be easily duplicated and disseminated. In the past, human distributed information verbally or physically, in which people needed to directly contact with each other. Later, they depended on paperwork to transmit information, such as letters, books, and newspapers etc. Nowadays, most information are distributed or transmitted digitally, which is more effective and efficient. The storage of information has become more convenient too. The storage devices take up little physical space, instead of a warehouse to store several million sheets of paper. Besides, they keep information longer, more reliable, and more secure. The time that spent on distributing the information has become shorter as well. Since the information distributes in digital format, it can be transmitted in variety of media, for instance, electromagnetic, optical fibers, and microwaves. Due to the rapid evolution of information technology, a vast amount of information can be easily obtained from different sources such as the Reuters and the Internet etc. We can obtain information very promptly and the time delay of receiving the information of an event from its occurrence tends to be zero. People have accustomed to share their ideas and information via e-mail and ICQ etc. The Internet has become a universal repository of human knowledge and cultures, which allows unprecedented sharing of ideas and information in a scale never been attained before.

Indeed, the problem of excessive information has occurred. Even if only one of the information sources is selected, we still cannot have adequate time to digest all the materials. Everyday, people take so much time to read through all their documents. Hardly ever people would completely read all the articles in the newspapers, but rather, they would selectively read articles according to how attractive the topic is. However, in the latter case the reader may miss some useful information, as they have not read it.

1.2 Problem Needs To Solve

Basically there are two problems: The first is the information overflow and the second is the inadequate utilization of the available information. The first problem is at the level of information management and the second one is at the knowledge management level. Our research is based on the assumption that processing information manually is so inefficient and ineffective such that it may become the bottleneck to the management.

Using artificial intelligence techniques to support human information processing has 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 the utilization of information (Everitt, 1980; Frawley et al, 1991; Salton, 1978). 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 databases, identifying patterns of spending of credit card holders, and examining corporate

intelligence reports (Addison, 1991; Chorafas and Steinman, 1990; Hayes and Weinstein, 1991).

However, most of these applications are limited to small and selected domains, such as, using symbolic pattern analysis to predict the changes and trends in prices of stocks (Kandt and Yuender, 1990) and in foreign exchange rates (Addison, 1991). In contrast to these heuristic approaches, the theory-based knowledge approach has led to the development of several model-based predictions systems, for example, predicting the fluctuation in the foreign exchange rate (Holsapple et al, 1998). Still, most of the above applications are too rigid, and it is possible that they could accidentally remove relevant information and generate recommendations that are not optimal.

1.3 Research Contributions

As we have illustrated how information technologies can lead to successful management of information overload and overflow, we proposed in this thesis, a new summarization approach, that is based on artificial intelligence and statistical techniques, to summarize articles or documents. The system is able to extract structures, patterns, heuristics or semantics to develop a set of concept spaces that represent the ideas of the selected articles. Then the system will generate a one or two pages summary for the users based on the concept spaces. Users normally only take few minutes to read through summary which is less than two pages. If users are interested in knowing the detailed about a particular issue, they can retrieve the related or associated paragraphs or articles from the summary.

1.3.1 Using Concept Space in Summarization

In this research, we have applied concept space approach in automated summarization. Automated summarization is not a new topic, some techniques have already been developed for solving this problem. Statistical approach is one of the commonly adopted methods, which is to identify important ideas of an article according to the frequency of keywords. Instead of using a single word, concept space approach extracts word phrases formed by adjacent words, as a word phrase can represent a meaning more specifically than a single word does. Sometimes, a single keyword, or even for a single word phrase, is not sufficient to stand for a particular topic. Therefore, we have employed the technique of Hopfield Net to group up different word phrases to form an "issue", based on the topic shared by different related articles from which the word phrases are extracted. An issue, consist of one or more word phrases, would describe a key topic of a set of articles. An automated summary would then be produced based on the system-generated issues. Since an issue may contain several word phrases, it would be rather comprehensive to represent a particular topic.

1.3.2 New Extraction Method

We have established a new extraction method to select sentences from the content of articles. A key issue, or a topic, generated by the system would be elaborated in one paragraph, and the length of that paragraph would be set by the user. The automated summary would be formed by grouping all the paragraphs together.

There should be a set of articles related to each particular topic. However, only a few of those articles would contain much detail about the topic, while most of them only contain very little information about the topic. Therefore, it is important to identify

those articles with very detailed information about the topic and then to extract sentences from those articles.

The extraction process would first find out the anchor documents of each key issue according to a formula. An anchor document is the most representing document for a key issue that includes the majority of relevant information about that key issue. After identifying the anchor documents, the system would extract the most representing sentence from each anchor document to form summary which would then be closely related to the key issue

1.3.3 Experiments on New System

We have carried out a set of experiments to evaluate the new system. The whole experiment would be divided into three parts according to following acpects: (1) System-generated summary compares to the source documents; (2) System-generated summary compares to the human-generated summary; and (3) System-generated summary compares to another system-generated summary.

The first part tests the coverage of the system-generated summary with respect to the source documents, which includes some measurements, such as, information loss, compression ratio, and the number of concept terms extracted.

The second part compares the system-generated summary to the human-generated one. Ten subjects would be invited to generate a set of word phrases to represent the key ideas of the source documents after they have read the source documents. And the

result would be used as the benchmark to make comparisons. We would use precision and recall measurement in the comparison.

The last part evaluates different system-generated summaries by testing their level of acceptance by human experts. Fifteen subjects would evaluate the three summaries generated by different methods regarding to their (1) readability, (2) coverage, and (3) time-effectiveness and helpfulness to the users. In addition, they would also evaluate the performance of diverge process in terms of the relevance, usefulness and representing features of each sentence extracted by our system.

1.4 Organization of This Thesis

The rest of the thesis is organized as follows: In chapter 2, it discusses some previous relevant research. Chapter 3 states the algorithm adopted in our summarization approach. Chapter 4 presents the process of text summarization. Details of the algorithm for generating concept spaces will be discussed in chapter 5. In chapter 6, it discusses about extraction method by identifying the anchor documents. Chapter 7 describes the experiment and user evaluation of the system. This thesis is concluded with a discussion about future research in chapter 8.

CHAPTER TWO

2. LITERATURE REVIEW

Summarization is to transform a source text to a summary text through content reduction by selecting and/or re-generalizing the important content in the source text (Sparck Jones, 1999). In the summarization process, we must characterize a source text as a whole and identify which is important content that need to be included in the summary. Basically there are three approaches of the text summarization method to achieve this: Classic Approach, Statistic Approach, and Nature Language Processing (NLP) Approach.

2.1 Classical Approach

Classical approach technique has developed 40 years before, it serves as a fundamental basis for summarization. Many contemporary summarization systems, even using the statistic approach and Natural Language Process approach, also mix with classical approach technique. Luhn is the pioneer work on automated summarization that tried to make up abstract for chemical literature (Luhn, 1959). Because the use of abstracts is an established practice in science and technology, and it has a standard style for academic literature, automated generation of abstract seems more desirable and easier to develop, but, this means the system is restricted on academic literature area only.

2.1.1 Luhn's Algorithm

Luhn's algorithm is to identify the "significant" sentence which carries out the key idea of the text. The "significance" factor of a sentence is derived from an analysis of its words. He proposed that the frequency of word occurrence in an article acts as a useful measurement of word significance, and the relative position within a sentence of words also acts as a useful measurement for determining the significance of sentences. In consequence, the significance factor of a sentence will be based on a combination of these two measurements. The advantage of this method is direct and not complex, it does not need to handle any linguistic matter such as grammar and syntax. The details of Luhn's algorithm (Luhn, 1959) presented in follow.

2.1.1.1 Optimal Word Frequency

At the first step of selecting "significant sentence", the words in the text will be ranked according its appearance frequency. Base on the fact that a writer normally repeats certain words when he advances or discusses his arguments and as he elaborates on an aspect of a subject. This means the emphasis is a useful indicator of significance. However, for some high frequency word will be described as too common word that would regard as "noise", these words are without any specific meaning. This kind of word will be ignored in the calculation process. Besides, the system will establishes a high frequency cutoff or confidence limits through statistical methods. When some high frequency words above this upper boundary will be ignored. And also there is a low boundary, the word below a certain frequency, it will also be cutoff. Therefore the optimum frequency is the range between the upper boundary and lower boundary.

After the cutoff by most common word list and optimum frequency, the remained words called "significant word".

2.1.1.2 Relative Position of Significant Word

The closer significant words are associated, the more specifically an aspect of the subject is being represented. Therefore, wherever the high frequently occurring of different significant words found in a close location, that means probability is very high that the information being conveyed is most representative of the article. Therefore, the "significance factor" can be derived which reflects the number of occurrences of significant words within a sentence and the linear distance between them due to the intervention of non-significant words. Then all sentences is ranked in the order of their significance according to its relative position of significant word, and one or several of the highest ranking sentences may be selected to serve as the automated-abstract.

We need to set a limit for the distance at which any two significant words shall be considered as being significant related. Two significant words have to within a close distance to reflect their relationship of this two word in a sentence. Useful limit is four or five non-significant words between significant words in a sentence.

The weighting equation for computing the significance factor as below.

$$\text{Significance Factor} = \frac{(\text{number of significant words in the bracketed})^2}{\text{total number of bracketed word}} \qquad (1)$$

For example, the follow sentence:

"Summarization is to [**transform** a source text to a summary text through content **reduction** by **selecting**] and/or re-generalizing the important content in the source text."

Where ***transform***, ***reduction*** and ***selecting*** are significant words in the article. Inside the bracket is the effective area of significant words. Thus,

$$\text{Significance Factor} = 3^2 / 13 = 0.692$$

2.1.2 Edumundson's Algorithm

For Luhn's method is using the word frequency and the position of the word in the sentence. In other work, Edmundson (1969) presented four factors for weighting the important sentences. The automatic extracting system proposed, is based on assigning to text sentences numerical weights that were functions of the weights assigned to certain machine-recognizable characteristics, or called **clues.** The four basic methods are cue base, key base, title base, and location base.

2.1.2.1 Cue Method

In the cue method the machine-recognizable clues are certain general characteristics of the corpus provided by the bodies of documents. The cue method is based on the hypothesis that the probable relevance of sentences is affected by the presence of pragmatic words such as "significant", "impossible", and "hardly". Such kind of

words is called *cue phrase*. This approach is through the cue words or phrases to identify the important sentences, which are then used to generate the summary. The cue dictionary in the system comprises three subdictionaries: Bonus words that are positively relevant; Stigma words that are negatively relevant; and Null words that are irrelevant. Null candidates – dispersion greater than a chosen threshold and selection ratio between two chosen thresholds; Bonus candidates – selection ratio above the upper threshold; Stigma candidates – selection ratio below the lower threshold. The final cue dictionary in the system contained 139 null words, 783 bonus words, and 73 stigma words. The final cue weight for each sentence is the sum of the cue weights of its constituent words. Frequently we apply cue phrase weighting schema in the summarization process to increase the overall performance.

2.1.2.2 Key Method

Key method is like the Luhn's algorithm for creating automatic extracts. Its machine-recognizable clues are certain specific characteristics of the body of the given document. It is based on the hypothesis that high-frequency content words are positively relevant or carrying important message. Thus, key words are assigned positive weights equal to their frequency of occurrence in the document. The final key weights of a sentence are the sum of its constituent words.

2.1.2.3 Title Method

The title method is based on the hypothesis that an author choose the title as circumscribing the subject matter of the document, also, when the author partitions the body of the document into different sections, the author will summarize each section

by choosing appropriate headings. Hence the words in the title or heading are positively relevant and high level significance. Thus, the clues are certain specific characteristics of the skeleton of the document, i.e. title and headings. In the system, it built the title glossary that consisting of all non-null words of the title, subtitle, and headings for that document. Words in the title glossary were assigned positive weights. The final title weight for each sentence is the sum of the title weights of its constituent words.

2.1.2.4 Location Method

The machine-recognizable clues in location method are provided by the skeletons of documents, i.e. heading and format. Two criteria are used for identifying the sentence: (1) Sentences occurring under certain headings are positively relevant and (2) topic sentences tend to occur very early or very late in a document and its paragraphs. This approach is weighting the sentences according to their ordinal position in the text, for example in first and last paragraphs, and as first and last sentences of paragraphs, it is more possibility to be extracted. The final location weight for each sentence is the sum of its heading weight and its ordinal weight.

The idea of location approach simply determines the importance of the sentence or paragraph according to its position in the article. In the recent research, Lin and Hovy (1997) further studies the location method, they try to find out where is the optimal position in the text, thus the sentence or word phrase in the optimal position can be extracted and used for presenting the topic of the text. They proposed Optimal Position Policy which according the list of topics keywords in the document to estimate where is the optimal position of the document. The experiment used the Ziff-

CHAPTER TWO

2. LITERATURE REVIEW

Summarization is to transform a source text to a summary text through content reduction by selecting and/or re-generalizing the important content in the source text (Sparck Jones, 1999). In the summarization process, we must characterize a source text as a whole and identify which is important content that need to be included in the summary. Basically there are three approaches of the text summarization method to achieve this: Classic Approach, Statistic Approach, and Nature Language Processing (NLP) Approach.

2.1 Classical Approach

Classical approach technique has developed 40 years before, it serves as a fundamental basis for summarization. Many contemporary summarization systems, even using the statistic approach and Natural Language Process approach, also mix with classical approach technique. Luhn is the pioneer work on automated summarization that tried to make up abstract for chemical literature (Luhn, 1959). Because the use of abstracts is an established practice in science and technology, and it has a standard style for academic literature, automated generation of abstract seems more desirable and easier to develop, but, this means the system is restricted on academic literature area only.

2.1.1 Luhn's Algorithm

Luhn's algorithm is to identify the "significant" sentence which carries out the key idea of the text. The "significance" factor of a sentence is derived from an analysis of its words. He proposed that the frequency of word occurrence in an article acts as a useful measurement of word significance, and the relative position within a sentence of words also acts as a useful measurement for determining the significance of sentences. In consequence, the significance factor of a sentence will be based on a combination of these two measurements. The advantage of this method is direct and not complex, it does not need to handle any linguistic matter such as grammar and syntax. The details of Luhn's algorithm (Luhn, 1959) presented in follow.

2.1.1.1 Optimal Word Frequency

At the first step of selecting "significant sentence", the words in the text will be ranked according its appearance frequency. Base on the fact that a writer normally repeats certain words when he advances or discusses his arguments and as he elaborates on an aspect of a subject. This means the emphasis is a useful indicator of significance. However, for some high frequency word will be described as too common word that would regard as "noise", these words are without any specific meaning. This kind of word will be ignored in the calculation process. Besides, the system will establishes a high frequency cutoff or confidence limits through statistical methods. When some high frequency words above this upper boundary will be ignored. And also there is a low boundary, the word below a certain frequency, it will also be cutoff. Therefore the optimum frequency is the range between the upper boundary and lower boundary.

After the cutoff by most common word list and optimum frequency, the remained words called "significant word".

2.1.1.2 Relative Position of Significant Word

The closer significant words are associated, the more specifically an aspect of the subject is being represented. Therefore, wherever the high frequently occurring of different significant words found in a close location, that means probability is very high that the information being conveyed is most representative of the article. Therefore, the "significance factor" can be derived which reflects the number of occurrences of significant words within a sentence and the linear distance between them due to the intervention of non-significant words. Then all sentences is ranked in the order of their significance according to its relative position of significant word, and one or several of the highest ranking sentences may be selected to serve as the automated-abstract.

We need to set a limit for the distance at which any two significant words shall be considered as being significant related. Two significant words have to within a close distance to reflect their relationship of this two word in a sentence. Useful limit is four or five non-significant words between significant words in a sentence.

The weighting equation for computing the significance factor as below.

$$\text{Significance Factor} = \frac{(\text{number of significant words in the bracketed})^2}{\text{total number of bracketed word}} \qquad (1)$$

For example, the follow sentence:

"Summarization is to [**transform** a source text to a summary text through content **reduction** by **selecting**] and/or re-generalizing the important content in the source text."

