EXPLOITING TEXTUAL STRUCTURES OF TECHNICAL PAPERS FOR AUTOMATIC MULTI-DOCUMENT SUMMARIZATION

ZHAN JIAMING

(B. Eng., University of Science and Technology of China)

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF MECHANICAL ENGINEERING

NATIONAL UNIVERSITY OF SINGAPORE

2008

Acknowledgements

Firstly, I am deeply grateful to my supervisor, Prof. Loh Han Tong, under whose guidance I chose this topic and began the thesis. His wide knowledge and logical way of thinking have been of great value to me. His understanding, encouraging and personal guidance have provided a good basis for this thesis. I would also like to thank the other panel members of my Ph.D. Qualifying Examination, Prof. Wong Yoke San, Prof. Ong Chong Jin and Prof. Poh Kim Leng, for their helpful and constructive comments in the initial stage of this research.

This work would not have been possible without the support and help of my senior colleagues, Dr. Rakesh Menon, Dr. Shen Lixiang and Dr. Liu Ying. Numerous fruitful discussions with them have created a lot of good ideas and have a direct impact on the final form and quality of this thesis. I would also like to appreciate Mr. Ivan Yap, for his kind help in some of the core codes in the experiments.

I cannot end without thanking my parents, on whose constant love I have relied throughout my Ph.D. study. Their love is a persistent inspiration for my journey in this life. It is to them that I dedicate this work.

Table of Contents

Summary

In today's knowledge-intensive engineering environment, information management is an important and essential activity. Existing research on engineering information management has mainly focused on structured numerical data such as computer models and process data. Textual data, such as technical papers, patent documents and customer reviews, which constitute a significant part of engineering information, have been somewhat ignored. Recently, with an explosive growth of textual information created and stored digitally, there has been an increasing demand to reduce the time in acquiring useful information from massive textual data. Automatic text summarization technology has proven to be very helpful in integrating the information from multiple documents and facilitating the process of information searching and management. Therefore, this thesis examines the challenging issues of automatically summarizing multiple technical papers.

Previous text summarization research has mainly focused on the domain of news articles. Compared to news articles, summarization of technical papers is different in terms of readers' information requirements and document genre. Existing Multi-Document Summarization methods cannot address the specialties of the technical paper domain and cannot reveal the internal textual structures of multiple papers. Therefore, it motivated the detailed investigation into the structures within multiple real-world documents and how these structures could help in Multi-Document Summarization.

Based on the analysis of the Document Understanding Conference (DUC) corpus of manual summaries, the notions of macrostructure and microstructure are proposed. These two structures are assumed to constitute important information within multiple documents that will affect the summarization performance. Macrostructure is defined as the significant topics shared among different input documents, while microstructure is defined as sentences that acted as elaborating information for macrostructure. Experimental results demonstrated that human summarizers heavily relied on the macrostructure in writing their summaries. Moreover, it was found that microstructure offered complementary information for macrostructure and both structures constituted the important information in summarization modeling and evaluation.

A multi-paper summarization framework based on macrostructure and microstructure is then proposed in this thesis. The factors in macrostructure generation were examined by ANOVA test and it was found that the topic extraction threshold and the topic ranking scheme could significantly affect the summarization performance. In the domain of technical papers, microstructure was defined as rhetorical structure within each single paper. The identification of microstructure was approached as a problem of automatically assigning rhetorical categories to every sentence in the paper document. The algorithms of Naïve Bayes and SVMs were experimented in building the rhetorical classification models, and SVMs outperformed Naïve Bayes in terms of

F-measure. The evaluation experiments showed that the summarization approach based on macrostructure and microstructure, compared with the peer systems of Copernic summarizer and clustering-summarization, could better identify the topical relationship among real-world papers and better recognize their similarities and difference.

Finally, two case studies are introduced to consolidate and extend this research in the sense of applying summarization within Engineering Information Management and text mining. One case study was to apply the proposed summarization framework in the domain of online customer reviews. The other case study examined the application of summarization to improve automatic text classification.

List of Tables

List of Figures

List of Abbreviations

AI	Artificial Intelligence
ANOVA	ANalysis Of VAriance
BLEU	Bilingual Evaluation Understudy
CAD	Computer Aided Design
CAE	Computer Aided Engineering
CAM	Computer Aided Manufacturing
CAPP	Computer Aided Process Planning
CRM	Customer Relationship Management
CST	Cross-document Structure Theory
DF	Degree of Freedom
DUC	Document Understanding Conference
EIM	Engineering Information Management
ERP	Enterprise Resource Planning
FSs	Frequent word Sequences
GMAT	Graduate Management Admission Test
HTML	HyperText Markup Language
IE	Information Extraction
IR	Information Retrieval
KDD	Knowledge Discovery in Databases
LSI	Latent Semantic Indexing
MCV1	Manufacturing Corpus Version 1

MDS	Multi-Document Summarization
MES	Manufacturing Execution System
MMI	Macrostructure- and Microstructure-level Information
MS	Mean of Squares
MUC	Message Understanding Conference
NLP	Natural Language Processing
PDM	Product Data Management
R&D	Research and Development
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
SCUs	Summarization Content Units
SS	Sum of Squares
SUMMONS	SUMMarizing Online NewS articles
SVD	Singular Value Decomposition
SVMs	Support Vector Machines
TRIZ	A Romanized acronym for Russian “Теория решения изобретательских задач” (Teoriya Resheniya Izobretatelskikh Zadatch), meaning "theory of inventive problem solving"
VSM	Vector Space Model
WWW	World Wide Web
XML	eXtensible Markup Language

Chapter 1

Introduction

Information management is an important and essential activity in today's knowledge-intensive engineering environment. Engineering information to be managed includes patent documents, design notes, computer models, process data, customer records, etc., produced in the processes of Research and Development (R&D), product design and manufacturing, e-Business and e-Commerce (Anderson and Kerr, 2001; Curtis and Cobham, 2000; Stark, 1992; Tanaka and Kishinami, 2006). Such information and data are of principal importance for engineering activities, and thus effective and efficient management of information is one of the key factors by which the industrial and engineering performance can be greatly improved (Chaffey and Wood, 2004; Hicks et al., 2006; Laudon and Laudon, 1996; Tirpack, 2000).

Existing research on Engineering Information Management (EIM) has mainly focused on the domain of numerical data (Anderson and Kerr, 2001; Stark, 2005; Tanaka and Kishinami, 2006). Textual data, such as technical papers, patent documents, e-mails and customer reviews, which constitute a significant part of engineering information, have been relatively ignored. Recently, with an explosive growth of textual information created and stored in the enterprise intranets and the World Wide Web (WWW), there has been an increasing demand of advanced techniques to reduce the time in acquiring useful information and knowledge from massive quantities of textual data.

Automatic text summarization technology has proven to be helpful in integrating the information from multiple documents and facilitating the process of information searching and management. Therefore, this thesis examines the summarization technology within an engineering domain. In particular, the challenging issues of summarizing multiple technical papers are investigated.

1.1 Information Management in Engineering Domain

Information management is the handling of information acquired from one or multiple sources in a way that optimizes access by all who have a share in that information or a right to that information (Chaffey and Wood, 2004; Curtis and Cobham, 2000). By the late 1990s, the increase in the volume of electronic data disseminated across personal computers and networks spawned the increasing need to make these data more accessible through the tools of information management.

As shown in Figure 1.1, information lies at the core of a modern engineering environment, comprising not only numerical data like computer models but also textual data such as patent documents, technical papers and customer e-mails. These data, produced and stored by the tools like computer-based systems (CAD, CAM, CAE, CAPP) and patent databases, are cycled in the engineering activities of R&D, design and production, e-Business and e-Commerce.

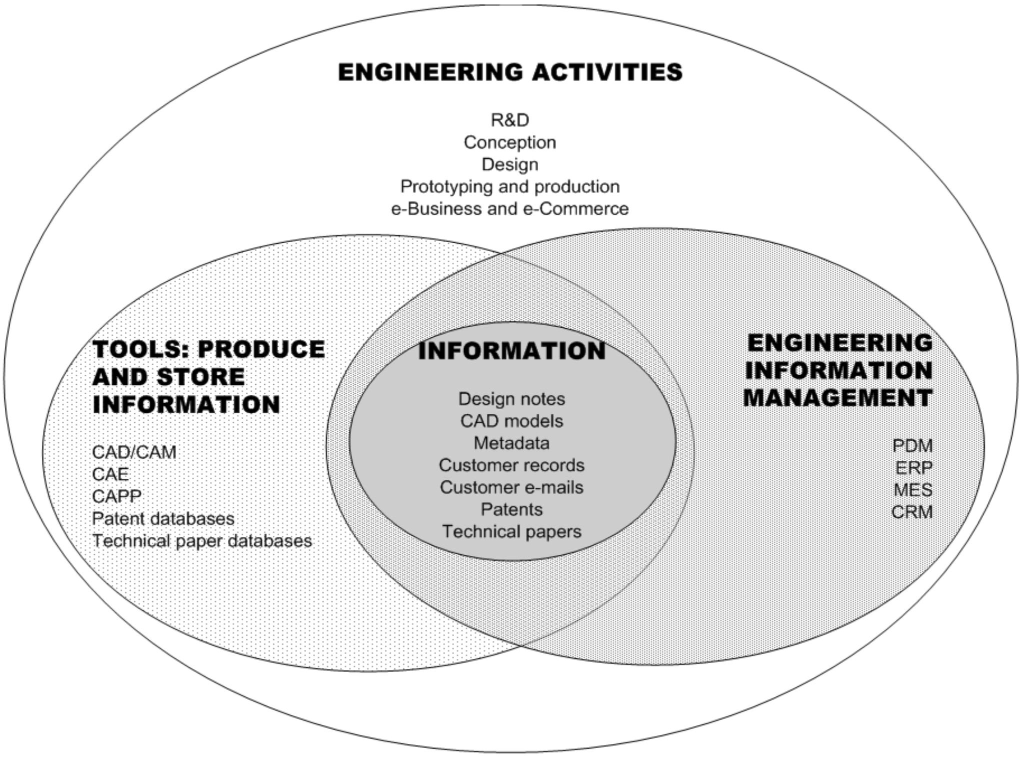


Figure 1.1 Information flow within modern engineering environment

The massive amount of data demands powerful EIM systems to help in improving the flow, quality and use of engineering information which is related to the processes of R&D, design, production and services. EIM systems should provide improved management of the engineering processes through better control of product data and configurations. Moreover, EIM systems manage the flow of work through those activities that create or use engineering information. EIM is also expected to provide support for the activities of product teams and for advanced organizational techniques such as concurrent engineering, which can help in reducing engineering costs and product development cycle.

Up to now, most EIM applications focus on allowing users to share information and on handling of numerical data. Some of them are briefly reviewed as follows.

1.1.1 Product Data Management

Product Data Management (PDM) is used to produce and handle relations among data that define a product throughout the product life cycle, from conception, through development, and production to distribution, and beyond (Leong et al., 2002; Liu and Xu, 2001; Tanaka and Kishinami, 2006). The information being stored and managed includes product data such as CAD models, drawings and their associated metadata, specifications, manufacturing and assembly plans, and test procedures. PDM enables people from all divisions to participate in different phases of the product throughout its life cycle. With the help from networks, it is possible to establish information connectivity across a world of immense geography and diverse platforms.

1.1.2 Enterprise Resource Planning

Enterprise Resource Planning (ERP) systems are designed to integrate all data and processes of an organization into a unified system and to help plan the utilization of enterprise-wide resources (Shafiei and Sundaram, 2004; Willcocks and Sykes, 2000). A key ingredient of most ERP systems is the use of a unified database to store data for the various system modules. ERP is sometimes confused with PDM. PDM is strongly rooted in the world of development and design, and therefore, it manages engineering and product design data and their relationships throughout a product life cycle, whereas ERP is a control system specifically for manufacturing and usually collaborates with Manufacturing Execution System (MES).

1.1.3 Manufacturing Execution System

A MES handles a variety of functions, all of which are connected to the flow of work in the manufacturing process. In a nutshell, MES helps manufacturing companies to manage the flow of manufacturing process, to collect and analyze data generated by and during the manufacturing process (Ake et al., 2004; Liu et al., 2006). As shown in Figure 1.2, MES bridges the gap between ERP and shop floor control systems by providing links among shop floor instrumentation, control hardware, planning and control systems, process engineering, production execution, sales force and customers.

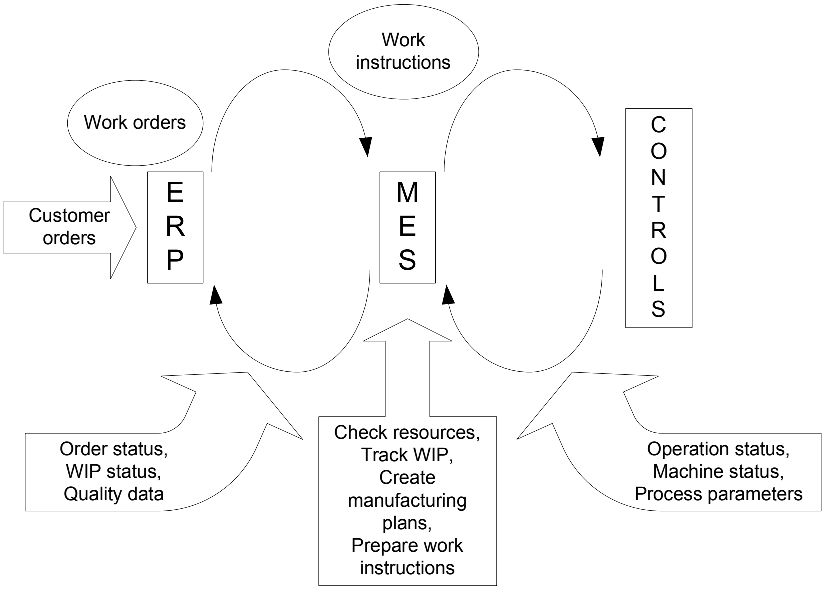

Work instructions
Work orders
Customer orders
ERP
MES
CONTROLS
Order status, WIP status, Quality data
Check resources, Track WIP, Create manufacturing plans, Prepare work instructions
Operation status, Machine status, Process parameters

Figure 1.2 A typical work flow model in a manufacturing plant

1.1.4 Customer Relationship Management

Customer Relationship Management (CRM) serves the identification of market needs.

It can be viewed as the process of constructing a detailed database of customer information and interactions, modeling customer behaviors and preferences using such a database, and turning the predictions and insights into marketing actions to achieve the strategic goals of identifying, attracting and retaining customers (Ganapathy et al., 2004; Yen et al., 2004). Typical CRM modeling tasks include product recommendation, personalization, and the analysis of factors driving customer retention and loyalty.

1.2 Motivation of the Study

As mentioned above, existing studies of EIM mainly focus on the handling and mining of numerical data and there has been a general lack of attention paid to the management of textual information within an engineering environment.

1.2.1 Mining of Numerical Data

Data mining is motivated by the situation of "information rich but knowledge poor" (Fayyad, 1996). The fast-growing, tremendous amount of data, collected and stored in large and numerous databases, has far exceeded our human ability for comprehension without powerful tools. Simply stated, data mining refers to extracting or "mining" useful knowledge from massive data. Many people treat data mining as a synonym for another popularly used term, Knowledge Discovery in Databases (KDD). Alternatively, others view data mining as simply an essential step in the process of KDD, as shown in Figure 1.3 (Fayyad, 1996; Han and Kamber, 2001). Data mining tools have been employed in some engineering applications such as market need analysis (Li and Yamanishi, 2001; Yan et al., 2001), product design (Ishino and Jin, 2001; Schwabacher et al., 2001), manufacturing (Gardner and Beiker, 2000; Lee and Park, 2001), and services (Fong and Hui, 2001; Tan et al., 2000).