Where ***transform***, ***reduction*** and ***selecting*** are significant words in the article. Inside the bracket is the effective area of significant words. Thus,

$$\text{Significance Factor} = 3^2 / 13 = 0.692$$

2.1.2 Edumundson's Algorithm

For Luhn's method is using the word frequency and the position of the word in the sentence. In other work, Edmundson (1969) presented four factors for weighting the important sentences. The automatic extracting system proposed, is based on assigning to text sentences numerical weights that were functions of the weights assigned to certain machine-recognizable characteristics, or called **clues.** The four basic methods are cue base, key base, title base, and location base.

2.1.2.1 Cue Method

In the cue method the machine-recognizable clues are certain general characteristics of the corpus provided by the bodies of documents. The cue method is based on the hypothesis that the probable relevance of sentences is affected by the presence of pragmatic words such as "significant", "impossible", and "hardly". Such kind of

words is called *cue phrase*. This approach is through the cue words or phrases to identify the important sentences, which are then used to generate the summary. The cue dictionary in the system comprises three subdictionaries: Bonus words that are positively relevant; Stigma words that are negatively relevant; and Null words that are irrelevant. Null candidates – dispersion greater than a chosen threshold and selection ratio between two chosen thresholds; Bonus candidates – selection ratio above the upper threshold; Stigma candidates – selection ratio below the lower threshold. The final cue dictionary in the system contained 139 null words, 783 bonus words, and 73 stigma words. The final cue weight for each sentence is the sum of the cue weights of its constituent words. Frequently we apply cue phrase weighting schema in the summarization process to increase the overall performance.

2.1.2.2 Key Method

Key method is like the Luhn's algorithm for creating automatic extracts. Its machine-recognizable clues are certain specific characteristics of the body of the given document. It is based on the hypothesis that high-frequency content words are positively relevant or carrying important message. Thus, key words are assigned positive weights equal to their frequency of occurrence in the document. The final key weights of a sentence are the sum of its constituent words.

2.1.2.3 Title Method

The title method is based on the hypothesis that an author choose the title as circumscribing the subject matter of the document, also, when the author partitions the body of the document into different sections, the author will summarize each section

by choosing appropriate headings. Hence the words in the title or heading are positively relevant and high level significance. Thus, the clues are certain specific characteristics of the skeleton of the document, i.e. title and headings. In the system, it built the title glossary that consisting of all non-null words of the title, subtitle, and headings for that document. Words in the title glossary were assigned positive weights. The final title weight for each sentence is the sum of the title weights of its constituent words.

2.1.2.4 Location Method

The machine-recognizable clues in location method are provided by the skeletons of documents, i.e. heading and format. Two criteria are used for identifying the sentence: (1) Sentences occurring under certain headings are positively relevant and (2) topic sentences tend to occur very early or very late in a document and its paragraphs. This approach is weighting the sentences according to their ordinal position in the text, for example in first and last paragraphs, and as first and last sentences of paragraphs, it is more possibility to be extracted. The final location weight for each sentence is the sum of its heading weight and its ordinal weight.

The idea of location approach simply determines the importance of the sentence or paragraph according to its position in the article. In the recent research, Lin and Hovy (1997) further studies the location method, they try to find out where is the optimal position in the text, thus the sentence or word phrase in the optimal position can be extracted and used for presenting the topic of the text. They proposed Optimal Position Policy which according the list of topics keywords in the document to estimate where is the optimal position of the document. The experiment used the Ziff-

Davis text corpus, which included 13,000 newspaper, for testing. Table 1 shows the result of optimal position found.

Table 1: Optimal position list

Importance level	Position
1	(T) (P_2,S_1) (P_3,S_1) (P_2,S_2)
2	(P_4,S_1) (P_5,S_1) (P_3,S_2)
3	(P_1,S_1) (P_6,S_1) (P_7,S_1) (P_1,S_3) (P_2,S_3)

Edmundson's system extracts clues may come from two sources – structural and linguistic. Title method and Location method belong to structural source. Key method and Cue method belong to linguistic source.

The basic technique for classical approach is using location, cue phrase and word frequency as indictor for computing the importance of the sentence. Then extracts the most important sentence to form abstract/summary directly. However, for the classical approach, Luhn's and Edmundson's algorithm, computing process is still without consider linguistic implication of the sentence, it just selects the sentence according to the physically factor of the sentence/word only. As the result, the computation of significance factor much depends on the author's writing style or the subject of the text.

2.3 Statistical Approach

Statistical approach used to find out the importance of the key words or sentences based on term frequency and document frequency. Each key word or key phrases are assigned a weighted score. Afterwards, using the term frequency and document frequency, keywords in different articles can be identified by statistical algorithms, and then using these keywords for information retrieval.

G. Salton had done a lot of work on this area. He applied vector space model for text processing to get a superior retrieval result, and used this technique in summarization (Salton et al, 1997). For each text/paragraph is represented by a vector of weighted term, and then by computing the pairwise similarity coefficients to measure the vocabulary overlap between the corresponding text. If the similarity between two vectors is large enough to be regarded as non-random, it can conclude that the vocabulary-matches between the corresponding texts is meaningful, thus, a pairwise links between the texts is generated. The link indicates that the linked texts are semantically related. Then, the process of automatic summary generation reduces to task of extraction, such that it can use heuristics based upon a detailed statistical analysis of word occurrence to identify the text-pieces (sentences, paragraph, etc.). Consequently the text-pieces are likely to convey the content of a text, and concatenate the selected pieces together to form the final extract – summary.

2.4 Natural Language Processing Approach

Many applications included computational linguistics process, such as machine translation, information retrieval, and speech synthesis and recognition. In multi-language social, such as Hong Kong society using both Chinese and English, every day

millions of words need to be translated from one language to another, and retrieves information from many bi-lingual documents. This result in an overwhelming need for some automated computing process to handling linguistics task, when done correctly, it can reduce many time-consuming and mentally demanding processes.

Natural Language Processing approach solves the lexical problem by identifying, the subject, verb, etc. But it is still difficult to generate good results if structures of sentences are complicated, or writing style varies. Besides different language has different lexical structure, there is still room for improvement, but it is very labour-intensive. Lehnert (1981) used lexical chain to summarize narrative articles. The algorithm is to capture the primitive and complex plotting units in order to identify the connectivity and symmetry of how characters or words interact. The summarization process identifies string lexical chains and extract significant sentences based on the identified lexical chains. However, this approach is only valid for generating summaries from narrative articles.

More general approach that based on lexical chain was proposed by Barzulay and Elhadad (1997). The procedure for constructing lexical chains included (1) Select a set of candidates; (2) For each candidate word, find an appropriate chain relying on a relatedness criterion among members of the chains; (3) If it is found, insert the word in the chain and update it accordingly. The system applied WordNet lexical database (contained more than 118,000 different word forms) for determining relatedness of the words, that used for calculation of lexical chains. The system will choose simple nouns and noun compounds as candidate words. This extends the set of candidate words to include noun compounds. After that, it builds chains in every text segment

according to relatedness criteria, and merges the chains from the different segments using much stinger criteria for connectedness.

Actually there is few pure NLP summarization system or pure statistical summarization system. Many systems apply both NLP and statistic techniques together to solve summarization problem. It can use statistically technique to analyze the content to figure out the important meaning. Then, by helping with Natural Language Processing, we are able to generate new sentence to form summary.

CHAPTER THREE

3. PROPOSED SUMMARIZATION APPROACH

The theme of the summarization is to present most of the ideas in the original document in a short space. The problem of information explosion, or information overloading, has indicated that summarization is extremely essential and helpful in saving time and efforts for the users. Therefore, during the past two decades many approaches have been proposed for summarization.

Generally, people would receive information from pushing and pulling actions. When a person is searching materials or retrieving information in the World Wide Web, if he has already had a target object before searching the information, he just looks for any suitable information matched with his target. As the result this person is doing a 'pulling' action, he requests the information in an active manner. Moreover, summarization process is likely to provide a 'pushing' action. Without specifying any criteria from the users, it automatically generates a summary for them. Before the users read any document, they can first read its summary to get the generalized idea of the document. Yet, it is still a process of retrieving information, but for one whose objectives are not clearly defined when he reads the summary, and whose purpose may change while reading. A good summary can provide an easy way to select which document they are interested in.

3.1 Direction of Summarization

There are two types of summarization – text extraction and fact extraction (Sparck Jones, 1999). First, if we do not have any target object, there is no prior presumption about what sort of content information is essential. Then the extraction is 'what you see is what you get', we just extract what is the view in the source text and transfers it to constitute the summary text. Text extraction is an open approach. Second, if we have already decided a sort of subject content to look for in the source documents, that is what you seek to extract. This fact extraction is 'what you know is what you get', it is a closed approach.

In the text extraction processing, it effectively merges the interpretation and generation stages. Key sentences in the text are identified by some mix of statistic, location, and cue word approach. The extracted sentences are taken as its own representation to form the summary, without any interpretation. And this representation is then subject to a generation stage which is simply extractive. The selected sentences together usually have some relationship to what would independently be judged as important source content, and allow the reader to infer what it might be. But the output summary text may be obscure.

With fact extraction the interpretation and generation stages are also essentially merged. The source text expressions that bear on the specified generic concept or concept relations, then it seeks for this particular factual instantiations. The primary character of this approach is to allow only one view of what is important in a source.

Advantage of text extraction is more generality. However, it generates low-quality summary output because the weak and indirect methods used are not very effective in identifying important material and in presenting it as well-organized text. The fact extraction approach can generate better quality summary output, in substance and presentation, but with disadvantages that required type of information has to be explicitly.

3.2 Overview of Summarization Algorithm

The requirements for summarization have become more demanding and the systems have also become more powerful and sophisticated. At the initial stage, text summarization could only deal with one single document with one specific domain. However, some systems developed during the past five years have been able to generate summaries from multiple documents with multiple domains. That is, the trend of text summarization has moved from a single document to multiple documents, from domain dependent to domain independent, and from a single language to multiple languages.

The system presented in this thesis is for summarizing multi-documents with general domain , that is, domain independent. First the system would calculate the term frequency and document frequency in order to identify the meaningful word phrases in the documents. Cluster analysis algorithm, Hopfield Net and Genetic Algorithm, would be used to generate the concept spaces. Then our system would use those terms in the concept spaces to represent the key ideas of the documents. Finally a summary would be generated using the sentence selection method. We would illustrate the main steps of summarization in the following sections:

3.2.1 Document Pre-processing

Before implementing any statistical algorithm, it needs to do some pre-processing work to assist the calculation. First, it should provide positional information of all the words in the document. A more sophisticated version of the inverted index would be applied for carrying such positional information. Instead of just listing the documents in which word appears, the positional information of all its occurrences in the document would be stored as well.

The second step is to form the word phrases. A word phrase is more representable than a single word as it can illustrate a meaning more specifically. For example, the word phrase 'interest rate' would have a more specific meaning than the single word 'interest'. There are several definitions for the term 'interest', (1) curiosity and concern; (2) advantage or benefit, and (3) money charged in financial operation. However, for the term 'interest rate', its meaning is confined to the third explanation only. In this step, usually it applies stop word list to identify the boundary of word phrases. An example of stop word list is shown in Figure 1. For the sentence: "A system menu is different from other components", it would yield three word phrases of "*system menu*", "*different*", and "*components*".

Third, it applies the stemming technique to simplify the extracted words. Stemming usually refers to a simplified form of morphological analysis consisting simply of truncating a word. For example, *laughing*, *laugh*, *laughs* and *laughed* would be stemmed to *laugh*.

a, also, an, and, as, at, be, but, by,

can, could, do, for, form, go,

I, if, in, into, it, its,

my, of, on , or, out, say, she,

that, the, their, there, therefore, they,

this, these, those, through, to, until,

we, what, when, where, which, while, who,

with, would, you, your

Figure 1: Stop word list

The performance of the key term extraction process can be enhanced with the use of a particular thesaurus which provides a set of technical terms in a specific area/field. By using a thesaurus, we can identify the meaning of a searched word as well as its synonym(s). It can also allow us to generate a new sentence with similar meaning in different wording.

Although using a thesaurus may obtain a rather good result, there are still limitations for this method: (1) Domain dependent – the system has to change the thesaurus in use for articles of different domains as it can only use one particular thesaurus at a time. Therefore, to summarize a set of articles in different domains, the system would have to frequently change and use different thesauruses and it would be very inefficient and time consuming. (2) Storage limitation – if the number of terms in a thesaurus increase, the storage capacity has to increase as well; it would thus increase the processing time. (3) New words update – the thesaurus has to keep up-to-date with those new terms emerged in the changing society. For example, the term "ftp" is not included in some

older version of dictionaries. Therefore, in our system, we would not adopt any thesaurus for the key term extracting process.

3.2.2 Vector Space Model

The vector space model is one of the techniques for ad hoc retrieval. It is widely adopted for its good performance in information retrieval. It has applied the concept of spatial proximity for semantic proximity. Documents and queries can be represented in a high-dimensional space, in which each dimension of the space corresponds to a word in a set of documents.

A vector in a concept space represents a word term which is used for a single word or a word phrase. We would concern term weight rather than word weights because dimension in the vector space model can correspond to word phrase as well as a single word. If the angle between two vectors is small, they would be closely related to each other. The term weight would be calculated with term frequency and document frequency.

The term frequency is used to reflect how salient a word is within a document. The higher the term frequency (the more often the word occurs), the better the word can represent the content and the more important the word is. The following equations are example of term weight derived from term frequency.

$$F(tf) = sqrt(tf) \quad , tf \text{ is term frequency, and} > 0 \qquad (2)$$

$$F(tf) = 1 + \log(tf) \quad , tf \text{ is term frequency, and} > 0 \qquad (3)$$

Document frequency can be interpreted as an indicator of informativeness. Usually, there would be proximity for the occurrence of a keyword in a document. In contrast, a non-keyword would spread throughout the whole document. Different weight schemes derived from document frequency would include inverse document frequency (*idf* weighting) and logarithmic document frequency.

Then, the fact that two or more terms occur in the same documents more often than chance. Co-occurrence technique is tried to capture any related terms or paragraph, and group them together. Co-occurring terms are projected onto the same dimensions, non-co-occurring terms are projected onto different dimension. We can look as a similarity metric that is an alternative to word overlap measures by tf.idf.

At last, we have employed the technique of Hopfield Net to group up different word phrases to form an "issue". An issue, consist of one or more word phrases, would describe an important idea.

3.2.3 Sentence Extraction

After identifying the important ideas, we would try to extract some sentences directly from the content to form the summary. We would find out the most important documents according to their weights, and from each document one or more sentence would be extracted to represent its topic. Those related sentences would be grouped into one single paragraph and the summary generated would not exceed two pages.

3.3 Evaluation Method

Finally, we have conducted a set of experiments to evaluate the new system. Oftentimes human analyses or human judge will introduce his own biases, so we need a standard for evaluating a system-generated summary. However serious questions remain concerning the appropriate methods and types evaluation. In generally, there are two main approaches to evaluate a computer-generated summary. The first is an extrinsic evaluation in which the quality of the summary is judged based on how it affects the completion of some other task. The second approach, an intrinsic evaluation, judges the quality of the summarization directly based on user judgements of informativeness, coverage, etc.