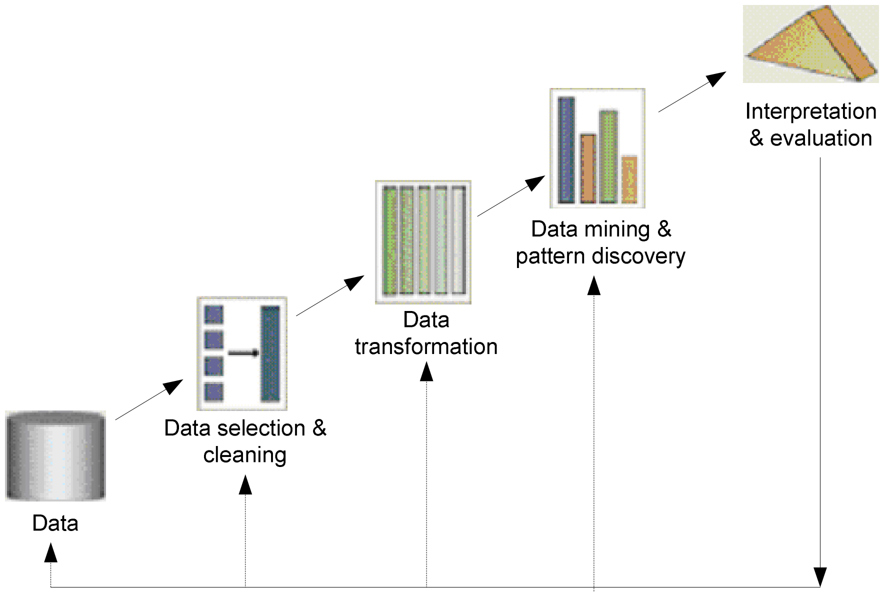

Interpretation & evaluation
Data mining & pattern discovery
Data transformation
Data selection & cleaning
Data

Figure 1.3 Data mining as an essential step in knowledge discovery process

1.2.2 Obstacles for Textual Information Processing

However, currently little attention has been paid to the mining of textual data within an engineering environment. There are probably three major reasons for this lack of attention:

- Numerical data are well structured and organized in databases, which makes them relatively easy to handle. There are already various established techniques for numerical data management and analysis. In comparison, textual data are usually stored as unstructured free texts or semi-structured data so that there is a greater level of difficulty in handling textual databases.

- Compared to the relatively clean numerical data, textual data contain a lot of noisy and redundant information. This characteristic creates an obstacle for further management of textual information.

- Most existing EIM applications have focused on design and manufacturing phases in which numerical information dominates. Textual information within an engineering environment is usually stored simply as archive for the purpose of information searching.

However, textual data offer a wealth of information in engineering activities and therefore motivate this study to investigate the challenging issues in textual information management.

1.2.3 Value of Textual Information

With the development of e-Engineering and e-Business, nowadays a huge amount of textual information is stored in enterprise intranets and the WWW, commonly appearing in e-mails, design notes, memos, notes from call centres and support operations, news, user groups, chats, reports, letters, surveys, white papers, marketing material, research, presentations and web pages (Blumberg and Atre, 2003). Just like numerical data, the textual data within the engineering environment possess a lot of valuable information. For example, technical papers and patent documents provide important references for R&D and product development (Liu, 2005; Loh et al., 2006; Menon et al., 2004); online customer reviews offer valuable comments for product design and manufacturing (Zhan et al., 2007).

Most textual information can be categorized into unstructured or semi-structured data. Such data lack a structure that is easily read and processed by a machine compared to structured data. Data with some form of structure may also be referred to as unstructured data if the structure is not helpful for the desired processing task. For example, a HyperText Markup Language (HTML) web page is structured by tags, but this structure is often oriented towards formatting, rather than performing more complex tasks with the content of the page. EXtensible Markup Language (XML) files can be viewed as semi-structured documents since they are formatted towards better indexing and searching. However, they are still far from fulfilling all the complex information needs in engineering environment, such as integrating information from multiple textual sources.

1.2.4 Management of Textual Information

Because of the wealth of information involved in textual data, how to utilize and how to discover knowledge from them effectively and efficiently is a concern. Unfortunately, only a few studies have been reported on textual information management within engineering domains, due to the obstacles that have been mentioned. The existing studies, focusing on making textual information more useful throughout the engineering process, can be divided into two major areas: information indexing & searching and automatic text classification.

1.2.4.1 Textual Information Indexing and Searching

Textual information indexing & searching focuses on developing methods to better index textual data and providing better searching experiences (Fong and Hui, 2001; Wood et al., 1998; Yang et al., 1998).

Wood et al. (1998) described a method based on typical Information Retrieval (IR) techniques for retrieval of design information. They created a hierarchical thesaurus of life cycle design issues, design process terms and component and system functional decompositions, so as to provide a context based IR. Within the corpus of case studies they investigated, it was found that the use of a design issue thesaurus could improve query performance compared to relevance feedback systems, though not significantly.

Yang et al. (1998) focused on making textual information more useful throughout the design process. Their main goal was to develop methods for search and retrieval that allow designers and engineers to access past information and encourage design information reuse.

Fong and Hui (2001) developed a data mining technique to mine unstructured, textual data from a customer service database for online machine fault diagnosis. In particular, neural networks were used within a case-based reasoning framework for indexing and retrieval of the most appropriate service records based on a user’s fault description.

1.2.4.2 Automatic Text Classification

Automatic text classification is to automatically classify textual data, like technical papers, patent documents, service records, to the predefined categories (Liu, 2005; Loh et al., 2006; Menon et al., 2004; Tan et al., 2000). The purpose is to provide better organization of textual databases and to facilitate effective and efficient IR tasks.

Tan et al. (2000) investigated service centre call records comprising both textual and fixed-format columns, to extract information about the expected cost of different kinds of service requests. They found that the incorporation of information from free-text fields provided for a better categorization of these records, thus facilitating better predictions of the cost of the service calls.

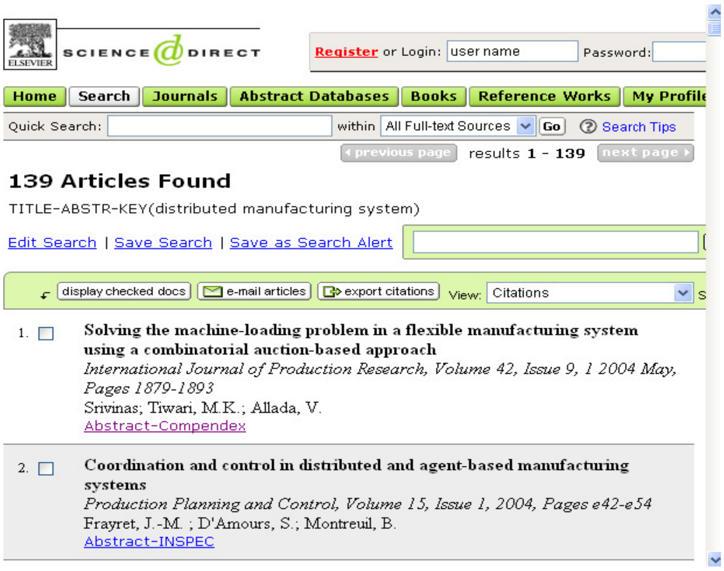
Menon et al. (2004) further established the needs and benefits of applying textual data classification within the product development process and presented successful implementations of textual data classification within two large multinational companies.

Recently, automatic text classification has been applied to different types of documents in engineering domain, such as automatic hierarchical classification of technical papers for manufacturing IR (Liu, 2005) and automatic patent document classification for TRIZ users (Loh et al., 2006).

1.2.5 Motivation for Text Summarization in Engineering Domain

As can be seen, existing studies on engineering textual information management were mainly focusing on the issue of organizing the huge amount of information and facilitating the process of information searching. On the other hand, another important issue, i.e. integrating information from multiple textual sources and extracting useful information to fulfill users' requirements, has not yet been addressed by previous studies.

The development of techniques like indexing, searching and classification has provided powerful tools for information seekers in engineering environment. However, due to the current overload of engineering information (such as technical papers, patent documents and customer reviews), even with these powerful tools, users may encounter a huge amount of retrieved documents for any given query. For example, when the query *distributed manufacturing system* is submitted to the ScienceDirect database (http://www.sciencedirect.com/), a total of 139 papers are retrieved, as shown in Figure 1.4. The user has to screen these documents manually, until suitable documents relevant to his specific purpose are identified. This process can be very time consuming.



results 1 - 139

139 Articles Found

TITLE-ABSTR-KEY(distributed manufacturing system)

1. □ **Solving the machine-loading problem in a flexible manufacturing system using a combinatorial auction-based approach**
International Journal of Production Research, Volume 42, Issue 9, 1 2004 May, Pages 1879-1893
Srivinas; Tiwari, M.K.; Allada, V.
Abstract-Compendex

2. □ **Coordination and control in distributed and agent-based manufacturing systems**
Production Planning and Control, Volume 15, Issue 1, 2004, Pages e42-e54
Frayret, J.-M. ; D'Amours, S.; Montreuil, B.
Abstract-INSPEC

Figure 1.4 ScienceDirect search result given the query “distributed manufacturing system”

In such context, a summarization system, which can integrate the information from retrieved documents and facilitate the searching process, is much needed. The retrieved documents, regarding the same query, must share much common information which is interesting to users. Besides, in some documents there must exist some unique information which is also useful for users to decide whether it is worthwhile to read the source documents. Therefore, the summarization system should be able to integrate the common information from all documents and point out the unique information for each single document. At the same time, this summarization system should be able to exclude the redundant and noisy information across the documents. The realization towards such summarization system is the focus of this study.

1.3 Objectives and Significance of the Study

The main objective of this study is to conduct a comprehensive investigation on the challenging issues in automatic summarization of multiple textual documents within the engineering domain, with an emphasis on the problem of summarizing multiple technical papers. Technical papers, as an important part of textual information within engineering domain, are essential for engineering research and knowledge management. Compared to other types of engineering texts such as customer e-mails and customer reviews, technical papers are more formally written and structured, homogeneous and knowledge-intensive. Therefore, we intended to apply technical papers as our study target and we started from here to build a framework of summarizing multiple engineering documents.

The research goals in this study could be outlined as follows:

- A preliminary investigation would be conducted, in order to figure out the significant issues in summarizing multiple technical papers and to provide a basement for further researches.

- An automatic summarization framework for multiple technical papers would be proposed. This summarization framework, addressing the specialties in the domain of technical papers, integrates information from multiple papers, extracts common knowledge and highlights the differences among different documents. The output summary of this summarization framework should be in a form of structured or semi-structured text.

- The proposed summarization framework would be tested under different parameterizations to discover factors that would affect the summarization performance. Moreover, it would be evaluated based on existing benchmark summarization systems.

- Case studies would be conducted to examine the application of automatic text summarization in facilitating other tasks within engineering information management and text mining.

This study aimed to provide a comprehensive examination of summarizing multiple technical papers and to enrich this infant research area. The significant issues addressed and the summarization framework proposed in this study should therefore contribute to a pioneer work in automatic summarization of multiple engineering documents. The exploration of applying summarization techniques in other textual information management tasks should provide useful knowledge for the application of summarization in EIM and establish a foundation for future research.

Summarization is a process to distill the most important information from source documents and at the same time remove irrelevant and redundant information. Moreover, the output of our summarization system would be a well structured text compared to the source documents. Therefore, this study could probably address the limitations for applying EIM to textual information that have been mentioned in

Section 1.2.2.

Although technical papers were the focus in this study, news articles were still widely applied in the experiments of this study because the standard corpora available for summarization research were based on news articles. Therefore, this study may also enhance our understanding of applying the proposed summarization methods to a broader domain of textual information.

1.4 Organization of the Thesis

The rest of the thesis is outlined as follows.

Chapter 2 provides a comprehensive literature review of automatic text summarization, with special focus on multi-document summarization and technical papers summarization because of their relevance to this study.

Chapter 3 conducts a preliminary investigation of the significant issues in multi-paper summarization, in order to provide a basement for further researches. Specifically, the chapter discusses the special characteristics of summarization task within the domain of technical papers. Moreover, a popular multi-document summarization method was experimented in summarizing multiple papers.

Chapter 4 studies the structure and relationship within multiple documents based on the analysis of real-world document sets. The notions of macrostructure and microstructure were proposed. Experiments were introduced to examine the influence of macrostructure and microstructure on summarization performance.

Chapter 5 proposes a multi-paper summarization framework based on macrostructure and microstructure. The discussion of macrostructure and microstructure in Chapter 5 was focused on the domain of technical papers.

The evaluation of multi-paper summarization system based on macrostructure and microstructure is discussed in Chapter 6. The evaluation task was designed to discover the factors within the system that would affect the summarization performance. Another purpose of the evaluation task was to compare the performance between the proposed summarization framework and other existing systems.

Two case studies are presented in Chapter 7 in order to further consolidate this research. One case study was to apply summarization in processing online customer reviews to help product designers, merchants and potential shoppers for their information seeking. The other case study was to utilize summarization to improve the performance of automatic text classification.

Chapter 8 concludes this study and offers suggestions for future work.

Chapter 2

Literature Review of Automatic Text Summarization

We benefit from various types of text summarization in our daily lives, e.g. BBC headlines, reviews of best-sellers and abstracts of scientific articles. Manually summarizing textual documents usually requires enormous human efforts, and this motivated the technology of automatic text summarization (Luhn, 1958; Mani, 2001). Research of automatic text summarization can be traced back to 1950s, with a renaissance of approaches from 1990s due to the development of computing technology and the explosive growth of electronic documents. This chapter presents a comprehensive review regarding the state-of-the-art researches on automatic text summarization. Since this thesis focuses on the task of summarizing multiple technical papers, the related studies of multi-document summarization and technical paper summarization are reviewed in Section 2.3 and 2.4.

2.1 Overview of Automatic Text Summarization

Summarization can be defined as the process of distilling the most important information from source documents to produce an abridged version for a particular user or task (Barzilay and Elhadad, 1997; Mani and Bloedorn, 1999; Sullivan, 2001; Visa, 2001). An alternative view is that summarization is to seek a trade-off between **condensing texts** and **preserving "important content" in source documents**. The "important content" in source documents varies with different requirements of users or tasks. Therefore, summarization is a user-oriented or task-oriented process.

2.1.1 Types of Text Summarization

The approach and the objective of summarization determine the type of a summary that is generated. The major types of summary are listed as follows:

- **Extract vs. Abstract**

  An extract consists wholly of portions extracted verbatim from the source document (they may be single words or whole passages), while an abstract consist of novel phrasings describing the content of the source document (which might be paraphrases or fully synthesized text) (Hovy and Lin, 1999). Abstraction aims to simulate manual summarization process which includes sentence compression and generation (Knight and Marcu, 2002; Mani et al., 1999). Existing summarization researches mainly focus on extraction since the development of abstraction is limited with the existing technologies of Artificial Intelligence (AI) and Natural Language Processing (NLP).

- **Indicative vs. Informative**

  An indicative summary aims to highlight the specialties for the document, helping a reader to decide whether it is worth reading the full document, while an informative summary synthesizes the important content in the document and the reader can acquire useful information from it without referring to the full document (Paice, 1990; Kan et al., 2001).

- **Generic vs. Query-biased**

Compared to a generic summary, a query-biased summary presents the content that is most closely related to user's queries (Goldstein et al., 1999; Tombros and Sanderson, 1998). This is often used in information searching services, in which the sentences relevant to user's queries are given more weights.

- **Just-the-news vs. Background**

  A just-the-news summary provides the newest facts given in the source document, assuming the reader is familiar with the topic, while a background summary offers certain background information regarding the topic (Hovy and Lin, 1999).

- **Evaluative vs. Neutral**

  An evaluative summary, or critical summary, offers a critique of the source document, while a neutral summary tries to be objective in summarizing the document (Hovy and Lin, 1999).

- **Single-document vs. Multi-document**

  In terms of the number of source documents to be summarized, summarization tasks can be categorized into single-document summarization and Multi-Document Summarization (MDS) (Mani and Bloedorn, 1999; Mckeown & Radev, 1995). Since MDS is the focus of this study, it is discussed in detail in Section 2.3.

2.1.2 General Architecture of Automatic Text Summarization System

Hovy and Lin (1999) described a general architecture of automatic text summarization system, as given in Figure 2.1. In this architecture, summarization is separated into three steps after pre-processing of input text: sentence selection, interpretation and sentence generation.

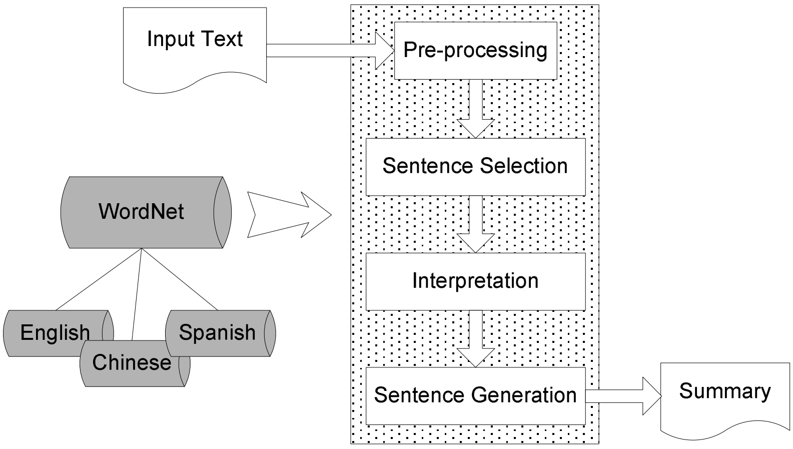


Figure 2.1 The architecture of summarization system

The first step of summarization is to filter the input text to retain only the most important information. Typical method is to extract the most important sentences which contain the topical information of the input text. The next two steps, i.e. interpretation and sentence generation, aim to make the output summary more coherent and readable. The goal of interpretation step is to fuse related topics into more general ones (e.g. *He ate oranges, durians, pineapples* → *He ate fruits*). The step of sentence generation is to rephrase and reorganize sentences into a coherent and new text.

Among these three steps, sentence selection is the core step since it deals with the key problem of summarization: **condensing source texts** and **preserving important content in source texts**, while the other two steps aim to make the output summary more coherent and readable. Therefore, most of existing summarization researches focus on the step of sentence selection. The methods for sentence selection are reviewed in Section 2.2.

2.2 Methods for Sentence Selection

In a typical process of sentence selection, a textual document is segmented into sentences first, scores are then assigned to each sentence according to a certain scoring function and finally the sentences with top scores are selected to be included in the summary until the predefined summary length is reached. In this process, sentence score can be calculated as a combination of various features, e.g. sentence position, indicator phrases, word frequency, discourse structure, etc. (Barzilay and Elhadad, 1997; Edmundson, 1969; Hovy and Lin, 1999; Kupiec et al., 1995; Marcu, 1999). Some of the popularly used features are listed in the following:

- **Frequent words**

  Frequent words are the words whose frequency in the source document is greater than a predefined threshold, but except the function words, such as *the*, *although*, *its*, etc. By using this feature, sentences which contain more frequent words are assumed to contain more topical information (Earl, 1970; Edmundson, 1969).

- **Title and heading words**

  The assumption here is that words except function words in title and headings of documents represent topical information. Sentences which contain these words should be given higher scores (Edmundson, 1969). It is worthwhile to point out that some headings in technical papers do not contain topical words, such as

Introduction, *Methodology*, *Results and Discussion*, etc.

- **Sentence position**

  Baxendale (1958) first stated that within a paragraph the first and last sentence are usually the most central to the theme of the article. Lin and Hovy (1997) utilized techniques of machine learning to identify the relationship between sentence importance and its position in the paragraph.

- **Indicator words and phrases**

  Indicator words and phrases, although not in themselves key words, provide an indication of whether the sentence contains topical content. Typical examples of indicator phrases are *in conclusion*, *this article*, *our work*, etc. Sentences which contain these phrases are assumed to contain significant information. Indicator phrases are dependent on the document genre. The list of indicator phrases for a certain document genre is usually constructed manually or by machine learning (Hovy and Lin, 1999).

- **Sentence length**

  This feature is based on the assumption that very short sentences tend not to contain topical information (Kupiec et al., 1995). Only sentences longer than a threshold are considered for including in the summary.

- **Query words**

  This feature is specifically set for query-biased summarization. Sentences in which query words (except function words) appear are given higher scores in sentence selection process (Tombros and Sanderson, 1998).

- **Lexical chains**

Lexical chains are sequences of related words grouped together by text cohesion relationships of repetition, synonymy, hypernymy (the semantic relation of being superordinate or generic, e.g., *plant* is a hypernym of *flower* and *tree*), antonymy and holonymy (the semantic relation that holds between a whole and its parts, e.g., *body* is a holonym of *arm* and *leg*), etc. These relations can be derived from the WordNet thesaurus (Miller, 1995). Barzilay and Elhadad (1997) identified strong lexical chains in source documents and added scores to those sentences attached with the strong lexical chains.

- **Discourse structure**

Discourse structure is used to describe the relationship among sentences and clauses, as shown in Figure 2.2 (Mann and Thompson, 1988; Marcu, 1999). Typical relationships among sentences include *elaboration*, *justification*, *contrast*, *condition*, etc. The salience of sentences and clauses can be computed based on the discourse structure.

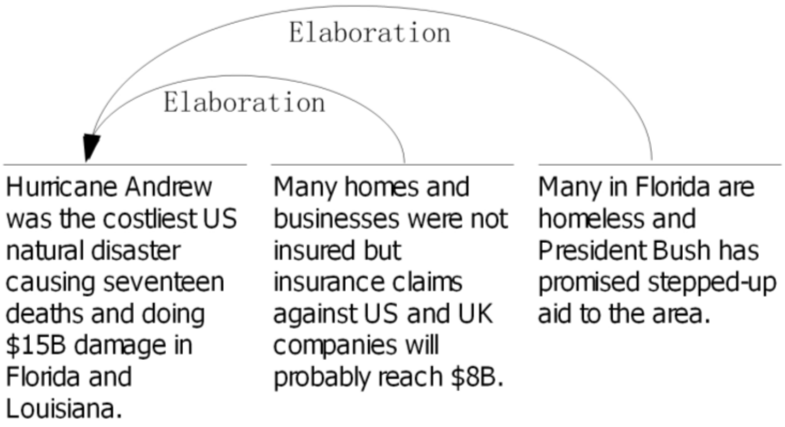

Elaboration
Elaboration
Hurricane Andrew was the costliest US natural disaster causing seventeen deaths and doing $15B damage in Florida and Louisiana.
Many homes and businesses were not insured but insurance claims against US and UK companies will probably reach $8B.
Many in Florida are homeless and President Bush has promised stepped-up aid to the area.

Figure 2.2 Discourse structure within sentences and clauses

In the sentence selection process, sentence score can be calculated as a linear combination of features (F_1, F_2, … F_k):

$$\text{Score} = \sum_{i=1}^{k} w_i F_i \tag{2.1}$$

where the coefficient w_i for each feature F_i is determined through analysis or training of the text corpus of "ideal" summaries along with their corresponding full texts (Edmundson, 1969).

An alternative method to combine features into sentence score is Bayes' rule, by calculating the probability for a sentence s to be included in the summary S given the k features F_i (i = 1, 2, …, k) with the assumption that features are statistically independent with each other (Kupiec et al., 1995):

$$\begin{aligned}\text{Score} &= P(s \in S \mid F_1, F_2, \ldots F_k) = \frac{P(F_1, F_2, \ldots F_k \mid s \in S)P(s \in S)}{P(F_1, F_2, \ldots F_k)} \\ &= \frac{\prod_{i=1}^{k} P(F_i \mid s \in S)P(s \in S)}{\prod_{i=1}^{k} P(F_i)}\end{aligned} \tag{2.2}$$

where $P(s \in S)$ is a constant, while $P(F_i | s \in S)$ and $P(F_i)$ can be known directly from the training corpus. The features chosen by Kupiec and his colleagues were all discrete, so this equation can be formulated in terms of probabilities rather than likelihoods.

2.3 Multi-Document Summarization

Initially, summarization research focused on summarizing a single document. Recently, as an outcome of the capability to collect large sets of documents online and the increasing demand for users to acquire knowledge from vast amount of information, there is a demand for more advanced technology to generate summary from a collection of documents, i.e. MDS.

Instead of focusing on single article, MDS deals with multiple documents which have relationship with each other (Mani and Bloedorn, 1999; Mckeown and Radev, 1995). The relationships among documents involve whole-part, differences in detail, differences in perspective, temporal trend, etc. (Mani and Maybury, 1999). The number of documents to be summarized can range from large gigabyte-sized collections to very small collections. Various MDS systems have been proposed in the past decade (Mani and Bloedorn, 1999; Moen et al., 2005; Radev et al., 2004).

2.3.1 Clustering-Summarization

Most of the existing MDS methods are based on the framework of clustering-summarization (Boros et al., 2001; Maña-López, 2004; Radev et al., 2004). Clustering-summarization first separates a set of documents into several non-overlapping groups of documents or sentences. Summarization is then performed separately within each group. The framework is shown in Figure 2.3.

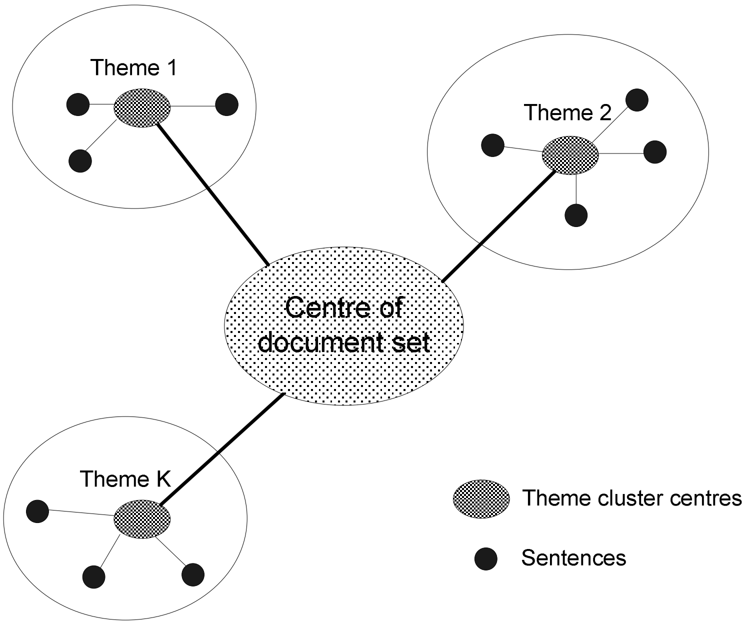


Figure 2.3 The framework of clustering-summarization

The framework of clustering-summarization is widely applied in the existing MDS studies because of its domain independence. The assumption of this framework is: a document set consists of several themes and a desired summary should cover as many of these themes as length constraint permits. Within this framework, each cluster represents a theme in the document set. Sentences within each cluster are ranked according to their distance from the cluster center, representing similarity of sentences and the theme. Similarly, clusters are ordered by their distance to document set, representing the importance of this theme.

However, there are two limitations to the clustering-summarization approach when applied to the domain of multi-paper summarization:

- The number of clusters, i.e. the number of themes is difficult to determine without prior knowledge regarding the set of papers. Inappropriately choosing this number will inevitably introduce noise and reduce effectiveness.

- In clustering-summarization, the document set is split into non-overlapping clusters and each cluster is assumed to discuss one theme. However, in a real-world paper set, themes often overlap with each other and are not perfectly distributed in non-overlapping clusters of papers. Each theme is associated with multiple papers. On the other hand, each paper in the set possibly discusses several themes instead of only one.

2.3.2 Examples of Domain Dependent MDS Systems

Clustering-summarization is a domain independent MDS framework. Unlike this framework, some MDS systems have been proposed specifically designed for certain document genres, e.g. news articles about terrorism. Some of these domain dependent MDS systems will be reviewed in the following.

SUMMONS (SUMMarizing Online NewS articles) system was proposed to generate summaries for multiple news articles (Mckeown and Radev, 1995). The input of this system is a set of templates generated by the MUC (Message Understanding Conference) system which operates on the terrorism domain and uses Information Extraction (IE) technique to fill 25 fields including *perpetrator*, *victim*, *type of event*, etc. Each template input to SUMMONS represents the information extracted from one or more articles. The MUC templates are then compared and merged using various planning operators. Each operator combines or synthesizes a pair of templates to a new template. For example, when two sources report conflicting information about the same event, the *contradiction* operator should be used. In the synthesis phase, the summarizer then uses text generation techniques to express the contradiction. There are seven operators in SUMMONS, including *agreement*, *addition*, *contradiction*, etc. The architecture of SUMMONS is shown in Figure 2.4. The summaries generated by SUMMONS are more coherent and informative compared to other peer systems. However, it can only deal with terrorism news articles because its input templates and planning operators are all dependent to this domain.

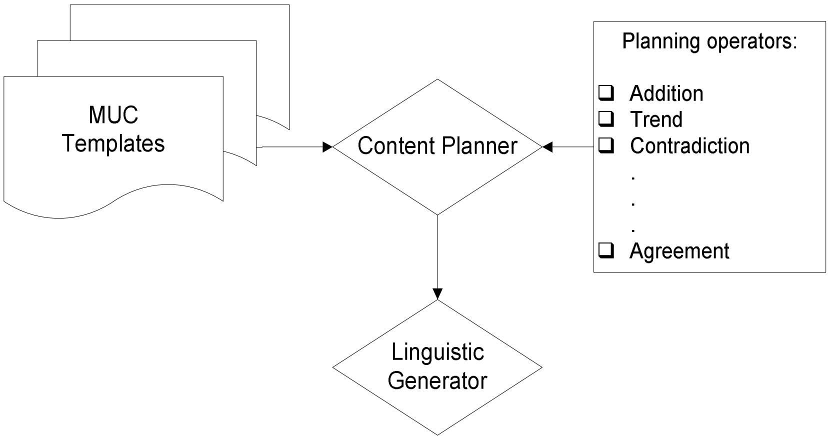


Figure 2.4 The architecture of SUMMONS

Radev (2000) proposed Cross-document Structure Theory (CST) which was a taxonomy of the information relationships among related documents. The concept of CST is similar to discourse structure within single document. These cross-document relationships can assist in MDS and some of them are direct descendents of those used in SUMMONS.

In the summarization system proposed by Mani and Bloedorn (1999), the assumption was that the more strongly connected a text unit was to other units, the more salient it was. For each document, they constructed a graph representation whose nodes are term occurrences and whose edges are cohesion relationships (proximity, repetition, synonymy, hypernymy and coreference) between terms. The architecture of this summarization system is shown in Figure 2.5. In the first phase, a graph for each document is built. Then salient nodes in each graph related to the *topic* are discovered and reweighted. The *topic* can be user's query for user-focused summary. The set of reweighted nodes for each graph are then compared and the result of this comparison is used in the final phase to extract sentences. This MDS system was restricted to summarization of only two documents.

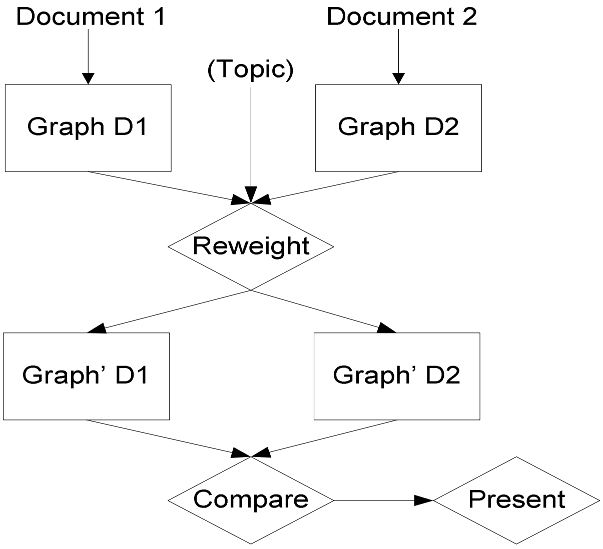


Figure 2.5 Summarization system of Mani and Bloedorn

2.4 Related Work of Technical Paper Summarization

Although there existed many studies regarding automatic text summarization, few of them have taken account of the domain of technical papers. The possible reason is that almost all technical papers have abstracts generated by the authors themselves. However, the author-generated abstracts are all based on single paper and cannot satisfy the information need raised by browsing multiple papers as has been mentioned in Chapter 1. This section will review the numbered existing studies regarding automatic paper summarization. All these existing studies were focused on generating summary automatically for single paper.

2.4.1 Existing Studies of Single Paper Summarization

Paice (1990) stated the challenges in technical paper summarization. The author highlighted the problem of anaphoric reference. Anaphoric reference, such as *this method*, *those experiments*, used to avoid repetition, is an inevitable problem in the domain of technical articles which causes the incoherence of the summary.

Paice and Jones (1993) focused on summarizing technical articles of crop agriculture and utilized IE techniques to instantiate semantic roles such as SPECIES, PEST, SOIL and CLIMATE. Their approach was restricted in a narrow domain of crop agriculture and highly structured articles and was difficult to be applied in the general domain of technical papers.