3.3.1 Recall, Precision and F-measure

Since the quality of many retrieval systems depends on how well they manage to rank relevant documents before non-relevant once, so researchers have developed evaluation measures specifically design to evaluate relevant document ranking. The measurement included recall and precision, which are almost the golden standard for measurement of category result. Furthermore, if both recall and precision are important, F-measure can be used for evaluation at fixed cutoffs. In the Equation 6, where P is the precision, R is the recall, and a determines the weighting of precision and recall.

$$\text{Recall} = \frac{\text{categories found and correct}}{\text{total categories correct}} \tag{5}$$

$$\text{Precision} = \frac{\text{categories found and correct}}{\text{total categories found}} \tag{4}$$

$$F = 1 / (a * 1/P + (1-a)*1/R) \quad (6)$$

3.4 Advantage of Concept Space Approach

In the summarization process, the key ideas of each document need to be identified. However, a vocabulary-switching problem exists in human writing or speaking. Different people use different vocabulary to describe the same idea. Early research has identified that, the possibility for two people to use the same term to describe one subject, is less than 0.2 (Furnas, 1987).

By using concept space approach, more than one words or phrases are used to represent one idea. It provides greater flexibility and representative than the approaches that based on single keyword. And also it is domain independent, no thesaurus is needed to support segmentation.

CHAPTER FOUR

4. SYSTEM ARCHITECTURE

First of all, we can take human summarization process as our reference. In human-quality summarization, it includes the following processes:

(1) Understand the contents of the documents.
(2) Identify the most important ideas of the information or the key concepts.
(3) Write up a summary that contains such ideas or key concepts.

There are three main steps of building a summary in human process. As researches try to simulate the human summarization process by using computer, generally computing summarization also includes three steps.

Sparch Jones (1999) defined summarization process:

Interpretation:

source text interpretation to source text representation

Transformation:

source representation transformation to summary text representation

Generation:

summary text generation from summary representation

Hovy and Lin proposed (1997),

summarization = topic identification + interpretation + generation

And according to Mani and Bloedorn (1999), summarization process can be characterized as ***analysis***, ***refinement***, and ***synthesis***.

Basically, the spirit of summarization is same among different definition. The first step is to identify the key topic in the content. Second, it needs analyses the key topic, actually what is the meaning of author presenting. Finally, it needs to write up a short passage as a summary, by re-generating sentence or extracting the important sentence directly from content. This thesis proposes that the summarization process can separate into two sub-processes —**Converge** and **Diverge**.

4.1 Converge Process

During the Converge Process, system identifies the key ideas from the documents. It condenses the source documents into words or word phrases. In the scanning process of human, while they have read the important sentence, they will underline this sentence or the key word. Then, these sentences or wordings would be used in the building process of summary. The converge process is exactly simulate the scanning process. After the converge process, it will convert a set of document into a list of word phrases which called concept term. The word phrase act as the intermediate product, the advantages of using word phrase are (1) concise – can easy understand by user, (2) squeezed – cut out non-important content, reduce the size and time for reading. The list of concept term is directly affecting the quality of summary, so we leave the room for refinement of the list by the user. At this stage user can alter the

list to cut out the non-representative concept term or to modify the wording of concept term.

Briefly, we use statistic technique to condense the contents by calculating the term frequency and document frequency of words or word phrases. The system will select the word phrases with frequency count over the threshold. With these words or word phrases, we use cluster analysis to determine it is make sense. For the word phrases without any specific meaning will be delete. In addition, Hopfield Net algorithm is applied to group the word phrases which have related ideas together. Each resulted group will be a key issue, then we use the generated set of key issue to represent the major ideas of the documents. Then, the word phrases included in the key issue are called concept term.

During this process, how to minimize the information loss is a critical issue. As volume of content will be cut out, there is limited information can be remained. The quality of the summary is much depended on the selection of key issue. If the control of the selection is poor, much important information will be cut out as garbage. If there are too few word phrases being selected, it is possible that part of the contents in the original documents will be lost. If the number of the issue increase, more concept terms (word phrases) will be included such that more information carried. Intuitively, we should increase the length of the key issue list to avoid the information loss. However, this action would reduce the precision simultaneously. There is a trade off of recall and precision

4.2 Diverge Process

After the list of key issue is produced, the next stage is the Diverge Process. Diverge Process is the process of developing a passage of summary by sentence extraction. The sentence extraction process is based on location method to extract important sentences from the original set of documents to develop the summary. A new sentence is very difficult to reinterpret or regenerate, as it requires special techniques to handle semantic and lexical complexities. For the domain dependent text summarization, the level of difficulty of sentence re-generation is lower. However, it is more difficult to generate coherent summary in domain independent documents. In contrast, by directly extract the sentence from the source document, that can make sure the sentences are without grammatical mistake and fluent.

We assumed that there should be a set of related document mentioning for each key issue. Therefore, we try to rank the relatedness of the document to each issue, and treat the most related documents of an issue to be the anchor document. The found anchor document should be included most information of an issue. After we identified which are anchor documents, then the system extract the most representative sentence from it according the location method. The original sentences, which contained concept term identified in the Converge Process, are picked out directly from the source documents. One or more meaningful sentences are grouped to form a paragraph for each issue, and then all the paragraphs are grouped together to produce the summary for the document set.

In the result, the summarization process of the system is divided into two steps. First the source document condenses into a set of concept term. And then using the concept

term as the intermediate product to burst out a summary. The flow of the summarization process of the system is shown in Figure 2.

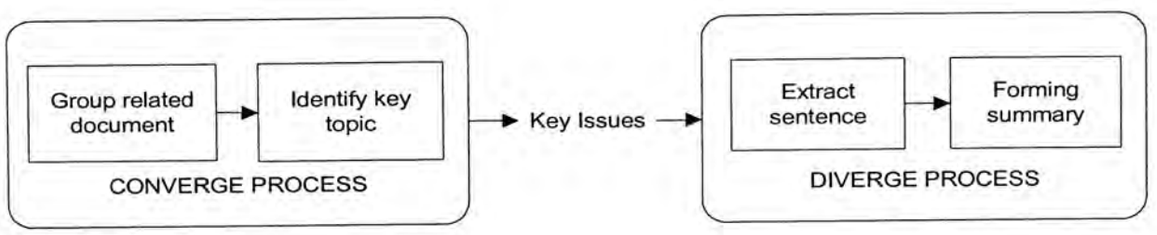


Figure 2: System Flow

4.3 Backward Search

Based on the summary, the users could retrieve more detail information. Each of the extracted sentences likes a button. When the user clicks the sentence, then the system will retrieve the paragraphs which contained the clicked sentence. When the user is interested at an issue, he can retrieve the related document of this issue to get more detail information.

CHAPTER FIVE

5. CONVERGE PROCESS

In this research, the proposed system used concept space approach to support the Converge Process. The system would generate a set of key terms to represent the key ideas of the documents through concept space approach (Chen et al, 1996; Yen et al, 1996). It first merged and indexed the documents to develop a formatted file. Then, it grouped the key words that were frequently mentioned to generate concept spaces by cluster analysis, for example, Hopfield net classification (Salton, 1989), and Machine Learning (Chen et al, 1992; Chen et al, 1993; Chen et al, 1994; Chen, 1995; Everitt, 1980).

5.1 Document Merging

In order to generate summary from multiple documents, this process merges all the source documents into a single file, and each document is partitioned into many paragraph-size units, which is called processing units. Size of the processing units is determined by the size of the original document. In our research we select a paragraph as a processing unit because a paragraph is contained sufficient contents, in terms of key words, to support the analysis that was based on manipulating the term frequency and the document frequency. Each sentence is marked and indexed with its position in the paragraph. Also each paragraph is marked and indexed with its position in the

document set. Thus, the output is a single file that contained indexed documents. Figure 3 shows the result after document indexing.

1.1	Speculators hang on for roller-coaster ride
1.2	Many speculators said they were happy with their investments despite the plunge in the Hang Seng Index.
1.3	Dozens of people in front of share price display screens at a Quarry Bay bank swore and shouted every time the prices of major blue chips fell, and cheered when the prices of red chips rose.
1.4	Veteran speculator Roman Hong, 41, who works for a bank but took a day off yesterday, said the fall in prices was good. "This is normal and healthy. It is a natural flow of the tide and is nothing special," he said.
1.5	"Years of experience tell me that I should make a big intake now, while the prices are low. I am taking a medium to long term strategy. What people discard, I take on."
1.6	Salesman Frankie Cheung, 37, agreed. Mr Cheung, who also took a day off work, said he rushed to buy shares in the morning after seeing prices drop. "I made a fortune this afternoon, as the shares I bought shot up, unlike others who have been less fortunate. I bought 20,000 blue chips and 10,000 technology shares, and in general I earned quite a lot," he said.
1.7	"I will buy even more this afternoon when share prices are low, and hopefully when prices rise again I can make more money."
1.8	Mrs Chan, a housewife who spent the morning at the bank with her mother and brother, said she had been less fortunate. She said the shares she bought weeks ago had risen in value but fallen back and although she had not lost money, she could not sell the shares.
1.9	"I have over 80,000 shares in my pocket and cannot sell them. The other 100,000 shares I bought yesterday are not doing very well today either," she said.
1.10	She said she was optimistic prices would rise.
1.11	But a 64-year-old retired man, Mr Cheung, clearly depressed, said he was "losing money like a dog".
2.1	Hi-tech stocks fuel 1,200-point drop
2.2	The Hang Seng Index plunged by more than 1,200 points yesterday - its biggest one-day decline in more than two years - as sharp losses on Wall Street on Tuesday rattled investors.
2.3	Technology and telecom stocks took a hammering, many dropping much more than the index's 7.18 per cent, following Tuesday's five per cent drop on the hi-tech US Nasdaq market. In early trading yesterday, the Nasdaq was down 1.9 per cent and the Dow Jones up 0.9 per cent.
2.4	"Nasdaq's tumble tells us that tech-stocks are extremely over-valued and that the bubble might be about to burst," said Core Pacific-Yamaichi head of research Alex Tang Yee-yuk.
2.5	The Hang Seng dropped 1,226.1 points, to close at 15,846.72, as part of a big sell-off in Asian markets. It was the largest single-day points loss since the index fell 1,438.31 pointson October 28. 1997 at the onset of the regional financial crisis.

Figure 3: Paragraph unit format

5.2 Word Phrase Extraction

We developed 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), such as *'calculate', 'articulate', 'teach',' listen'* etc. When words in an article that matches with any of the stop words, the words would be removed from further analysis. However, any user would be able to modify the stop-word list to help sharpen the result.

We then use adjacent words to form phrases. After examining similar documents, we decide to form phrases that contained up to three words because most subject descriptors are less than four words. Our system would generate 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". (A threshold process was then adopted to remove some of the noise produced as the result of term-phrase formation, e.g., "Street firm".)

5.3 Automatic Indexing

In order to represent each term, we proposed an indexing method. All terms are represented in upper case. For each term, we use six numbers to store its information. The first number indicates a unique identification number for document. Second umber is paragraph number in a document. We reserved third number for future use, it is dummy and set to 1. Forth number and fifth number represent the sentence number and word number in the sentence, respectively. The last number represents the word count, i.e., number of words in a term. A sample is shown in Figure 4. All of the extracted word phrases are stored in a text file.

```
1.4 1 1 11 1 CHINADOTCOM
1.4 1 1 11 2 CHINADOTCOM CORP'S
1.4 1 1 12 1 CORP'S
1.4 1 1 13 1 CHINA
1.4 1 2 15 1 INTERNET
1.4 1 2 15 2 INTERNET PORTAL
1.4 1 2 15 3 INTERNET PORTAL ARM
1.4 1 2 16 1 PORTAL
1.4 1 2 16 2 PORTAL ARM
1.4 1 2 16 3 PORTAL ARM HONGKONG
1.4 1 2 17 1 ARM
1.4 1 2 17 2 ARM HONGKONG
1.4 1 2 18 1 HONGKONG
... ....
```

Figure 4: Indexed word-phrase file

5.4 Cluster Analysis

Base on the vector model, we used *document frequency* and *term frequency* to determine the relationships among terms and hence identify the key terms. The term frequency, tf_{ij}, is the number of a term j that appears in a document i. The document frequency, df_j, is the number of document i that contains this term j. Generally, the term, which appears more times, is more important. And the term appears in fewer documents has more specific meaning. Simply, we can assign a weighted score to each term according to the application.

We determine the threshold of term frequency for removing the uncommon terms (e.g., typos and unique abbreviations) and for generating a set of terms (vocabulary) for further analysis. Only terms with term frequency past the threshold are considered in the following concept classification process. For example, if the threshold is set to be five, only those terms which appear more than five times in a document will be considered. Users can input this threshold for adjustment. A paragraph (document) is

considered "lost" if all the terms appeared in it are less than the specified term frequency threshold.

After the "lost" data is removed, the remaining terms are ranked according to the decreasing order of their document frequency. However, the most frequently occurring terms may be too general to convey any specific idea, such as, "BUSINESS", so once they have appeared in the stop-word list they are removed.

Next, we compute the combined weight of term j in document i, d_{ij}, based on the product of *term frequency* and *document frequency* as follows:

$$d_{ij} = tf_{ij}^{2} \times \ln df_j \qquad (6)$$

Notice that this computation is based on *document frequency* instead of the more conventional *inverse document frequency* used in solving large-scale automatic indexing problems. This change is made because of the need to identify common key issues. We adopted *document frequency* in order to increase the weights associated with the more frequently used terms. In the conventional automatic indexing environment, however, the practice is to assign higher weights to more specific or unique indexes through *inverse document frequency*.

$$\textit{Inverse document frequency} = (\log \frac{n}{df_j}) \qquad (7)$$

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

$$ClusterWeight(T_k, T_j) = \frac{\sum_{i=1}^{n} d_{ijk}}{\sum_{i=1}^{n} d_{ik}} \quad (8)$$

$$ClusterWeight(T_j, T_k) = \frac{\sum_{i=1}^{n} d_{ijk}}{\sum_{i=1}^{n} d_{ij}} \quad (9)$$

These two equations compute the similarity weights from term T_j to term T_k (the first equation) and from term T_k to term T_j (the second equation), where d_{ij} represents the combined weight of term T_j in document i, d_{ik} represents the combined weight of term T_k in document i, and d_{ijk} represents the combined weight of both descriptors T_j and T_k in document i. d_{ijk} is computed by:

$$d_{ijk} = tf_{ijk}^2 \times \ln df_{jk} \quad (10)$$

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. Figure 5 shows the sample file of cluster weighting. Besides, in order to obtain a reasonable number of co-occurring terms for each term, we experimentally set the weight threshold to be 0.3. This threshold ensures that only the most relevant terms are represented in our final concept space.

5.5 Hopfield Net Classification

We adopted a variant of the Hopfield network (Hopfield 1982) and its parallel relaxation procedure to identify *clusters* of relevant terms.

WIRELESS : CABLE : 1.000000
WIRELESS : CYBERWORKS : 0.584064
WIRELESS : HKT : 0.498036
WIRELESS : WIRELESS HKT : 0.449475
WIRELESS : MERGER : 0.266564

RICHARD LI : LI : 0.775756
RICHARD LI : RICHARD : 0.775756
RICHARD LI : CYBERWORKS : 0.677202
RICHARD LI : RICHARD LI TZAR-KAI : 0.581593
RICHARD LI : LI TZAR-KAI : 0.581593

CONFIDENTIAL GOVERNMENT DOCUMENT : GOVERNMENT : 0.654313
CONFIDENTIAL GOVERNMENT DOCUMENT : CONFIDENTIAL GOVERNMENT : 0.654313
CONFIDENTIAL GOVERNMENT DOCUMENT : BUSINESS : 0.654313
CONFIDENTIAL GOVERNMENT DOCUMENT : BUSINESS CONTACT : 0.654313
CONFIDENTIAL GOVERNMENT DOCUMENT : CONFIDENTIAL : 0.654313
CONFIDENTIAL GOVERNMENT DOCUMENT : GOVERNMENT DOCUMENT : 0.654313
CONFIDENTIAL GOVERNMENT DOCUMENT : DOCUMENT : 0.654313
CONFIDENTIAL GOVERNMENT DOCUMENT : CHENG : 0.316514
CONFIDENTIAL GOVERNMENT DOCUMENT : PUBLIC : 0.218104
CONFIDENTIAL GOVERNMENT DOCUMENT : FORMER : 0.218104

... ...