Teufel and Moens (2002), on the other hand, proposed the approach of rhetorical analysis in order to clearly define the function of various parts in a technical article, such as *Introduction*, *Experiments*, *Results* and *Conclusion*. The output of rhetorical analysis could be utilized as the starting material for further summarization processes.

2.4.2 Limitations of Existing Studies

There are two major limitations for the existing work on technical paper summarization which will be addressed in this study:

- **Most of existing studies ignored the special characteristics of technical papers.** Compared with other types of text documents, technical papers possess some special characteristics which can be utilized in summarization. For example, technical papers have a title, key words list, references list which builds linkage in various papers. Authors of technical papers tend to apply a lot of indicator phrases to organize their ideas, e.g. *in this paper*, *to summarize*, etc. Moreover, technical papers usually have a clear structure, started with *Introduction*, then *Methodology* or *Experiments*, finally *Conclusion*. Up to now, few summarization techniques have been proposed to address these special characteristics of technical papers.

- **No work has been done on automatic summarization of multiple technical papers**, although some work has been described to support authors in writing a review article for multiple technical papers. Nanba and Okumura (1999) proposed a supporting system that could classify the citation areas into three types in order to let review writers understand the relationship among papers. However, this work focused on the system which could support the process of manual summarization and did not offer a solution for automatic summarization. Existing MDS approaches focus mainly on news articles (Boros et al., 2001; Maña-López, 2004; Mckeown and Radev, 1995; Moen et al., 2005). Compared with news articles, summarization of multiple technical papers requires different approaches since there are a lot of differences in terms of document genre and readers' requirements.

2.5 Conclusion of the Chapter

This chapter has offered a comprehensive review of the state-of-the-art techniques in automatic text summarization, with the focus on multi-document summarization and technical paper summarization, since the focus of this study is multi-paper summarization.

Most existing MDS systems were based on the framework of clustering-summarization which has limitations when applied into the domain of technical paper summarization: the number of clusters is difficult to be predefined and the themes within a technical paper set are not perfectly distributed into non-overlapping clusters of papers. These limitations are further discussed and addressed in the following chapters of this thesis.

There are only a few existing studies for technical paper summarization and none of them focused on automatic summarization of multiple papers. Moreover, in existing work, the special characteristics of technical paper domain have not been successfully exploited. These limitations of the existing work motivate the research efforts in this study.

Chapter 3

Preliminary Investigation into Multi-Paper Summarization

The problem of summarizing multiple technical papers is the focus of this study. As has been reviewed in the previous chapter, existing research regarding summarization mainly focused on the domain of news articles and few of them have taken into account the domain of technical papers. Compared to news articles, technical papers possess special features, and moreover, readers may have special requirements for the summary of technical papers. Therefore, different approaches are demanded regarding the task of summarizing technical papers.

In this chapter, the special characteristics of summarization task within the domain of technical papers will be discussed in detail. Moreover, the popular MDS framework of clustering-summarizing will be applied into summarizing technical papers. This chapter reports a preliminary investigation of the challenging issues in summarizing multiple technical papers.

3.1 Special Characteristics of Technical Paper Summarization

Compared to news articles, summarization of technical papers demands different approaches, mainly due to the differences in terms of readers' information requirements and special characteristics of document genre.

3.1.1 Special Characteristics of Readers' Information Requirements

In a set of news articles about an event or a person, it is possible that similar contents occur repetitively in these articles. Such recurrent information is assumed to be the most significant information for readers. Therefore, from readers' perspectives, a summarization system is expected to utilize information fusion techniques to extract such significant information, while the uniqueness of each single article is not so important. A good summary should help in distilling important information and in turn save time for readers. In this sense, the summary for news articles should be "informative", i.e. readers expect to acquire enough information from the summary without having to refer to source articles.

Figure 3.1 shows a manually written summary for seven news articles regarding *Hurricane Andrew* from the corpus of Document Understanding Conference (DUC) (http://duc.nist.gov). DUC corpus is a standard corpus collected in the annual Document Understanding Conference and used as a benchmark in summarization research. This corpus contains documents with

- Manually created summaries
- Automatically created baseline summaries
- Submitted summaries created by the participating groups' systems
- Tables with the evaluation results

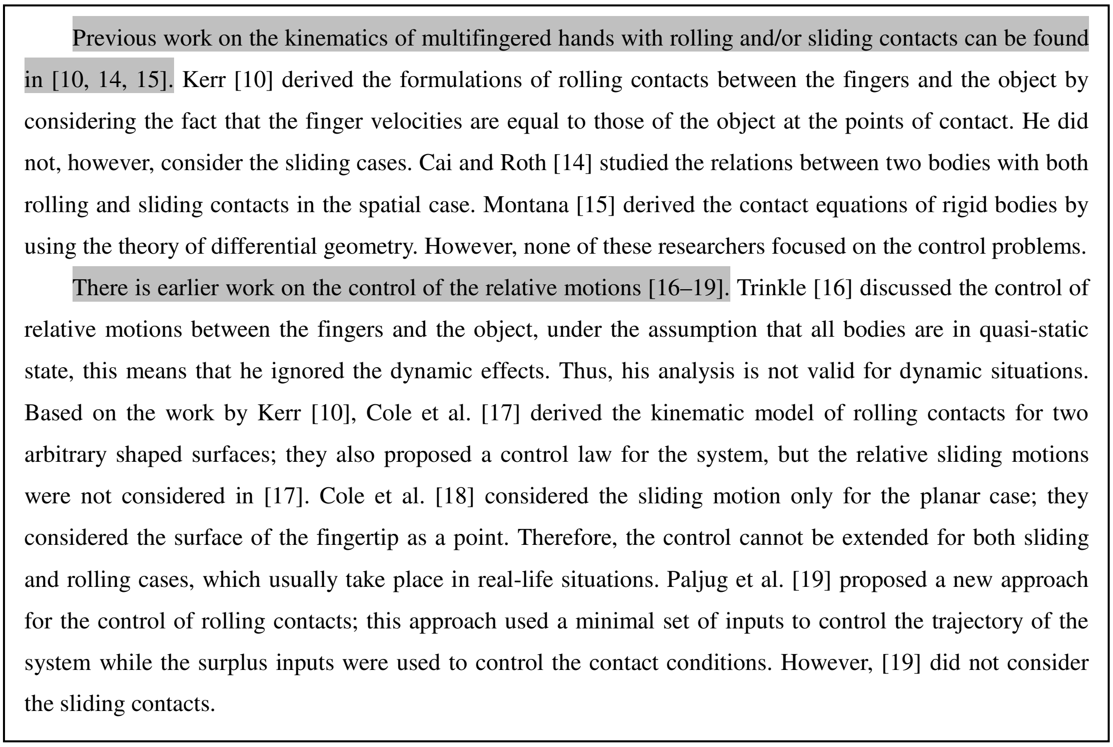
The summary in Figure 3.1 includes the significant content shared within the seven source articles and excludes the irrelevant and redundant information. It provides a short text with wealth of information.

> Hurricane Andrew, the costliest natural disaster in US history, killed at least 17 people. Southern Florida, in particular, Dade County was the scene of greatest damage. One in every eight homes was destroyed. In Florida overall, 150,000 persons were left homeless, and a week after the storm, 275,000 homes and businesses were still without electricity. Louisiana was also severely damaged by Andrew. It was initially feared that the storm might hit New Orleans which, because it is below sea level would be especially vulnerable. However, Andrew made landfall 60 miles to the west and most of the extensive damage was to rural areas with the oil refining industry left mostly untouched.
>
> US insurers expected Andrew claims could reach $8B. Claims against British companies could reach $1B. Total losses could be $15B with much of the damage to uninsured homes and businesses.
>
> On-site officials in Florida were critical of delays in getting food, drinking water, and other needed supplies to the area. Federal officials admitted problems and President Bush ordered troops to the area. The Federal Emergency Management Agency, saddled with many political appointees, had no plan to deal with the disaster. President Bush made a second trip to Florida and promised to rebuild Homestead Air Base.

Figure 3.1 A manual summary of seven news articles talking about "Hurricane Andrew" from DUC corpus

However, for a set of technical papers sharing the same topic, paper authors tend to distinguish their researches by concentrating on their own contributions. The uniqueness of each paper is interesting to readers in addition to the common knowledge across papers. Moreover, for users who query information in the technical paper databases, their information needs are diverse. Some users are interested in methodology, while some others may only look for particular experimental results or equipments. Unless we present the full papers to readers, it is unlikely to satisfy such diverse information needs with a short summary. In this sense, presumably, the summary of multiple technical papers is an "indicative" one whose purpose is to provide clues for further reading. Hence, the summary of multiple technical papers is expected to include the background knowledge and also to indicate different papers' unique ideas, approaches and contributions.

A literature review section in a technical article is presented in Figure 3.2. This short text summarizes seven papers about *multifingered robot hands*, although it could be biased according to the author's own understanding. In this summary, each paragraph describes one subtopic indeed. The first paragraph focuses on *kinematics of multifingered hands with rolling and/or sliding contacts* and summarizes the individual contributions of three papers with respect to this topic. If researchers are interested in this topic, they can choose the relevant papers for further reading. The second paragraph is more concerned with *control of relative motions*, both in *sliding* and *rolling*. Readers interested in this topic will then concentrate on the papers mentioned.



Previous work on the kinematics of multifingered hands with rolling and/or sliding contacts can be found in [10, 14, 15]. Kerr [10] derived the formulations of rolling contacts between the fingers and the object by considering the fact that the finger velocities are equal to those of the object at the points of contact. He did not, however, consider the sliding cases. Cai and Roth [14] studied the relations between two bodies with both rolling and sliding contacts in the spatial case. Montana [15] derived the contact equations of rigid bodies by using the theory of differential geometry. However, none of these researchers focused on the control problems.

There is earlier work on the control of the relative motions [16–19]. Trinkle [16] discussed the control of relative motions between the fingers and the object, under the assumption that all bodies are in quasi-static state, this means that he ignored the dynamic effects. Thus, his analysis is not valid for dynamic situations. Based on the work by Kerr [10], Cole et al. [17] derived the kinematic model of rolling contacts for two arbitrary shaped surfaces; they also proposed a control law for the system, but the relative sliding motions were not considered in [17]. Cole et al. [18] considered the sliding motion only for the planar case; they considered the surface of the fingertip as a point. Therefore, the control cannot be extended for both sliding and rolling cases, which usually take place in real-life situations. Paljug et al. [19] proposed a new approach for the control of rolling contacts; this approach used a minimal set of inputs to control the trajectory of the system while the surplus inputs were used to control the contact conditions. However, [19] did not consider the sliding contacts.

Figure 3.2 The literature review part in "Chen, J. and M. Zribi. Control of multifingered robot hands with rolling and sliding contacts. International Journal of Advanced Manufacturing Technology, 16(1), pp. 71–77. 2000" (Topical sentence in each paragraph is highlighted.)

3.1.2 Special Characteristics of Document Genre

In addition to the special information requirements from readers, technical papers possess some special characteristics in terms of document genre which may be considered in summarization process:

- Technical papers usually have a kind of rhetorical structure, e.g. starting with an introduction section, followed by literature review, experiments and results, and finally conclusion. Such rhetorical structure has not yet been effectively addressed in technical paper summarization (Teufel and Moens, 2002).

- Technical papers include a relatively fixed set of indicator phrases, such as "in conclusion", "this paper", "our work", etc. Sentences which contain these phrases are usually containing significant information and should be given more weight in summarization.

- Each technical paper usually has a title and a list of key words, which are assumed to highlight the most important points from the authors' point of view. The terms in the title and key words, except function words, should be given more weights and accordingly, sentences containing these words should be assigned higher scores.

- Technical papers usually have citation sections which indicate the relationship among each other. Some studies have been reported to automatically classify the citation areas in the source paper into different types in order to let review writers better understand the relationship among papers (Nanba and Okumura, 1999).

To sum up, the major differences regarding MDS in the domain of news articles and technical papers are listed in Table 3.1, which may provide principles for further summarization research in the domain of technical papers.

Table 3.1 Differences regarding MDS in the domain of news articles and technical papers

	News articles	**Technical papers**
Document genre	Free writing styles, usually starting with topical sentences and followed by elaboration information	Follow certain writing styles, with relatively fixed rhetorical structures, frequently occurring indicator phrases, citations, etc.
Relationship among documents	Multiple news articles talking about a same event or a same person	Multiple papers discussing a same topic, or from a same author, or retrieved with a same query
Content overlap among documents	A lot of recurrent information and content overlap	Some commonalities in the part of *introduction* and *literature review*, few commonalities in other parts which focus on authors' own contributions
Information requirements of readers	Readers want to have an essential idea about the event, person or topic. Details and differences among articles are not that important. Summary should substitute the source articles to some extent.	Readers are interested in relationships among papers (what is the topic about and what are different authors' contributions), so that they can further choose the papers most interesting to them. Summary provides a clue for further reading of the full papers.
Goal of summarization	To present the distilled, most important and common information	To present the background knowledge and common topics, and indicate the different contributions of authors as well

3.2 Pre-Processing of Textual Documents

Previous studies have demonstrated that pre-processing steps can affect the performance of text retrieval, classification and summarization (Salton et al., 1997; Yang and Chute, 1994). Typical steps include stop words removal and word stemming. Moreover, in the domain of technical papers, acronyms are very popular and need to be addressed in the pre-processing steps. Therefore, these three pre-processing steps are applied for the experiments in the rest of the thesis.

3.2.1 Stop Words Removal

Stop words are those words which rarely contribute useful information in terms of document relevance. Most stop words are functional words which do not carry meaning, including articles, prepositions, conjunctions and some other high-frequency words, such as *a*, *the*, *of*, *and*, *I*, *it* and *you*. The assumption is that, when assessing the contents of natural language, the meaning can be conveyed more clearly, or interpreted more easily, by ignoring the functional words. Removal of non-informative stop words has been a common technique in text indexing, retrieval and classification to reduce the noisy information and to improve the accuracy (Van Rijsbergen, 1979).

3.2.2 Word Stemming

Another common pre-processing step in dealing with textual data is word stemming. Stemming is the process of reducing inflected or derived words to their stem, base or root form. For example, a stemming algorithm for English should stem the words *fishing*, *fished*, and *fisher* to the root word, *fish*.

The most popularly used stemming algorithm in text mining is suffix stripping algorithm, since it does not rely on a lookup table that consists of inflected forms and root form relations (Lovins, 1968; Porter, 1980). In suffix stripping algorithm, a set of rules are stored which provide a path for the algorithm, given an input word form, to find its root form. Some examples of the rules include:

- If the word ends in *ed*, remove the *ed*
- If the word ends in *ing*, remove the *ing*
- If the word ends in *ly*, remove the *ly*

Although suffix stripping algorithm is sometimes regarded as crude given the poor performance when dealing with exceptional relations (like *ran* and *run*), it is still widely applied due to its easy implementation in automatic text processing systems and has shown the capability to reduce the redundancy and dimension of the document space representation (Scott and Matwin, 1999; Sullivan, 2001). In this study, Porter's Algorithm (Porter, 1980) is applied for word stemming.

3.2.3 Acronyms Identification and Replacement

Acronyms are ubiquitous in technical papers, for example:

Automated Guided Vehicle = AGV

Database and Network Support Subsystems = DNSS

Small and Medium Sized Enterprises = SMEs

If not matched with their expansions, acronyms can be a significant obstacle for readers to understand the texts and also introduce noise into the summarization system. When acronyms first occur in articles, authors usually give the full expressions and enclose the acronyms in the following parentheses, making them easy to be identified and matched with their full expressions, for example:

A key component of computer integrated manufacturing (CIM) is computer aided process planning (CAPP).

In our system, acronyms and their preceding word sequences are detected. If they can be matched, a link will be assigned between them and they will be added into the library of acronyms. Later when an acronym reoccurs, we can locate its expansion in the library. In processing acronyms, we usually omit the last lower cased *s*, e.g. *SME* and *SMEs* are treated as the same thing. Both of them refer to *Small and Medium Sized Enterprises*.

3.3 Clustering-Summarization of Multiple Papers

Clustering-summarization, as a popular and domain-independent framework for MDS, was applied in the preliminary investigation of multi-paper summarization in order to build a benchmark for further research. Figure 3.3 demonstrates the process of multi-paper summarization based on the framework of clustering-summarization. The paper set is first divided into several clusters and each cluster represents one theme in the paper set. Summarization is then performed in each individual cluster of papers and output to readers. By reading the summary, readers are expected to decide which cluster or theme is of interest to them.

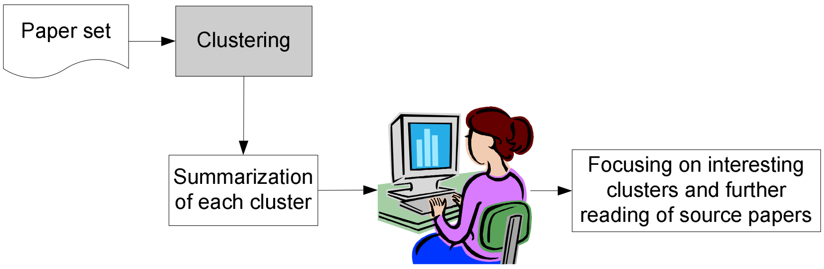

Paper set
Clustering
Summarization of each cluster
Focusing on interesting clusters and further reading of source papers

Figure 3.3 Clustering-summarization of multiple papers

Within the framework shown in Figure 3.3, clustering of technical papers is the core step and will influence the performance of the following steps. Document clustering is the process of grouping a set of textual documents into groups of similar documents (Sullivan, 2001; Tkach, 1997; Visa, 2001). It can be utilized to offer an overview of the content and structure of a document set, and facilitate the process of browsing to find relevant information. For example, when we submit a query *java* to a search engine, hundreds or thousands of documents might be retrieved. Ideally, we would like the search engine to automatically cluster the retrieved documents to several groups, such as an Indonesian island, a kind of coffee and a high-level programming language, so that we can quickly identify the relevant documents. Typical clustering methods include *K*-means (Bishop, 1995), agglomerative clustering (Voorhees, 1986), self-organizing maps (Kohonen, 1997), etc.

Traditionally most clustering studies were based on the indexing scheme of Vector Space Model (VSM) which has a few limitations such as high dimensionality and weakness in handling synonymous and polysemous problems. Latent Semantic Indexing (LSI) is able to deal with such problems to some extent and therefore has been applied as the indexing scheme for clustering in this study. VSM and LSI will be discussed in detail in the following section and an experiment to compare the performance between VSM-based clustering and LSI-based clustering is also presented.

3.4 Indexing Scheme in Document Clustering

The objective of document clustering is to maximize the intra-cluster similarity and the inter-cluster dissimilarity. The similarity between two technical papers can be measured based on a set of features, e.g. author (papers from the same author may be more similar), publication journal or conference (papers from the same source may be more similar), citation structure (papers which have citation links may be more similar). In addition to these features, the most essential and robust way to measure the similarity of two documents is to calculate the cosine-based "distance" between the vectors of the two documents. For example, the similarity between two document vectors v_i and v_j is defined by the cosine of the angle between them (Sullivan, 2001):

$$Sim(v_i, v_j) = \cos(v_i, v_j) = \frac{v_i \bullet v_j}{\|v_i\| \cdot \|v_j\|} \tag{3.1}$$

where $\bullet$ indicates the dot product of vectors v_i and v_j, $\|v_i\|$ and $\|v_j\|$ are the Euclidean lengths of vectors v_i and v_j.

3.4.1 Vector Space Model

Traditionally, the document vector is often modeled as a vector of index *terms* (a list of words after preprocessing such as stop words removal and word stemming) in the document (Beil et al., 2002):

$$v_i = (w_{i1}, w_{i2}, ..., w_{in}) \tag{3.2}$$

where w_{ij} is the weight of term *j* in document *i*.

This representation scheme is called vector space model (VSM) which is widely used in IR, clustering, classification and other text mining tasks (Salton et al., 1975). If a document set contains d documents and t index terms, we can build a $d \times t$ document-by-term matrix:

$$(v_1, v_2, ..., v_d)^T = [w_{ij}] \quad (i\text{=1~}d,\ j\text{=1~}t) \tag{3.3}$$

Vector space model has two major limitations (Sullivan, 2001):

- By using VSM, we have to deal with thousands or even tens of thousands of distinct terms in the document-by-term matrix. Many terms only appear in one or two documents, making the document-by-term matrix extremely sparse. Even with preprocessing like stop words removal and stemming, we may still have an extremely high number of dimensions to deal with, which will reduce the efficiency and accuracy for clustering process.

- The second problem is the natural language related problems of synonymy and polysemy. Synonymy refers to the fact that multiple words can have the same or similar meaning. For example, we may use *motorcar*, *car*, *automobile*, etc. to refer to a same object: a self-propelled passenger vehicle that usually has four wheels and an internal-combustion engine, used for land transport. When we submit a query which contains any of these words in a web search engine, we would like the search engine to treat them as the same thing. On the other hand, polysemy refers to the fact that a single word may have multiple meanings. A typical example is the word *java* which could mean an Indonesian island, a kind of coffee or a high-level programming language.

An alternative document indexing scheme, i.e. LSI, is able to deal with the two limitations of VSM to some extent and therefore was investigated in this study (Deerwester et al., 1990).

3.4.2 Latent Semantic Indexing

Unlike VSM which indexes documents with words, LSI tries to extract the latent concepts of text documents by identifying the pattern of word co-occurrence, e.g. *computer aided design*, *data mining* (Deerwester et al., 1990). Instead of using manually constructed dictionaries, knowledge bases or syntactic parsers, LSI utilizes purely statistical techniques to find co-occurrence of words.

The mathematical method underlying LSI is Singular Value Decomposition (SVD) (Gentle, 1998). Through SVD, the document-by-term matrix $X_0\,(d \times t)$ can be decomposed into the product of three matrices:

$$X_0 = D_0 \cdot S_0 \cdot T_0^T \qquad (3.4)$$

$D_0(d \times m)$ and $T_0(t \times m)$ have orthonormal columns. Column vectors in D_0 are called left singular vectors and column vectors in T_0 are called right singular vectors. $S_0(m \times m)$ is a diagonal matrix, in which $m = \min(d,t)$. Singular values are ordered by size along the diagonal of matrix S_0.

If we only keep the first k ($k \leq m$) largest singular values and set the rest to zeros, and only keep the first *k* columns of D_0 and T_0 accordingly, there will be three new matrices D ($d \times k$), S ($k \times k$), T ($t \times k$). The product of these three matrices is:

$$X = D \cdot S \cdot T^T \qquad (3.5)$$

The new matrix X is one of rank *k* which is closest in the least squares sense to X_0. Hence, the vector space is reduced from *m* dimension to *k* dimension. In the rest of this chapter, LSI-*k* is used to denote such dimension reduction process.

Through LSI-*k*, all documents and terms are mapped into *k*-dimensional space, as can be seen from document matrix D ($d \times k$) and term matrix T ($t \times k$). Document clustering can be performed in the reduced vector space of *k* dimensions. Previous research has proven that LSI can reduce clustering time and improve clustering efficiency on large-sized document sets (Schutze and Silverstein, 1997). In this study, since small-sized document sets are the focus in the MDS process, the accuracy of clustering is of more concern. Therefore, it is necessary to investigate whether LSI can improve the clustering accuracy of small-sized document sets and the optimal dimensions of LSI in clustering task.

3.4.3 Design of Experiment to Compare VSM and LSI

The purpose of the experiment is two folds:

- To examine whether LSI can improve the clustering accuracy of small-sized document sets compared to VSM
- To examine the optimal dimensions of LSI in clustering task, i.e. to decide the optimal *k* in LSI-*k*

Thirty document sets from the corpus Reuters-21578, Distribution 1.0 (http://www.daviddlewis.com/resources/testcollections/reuters21578/) were applied in this experiment. Reuters-21578 is currently the most widely used benchmark collection for text classification and text clustering research. The data was originally collected and labeled by Carnegie Group, Inc. and Reuters, Ltd.

The sample size and number of classes for all 30 document sets are listed in Table 3.2. Ten document sets have either two or three classes, and another ten sets have six or seven classes, while the remaining ten sets have 11 or 12 classes.

Table 3.2 Thirty document sets from Reuters-21578

Size of document set (number of documents)	Number of classes	Size of document set (number of documents)	Number of classes	Size of document set (number of documents)	Number of classes
49	3	50	6	51	12
54	3	70	6	65	12
66	3	92	6	71	11
72	3	105	6	86	12
90	2	130	6	91	12
108	3	149	7	110	12
129	3	151	6	147	12
152	3	182	6	180	12
190	3	222	6	220	12
231	3	261	6	297	12

K-means was applied as the clustering algorithm in this experiment. The algorithm starts by partitioning the input points (documents) into *K* initial sets. It then calculates the mean point, or centroid, of each set. It constructs a new partition by associating each point with the closest centroid. Then the centroids are recalculated for the new clusters. The algorithm repeats by alternate application of these two steps until convergence, which is obtained when the points no longer switch clusters, or alternatively centroids are no longer changed (Bishop, 1995).

The clustering quality is evaluated by *F*-score, which is the combination of recall (*R*) and precision (*P*) (Steinbach et al., 2000). After clustering of a document set, a cluster C has a counterpart of a predefined class T in the document set. C is treated as the retrieved set of documents for a query and T as the desired set of documents for the query. The recall and precision for each cluster are defined as follows:

$$R(\mathrm{C,T}) = \frac{N_1}{N_2}, \quad P(\mathrm{C,T}) = \frac{N_1}{N_3} \tag{3.6}$$

where

N_1 = number of documents in class T which are assigned to cluster C

N_2 = total number of documents in class T

N_3 = total number of documents in cluster C

The *F*-score is to seek the balance between recall and precision:

$$F = \frac{2RP}{R + P} \tag{3.7}$$

The average *F*-score is calculated across all clusters to measure the clustering accuracy.

For all the 30 document sets, clustering was performed based on LSI-*k*, with *k* ranging from two to the document size subject to a maximum of 100. The clustering quality of LSI-*k* was compared with VSM based on average *F*-score.

3.4.4 Experimental Results

Table 3.3 and Figure 3.4 show the experimental result of the document set with 261 documents and six classes. Recall and precision for each cluster are presented in this table. The last column gives the average *F*-score across six classes. It can be found that the average *F*-score of LSI-10 is the highest, reaching 0.718, and much higher than original VSM, in which the *F*-score is only 0.429.

Table 3.3 Clustering results of the document set with 261 documents

Class label		1	2	3	4	5	6	Average *F*-score
LSI-100	Recall	0.444	0.289	0.122	0.283	0.357	0.405	0.332
	Precision	0.244	0.245	0.185	0.232	0.938	0.630	
LSI-50	Recall	0.378	0.311	0.317	0.283	0.452	0.571	0.400
	Precision	0.362	0.226	0.317	0.342	0.388	1.000	
LSI-40	Recall	0.356	0.222	0.488	0.326	0.405	0.667	0.422
	Precision	0.250	0.233	0.392	0.484	0.395	0.966	
LSI-30	Recall	0.422	0.267	0.634	0.196	0.452	0.190	0.353
	Precision	0.311	0.218	0.520	0.250	0.373	1.000	
LSI-20	Recall	0.444	0.400	0.585	0.522	0.143	0.810	0.490
	Precision	0.444	0.353	0.558	0.500	0.150	1.000	
LSI-10	Recall	0.667	0.667	0.780	0.326	0.857	0.929	0.718
	Precision	0.469	0.811	0.865	0.313	1.000	1.000	
LSI-5	Recall	0.578	0.667	0.829	0.500	0	0.524	0.506
	Precision	0.426	0.612	0.872	0.377	0	0.786	
LSI-2	Recall	0.422	0.311	0.366	0.065	0.476	0.762	0.399
	Precision	0.288	0.259	0.259	0.214	0.556	0.970	
VSM	Recall	0.956	0.089	0.463	0.087	0.167	0.952	0.429
	Precision	0.269	0.143	1.000	0.571	1.000	1.000	

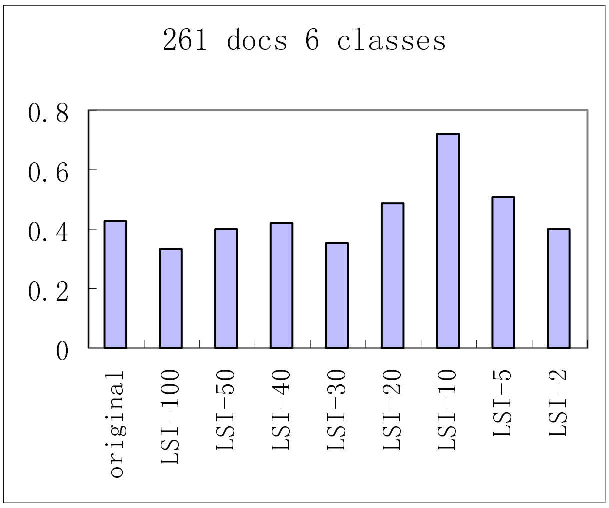


Figure 3.4 Comparison for clustering results using LSI-k of the document set with 261 documents (Vertical axis is average F-score.)

The experimental results of some other document sets are shown in Figure 3.5. The

LSI-*k* which achieved the best clustering performance for each document set is listed in the Table 3.4. These results show that by using an appropriate LSI-*k* we can greatly improve the clustering accuracy of small document sets.

It can be found that the optimal *k* for LSI-*k* is highly related to number of clusters in the document set. As can be seen in Figure 3.5 and Table 3.4, when the number of classes is two or three, *k* around five can achieve the best accuracy. When there are six or seven classes in a document set, *k* around ten gives the best results. As the number of classes reaches 11 or 12, *k* greater than or equal to 20 are needed. One possible reason is that each dimension in LSI represents a latent concept. The more classes a document set has, the more dimensions are needed to model this document set.

Moreover, it can be found that the optimal LSI-*k* has little relationship with sample size in small document set clustering. Table 3.4 shows that LSI-5 is the optimal for sample size of 49 and sample size of 231 (both have three classes); LSI-10 is the optimal for sample size of 50 and sample size of 261 (both have six classes); LSI-20 is the optimal for sample size of 51 and sample size of 297 (both have 12 classes).

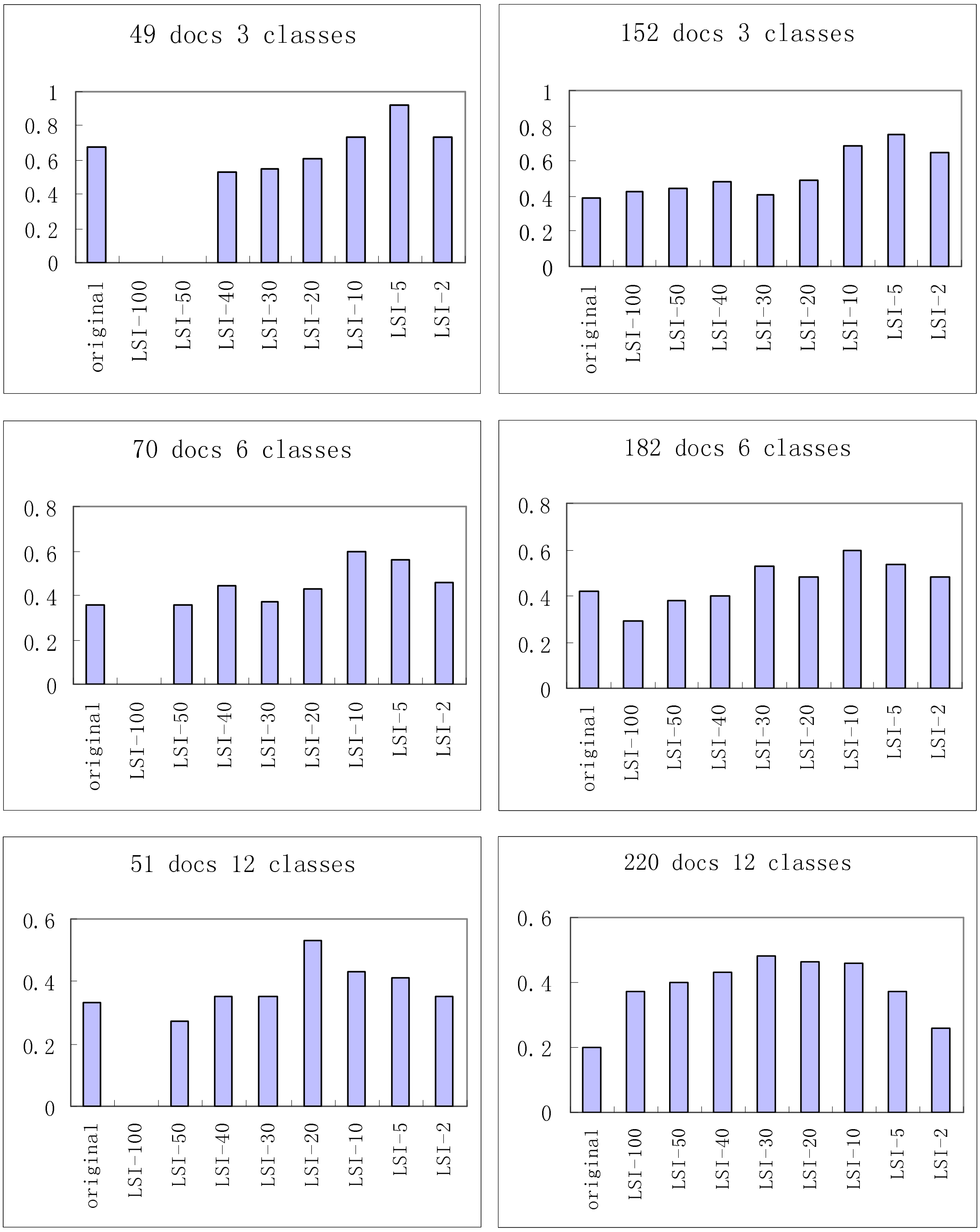


Figure 3.5 Clustering results for some document sets (Vertical axis is average F-score.)