Figure 5: Cluster weighting

Each term in the co-occurrence analysis results is treated as a neuron and the asymmetric weight between any two terms is taken as the unidirectional, weighted connection between neurons. Hopfield algorithm activates its neighbours, combines weights from all associated neighbours, and repeates this process for all terms, according to the decreasing order of their term frequencies, until the output pattern has converged.

The resulting output reveales all concepts that are semantically relevant to the input terms. The clusters with strongly related terms forms the concept spaces. Therefore the result list of concept spaces given is the key ideas of the documents.

The final output from the above process is a network of terms and their weighted relationships. It is similar to a neural network of nodes and weighted links.

A sketch of the *Hopfield net concept classification* procedure follows:

- ***Assigning Connection Weights:***

  t_{ij} represents the synaptic weight from node i to node j.

- ***Initialization with Unknown Input Pattern:***

$$\mu_i(0) = x_i, 0 \le i \le n-1 \tag{11}$$

  $\mu_i(t)$ is the output of node i at time t and x_i has a value between 0 and 1.

- ***Parallel Activation and Iteration:***

$$\mu_j(t+1) = f_s[\sum_{i=0}^{n-1} t_{ij}\mu_i(t)], 0 \le j \le n-1 \tag{12}$$

  f_s is the continuous Sigmoid transformation function (Frawley, 1991) as shown below:

$$f_s(net_j) = \frac{1}{1+\exp[\frac{-(net_j-\theta_j)}{\theta_0}]} \tag{13}$$

  where $net_j = \sum_{i=0}^{n-1} t_{ij}\mu_i(t)$, θ_j serves as a threshold or bias and θ_0 is used to modify the shape of SIGMOID function. This formula shows the parallel relaxation property of the Hopfield net.

- ***Convergence***

  This above process is repeated until there is no longer a change between two iterations in terms of output, which is accomplished by checking:

$$\sum_{j=0}^{n=1}[\mu_j(t+1)] - \mu_j(t)]^{2} \leq \varepsilon \qquad (14)$$

  where ε is the maximal allowable error determined empirically.

At finally, it will identify a list of key topic by the Hopfield Net, the activation procedure are shown in Figure 6. Figure 7 shows the resulted system-generated key issue.

```
neuron 0: 0 1 2
neuron 1: 0 1 2
neuron 2: 0 1 2
neuron 3:
neuron 4:
neuron 5:
neuron 6:
neuron 7: 5 7 41 49
neuron 8:
neuron 9:
neuron 10:
neuron 11:
neuron 12:
neuron 14:
neuron 15:
neuron 16:
neuron 17:
neuron 18:
neuron 19:
neuron 20:
neuron 21:
neuron 22: 0 1 2
neuron 24:
neuron 25:
neuron 26:
neuron 27:
neuron 28:
neuron 29: 0 1 2
neuron 30:
```

Figure 6: Weighting between different concept terms

1. INTERNET/HUTCHISON/
2. SINGAPORE/MERGER/TELECOMMUNICATIONS/SINGTEL
3. FREE TRADE/
4. TALKS/EU/MAINLAND/
5. INVESTMENT/SOFTBANK/INDEX/
6. WORLD TRADE ORGANISATION/
7. MANAGING DIRECTOR/CHEUNG KONG/JOINT VENTURE/
8. BNP SECURITIES|SHORT-SELLING|
9. ANALYSTS|JAPAN'S|

Figure 7: System-generated key issue

CHAPTER SIX

6. DIVERGE PROCESS

As illustrated above, the converge process has identified a set of concept terms from the documents. There terms would be converted stepwise into a complete summary which covers more detailed information. The conversion process would simulate the "copy and paste" method of human summarization. Each of the concept terms is represented by a word or a word phrase, and is contained in an anchor sentence extracted from the original documents.

6.1 Concept Terms Refinement

The diverge process is based on the concept terms to generate a summary. Since the issue generated by the system may not accomplish their requirement or interest, the user is able to refine the concept terms in the issue optionally, or they may skip that refinement process. The user can then add, modify, delete or combine the concept terms before the sentence selection. The user-edited concept terms would be used in the next step. If an issue is meaningless, or the user is not interested in it, the user can get rid of this issue – user can delete the issue from the textarea in the system interface, and then sentence selection process would not consider this issue. However, the idea of the deleted issue would be missing in the summary. In case there are two issue

simultaneously representing the same event, the user can merge them into a single one. The refinement process can make the summary more concise.

6.2 Sentence Selection

Several anchor documents would be found for each key issue. An anchor document is the most representing document that includes the majority of relevant information about the key issue. An anchor sentence, which is the most representing sentence, would be extracted from each anchor document and thus closely related to the key issue. The anchor sentences of different key issues would be combined to form the summary for the document sets.

To identify an anchor document, we have to calculate the score for each document. The document score would be determined by the sentence score, which is based on the number of occurrence of the concept terms in the sentence and the location of the sentence in the document. Edmundson (1969) has defined the Location Method as follows:

> *"... ... the machine-readable cues are certain general characteristics of the corpus provided by the skeletons of documents, i.e. headings and format. The location method is based on the hypothesis that: (1) sentences occurring under certain heading are positively relevant; and (2) topic sentences tend to occur very early or very later in a document and its paragraphs."*

In our system, there are two scoring schemes for the location method. (1) If the sentence is located at the beginning of the document, the sentence would obtain a higher score. (2) If the sentence is located at both the beginning and the end of the document, the sentence would obtain a higher score. The first scheme is more suitable for news articles, while the second for documents in different styles.

Top-down Approach:

$$SentenceScore_{ij} = \frac{1}{\sqrt{s}} \times \frac{1}{p} \tag{15}$$

Head-tail Approach:

$$SentenceScore_{ij} = \begin{cases} \frac{1}{\sqrt{s}} \times \frac{1}{n-p+1} & \text{otherwise} \\ \frac{1}{\sqrt{s}} \times \frac{1}{p} & \text{if } p \le \frac{n}{2} \end{cases} \tag{16}$$

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 is calculated as the square root of the sum of all sentence scores in a document.

$$DocumentScroe = \sqrt{\sum SentenceScore_{ij}} \tag{17}$$

This score reflects the degree of relevance between a document and the concept space. After ranking the documents based on the document score for each concept term, we would extract the most important sentences from the most relevant documents to form a summary, an example of system-generated summary is shown in Figure 8. Consequently, each issue has one or more anchor sentences, and then combining these sentences produced a summary.

In this system, users can adjust the number of maximum anchor sentences extracted from the document set. If the user wants to have a more detailed summary, he can increase the number of maximum anchor sentence. In contrary, he can decrease the number to obtain a more concise summary. However, there is a trade-off between precision and recall. If the length of summary increases, recall is also increased, but precision may be sacrificed (Jing et al, 1998).

From the converge process to the diverge process, some information may be lost. In the converge process, some ideas that are not conveyed in the set of concept terms could not be retrieved in the diverge process and hence they are lost.

In order to increase the quality of the summary, we try to enlarge the information coverage. Information coverage that means how many topic of the source document has been mentioned in the summary. Because of improving the coverage of the summary, in the sentence selection process, anchor sentence would not duplicate in the summary. No summary will contain duplicated sentence which does not carry out extra information, and the summary is inconsistency while two same sentences included in it.

Hutchison ties up US e-commerce firm. Blue chips such as Hutchison Whampoa could be a more viable bet than pure Internet or high-technology counters, institutional research vice-president Ng Kong Yong said yesterday after releasing his report, Winners and Losers in the New Technology Era. Hutchison Whampoa has concluded the latest in a string of information-technology investments by forming an alliance with US-based on-line shopping firm Priceline.com.

Let's have a little of the flavour of Singapore Telecom now that we are looking at the prospect of the Singapore Government becoming the largest shareholder in our biggest telecommunications operator. Lawmakers have voiced fears that Singapore would control Hong Kong telecommunications following a merger of Cable and Wireless HKT and Singapore Telecommunications. Cable & Wireless (C&W) is likely to bring in an international telecommunications company as a strategic partner in the proposed merger of its Hong Kong subsidiary with Singapore Telecommunications (SingTel), according to sources.

Trade negotiators from Beijing and Brussels have failed to agree on final terms for mainland entry into the World Trade Organisation, setting the stage for a further round of negotiations next month. Sino-Europe WTO talks at key stage. Neither can be judged enthusiastic responses to news this week that the two largely island-bound telecommunications' providers were in talks aimed at merging their operations as a forerunner to expanding more aggressively into the region.

Softbank uses Cheung Wah for China push. Hong Kong blue chips rebounded 2.15 per cent yesterday as renewed investment interest in recent favourites Hutchison and China Telecom brought an end to several days of lacklustre trade. The Hang Seng Index ended 59.14 points higher at 15,167.55.

The Hong Kong General Chamber of Commerce (HKGCC) yesterday warned that Hong Kong companies, being smaller, may be squeezed out by multinationals once China becomes a member of the World Trade Organisation. Trade negotiators from Beijing and Brussels have failed to agree on final terms for mainland entry into the World Trade Organisation, setting the stage for a further round of negotiations next month. "Other advantages would be the expansion of a regional presence for each of the companies, and in particular the ability to compete with global players as the China market is progressively opened under agreements reached for its entry into the World Trade Organisation," she said.

... ...

Figure 8: System-generated Summary

6.3 Backward Searching

Our system provides a "backward search" function. It allows the users to retrieve the contents of the original documents. Each of the extracted sentence like a button,

which the user clicks the sentence, then it can retrieve the paragraphs which contained this sentence. When you are interested in the content behind a sentence, you can click the sentence, and the system will perform a "backward search" to retrieve the paragraphs or documents that contain the concept terms in this anchor sentence.

Therefore, after going through the summary, the user can decide to query in more detailed about the concept terms that deemed important. Actually this is an essential function. As summary just helps the user to get a overall picture of the source document in a short time. After completed reading the summary, the user will decide which ideas are important and interested for them. While the user discover their target – important and interest idea, they will process a pull action that asks for more detail information. Then, the summarization process acts as the push action.

CHAPTER SEVEN

7. EXPERIMENT AND RESEARCH FINDINGS

In early days, evaluation of summarization mainly relied on human-generated summary. That supposed to be an ideal summary for benchmarking and comparing with that generated by the system. However, it is difficult to create an ideal summary. There is evidence of low agreement among humans of which sentences are good or suitable to be included in a summary. In the experiment conducted by Rath et. Al. (1961), on average, only fifty five (55) percent of sentences been selected in two consecutive trials of summary generation. They also compared the contents and styles of system-generated abstract with that generated by human. The result indicated that the human-selected sentences and the system-selected sentences differ significantly. For human generated summaries, the background of the subjects such as education and training greatly affected the outcome, resulted in wide variety of styles and sentences selection. In contrast, system generated summaries are more consistent.

Evidently, human and system used different approaches in selecting the "representative" sentences. Therefore, the comparison between the sentences selected by the human and those selected by the system is insufficient for evaluating the quality of the system-generated summary. Rather, such approach should only be considered part of the whole evaluation of summarization system.

Recently, many researchers used precision and recall to evaluate summarization systems, for example, Miike et al (1994), Hovy and Lin (1997), Mani and Bloedorn (1997), and Jing et al (1998). Precision and recall are two key measurements to the quality of information retrieval. Which provide good measure to the relevance between system-generated summary and human-generated summary.

In order to achieve more reliable results, our evaluation is divided into three comparisons: 1. System-generated summary to the source documents; 2. System-generated summary to the human-generated summary; and 3. System-generated summary to another system-generated summary. The first one tests the coverage of the system-generated summary with regard to the source documents. Which includes some measurements, such as, information loss, compression ratio, and the number of concept terms extracted. The second part compares the system-generated summary to the human-generated one. Subjects were invited to generate a set of word phrases to represent the key issues or ideas of the source documents, and be used as the benchmark to make comparison. The last subjects would evaluate different system-generated summaries by testing their level of acceptance by human experts.

To support our experiments, on-line news articles collected from the websites of *South China Morning Post* and *Hong Kong Standard,* which are two major English on-line news websites in Hong Kong. These news articles covered different areas, such as financial, politics, sports, etc. These articles were captured within six months from our experiment and some articles reported same events.

Figure 9 shows the graphical interface of our summarization system. Before starting the summarization process, users need to input some parameters. "Document

Frequency" is the threshold of term frequency, which affects the processing time. Approximately, the processing time increases exponentially as "Document Frequency" increases. "Maximum sentences per issue" is the maximum number of sentence used to form a paragraph for related issues. It affects the length of the system-generated summary. If a user would like to get a more detailed summary, just increase this number. "Co-Occurrence Weight" determines how easily key terms can be grouped together to form concept spaces.

In addition, there are two "Retrieval Schemes" for chosen: "Top-Down" approach is more suitable for newspaper articles or home pages, in which key sentences are always appeared at the beginning, such as, headlines. "Head & Tail" is more suitable for longer or more formal articles where introduction and conclusion appeared.

Also there are three functionality provided to the users: 1. Show the merged documents; 2. Generates concept spaces or key issues; and 3. Develops a text summary in point-form and passage-form.

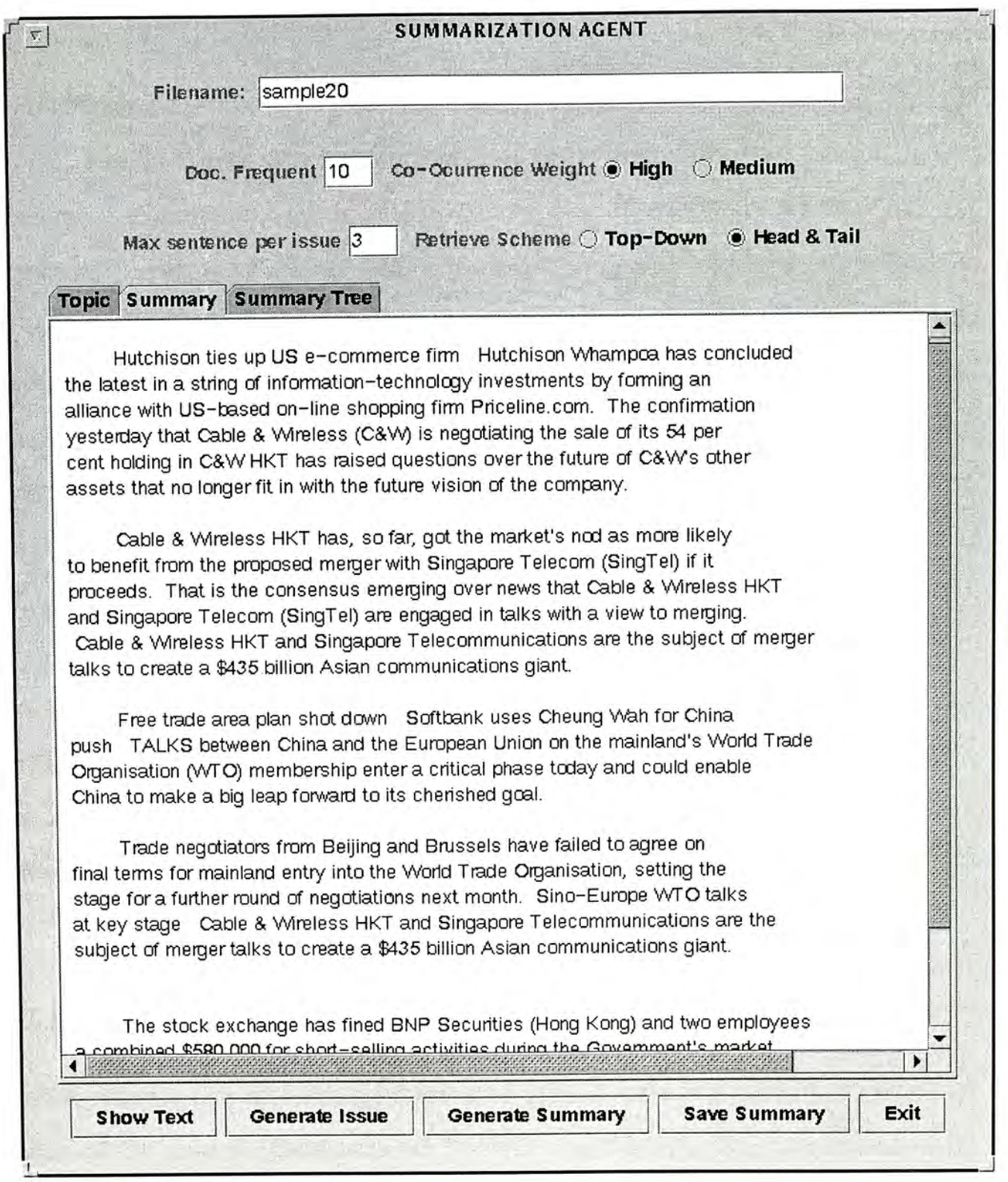


Figure 9: System GUI

7.1 System-generated Summary v.s. Source Documents

Summarization process generates a summary from the contents of a set of documents to users. Which helps them quickly and more easily grasp the key ideas or important insights. In addition, users are able to determine which documents are more relevant or more important that need further reading. In general, detailed information or background about those captured ideas are buried in the original documents, with such system, users are able to select the documents to read in order to get a more complete picture. Also, the importance of a key idea is determined by the scope of events that it associates. Therefore, such system is able to connect different documents together and allows users to create maps or trees of related events. These are major objectives of building the system.

In part one of the experiment, we investigated two factors: How compression rate affects the length of summary and its reading time as well as recall and information loss. Information loss and recall determine how much information (paragraphs) are lost after converging process.

7.1.1 Compression Ratio

Early research indicated that the summary length should not depend on the length or size of the original documents (Goldstein et al, 1999). The compression ratio decreases when the size of documents increases. Summary should be maintained in a reasonable length, which achieves the same objective as the executive summary of a report or abstract of a scientific paper. Therefore, the length of a summary of a one-hundred-page report should be the same as that of a ten-page article. Our system was

designed allowed the users select either a one-page or two-page summary to be generated from source documents. In our experiment, system generated summary from different sizes of source documents, started from 20 articles and scaled up to 200 articles. For each sample size, the output summary was bounded to two-page. As users be able to capture key ideas and their background information easily and quickly, that created significant advantages for them. The compression ratios of different size of documents are shown in Figure 10.

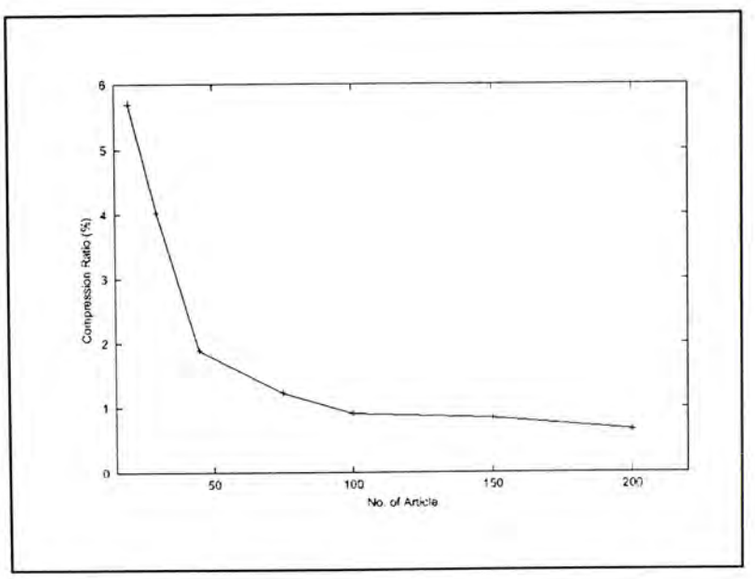


Figure 10: Compression Ratio

$$Compression\ \ Ratio = \frac{Words\ of\ Summary}{Words\ of\ Source\ Document} \tag{15}$$

The compress ratio is the number of words in summary generated by system divided by the number of words in source documents. The size of source documents increases, then the compression rate decreases.

7.1.2 Information Loss

Figure 11 shows the number of concept spaces covered in summary. Number of concept terms used in each summary is shown in Figure 12. Note that the number of concept spaces increased when the sample size increased. Besides, as the number of concept spaces increased, the number of total concept terms also increased. Figure 13 indicates that, on average, each concept space contains 2.2 concept terms.

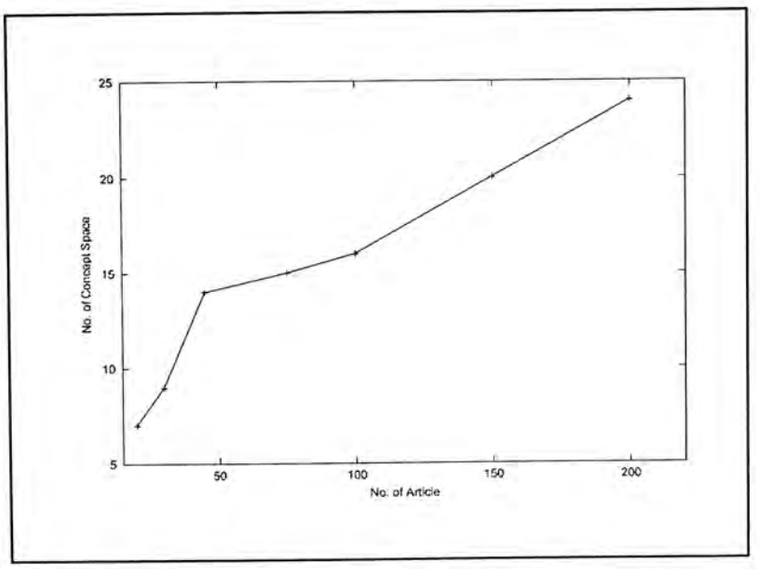


Figure 11: The number of concept space used in each summary

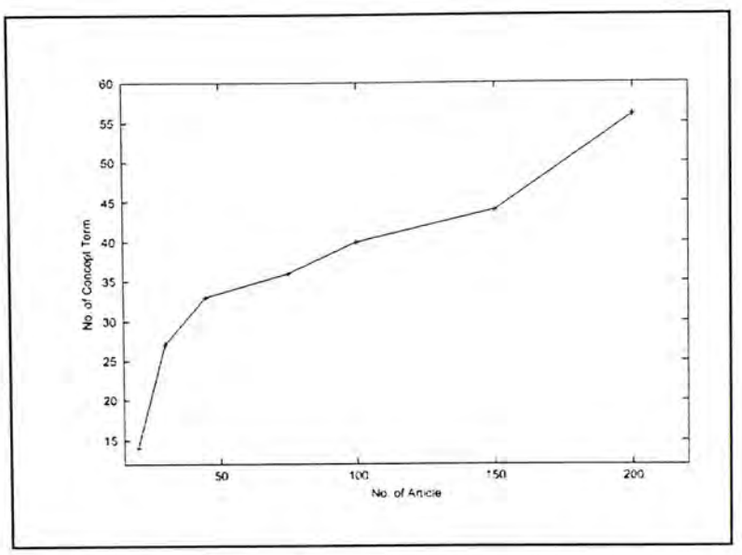


Figure 12: The number of concept term used in each summary

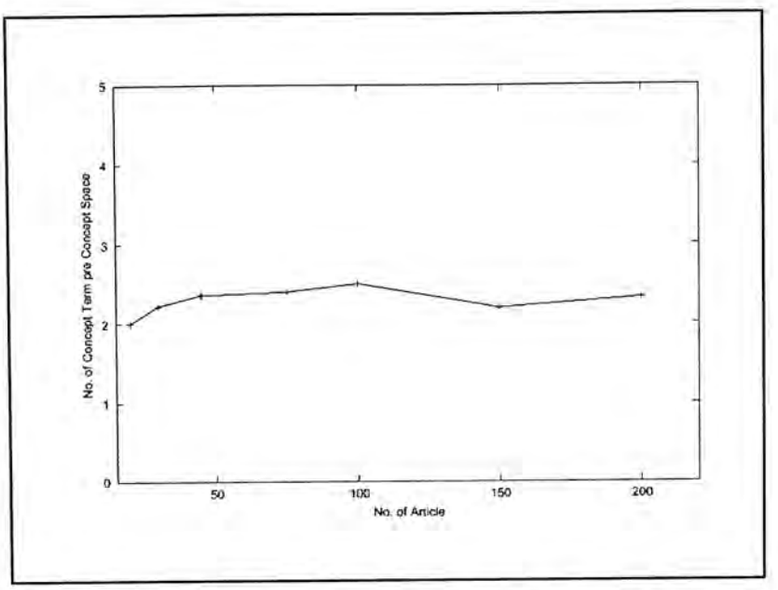


Figure 13: The number of concept term used in each concept space

This system provides a "backward search" function that allows users to retrieve related contents from the original documents according to the concept terms in each sentence been clicked. Users can retrieve the original paragraphs to get more detailed information by the "backward search" function. However, it only retrieves the paragraphs which contain at least one concept term appeared in the clicked sentence. A paragraph is considered "lost" if none of the concept terms appear in that paragraph. Therefore, that paragraph is no longer be able to be retrieved. We defined the information loss as the number of lost paragraphs divided by the total paragraphs. Figure 14 shows the information loss with respect to different sizes of source documents. In Figure 14, the minimum information loss is around twenty percent (20%), that means, twenty percent of paragraphs are unable to be retrieved.

As the sample size increases, the difficulty of converging all information into limited concept terms increases. The information loss would become serious when the size of

source documents reaches a certain level. In order to reduce the information loss, the system has to generate more concept terms/spaces and allow the summary to include most of them. Through such adjustment, the information loss can be limited to square root of the sample size. A linear increase is unacceptable to the users.

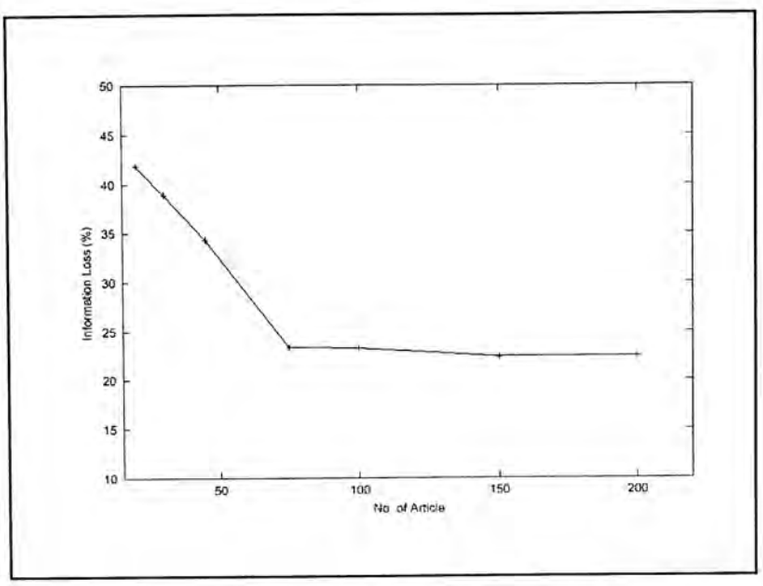


Figure 14: The information loss against number of article

Compare with the other techniques used in generating summary, our contribution is in the use of concept spaces and concept terms to determine the sentences to be included in summary. It is more flexible and domain-independent, which provides promising performance when handling large size of documents. In Figure 15, the information loss decreases as the number of concept spaces increases. Figure 16 shows the information loss against the number of concept terms used.

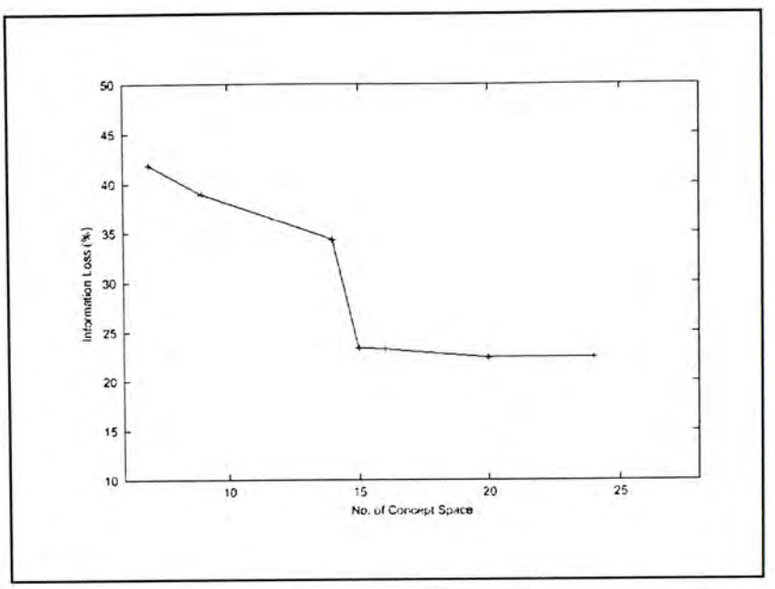


Figure 15: The information loss against concept space

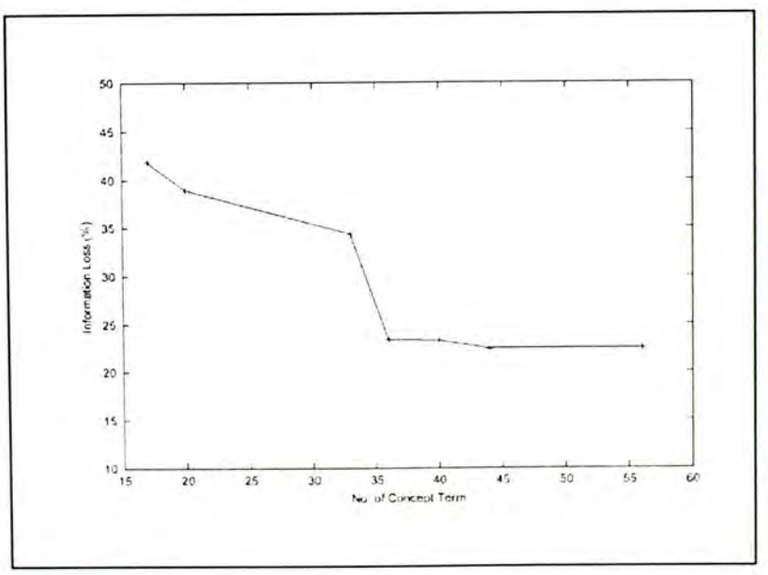


Figure 16: The information loss against concept term

Yet, it is impossible to infinitely increase the number of concept spaces for the following reasons. First, when the number of concept spaces increased, the size of summary also increased. Our aim is to keep the summary in one or two pages, which is most suitable for users, such as, executives who seldom read documents larger than

two pages. Secondly, if we increase the number of concept spaces by accepting those that ranked lower or with lower significance, the quality of the summary would be degraded.

Figure 17 shows the information loss against the number of concept spaces based on the same sample size - 100 news articles. The information loss almost reached the lower bound when system used more than 15 concept spaces. Therefore, the optimal number of concept space is around 15 based on this particular sample size.