Table 3.4 LSI-k which can achieve the best clustering performance for each document set

Size of document set (number of documents)	Number of classes	Optimal *k* for LSI-*k*
49	3	5
54	3	5
66	3	5, 10
72	3	5
90	2	5
108	3	5
129	3	5
152	3	5
190	3	5, 10
231	3	5
50	6	10
70	6	10
92	6	10
105	6	10
130	6	5, 10
149	7	10, 30
151	6	5, 10
182	6	10
222	6	10
261	6	10
51	12	20
65	12	20
71	11	20, 30
86	12	20
91	12	20
110	12	20
147	12	20, 30
180	12	10, 20
220	12	20, 30
297	12	20

3.4.5 Discussion

Document clustering is a key step for the clustering-summarization method. The purpose of the experiment in the preceding section is to compare the performance of VSM and LSI in clustering. The experimental results show that LSI can improve the clustering performance for the document sets with the size of tens to hundreds documents. Therefore, LSI is applied in the clustering-summarization benchmark system for this study.

3.5 Output of Clustering-Summarization

An output example of clustering-summarization is shown in Figure 3.6. As can be seen, the output is divided into clusters. These three clusters are non-overlapping, i.e. one paper belongs to only one cluster. This example shows the clustering-summarization output for 25 papers, with only cluster 1 presented in detail.

25 papers to be summarized

Cluster 1 (10 papers)

In the metal cutting, cutting oil is generally used for lubrication, cooling, chip disposal.

Dry, MQL cutting were carried out as a cutting mode for the comparison, and cutting force, tool wear, surface roughness, cutting mechanism such as the chip shape were compared and were examined.

As the result, the cutting force lowered in comparison with the dry-type cutting, and it was equivalent to the MQL cutting.

It is widely required not to use cutting oils containing surface reactive chlorine compounds in metal cutting for conservation of the global environment.

In the usual case of MQL cutting, the oil mist is supplied to the cutting area by external supply nozzles.

…

Cluster 2 (7 papers)

…

Cluster 3 (8 papers)

...

Figure 3.6 Output of clustering-summarization on 25 papers

3.6 Conclusion of the Chapter

This chapter has discussed the differences of MDS between the domain of news articles and technical papers in terms of readers' information requirements and document genres. Since existing MDS work mainly focused on the domain of news articles, this discussion is helpful to guide future researches on multi-paper summarization.

Moreover, the popular MDS framework of clustering-summarization has been implemented in multi-paper summarization. LSI was applied as the indexing scheme for clustering process and was demonstrated successful in achieving higher clustering accuracy. The optimal dimension of LSI in clustering process was also investigated. The clustering-summarization framework can be applied as a benchmark for further researches.

The output summary in Figure 3.6 reveals the two limitations of the clustering-summarization method which has been discussed in Chapter 2: it is hard to decide the number of clusters without the prior knowledge and the themes within a document collection are not perfectly distributed into non-overlapping clusters. This motivates the author to look into the textual structures within multiple technical papers and how these structures can help in summarization.

Chapter 4

Macrostructure and Microstructure within Multiple Documents

Through the preliminary investigation of multi-paper summarization, it was found that the clustering-summarization method did not address the special characteristics of the technical paper domain and could not reveal the internal structure of multiple documents, e.g. the topics within a set of documents are not easily distributed into non-overlapping clusters of documents. Therefore, it motivates the detailed investigation into the structures within multiple documents and how these structures can help in multi-document summarization.

In this chapter, qualitative analysis is conducted on the corpus of manual summaries, in order to find out the routines for human authors in writing summaries. Based on the analysis, the notions of macrostructure and microstructure are proposed and these two structures are believed to cover the most important information in summarizing multiple documents. This assumption is validated by further experiments. The document sets used in the analysis and experiments in this chapter are mainly chosen from the DUC corpus (http://duc.nist.gov/), due to its availability, popularity and credibility.

4.1 Analysis of DUC Corpus

Summarization can be treated as the process of extracting important information from input documents. The purpose of this analysis is to find out the routines for human authors in writing summaries, so that we can acquire a better understanding about how human authors define the "important information" in multi-document summarization.

A well-known challenge for summarization modeling and evaluation is that no single best or "gold standard" summary exists, which means, for a document collection there is often little consensus among summaries generated by different human authors, as reported by previous researches (Halteren and Teufel, 2003; Nenkova and Passonneau, 2004; Schlesinger et al., 2003). Halteren and Teufel (2003) reported that a stable consensus summary could only be expected if a large number of human-generated summaries were collected (at least 30-40 summaries). Their observation was based on summaries for single document. Other researchers also found that in the case of MDS, this problem still existed and was probably more acute due to the diversified topics and structures among documents (Schlesinger et al., 2003).

Nevertheless, previous researchers compared the overlap among summaries based on the word-match or sentence-match instead of on the structure-match (Halteren and Teufel, 2003; Nenkova and Passonneau, 2004). In this study, qualitative analysis was conducted to compare the overlap among summaries based on structure-match. The purpose was to investigate whether summaries very different in terms of words and sentences could still share a similar structure. If this is true, it will probably provide summarization research a guideline as to what a “gold standard” summary would be composed of.

4.1.1 DUC Corpus

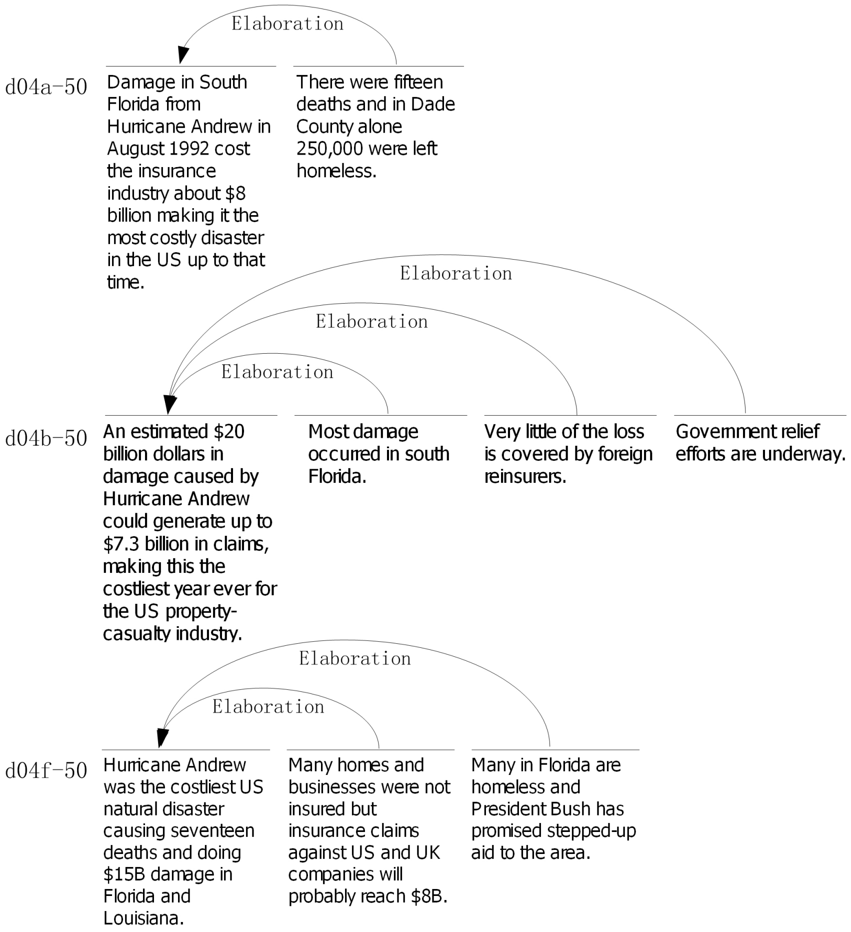
The analysis was based on the DUC corpus. The DUC corpus was built in the Document Understanding Conference (DUC) for summarization research and evaluation (http://duc.nist.gov/). It is one of the most popular benchmark in the summarization research community.

In our analysis, 30 document sets were chosen from the DUC-2001 corpus. Each document set contains ten documents on average (ranging from six to 16 documents per set). For every set, each of three professional authors wrote four summaries of length 50-word, 100-word, 200-word and 400-word respectively. The 400 word summary was produced first, and then a 200, 100, and 50 word summary produced using this summary (with references to the documents if necessary). Thus, there are 3 ×4=12 summaries for each of the 30 document sets.

4.1.2 Results of Analysis

Through the analysis of 30 document sets from DUC-2001, it was found that although there were great variations in different manual summaries at the word-level or sentence-level, they still followed a similar structure. It was observed that for most document sets, given a sufficient length, e.g. 200-word, summaries from different authors shared a similar structure. Therefore, it was believed that within any document set, there existed a macrostructure which could guide different authors to apply in their summaries.

It was noted that some sets were very cohesive, talking about one specific event, and their macrostructures could be easily identified. Such set shares a similar discourse structure among different manual summaries even with a very short length of 50-word, e.g. set d04 regarding the specific disaster of *Hurricane Andrew*. Figure 4.1 shows that the three summaries of 50-word have the similar discourse structure. They share the same meaning in their topical sentences (macrostructure-level information): *Hurricane Andrew was the costliest natural disaster in US up to that time*, followed by elaborating sentences (microstructure-level information). The discourse structures in these three summaries are largely consistent, although the elaboration sentences in different summaries emphasize different aspects of the disaster and are presented in different orders, mainly due to authors' own understanding of the topic and their preferences in generating the summary.



Damage in South Florida from Hurricane Andrew in August 1992 cost the insurance industry about $8 billion making it the most costly disaster in the US up to that time.

There were fifteen deaths and in Dade County alone 250,000 were left homeless.

An estimated $20 billion dollars in damage caused by Hurricane Andrew could generate up to $7.3 billion in claims, making this the costliest year ever for the US property-casualty industry.

Most damage occurred in south Florida.

Very little of the loss is covered by foreign reinsurers.

Government relief efforts are underway.

Hurricane Andrew was the costliest US natural disaster causing seventeen deaths and doing $15B damage in Florida and Louisiana.

Many homes and businesses were not insured but insurance claims against US and UK companies will probably reach $8B.

Many in Florida are homeless and President Bush has promised stepped-up aid to the area.

Figure 4.1 Discourse structures of three manual summaries (50-word) for a cohesive document set d04

However, some sets are not so cohesive, e.g. set d11 in DUC-2001, with eight articles talking about different aspects of *tornadoes* as follows:

- *Correct response when a tornado comes*
- *A series of tornadoes in Madison, Florida*
- *Tornadoes in 1988*
- *Some general facts about tornado*
- *A tornado in Huntsville, Alabama*
- *A series of tornadoes and severe thunderstorms including the one in Huntsville, Alabama*
- *Tornadoes in 1990*
- *Some research work of Professor Tetsuya Theodore about tornado*

As can be seen, in this set, some articles talk about the general facts of tornadoes or specific tornadoes, some focus on the correct response for a tornado, and lastly, one article introduces the research work of a professor about tornadoes. The topics of these articles vary greatly, making it difficult for any author to summarize them into a short paragraph of 50 words. Figure 4.2 shows that the three summaries with 50-word share little overlap and follow different structures.

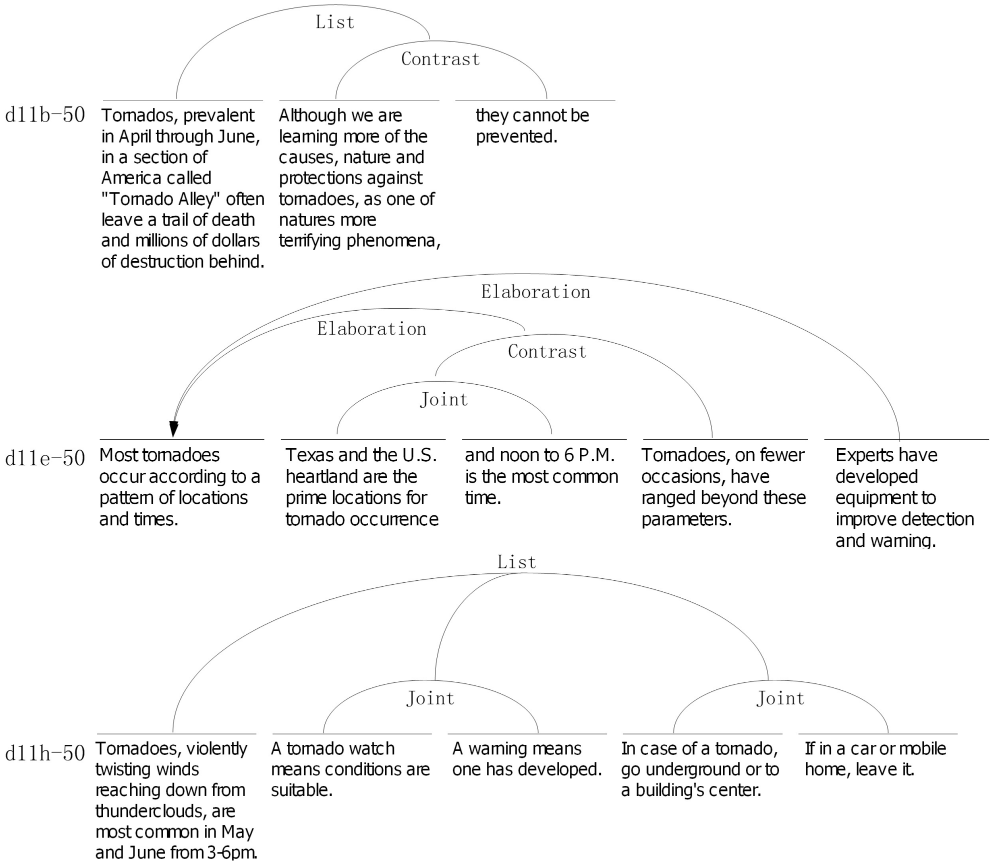

List
Contrast
d11b-50
Tornados, prevalent in April through June, in a section of America called "Tornado Alley" often leave a trail of death and millions of dollars of destruction behind.
Although we are learning more of the causes, nature and protections against tornadoes, as one of natures more terrifying phenomena,
they cannot be prevented.
Elaboration
Elaboration
Contrast
Joint
d11e-50
Most tornadoes occur according to a pattern of locations and times.
Texas and the U.S. heartland are the prime locations for tornado occurrence
and noon to 6 P.M. is the most common time.
Tornadoes, on fewer occasions, have ranged beyond these parameters.
Experts have developed equipment to improve detection and warning.
List
Joint
Joint
d11h-50
Tornadoes, violently twisting winds reaching down from thunderclouds, are most common in May and June from 3-6pm.
A tornado watch means conditions are suitable.
A warning means one has developed.
In case of a tornado, go underground or to a building's center.
If in a car or mobile home, leave it.

Figure 4.2 Discourse structures of three manual summaries (50-word) for a loose document set d11

However, it does not mean that the set d11 lacks a macrostructure. The length restriction (50-word) forced summary authors to exclude a lot of important contents and only select one or two topics. The selection process could be somewhat random. Therefore, for a set like d11, short summaries may not be consistent, even written by the same author at different times. However, when authors are allowed to include more words in summaries, making it possible for them to cover more topics, they may then show a similar structure in their summaries. As shown in Figure 4.3, when authors were given 200-word length restriction, the summaries all include the following three major topics, although they are emphasized differently, depending on authors' background knowledge, perspectives and composition techniques:

- *Some general facts about tornadoes, e.g. season, occurring places, etc.*
- *Fujita Scale measures for tornadoes proposed by Professor Tetsuya Theodore Fujita*
- *Safety plans before tornadoes strike and the meaning of tornado watch and warning*

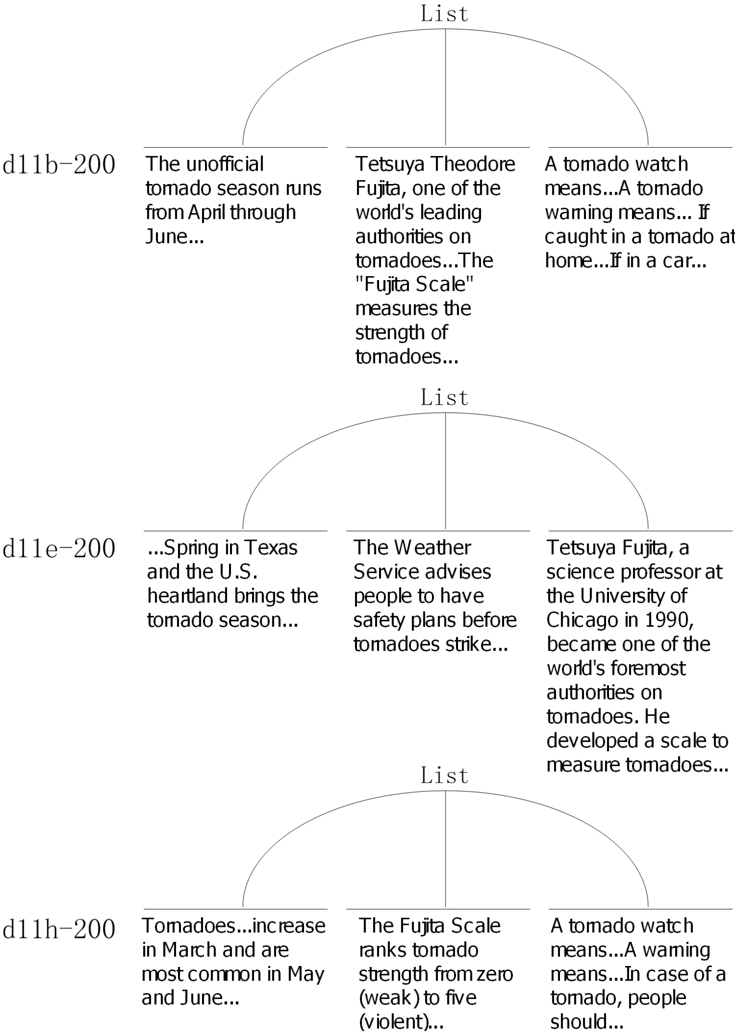

List
d11b-200
The unofficial tornado season runs from April through June...
Tetsuya Theodore Fujita, one of the world's leading authorities on tornadoes...The "Fujita Scale" measures the strength of tornadoes...
A tornado watch means...A tornado warning means... If caught in a tornado at home...If in a car...
List
d11e-200
...Spring in Texas and the U.S. heartland brings the tornado season...
The Weather Service advises people to have safety plans before tornadoes strike...
Tetsuya Fujita, a science professor at the University of Chicago in 1990, became one of the world's foremost authorities on tornadoes. He developed a scale to measure tornadoes...
List
d11h-200
Tornadoes...increase in March and are most common in May and June...
The Fujita Scale ranks tornado strength from zero (weak) to five (violent)...
A tornado watch means...A warning means...In case of a tornado, people should...

Figure 4.3 Discourse structures of three manual summaries (200-word) for document set d11

4.2 Textual Structures within Multiple Documents

Based on the qualitative analysis, we propose that the “important information” within a set of documents can be divided into two parts: macrostructure and microstructure.

- **Macrostructure** is defined as the significant topical information shared among input documents. This information can guide different summary authors to adopt a similar structure in their summaries.
- **Microstructure** consists of the sentences or clauses acting as the elaborating or complementary information for macrostructure. Based on microstructure, different summary authors might include different details to elaborate the topics in summaries due to their own background knowledge, composition skills and unique understanding of the input documents.

In traditional linguistics, macrostructure refers to structure in a single document and represents relations between blocks of sentences (Hutchins, 1987). This work actually extends the definition of macrostructure to a structure consisting of important topics across multiple documents and revealing the topic links in these documents.

We believe that macrostructure and microstructure represent the actual structure within multiple documents that will affect summarization modeling and evaluation. The analysis results of DUC corpus implied that macrostructure within documents could guide different authors to adopt a set of similar topics in their summaries. On the other hand, different authors might include different details to elaborate the topics in summaries due to their own background knowledge, composition skills and unique understanding of the input documents. It is worthwhile to point out that macrostructure is not always apparent enough to be identified, especially for document sets loosely structured like d11 in Figure 4.2. In such cases, different authors might not be able to generate consistent content in summaries of short length because they have to discard a lot of important information to comply with the length restriction. However, when the authors are given a more liberal length limits, e.g. 200 words, their summaries are more likely to achieve high agreement in terms of macrostructure, as can be seen in Figure 4.3.

4.3 Identification of Macrostructure and Microstructure

To automatically identify macrostructure, the important topics in a document set are extracted and ranked according to their significance. The sentences in which these topics appear are selected and comprise the candidate sentences for microstructure. The detailed approaches for macrostructure and microstructure identification are given as follows.

4.3.1 Macrostructure

Different approaches of topic identification have been reported in previous work (Choi, 2000; Clifton et al., 2004; Hearst, 1997; Moens and De Busser, 2001). The typical method for topic identification in single document is text segmentation, which is to segment the text by similarity of adjacent sentences and detect the boundary of subtopics (Choi, 2000; Hearst, 1997; Moens and De Busser, 2001; Ponte and Croft, 1997). A popular method for topic identification in multiple documents is text clustering, i.e. to split the whole set into several non-overlapping groups (Clifton et al., 2004; Radev et al., 2004). Each group of documents is assumed to discuss one topic. However, it is usually difficult a priori to determine the optimal number of clusters. Moreover, in a real-world document set, topics often overlap with each other across documents and are not perfectly distributed in non-overlapping groups of documents, as discussed in Chapter 3.

Our process of topic identification is similar to Liu (2005), using Frequent word Sequences (FSs) to handle concepts in text classification. A FS is a sequence of words that appears in at least σ documents in a document set (σ is the pre-specified threshold for supporting documents). Algorithm 4.1 demonstrates the process to extract all the FSs in a document set. The process starts with collecting all the frequent word pairs, i.e. FSs with two words. These FSs are then expanded with one more word and therefore form a set of word sequences with length three. All the FSs with length three are then expanded. This process is iteratively performed until there is no FS left for expansion. The threshold for supporting documents is chosen according to the size of the document set.

In order to reduce noisy information, pre-processing is performed before the document set is sent for topic identification. Particular steps include stop words removal and word stemming (Porter, 1980).

Algorithm 4.1 Discovery of all FSs in a set of documents

```
Input:   $D$: a set of pre-processed documents, $\sigma$: a frequency threshold
Output:  $Fs$: a set of frequent word sequences
    // Initial phase: collecting all frequent pairs
    For all the documents $d \in D$
        Collect all the ordered pairs and occurrence information within $d$
    End For
    $Seq_2$ = all the ordered word pairs that appear in at least $\sigma$ documents in $D$
    // Discovery phase: building longer word sequences
    $k := 2$
    $Fs := Seq_2$
    While $Seq_k$ is not void
        For all phrases $s \in Seq_k$
            Let $l$ be the length of the sequence $s$
            Find all the sequences $s'$ such that $s$ is a subsequence of $s'$ …
             and the length of $s'$ is $l$+1
            For all $s'$
                If $s'$ appears in at least $\sigma$ documents in $D$
                    $S := S \cup \{s'\}$
            End For
            $Fs := Fs \cup S$
            $Seq_{k+1} := Seq_{k+1} \cup S$
        End For
        $k := k + 1$
    End While
    Return $Fs$
```

A FS is considered as the representative of one topic in a document set. Topics are ranked based on their scores. In our experiment, the score of a topic is calculated in three forms, as given in Table 4.1.

Table 4.1 Scoring schemes of topics

Scoring scheme	Description
tf	f
tf.df	$f/\log_2 \frac{D+1}{d}$
tf.idf	$f \cdot \log_2 \frac{D+1}{d}$

where f is the frequency of the topic in the document set, D is the total number of documents in the set, d is the number of documents in which the topic occurs. These scoring schemes are widely used in information retrieval and text mining as statistical measures to evaluate how important a word or a phrase is to a document in a document set (Salton and Buckley, 1988).

Topics are ranked based on the three scoring schemes in Table 4.1. The top ranked topics constitute the macrostructure for a document set. The performance of these three scoring schemes will be compared in the following experiments.

4.3.2 Microstructure

As has been defined, microstructure consists of the information that acts as the elaborating and complementary parts for macrostructure. Therefore, we highlight the sentences in which the top ranked significant topics appear. Sentences of microstructure will then be selected from these highlighted sentences. The method of Maximal Marginal Relevance (MMR) (Carbonell and Goldstein, 1998) is applied in the sentence selection process in order to reduce the redundancy. MMR intends to balance the trade-off between the centrality of a sentence with respect to the topic and its novelty compared to the sentences already selected, which is actually to maximize the marginal relevance in the following form:

$$MR(s_i) = Sim(s_i, C) - \max_{s_j \in S} Sim(s_i, s_j) \tag{4.1}$$

where *C* is the set of all candidate sentences to be selected, *S* is the set of sentences already included in the microstructure. With regard to *Sim*, we adopt a cosine similarity measure between sentence vectors. As defined in Equation 3.1, cosine similarity is a measure of similarity between two vectors of *n* dimensions by finding the angle between them. It is often used to measure the similarity of documents, paragraphs and sentences in text mining. Each element of a sentence vector represents the weight, i.e. appearing frequency, of a word-stem in the sentence after removing stop words.

After calculating the marginal relevance in Equation 4.1 for all candidate sentences, the sentence s_i with the highest marginal relevance will be included in the microstructure. This sentence selection process iterates until the expected number of sentences is reached.

4.4 Influence of Macrostructure and Microstructure on MDS

In order to investigate the influence of macrostructure and microstructure on MDS process and summarization performance, two experiments were performed in this study.

The first experiment focused on the influence of macrostructure on the manual MDS. As we have surmised, the macrostructure within multiple input documents could guide different human summarizers to apply a similar structure in their summaries. In this experiment, we intended to quantitatively measure the influence of macrostructure on manual summarization. Specifically, we examined the association between the significance of a macrostructure-level topic and the likelihood that this topic would be selected by a human summarizer for inclusion in the summary. Moreover, we investigated the inter-agreement among human summarizers in terms of applying macrostructure in summaries.

In the second experiment, we proposed a summarization evaluation framework based on macrostructure- and microstructure-level information. This evaluation framework was compared to the existing summarization evaluation methods like ROUGE (Lin, 2004) and Pyramid (Nenkova and Passonneau, 2004). We intended to examine whether using macrostructure and microstructure in summarization evaluation can generate consistent results with existing evaluation methods based on human-generated summaries. Moreover, we investigated the influence of different proportions of macrostructure and microstructure.

4.4.1 Experiment 1: Consensus on Macrostructure from Different Human Summarizers

The purpose of this experiment was to investigate whether human summarizers relied on macrostructure in the process of manual MDS and to examine the inter-agreement among human summarizers in terms of applying macrostructure. Specifically, we focused on the following two questions:

- Are more significant topics in input documents more likely to appear in a manual summary?
- Do different human summarizers have more agreement in terms of applying more significant topics in their summaries?

The data used in this experiment were 30 document sets from the DUC-2001 corpus. There are ten documents on average in each document set. For each document set, each of three summary authors wrote four summaries with 50-word, 100-word, 200-word and 400-word. The 400-word summaries were produced first, and then 200, 100, and 50 word-summaries were produced based on 400-word summaries (with references to the input documents if necessary). The detailed description of this corpus can be found at http://duc.nist.gov/.

For each document set, topics were extracted and ranked based on *tf*, *tf.df, tf.idf*. Table 4.2 shows the percentage of *N* top ranked topics from the input documents that appear in at least one of the three 400-word manual summaries across the 30 document sets, for *N* = 1, 5, 10, …, 30. Table 4.3 shows the average number of summarizers that agree with each other in choosing the *N* top ranked topics across the 30 document sets.

Table 4.2 Percentage of the top ranked topics that appear in at least one of the three 400-word manual summaries (average across 30 document sets)

	tf	*tf.df*	*tf.idf*
1 top ranked topic	96.7%	96.7%	83.3%
5 top ranked topics	89.3%	90.0%	78.0%
10 top ranked topics	81.7%	81.7%	67.3%
15 top ranked topics	78.2%	78.9%	65.3%
20 top ranked topics	75.7%	74.5%	64.3%
25 top ranked topics	71.2%	70.7%	62.0%
30 top ranked topics	67.9%	67.7%	60.9%

Table 4.3 Average number of summarizers (out of three) that agree with each other in choosing the top ranked topics across 30 document sets (400-word summaries)

	tf	*tf.df*	*tf.idf*
1 top ranked topic	2.67	2.60	1.83
5 top ranked topics	2.24	2.26	1.83
10 top ranked topics	2.01	1.99	1.58
15 top ranked topics	1.84	1.82	1.49
20 top ranked topics	1.73	1.70	1.46
25 top ranked topics	1.61	1.61	1.40
30 top ranked topics	1.55	1.54	1.36

Two significant observations can be made from Table 4.2 and 4.3:

- The highly ranked topics from the input documents are very likely to appear in the manual summaries. The more highly ranked a topic is in the input documents, the more likely it will appear in a manual summary and the more likely it will be agreed by different human summarizers. For example, the tables show that, across the 30 sets, 89.3% of the top five topics ranked by *tf* in the input documents were used in at least one of the three summaries and averagely there are 2.24 summarizers out of three agreed with each other by using these topics. However, for the 20 top ranked topics, the percentage of these topics appearing in the manual summaries reduced to 75.7% and only 1.73 out of three summarizers agreed with each other by adopting these topics.

- The ranking schemes *tf* and *tf.df* achieved better performance than *tf.idf*. The possible reason is that *tf.idf* tends to give high scores to the topics which differentiate one document from others. However, in MDS, the human authors mainly focus on the commonalities among documents, i.e. they tend to use those topics that appear frequently across documents. Therefore, using *tf.df* or *tf* to rank topics can achieve better performance.

Next, we fixed the ranking scheme as *tf* and investigated the appearance of topics in summaries with different lengths (400-word, 200-word, 100-word and 50-word). The results are presented in Tables 4.4 and 4.5 and Figures 4.4 and 4.5.

Tables 4.4 and 4.5 and Figures 4.4 and 4.5 clearly show that when summary authors are allowed more words for summarization, they are able to include more significant topics in their summaries. Moreover, more agreement is achieved among different summary authors for longer summaries. For example, on average 89.3% of the five top ranked topics appeared in at least one of the three 400-word summaries. This appearance frequency decreased to 56% for the summaries of 50-word. The average number of summary authors that agreed with each other in choosing the top five topics increased from 1.22 to 2.24 when summary length increased from 50-word to 400-word. This finding proved our supposition in Section 4.1 that, given more summary length, summary authors were able to include more significant topics in their summaries and they could agree more in choosing the significant topics.