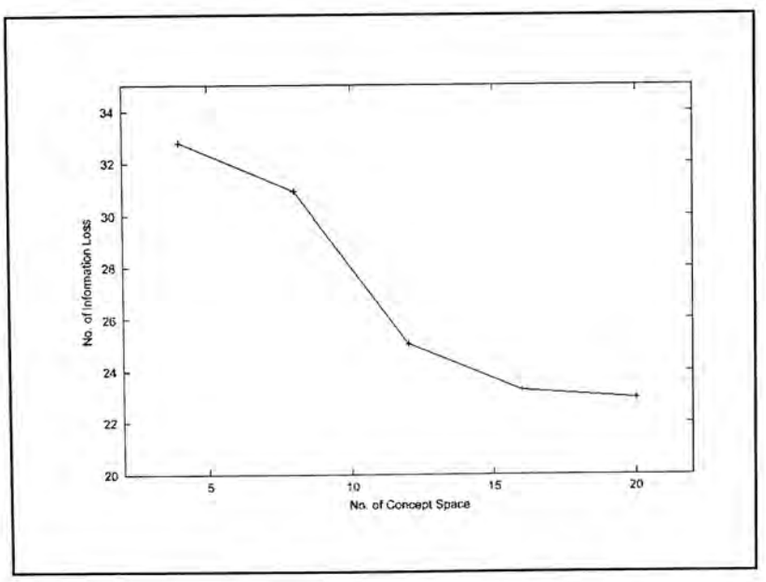


Figure 17: The information loss of sample size 100

7.2 System-generated Summary v.s. Human-generated Summary

To make the evaluation more comprehensive and convincing, we compared the performance of our system with another automated summarization system. Human-generated summaries acted as the benchmark for such comparison. Since summarization includes both converge process and diverge process, we divided the

system evaluation into two parts: Part 2 and Part 3. Part 2 compared the concept terms that generated by the system with those selected by users. Part 3 compares the contents of summaries that generated by our system and that generated by EXTRACTOR.

7.2.1 Background of EXTRACTOR

EXTRACTOR generated a requested number of key phrases, from three to thirty depends on user's selection - the default value is seven, from one document. The number of key phrases actually generated may be slightly different from the requested number, which depends on the length of the input document. It extracts sentences from input document to form summary according to the generated key phrases. A "key phrase" for EXTRACTOR equals "word phrase" defined in our system, both refer to two or more words linked together as a phrase.

In many information retrieval and summarization systems, the processes involve finding particular information that specified by users and that is called a 'query'. However, both EXTRACTOR and our system, the word phrases representing the key issues of the document, which are produced without any specification in advance. The output of the word phrases is based on the type of factual and prosaic ideas of input document rather than restricted by a query.

The prior version of EXTRACTOR used the C4.5 Decision Tree Induction Algorithm (Quinlan, 1993) to identify key phrases. In the latest version, it combined the Genitor Genetic Algorithm (Whitley, 1989) and the original key phrase extraction algorithm resulted in the GenEx (Genitor plus Extractor) algorithm. The use of the Genitor

Genetic Algorithm is to maximize the performance (fitness) on training the data and tuning twelve parameters of EXTRACTOR. These twelve parameters are listed in Table 2. According to the experiments done by Turney (2000), performance of EXTRACTOR after being tuned by the genetic algorithm, has been proved better than the original with C4.5 alone. The precision of EXTRACTOR has also been improved after implementing the genetic algorithm.

There are ten steps to the EXTRACTOR algorithm (Turney, 2000). These steps and the system flow of EXTRACTOR can be found in Appendix A.

Table 2: The twelve parameters of EXTRACTOR

Parameter Number	Parameter Name	Parameter Type	Parameter Range
1	NUM_PHRASES	Integer	[3, 30]
2	NUM_WORKING	Integer	[15, 150]
3	FACTOR_TWO_ONE	Real	[1, 3]
4	FACTOR_THREE_ONE	Real	[1, 15]
5	MIN_LENGTH_LOW_RANK	Real	[0.3, 3.0]
6	MIN_RANK_LOW_LENGTH	Integer	[1, 20]
7	FIRST_LOW_THRESH	Integer	[1, 1000]
8	FIRST_HIGH_THRESH	Integer	[1, 4000]
9	FIRST_LOW_FACTOR	Real	[1, 15]
10	FIRST_HIGH_FACTOR	Real	[0.01, 1.0]
11	STEM_LENGTH	Integer	[1, 10]
12	SUPPRESS_PROPER	Boolean	[0, 1]

7.2.2 Evaluation Method

Ten subjects were invited to generate word phrases from the testing sample. We then compared their results with the system-generated word phrases from both our system and EXTRACTOR. These subjects are under-graduate and post-graduate students. We used the same testing sample for our system and EXTRACTOR to generate lists of word phrases. The sample contained twenty articles and the total number of words in the document set was 10,862. Therefore, the average number of words per article was 543.1. Subjects were asked to generate key words or key phrases after they read all the sample documents.

The comparison was based on three factors: recall, precision, and F-measure. In the matching process, we defined a human-generated word phrase matches a machine-generated word phrase if both correspond to the same sequence of stems. A stem is the remaining part of a word when its suffix is removed. Such principal is also applicable to the short form of a word and it is case insensitive. For example, "stock" would match "stocks", "Hong Kong" would match "HK", "SingTel" would match "Singapore Telecommunication", but "Hong" would not match "Hong Kong".

The above three measures were used to evaluate how comprehensive the set of concept terms covered the ideas in the news articles. The experimenters carefully reviewed and compared all the lists, details of the experiment results are attached in Appendix E.

$$Recall = categories\ found\ and\ correct\ /\ total\ categories\ correct \quad (16)$$

$$Precision = categories\ found\ and\ correct\ /\ total\ categories\ found \quad (17)$$

$$F\text{-}measure = 2*R*P\ /\ (R + P) \quad (18)$$

In the experiment, the average number of key phrases identified by the subjects was 14.3. For the EXTRACTOR system, as it would not automatically determine the number of key phrases to be generated, we had to input the number. For a fair comparison, we assumed the number of key phrases generated from EXTRACTOR be equal to the average number of word phrases identified by subjects. Therefore, after rounding up to an integer, we entered fifteen (15). Our system generated seventeen word phrases, which was very close to the average number of word phrases (14.3) identified by subjects. Average word phrase identified by users was:

$$(11+14+12+14+20+15) / 6 = 14.3$$

Table 3: Statistic Result

	Precision	***Recall***	***F-measure***
EXTRACTOR	0.2067	0.2280	0.2168
System	0.3294	0.4150	0.3673

The experiment results indicated that the performance of our system was better than that of EXTRACTOR. Table 3 indicates that the precision, the recall, and the F-measure of our system are higher than those of EXTRACTOR.

7.3 Evaluation of different System-generated Summaries by Human Experts

This part included two evaluations, that tested the expert acceptance of the system-generated summary, especially, whether it assists human experts in their work or not. The summarization technique ultimately aims to increase the working efficiency of human. The first one asked each subject to grade three summaries generated by three different summarization systems, which are: our system, EXTRACTOR, and Microsoft's Word 2000. The second evaluation focused on the diverge process. Subjects were asked to assign a grade to the extracted sentences of our system for each performance attribute: *Relevance*, *Usefulness*, and *Representative*.

In the first evaluation, we added one more existing system in the comparison. Microsoft's Word 2000 provides an AutoSummarize feature, which identifies important sentences in the document being edited. The sentences identified would be either highlighted in the original document or extracted to form an individual paragraph. Users are allowed to specify the compression ratio of the summary from the original document.

We evaluated the following three features of the summary by a questionnaire: readability, coverage, as well as time-effectiveness and helpfulness to the users. A sample of the questionnaire is included in Appendix G. The first two questions evaluated readability, questions 3 and 4 evaluated coverage, and the last question evaluated the last feature - time-effectiveness and helpfulness to the users. The evaluation results are shown in Table 4 and the source data are included in the Appendix H.

The results indicated that our system received the highest score on four questions and the highest average index as well. Such results indicate that our system is generally considered to be better than the other two by the users. The summary generated by our system is the most informative regarding to the coverage of the original contents. It helps users to save time and effort in understanding the contents. The result of question 2 indicated that the summary generated by MS Word 2000 is the most concise among all for it is in point-form and contains fewer words. The length of the summary generated by MS Word 2000 is smaller than that of the summary from our system by 32.7 percent. The detailed data are shown in Table 5.

Table 4: Average grading of the questionnaire

	MS Word 2000	***Extractor***	***System***
Q1	2.538	3.462	3.923
Q2	3.538	2.923	3.077
Q3	2.308	3.538	3.846
Q4	2.615	3.154	3.615
Q5	3.308	3.231	3.692
Average Index	2.862	3.262	3.631

Table 5: Word count of the three summaries

	MS Word 2000	***Extractor***	***System***
Number of Words	327	527	486

As the summarization process of our system is divided into two sub-processes, both of the processes take an important role, the output of these sub-processes have been tested in our evaluations. So far, we have experimented on converge process, the evaluation mainly tests the accurate and comprehensiveness of the system-generated concept terms using precision and recall measurement. Therefore, this part tests the diverge process. In the evaluation, subjects were asked to assign a grade to the summary generated by the system for each performance attribute: *Relevance*, *Usefulness*, and *Representative*. Subjects need to grade each extracted sentence for *Relevance* and *Usefulness,* and gives a *Representative* grading for each key issue. A Likert scale of 5 was used for the subjects to choose. For example, relevance has the choices: Most Relevant (5), Moderate Relevant (4), Little Relevance (3), Too General (2), and Not Relevant (1). The number in the bracket is the score for each choice. There are two sample sets, one contained twenty articles, and the other sample set contained thirty articles. The evaluation form is included in Appendix I.

Intuitively, *Relevance* asks whether a particular sentence has given something about a key issue. *Usefulness* measures how useful the information contained in the sentence is helping the user to understand a key issue. *Representative* is the degree of the sentence to represent all the important ideas about a particular issue.

The results of these three criteria are very close as shown in Table 6. The results are promising, as the average is around the level of Moderate Relevant.

Table 6: Average score of Relevance, Usefulness, and Representative

Relevance	Usefulness	Representative
3.64	3.61	3.70

In addition, Table 7 shows the scores according to the ranked anchor sentences. Our system ranked the anchor sentences according to their scores obtained by the equations described in Section 4. Intuitively, ranked high anchor sentences should have better results in the evaluation, this is supported by the result: the higher the sentence ranked, the more it can be considered relevant. However, the result of the usefulness is reverse, maybe as the sentence selection scheme in the diverge process aim at extract the sentences which are most related to particular key issue, so anchor sentence score does not direct proportion to usefulness.

According to the sentence selection algorithm, sentences located at the beginning of a paragraph or document gain higher scores. Hence, they are more likely to be picked by the system as the anchor sentences. Consequently, titles or introductory sentences are usually selected. They often carry some key concepts of the contents of the documents. So certainly they are more relevant to the main ideas than other sentences in the documents. The sentences after the title or introduction elaborate the idea of the title. As a result the second and third anchor sentences look useful for the user to understand the ideas and hence the result came as mentioned. Based on intuition, usefulness should correlate relevance, but our experiment did not support such intuition and we will have a more complete study in the future.

Table 7: Relevance and Usefulness scores for ranked sentences

Rank	Relevance	Usefulness
1	3.70	3.57
2	3.64	3.58
3	3.54	3.70

CHAPTER EIGHT

8. CONCLUSIONS AND FUTURE RESEARCH

8.1 Conclusions

In this thesis, we used cluster analysis to group concept terms into concept spaces as the core to generate summary from multiple documents, which should be more flexible than other approaches, such as statistical approaches that based on key words or location. The results of evaluations indicated that such approach has good performance in capturing insights of the documents.

Through concept spaces, system connected related documents together and used concept terms to represent their major ideas. Then, a set of sentences were selected from the source documents based on the scores that determined by the combination of both term frequency and positions to generate a one- or two-page summary. In addition, users are able to retrieve related contents by clicking each sentence from the summary.

We have conducted some experiments to evaluate the proposed system. First, we tested the changing of compression ratio with different document size, and studied the relationship of information loss with the number of concept terms used. Then, we compared our system with an existing automated summarization system. Subjects

would generate sets of word phrases act as the benchmark, and we used recall and precision measurement for evaluating the performance. At last, subjects would evaluate different system-generated summaries.

In the first part of user evaluation, subjects needed to evaluate three automated generated summaries, one was generated by our system, the others were generated by EXTRACTOR software and MS Word 2000. That aimed to test the expert acceptance of the system-generated summary. In second part, subjects graded each sentence in the summary, as it tried to test the relevance, usefulness, and representativeness of summary that generated by our system.

The results of the experiments indicated that our approach could extract essential insights for the users, which significantly saved the time and effort for them to quickly grasp the contents of the documents. The statistical data supported our system is better than other automated summarization systems which tested in the experiments.

8.2 Future Work

However, there is still limitation on the summarization process and several research issues need to be further explored.

In the diverge process, the system selected the sentences based on the term frequency and position. The extracted sentences would not been modified. In the human-quality summarization process, summarizers usually modify or restructure those sentences. Thus, the summary that generated by the system may not be as flexible, smooth, or coherent as that generated by human being. In future work, it should use NLP

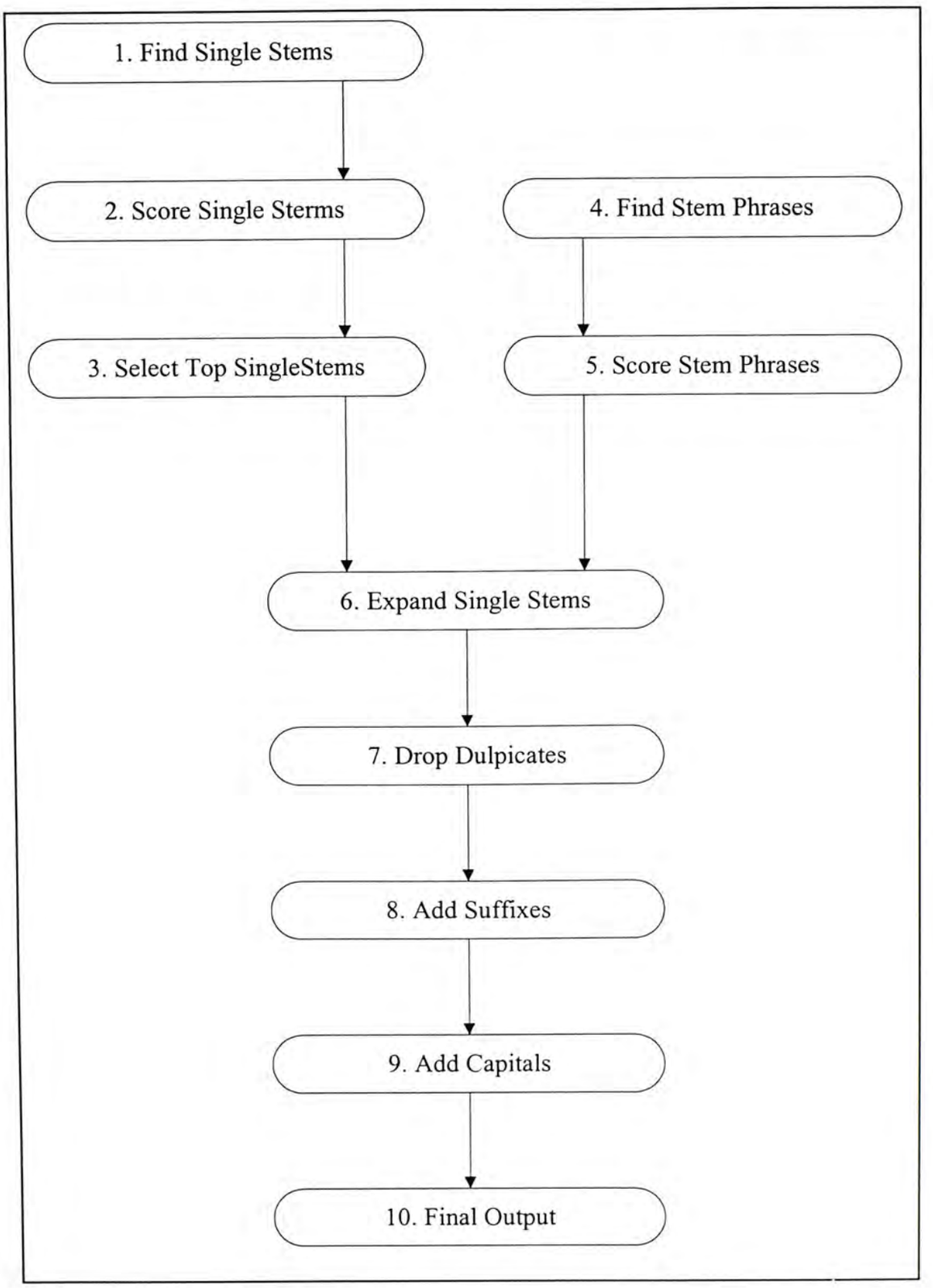
technique to let the summary be more fluent. As a paragraph of summary contained several sentences extracted from content, in order to generate a coherent text, we can use conjunction skill to combine separate facts into a single paragraph.

Information loss was unavoidable even the most flexible concept space technique was selected. It needs further exploring how to improve the coverage but not sacrifice the quality of summary. The converge and diverge processes could be further fine-tuned to minimize the information loss, such that limited the information loss to the square-root of the sample size. Besides, we could modify the sentence selection formula to increase the performance, and consider more factors in the selection process, such as the usefulness for user.

Our work has successfully used the concept space technique on the summarization process for domain-independent multiple documents for English. Then, we could apply the summarization technique for oriented language, e.g. Chinese. Base on the architecture of summarization system, it is able to summarize Chinese content by just replacing the pre-processing and segmentation modules. The equations in converge and diverge processes do not need to change. Our next step is to build a multi-language text summarizer, which can handle English and Chinese contents simultaneously

APPENDIX

A. EXTRACTOR System Flow and Ten-Step Procedure



Step 1: Find Single Stems

Make a list of all the words in the input text. Drop words with less than three characters. And drop stop words, using a given stop word list.

Step 2: Score Single Stems

For each unique stem, count down how the stem appears in the text and note when it first appears. The score is the number of times the stem appears in text, multiplied by a factor. If the stem first appears before FIRST_LOW_THRESH, then multiply the frequency by FIRST_LOW_FACTOR. If the stem first appears after FIRST_HIGH_THRESH, then multiply the frequency by FIRST_HIGH_FACTOR.

Step 3: Select Top Single Stems

Rank the stems in order of decreasing score and make a list of the top NUM_WORKING single stems. Cutting the list at NUM_WORKING, as opposed to allowing the list to have an arbitrary length, improved the efficiency of EXTRACTOR. It also acts as a filter for eliminating lower quality stems.

Step 4: Find Stem Phrases

Make a list of all phrased in the input text. A phrase is defined as a sequence of one, two, three words that appear consecutively in the text.

Step 5: Score Stem Phrases

For each stem phrase, count how often the stem phrase appears in the text and note when it first appears. Adding a score to each phrase, exactly as in step 2, using the parameters FIRST_LOW_FACTOR, FIRST_LOW_THRESH, FIRST_HIGH_FACTOR, and FIRST_HIGH_THRESH. If there is only one stem in the phrase, do nothing. If there are two stems in the phrase, multiply the score by FACTOR_TWO_ONE. If there are three stems in the phrase, multiply the score by FACTOR_THREE_ONE. Typically FACTOR_TWO_ONE and FACTOR_THREE_ONE are greater than one, since a phrase of two or three stems is necessary never more frequent than the most frequent single stem contained in the phrase, this factor can compensate for the fact that longer phrase are expected.

Step 6: Expand Single Stems

For each stem in the list of the top NUM_WORKING single stems, find the highest scoring stem phrase of one, two, or three stems that contains the given single stem. Keep this list ordered by the scores calculated in step 2. Now that the single stems have been expanded to stem phrases, we no longer need the scores that were calculated in step 5. That is, the score for a stem phrase (step 5) is now replaced by the score for its corresponding single stem (step 2).

Step 7: Drop Duplicates

The list of the top NUM_WORKING stem phrases may contain duplicates. For example, two single stems may expand to the same two-

word stem phrase. Delete duplicates from the ranked list of NUM_WORKING stem phrases, preserving the highest ranked phrase.

Step 8: Add suffixes

For each of the remaining stem phrases, find the most frequency corresponding whole phrase in the input text.

Step 9: Add Capitals

For each of the whole phrases (phrases with suffixes added), find the best capitalization. The best capitalization is, for each word phrase, find the capitalization with the least number if capitals.

Step 10: Final Output

We now have an ordered list of mixed-case (upper and lower case, if appropriate) phrases with suffix added. The list is ordered by scores calculated in step 2.

B. Summary Generated by MS Word2000

The merger would create the biggest Asian telecommunications company outside Japan and end British control of one of Hong Kong's biggest companies.
The British company owns 54.4 per cent of C&W HKT.

Lawmakers have voiced fears that Singapore would control Hong Kong telecommunications following a merger of Cable and Wireless HKT and Singapore Telecommunications.

The purchase will give Softbank a 61 per cent stake in the Hong Kong company.

Free trade area proposed

THE Hong Kong General Chamber of Commerce (HKGCC) yesterday warned that Hong Kong companies, being smaller, may be squeezed out by multinationals once China becomes a member of the World Trade Organisation.

The busines group yesterday issued its report on the impact on Hong Kong business of the mainland's accession to the world trade body.
``Hong Kong companies are relatively small in number.

THE government yesterday said concerns of a blurring in Hong Kong's status as a customs territory separate from China could make it difficult to implement a proposal by the Hong Kong General Chamber of Commerce for a free trade agreement between Hong Kong and the mainland.

Besides, she added, Hong Kong has always subscribed to multilateral trading arrangements.

The powerful Hong Kong General Chamber of Commerce in a report released on Tuesday urged the creation of a free trade arrangement between Hong Kong and China for a specific number of industries to enable Hong Kong companies to have better chances of competing with bigger multinational companies.

Hong Kong blue chips rebounded 2.15 per cent yesterday as renewed investment interest in recent favourites Hutchison and China Telecom brought an end to several days of lacklustre trade.

China Telecom bounced 3.33 per cent higher to $46.50, while Cheung Kong gained 3.7 per cent to $98.

"You can target the pure Internet companies . . .

Regulators said BNP Securities (Hong Kong) failed to immediately report a $1.5 billion short-selling transaction.

C. Summary Generated by Extractor Software

Cable & Wireless HKT has, so far, got the market's nod as more likely to benefit from the proposed merger with Singapore Telecom (SingTel) if it proceeds.

Neither can be judged enthusiastic responses to news this week that the two largely island-bound telecommunications' providers were in talks aimed at merging their operations as a forerunner to expanding more aggressively into the region.

Jardine Fleming Research advised clients the merger should result in a "significant upside" for HKT, though it maintained its six-month price target of HK$25.

Credit Lyonnaise Securities Asia flagged a "buy" on both stocks as a result of the news though it believed SingTel had the most to gain.

Hurdles ahead include who gets the top jobs, where the new entity will be located, and how to distribute ownership between major shareholders - chiefly what share the Singapore Government will walk away with and whether Deutsche Telecom will be cut into the deal.

A combined market capitalisation of about US$59 billion would be dwarfed by the giants of the industry, led by Nippon Telegraph and Telephone which is capitalised at about $239 billion.

Deep pockets are crucial in the highly capital-intensive telecoms business, and in an environment in which its sectors are being deregulated and investment opportunities are coming to the market this cash could be put to work to improve earnings rather than languishing in interest-bearing deposits.
1.20 Daniel Widdicombe, an analyst at Bear Stearns in Singapore, expects a 20 per cent fall in HKT's core earnings for the year to March 31, to 77 cents from earnings per share (eps)of 96 cents previously.

C&W HKT is facing increasing competition as the Government opens up the telecommunications industry to competition in an attempt to turn Hong Kong into a regional communications and information-technology hub.

Merging would help both companies withstand the effects of deregulation and pursue development of high-growth Internet and interactive services.

Phone, Internet and pay-TV users would get "more bang for their buck" if Cable & Wireless HKT merged with its Singapore counterpart, top analysts said yesterday.

Democrat Sin Chung-kai said the potential deal had sparked concern over telecommunications autonomy in the SAR.

Although there is uncertainty about the possibility of lay-offs, some unionists voiced concern over future management, which could wield the axe to cut costs.
These would include a regional pure cellular play, with interests in Thailand and the Philippines as well as Hong Kong and Singapore, and an Internet company, including the HKT/Star TV joint venture.

Such agreements would be in keeping with WTO rules, he said, adding that the idea had not yet been discussed with other parties, including mainland authorities.

Trade negotiators from Beijing and Brussels have failed to agree on final terms for mainland entry into the World Trade Organisation, setting the stage for a further round of negotiations next month.

Despite Beijing's decision to send a high-powered negotiating team for what was expected to be final talks on reaching a WTO accord with the European Union, the three-day meeting ended yesterday with the EU saying more needed to be done.

D. Summary Generated by Our System

Cable & Wireless HKT has, so far, got the market's nod as more likely to benefit from the proposed merger with Singapore Telecom (SingTel) if it proceeds. Cable & Wireless HKT and Singapore Telecommunications are the subject of merger talks to create a $435 billion Asian communications giant. Phone, Internet and pay-TV users would get "more bang for their buck" if Cable & Wireless HKT merged with its Singapore counterpart, top analysts said yesterday.

Hutchison ties up US e-commerce firm. Hutchison Whampoa has concluded the latest in a string of information-technology investments by forming an alliance with US-based on-line shopping firm Priceline.com. Blue chips such as Hutchison Whampoa could be a more viable bet than pure Internet or high-technology counters, institutional research vice-president Ng Kong Yong said yesterday after releasing his report, Winners and Losers in the New Technology Era.

Despite Beijing's decision to send a high-powered negotiating team for what was expected to be final talks on reaching a WTO accord with the European Union, the three-day meeting ended yesterday with the EU saying more needed to be done. Sino-Europe WTO talks at key stage. THE government yesterday said concerns of a blurring in Hong Kong's status as a customs territory separate from China could make it difficult to implement a proposal by the Hong Kong General Chamber of Commerce for a free trade agreement between Hong Kong and the mainland.

Free trade area proposed. Free trade area plan shot down.

Softbank uses Cheung Wah for China push. On paper, the team of 40 investment bankers working round the clock in Goldman Sachs' Hong Kong office to stitch up the deal have some compelling arguments in their favour. The venture will have an initial investment of US$20 million.

The stock exchange has fined BNP Securities (Hong Kong) and two employees a combined $580,000 for short-selling activities during the Government's market intervention in 1998. BNP censured for breaking regulations on short-selling. Analysts said the rise in turnover yesterday was a result of the $2.97 billion Pacific Century CyberWorks share placement on Tuesday to finance its 50-50 joint venture with CMGI, which is listed on the US Nasdaq market.

THE Hong Kong General Chamber of Commerce (HKGCC) yesterday warned that Hong Kong companies, being smaller, may be squeezed out by multinationals once China becomes a member of the World Trade Organisation. Trade negotiators from Beijing and Brussels have failed to agree on final terms for mainland entry into the World Trade Organisation, setting the stage for a further round of negotiations next month. "Other advantages would be the expansion of a regional presence for each of the companies, and in particular the ability to compete with global players as the China market is progressively opened under agreements reached for its entry into the World Trade Organisation," she said.

E. System-generated Word Phrases from Test Sample

EXTRACTOR	System
1. Internet 2. Singapore 3. HKT 4. Market 5. Merger 6. Analysts 7. Telecommunications 8. Securities 9. Hong Kong 10. Business 11. Telecom 12. Investment 13. SingTel 14. Negotiations 15. Singapore Government	1. HKT/ 2. SINGTEL/ 3. SINGAPORE GOVERNMENT/ 4. INTERNET/ 5. HUTCHISON/ 6. COMPANY/ 7. CHINA/ 8. TALKS/ 9. EU/ 10. FREE TRADE/ 11. INVESTMENT/ 12. SOFTBANK/ 13. SHARES/ 14. MARKET/ 15. SHORT-SELLING/ 16. STOCK/ 17. WORLD TRADE ORGANISATION/

F. Word Phrases Identified by Subjects

F.1 Subject One

Keyphrase found by user	*EXTRACTOR*	*System*
1. Softbank 2. Cheung Wah Development 3. HKGCC 4. China 5. WTO 6. EU 7. Hutchison 8. C&W 9. HKT 10. Merger 11. SingTel	P = 3 /15 = 0.2 R = 3 /11 = 0.2727	P = 7/17 = 0.4118 R = 7/11 = 0.6363

F.2 Subject Two

Keyphrase found by user	*EXTRACTOR*	*System*
1. Softbank 2. Cheung Wah Development 3. China 4. WTO 5. Hutchison 6. Alliance 7. Priceline.com 8. Hang Seng Index 9. Technology 10. TOM.com 11. Stock Exchange of HK 12. CWHKT 13. SingTel 14. Merger	P = 2/15 = 0.1333 R = 2/14 = 0.1429	P = 5/17 = 0.2941 R = 5/14 = 0.3571

F.3 Subject Three

Keyphrase found by user	*EXTRACTOR*	*System*
1. Backdoor listing 2. Softbank 3. WTO 4. Local economy 5. Joint venture 6. Hutchison 7. Price.com 8. HK Telecom 9. Pacific Century Works 10. Cable & Wireless HK 11. SingTel 12. Merger	P = 3/15 = 0.2 R = 3/12 = 0.25	P = 5/17 = 0.2941 R = 5/12 = 0.4167

F.4 Subject Four

Keyphrase found by user	*EXTRACTOR*	*System*
1. SoftBank 2. Cheung Wah Development 3. Beijing 4. European Union 5. World Trade Organization 6. Hutchison 7. Alliance 8. Priceline.com 9. Hang Seng Index 10. TOM.com 11. Short-selling 12. Merger 13. CWHKT 14. SingTel	P = 2/15 = 0.1333 R = 2/14 = 0.1429	P = 5/17 = 0.2941 R = 5/14 = 0.3571

F.5 Subject Five

Keyphrase found by user	*EXTRACTOR*	*System*
1. Softbank 2. Cheung wah development 3. Trade negotiators 4. Beijing 5. Brussels 6. Mainland entry 7. WTO 8. Hutchison 9. Alliance 10. On-line shopping 11. Priceline.com 12. Blue Chips 13. BNP Securities 14. Cable & Wireless HKT 15. Merging 16. Singapore Telecom 17. Hong Kong 18. Singapore 19. Future Mangagement 20. Cut cost	P = 4/15 = 0.2667 R = 4/20 = 0.2	P = 4/17 = 0.2353 R = 4/20 = 0.2

F.6 Subject Six

Keyphrase found by user	*EXTRACTOR*	*System*
1. Softbank 2. Cheung Wah Development 3. China 4. European 5. WTO 6. Hang Seng Index 7. Hutchison 8. New Technology 9. BNP Security 10. SEHK 11. C&W HKT 12. Negotiation 13. Merging 14. HK Telecommunication 15. SingTel	P = 3/15 = 0.2 R = 3/15 = 0.2	P = 7/17 = 0.4117 R = 7/15 = 0.4667

F.7 Subject Seven

Keyphrase found by user	*EXTRACTOR*	*System*
1. HSBC 2. Cheung Kong 3. Hutchison 4. Hang Seng Index 5. Internet 6. Joint Venture 7. Softbank 8. WTO 9. EU 10. Free Trade 11. Hong Kong 12. US 13. China 14. Priceline.com 15. Singapore Telecommunication 16. Cable & Wireless 17. HKT	P = 4/15 = 0.2667 R = 4/17 = 0.2353	P = 8/17 = 0.4706 R = 8/17 = 0.4706

F.8 Subject Eight

Keyphrase found by user	*EXTRACTOR*	*System*
1. Softbank 2. Hong Kong 3. Singapore 4. WTO 5. HKT 6. C&W 7. Technology 8. Hutchison 9. Finance 10. Telecom 11. Government	P = 4/15 = 0.2666 R = 4/11 = 0.3636	P = 4/17 = 0.2353 R = 4/11 = 0.3636

F.9 Subject Nine

Keyphrase found by user	*EXTRACTOR*	*System*
1. Cheung Kong 2. Cable and Wireless 3. HKT 4. SingTel 5. Hong Kong 6. Free Trade 7. WTO 8. Softbank 9. Cheung Wah Development 10. Trade negotiator 11. Beijing & Brussels 12. Hutchison 13. On-line shopping 14. BNP Securities 15. Short-selling 16. Internet	P = 4/15 = 0.2667 R = 4/16 = 0.25	P = 7/17 = 0.4118 R = 7/16 = 0.4375

F.10 Subject Ten

Keyphrase found by user	*EXTRACTOR*	*System*
1. Blue chip 2. e-commerce 3. HKGCC 4. WTO 5. Hutchison 6. Information technology 7. BNP Securities 8. HKT 9. Singapore Telecommunication	P = 2/15 = 0.1333 R = 2/9 = 0.2222	P = 4/17 = 0.2353 R = 4/9 = 0.4444

G. Sample of Questionnaire

Summarization System Project

-Department of Systems Engineering And Engineering Management

This questionnaire is to evaluate three system-generated summaries. Base on a testing sample contained 20 articles, we generated three summaries by different summarization system. Then we try to make a comparison among three summaries.

Answer the following four questions after completely read the summaries.

	Strongly Agree				Strongly Disagree
	5 ←----------				----------→ 1
Q1. The summary is readable and understandable (可以理解的)					
Summary One	5	4	3	2	1
Summary Two	5	4	3	2	1
Summary Three	5	4	3	2	1
Q2. The summary is concise (簡潔)					
Summary One	5	4	3	2	1
Summary Two	5	4	3	2	1
Summary Three	5	4	3	2	1
Q3. The summary provides a general picture (提供了大概的內容)					
Summary One	5	4	3	2	1
Summary Two	5	4	3	2	1
Summary Three	5	4	3	2	1
Q4. The summary is a valuable reference (俱參考價值)					
Summary One	5	4	3	2	1
Summary Two	5	4	3	2	1
Summary Three	5	4	3	2	1
Q5. The summary can help me to save time and effort rather than read through 20 articles (節省閱讀時間和功夫)					
Summary One	5	4	3	2	1
Summary Two	5	4	3	2	1
Summary Three	5	4	3	2	1

Thanks You

H. Result of Questionnaire

Question /summary	# 1	# 2	# 3	# 4	# 5	# 6	# 7	# 8	# 9	# 10	# 11	# 12	# 13
1.1	2	2	4	1	2	4	2	3	4	3	2	2	2
1.2	2	4	3	3	3	5	4	4	4	4	3	3	3
1.3	4	4	3	4	4	5	4	3	5	4	4	4	3
2.1	2	3	5	4	5	5	3	4	4	4	3	1	3
2.2	3	4	2	3	4	4	3	2	3	3	3	2	2
2.3	4	3	2	2	3	5	3	3	3	2	4	3	3
3.1	2	3	4	1	2	4	2	3	3	1	2	1	2
3.2	3	4	4	3	4	5	4	4	2	5	3	2	3
3.3	4	4	4	4	5	5	4	3	4	4	3	3	3
4.1	2	3	4	1	3	4	2	3	4	3	2	1	2
4.2	3	3	3	3	4	3	4	4	3	4	2	1	4
4.3	4	3	3	4	4	5	4	3	4	4	3	2	4
5.1	2	4	5	1	4	5	4	4	3	2	4	2	3
5.2	2	4	2	3	4	3	5	4	3	3	4	3	2
5.3	3	4	3	4	4	4	5	4	4	2	4	4	3

is the sample number.

	Word97	Extractor	System
Question 1	2.538	3.462	3.923
Question 2	3.538	2.923	3.077
Question 3	2.308	3.538	3.846
Question 4	2.615	3.154	3.615
Question 5	3.308	3.231	3.692
Average Score	2.862	3.262	3.631

I. Evaluation for Diverge Process

Evaluation for the Multi-Document Summarization System

Ref. No: Sample30

Name: _____

E-Mail _____

Key Issues generated by System

1. INTERNET/HUTCHISON/
2. SINGAPORE/MERGER/TELECOMMUNICATIONS/SINGTEL|GOVERNMENT|TELECOMS
3. FREE TRADE/
4. TALKS/EU/MAINLAND/
5. INVESTMENT/SOFTBANK/INDEX/
6. WORLD TRADE ORGANISATION/
7. MANAGING DIRECTOR/CHEUNG KONG/JOINT VENTURE/
8. BNP SECURITIES|SHORT-SELLING|
9. ANALYSTS|JAPAN'S|

Grading Method

Please evaluate the summary in aspects — Relevance, Usefulness, and Representative. The grading for the sentence is separated to five levels.

Relevance is whether this sentences telling something about its key topic (generated by the system). 1- Relevance, 2 – Moderate Relevance, 3 – Little Relevance, 4 – Too General, 5 – Not Relevance.

Usefulness is how useful of the information provided by this sentence for reader. 1- Usefulness, 2 – Moderate Usefulness, 3 – Little Usefulness, 4 – Too General, 5 – Not Usefulness.

Representative is the grade for the paragraph that can represent the whole idea/information of the topic. 1- Representative, 2 – Moderate Representative, 3 – Little Representative, 4 – Too General, 5 – Not Representative.

Extracted Sentence

	Contents	A	B	C
[1-1]	Hutchison ties up US e-commerce firm.			
[1-2]	Blue chips such as Hutchison Whampoa could be a more viable bet than pure Internet or high-technology counters, institutional research vice-president Ng Kong Yong said yesterday after releasing his report, Winners and Losers in the New Technology Era.			
[1-3]	Hutchison Whampoa has concluded the latest in a string of information-technology investments by forming an alliance with US-based on-line shopping firm Priceline.com.			

P.1

[2-1]	Let's have a little of the flavour of Singapore Telecom now that we are looking at the prospect of the Singapore Government becoming the largest shareholder in our biggest telecommunications operator.			
[2-2]	Lawmakers have voiced fears that Singapore would control Hong Kong telecommunications following a merger of Cable and Wireless HKT and Singapore Telecommunications.			
[2-3]	Cable & Wireless (C&W) is likely to bring in an international telecommunications company as a strategic partner in the proposed merger of its Hong Kong subsidiary with Singapore Telecommunications (SingTel), according to sources.			
[3-1]	Free trade area proposed.			
[3-2]	Free trade area plan shot down.			
[3-3]	Because of this, the value of an Internet company may be over-estimated if the company exaggerates its hit volume.			
[4-1]	Trade negotiators from Beijing and Brussels have failed to agree on final terms for mainland entry into the World Trade Organisation, setting the stage for a further round of negotiations next month.			
[4-2]	Sino-Europe WTO talks at key stage.			
[4-3]	Neither can be judged enthusiastic responses to news this week that the two largely island-bound telecommunications' providers were in talks aimed at merging their operations as a forerunner to expanding more aggressively into the region.			
[5-1]	Softbank uses Cheung Wah for China push.			
[5-2]	Hong Kong blue chips rebounded 2.15 per cent yesterday as renewed investment interest in recent favourites Hutchison and China Telecom brought an end to several days of lacklustre trade.			
[5-3]	The Hang Seng Index ended 59.14 points higher at 15,167.55.			
[6-1]	THE Hong Kong General Chamber of Commerce (HKGCC) yesterday warned that Hong Kong companies, being smaller, may be squeezed out by multinationals once China becomes a member of the World Trade Organisation.			
[6-2]	Trade negotiators from Beijing and Brussels have failed to agree on final terms for mainland entry into the World Trade Organisation, setting the stage for a further round of negotiations next month.			
[6-3]	"Other advantages would be the expansion of a regional presence for each of the companies, and in particular the ability to compete with global players as the China market is progressively opened under agreements reached for its entry into the World Trade Organisation," she said.			
[7-1]	Cheung Kong joins HSBC on Internet.			
[7-2]	Hong Kong blue chips rose 0.39 per cent yesterday on the back of the formation of a $3 billion e-commerce joint venture by four blue-chip companies but finished off their intraday highs as interest rate fears kept traders wary.			
[7-3]	Two months ago, it had a share swap with Japan's KDD Corp and said the two would set up a joint venture in April to offer full line services to multinational corporations.			
[8-1]	The stock exchange has fined BNP Securities (Hong Kong) and two employees a combined $580,000 for short-selling activities during the Government's market intervention in 1998.			
[8-2]	BNP censured for breaking regulations on short-selling.			
[9-1]	The Group of Seven leading industrialised nations might have done enough to keep the Japanese yen from surging in the near term by saying at the weekend they shared Japan's concerns about a strong yen, but their statement broke no new ground, analysts said.			
[9-2]	The lukewarm response comes despite a general vote of approval from analysts to the prospects of a merger.			
[9-3]	Analysts said bringing the Priceline model to Asia makes business sense, since there's an abundant supply of price-sensitive consumers.			

BIBLIOGRAPHY

[1] Addison, E. R. 1991. "Using news understanding and neural networks in foreign currency options trading," *Proc. Int'l Conf. Artificial Intelligence Applications*, pp. 319-323.

[2] Barzilay, R. and Elhadad, M. 1997. "Using Lexical Chains for Text Summarization," *Proc. ACL Workshop on Intelligent Scalable Text Summarization*, pp. 10-17.

[3] Barzilay, R., McKeown, K.R., and Elhadad, M. 1999. "Information Fusion in the Context of Multi-Document Summarization," Processing of the 37th Annual Meeting of the Assoc. of Computational Linguistics.

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

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

[6] Chen, H. and Lynch, K. J. 1992. "Automatic construction of networks of concepts characterizing document database," *IEEE Transactions on Systems, Man and Cybernetics,* vol.22, no.5, pp. 885-902.

[7] Chen, H., Lynch, K.J., Basu, K. and Ng, T. 1993. "Generating, integrating, and activating thesauri for concept-based document retrieval," *IEEE EXPERT, Special Series on Artificial Intelligence in Test-based Information Systems,* vol. 8, no.2, pp. 25-34.

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

[9] Chen, H. 1995. "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.

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

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

[12] Dalton, J. and Deshmane, A. 1991. "Artificial neural networks," *IEEE Potentials*, vol. 10, no.2, pp. 33-36.

[13] Edmundson, H.P. 1969. New methods in automatic extracting. Journal of the *ACM*, 16(2):264-285, April.

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

[15] Firmin,T and Chrzanowski, M. 1999. An Evaluation of Automatic Text Summarization Systems. "*Advances in Automatic Text Summarization*", edited by Inderjeet Mani and M.T. Maybury. The MIT Press, Cambridge, Massachusetts, London, England. pp.325-336.

[16] Frawley, W. J., Pietetsky-shapiro, G. and Matheus, C. J. 1991. *Knowledge discovery in database: an overview. In Knowledge Discovery in Database,* The MIT Press, Cambridge, MA, pp. 1-30.

[17] Furnas, G. W., Landauer, T. K. and Dumais, S. T. 1987. "The vocabulary problem in human-system communication," *Communication of the ACM*, 30(11): pp. 964-971.

[18] Goldstein, J., Kantrowitz, M., Mittal, V. and Carbonell, J. 1999. "Summarizing Text Documents: Sentence Selection and Evaluation Metrics," *SIGIR*, pp. 121-128.

[19] Hayes, P. J. and Weinstein, S. P. 1991. "Adding value to financial news by computer," *Proc. the First Int'l Conf. Artificial Intelligence Applications*, pp. 2-8.

[20] Holsapple, C., Tam, K. Y. and Whinston, A. 1998. "Adapting expert system technology to financial management," *Financial Management,* vol. 16, no. 3: pp. 12-22.

[21] Hopfield, J. J. 1982. Neural Network and physical systems with collective computational abilities. *Proc. National Academy of Science*, USA, vol. 78, no.8: pp. 2554-2558.

[22] Hovy, E. and Lin, C.Y. 1997. "Automated Text Summarization in SUMMARIST," *Proc. Workshop of Intelligent Scalable Text Summarization.*

[23] Jing, H., Mckeown, K., Barzilay, R. and Elhadad, M. 1998. "Summarization Evaluation Methods: Experiment and Analysis," *AAAI'98 Symposium*, Stanford Univeristy CA.

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

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

[26] Kolb, R. W. 1989. *Investment*, Foresman and Company, Glenview, Illinoise.

[27] Kupice, J.M., Pedersen, J. and Chen, F. 1995. "A Trainable Document Summarizer," *ACM SIGIR*, pp. 68-73.

[28] Lehnert, W. G. 1981. Plot Units: A Narrative Summarization Strategy, in *Cognitive Science* 4, pp. 293-331.

[29] Lehnert, W.G. 1983. "Narrative Complexity Based on Summarization Algorithms," *IJCAI*, 2(10): pp. 713-716.

[30] Lin, C.Y. and Hovy, E. 1997. "Identifying Topics by Position," *Proc. the 5th Applied Natural Language Conference on Applied Natural Language Processing*, pp. 283-290, Washington, D.C.

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

[32] Mani, I. and Bloedorn, E. 1999. "Summarizing Similarities and Differences Among Related Documents," *Information Retrieval,* vol.1 ch. 2, pp. 35-67.

[33] McKeown, K.R., Udith, J., Klavans, L., Hatzivassiloglou, V., Barzilay, R. and Eskin, E. 1999. "Towards Multidocument Summarization by Reformulation: Progress and Prospects," *AAAI.*

[34] Miike. S., Itoh, E., and Sumita, K. 1994. A Full-Text Retrieval System with a Dynamic Abstract Generation Function. In SIGIR '93, 69-77.

[35] Parsaye, K., Chignell, M., Khoshafian, S. and Wong, H. 1990. "Intelligent databases", *AI Expert*, vol. 5, no. 3, pp. 38-47. 1990.

[36] Qulinlan, J.R.1993. "*C4.5: Programs for machine learning,*" California: Morgan Kaufmann.

[37] Rath, G.J., Resnick, R., and Savage, T. R. 1961. The Formation of Abstracts By the Selection of Sentences: Part 1: sentence selection by man and machines". *American Documentation* 12 (2) pp 139-141.

[38] Ricardo, B.Y. and Berthier, R.N. 1999. *Modern Information Retrieve,* Addison Wesley Longman Limited.

[39] Salton, G. 1978. "Generation and search of clustered files," *ACM Transactions on Database Systems*, vol 3, no. 4, pp. 321-346, December.

[40] Salton, G. 1989. *Automatic Text Processing*, Addison-Wesley Publishing Company, Inc.

[41] Salton, G., Buckley, C., Singhal, A. and Mitra, M. 1997. "Automatic Text Structuring and Summarization," Information Processing and Management, 33(2): pp. 193-207, March.

[42] Sparck Jones, K. 1999. "Automatic summarizing : factors and directions" , in *Advances in Automatic Text Summarization*, edited by Mani, I. and Maybury, M.T., The MIT Press, Cambridge, Massachusetts, London, England. pp 1-14.

[43] Turney, P.D. 2000. "Learning algorithm for Keyphrase Extraction," *Information Retrieval*, 2(4), 303-336

[44] Whitley, D. 1989. "The GENITOR algorithm and selective pressure," *Proceedings of the Third International Conference on Genetic Algorithms (ICGA-89)*, pp. 116-121. Californa: Morgan Kaufmann.

[45] Wyle, M. F. 1991. "A wide area network information filter," *Proc. the First Int'l Conf. on Artificial Intelligence Applications,* Oct. 9-11, 1991, pp. 10-15.

[46] Yen, J., Ma, P.C., Sivakumar, V. and Chen, H. 1996. "A Software Agent for Analyzing Financial," IEEE Computer, 29(8): pp. 62-70, August.