Table 4.4 Percentage of the topics that appear in at least one of the three manual summaries (average across 30 document sets, topics are ranked by tf)

	400-word	200-word	100-word	50-word
1 top ranked topic	96.7%	83.3%	73.3%	70.0%
5 top ranked topics	89.3%	74.7%	60.0%	56.0%
10 top ranked topics	81.7%	66.7%	53.3%	45.3%
15 top ranked topics	78.2%	62.4%	49.1%	40.0%
20 top ranked topics	75.7%	58.7%	45.5%	35.7%
25 top ranked topics	71.2%	54.5%	41.6%	32.7%
30 top ranked topics	67.9%	52.1%	38.7%	30.6%

Table 4.5 Average number of summarizers that agree with each other in choosing the topics across 30 sets (topics are ranked by tf)

	400-word	200-word	100-word	50-word
1 top ranked topic	2.67	2.23	1.90	1.67
5 top ranked topics	2.24	1.79	1.41	1.22
10 top ranked topics	2.01	1.56	1.22	1.02
15 top ranked topics	1.84	1.42	1.13	0.93
20 top ranked topics	1.73	1.32	1.05	0.86
25 top ranked topics	1.61	1.23	0.98	0.81
30 top ranked topics	1.55	1.19	0.93	0.78

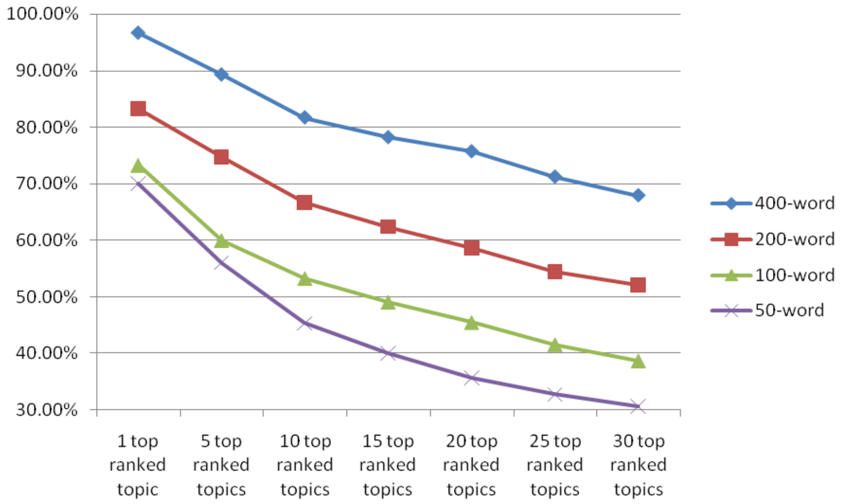


Figure 4.4 Percentage of the topics that appear in at least one of the three manual summaries (average across 30 document sets, topics are ranked by tf)

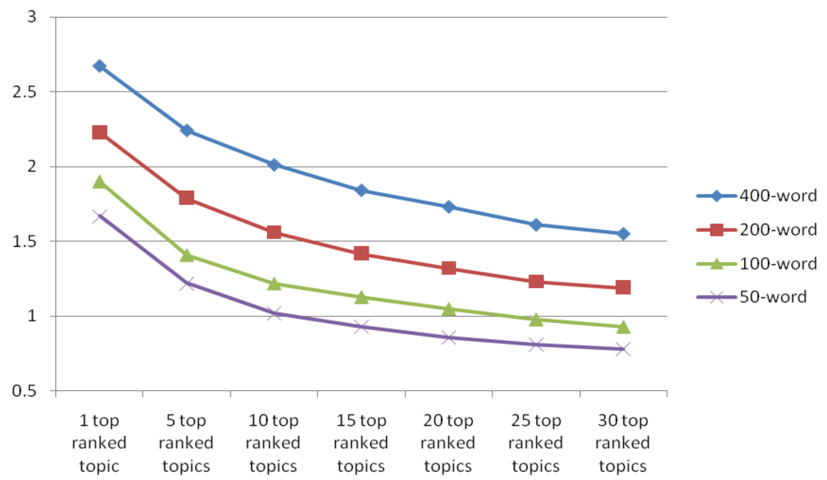


Figure 4.5 Average number of summarizers that agree with each other in choosing the topics across 30 document sets (topics are ranked by tf)

These results demonstrate that macrostructure extracted by the method of FSs is one of the important factors that influence the manual summarization process.

4.4.2 Experiment 2: Influence of Macrostructure and Microstructure on Summarization Performance

Summarization evaluation serves for a twofold purpose: to offer a benchmark of measuring summarization performance, and to provide clues of discovering important elements that would affect summarization performance. In this experiment, we built a summarization evaluation framework based on macrostructure- and microstructure-level information. Through comparison with existing evaluation methods, we intended to investigate the contribution of macrostructure and microstructure in the summarization performance.

Existing methods to evaluate summarization performance, like ROUGE and Pyramid, compare candidate summaries with one or more human-generated reference summaries (Jing et al., 1998; Lin, 2004; Nenkova and Passonneau, 2004). The general scheme of these methods is shown in Figure 4.6. Candidate and reference summaries are first fragmented into elements to be compared. The overlap between candidate and reference summaries with regards to the elements is recorded as the score of the candidate summary.

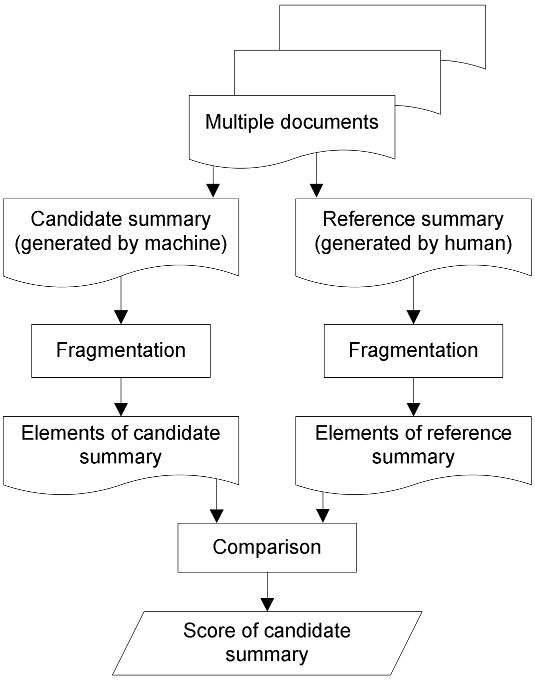

Multiple documents
Candidate summary (generated by machine)
Reference summary (generated by human)
Fragmentation
Fragmentation
Elements of candidate summary
Elements of reference summary
Comparison
Score of candidate summary

Figure 4.6 General framework for existing summarization evaluation methods

Instead of comparing candidate summary with manual summary, we built an evaluation framework based on Macrostructure- and Microstructure-level Information (MMI), as shown in Figure 4.7. In this framework, macrostructure-level information (topics) and microstructure-level information (elaboration sentences) are extracted from input documents and used to evaluate candidate summaries. The score of a candidate summary is a linear combination of two parts, macro- and micro-score, with a parameter λ ranging from 0 to 1, as defined in Equation 4.2. The macro-score indicates how much macrostructure-level information is covered by a candidate summary, i.e. how many significant topics appear in the candidate summary. Topics are ranked based on *tf* ranking scheme which achieved good performance in Experiment 1. The micro-score calculates the similarity between the candidate summary and the sentences that constitute microstructure-level information.

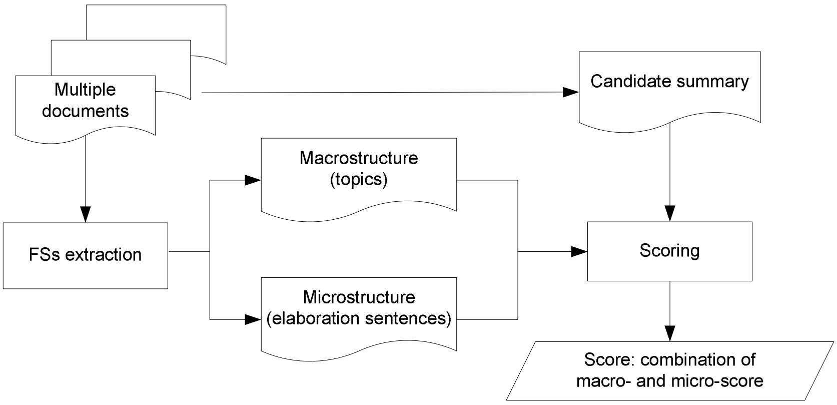

Multiple documents
Candidate summary
Macrostructure (topics)
FSs extraction
Scoring
Microstructure (elaboration sentences)
Score: combination of macro- and micro-score

Figure 4.7 Summarization evaluation based on Macrostructure- and Microstructure-level Information (MMI)

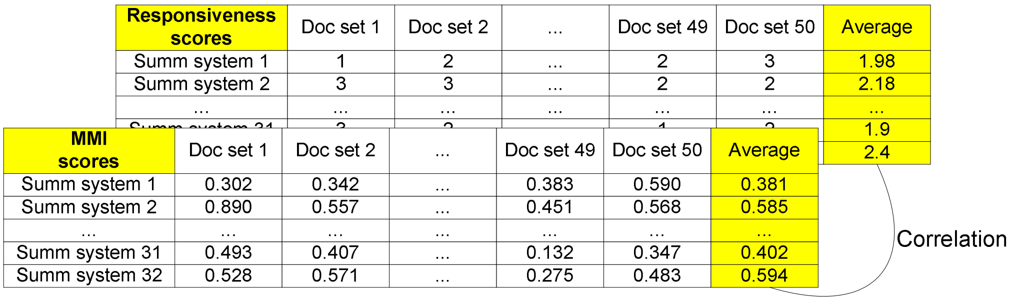
The parameter λ in Equation 4.2 is to tune the weights of macro- and micro-score in the total score so that we can investigate the contributions of macrostructure and microstructure information in summarization performance. When λ is set to 0, the total score is equal to micro-score; when λ is set to 1, only the macro-score affects the total score.

$$\text{Score} = \lambda \cdot \text{Macro\_score} + (1-\lambda) \cdot \text{Micro\_score} \tag{4.2}$$

We applied MMI framework to evaluate summaries in the DUC corpus. The purpose here was to investigate the influence of macrostructure- and microstructure-level information in the summarization performance and whether these two levels of information contributed as the important factors in MDS process. We computed the correlation between MMI assigned summary scores and summary scores given by human assessors (responsiveness scores). The responsiveness scores assigned by human assessors were used here as benchmark.

The data used in this experiment were 50 document sets from the DUC-2005 corpus. For each of the 50 document sets, there are 32 multi-document summaries which were generated by 32 summarization systems. Responsiveness scores assigned by human assessors are available for each of the 50×32 summaries. We computed the correlation between 32 systems' average MMI scores and their average responsiveness scores across all the 50 document sets, as shown in Figure 4.8. Pearson's product-moment correlation coefficient (Edwards, 1976), one of the most popular correlation coefficients, was used to computer the correlation in this experiment. Given two random variables *X* and *Y*, Pearson's product-moment correlation coefficient is obtained by dividing the covariance of the two variables by the product of their standard deviations, as defined in Equation 4.3.

$$\rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \cdot \sigma_Y} = \frac{E((X-\mu_X)(Y-\mu_Y))}{\sigma_X \cdot \sigma_Y} \tag{4.3}$$



Responsiveness scores	Doc set 1	Doc set 2	...	Doc set 49	Doc set 50	Average
Summ system 1	1	2	...	2	3	1.98
Summ system 2	3	3	...	2	2	2.18
...	...	...	...	...	...	...
Summ system 31	3	2		1	2	1.9
						2.4

MMI scores	Doc set 1	Doc set 2	...	Doc set 49	Doc set 50	Average
Summ system 1	0.302	0.342	...	0.383	0.590	0.381
Summ system 2	0.890	0.557	...	0.451	0.568	0.585
...	...	...	...	...	...	...
Summ system 31	0.493	0.407	...	0.132	0.347	0.402
Summ system 32	0.528	0.571	...	0.275	0.483	0.594

Figure 4.8 Correlation between MMI scores and responsiveness scores (The scores in this figure are only shown as examples)

To examine the influence of number of topics (macrostructure) in MMI performance, we built the MMI framework with 10, 20, 50, 100, 150 top ranked topics and all topics respectively. Moreover, we ranged the parameter λ from 0 to 1 in order to investigate the contributions of the two parts in Equation 2. Therefore, we had six (10_, 20_, 50_, 100_, 150_, all_topics) × 7 (λ = 1, 0.8, 0.6, 0.5, 0.4, 0.2, 0) = 42 variations of MMI evaluation, as given in Table 4.6. Table 4.6 and Figure 4.9 show the correlations between the 42 variations of MMI and responsiveness scores in evaluating DUC-2005 summaries. For the purpose of comparison, the correlations between traditional evaluation methods (ROUGE, Pyramid) and responsiveness scores are listed in Table 4.7.

Table 4.6 Correlations between MMI scores (42 variations) and responsiveness scores

	λ						
	1	0.8	0.6	0.5	0.4	0.2	0
10_topics	0.804	0.863	0.884	0.884	0.876	0.836	0.771
20_topics	0.818	0.865	0.888	0.891	0.885	0.852	0.793
50_topics	0.846	0.876	0.909	**0.915**	0.911	0.874	0.806
100_topics	0.835	0.871	0.896	0.896	0.886	0.837	0.759
150_topics	0.829	0.869	0.890	0.888	0.875	0.819	0.736
All_topics	0.825	0.862	0.870	0.857	0.832	0.750	0.644

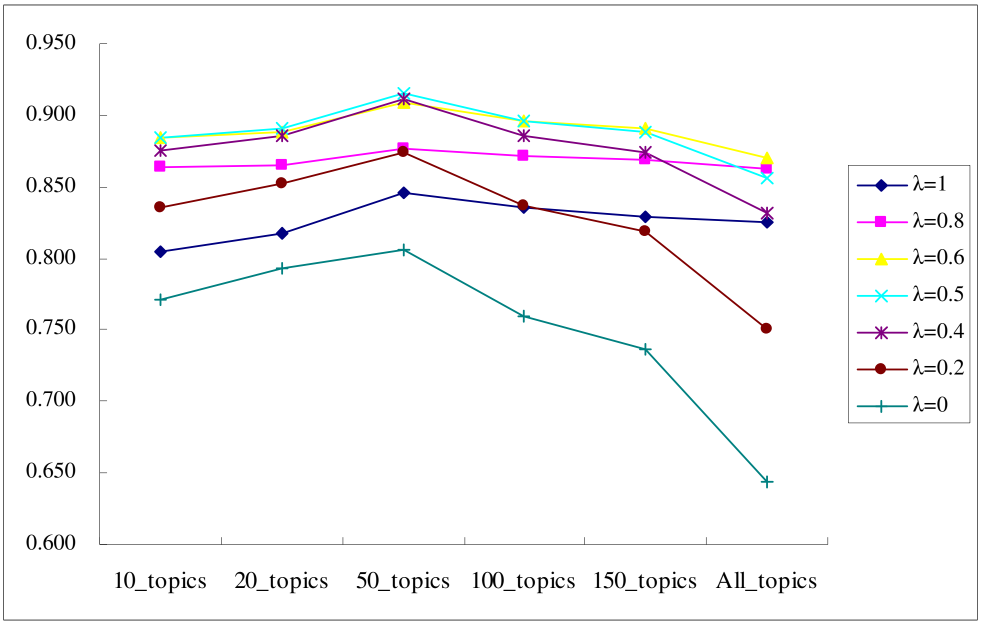


Figure 4.9 Correlations between MMI scores and responsiveness scores

Table 4.7 Correlations between existing evaluation methods (ROUGE, Pyramid) and responsiveness scores

	Correlation with Responsiveness
ROUGE-1	**0.926**
ROUGE-2	**0.910**
ROUGE-3	0.817
ROUGE-4	0.714
ROUGE-L	0.889
ROUGE-SU4	**0.922**
ROUGE-W-1.2	0.891
Pyramid	0.829

The following significant findings can be made based on the experimental results:

- Macrostructure and microstructure constitutes the important information in evaluating summarization performance. As can be seen in Table 4.6 and Figure 4.9, summary scores assigned by MMI were highly correlated with responsiveness scores assigned by human assessors. The highest correlation achieved by MMI (50_topics, λ=0.5) and human assessors reached 0.915, which was comparable with the highest correlation achieved by ROUGE and responsiveness scores (0.926, as shown in Table 4.7).

- Information from macrostructure and microstructure are both important in the summarization process. Microstructure offer complementary information for macrostructure. As shown in Figure 4.9, the performance of MMI evaluation was not good when only macrostructure-level information (λ=1) or microstructure-level information (λ=0) was considered. The best performance of MMI evaluation was achieved when λ was set to between 0.5 and 0.6.

- It was also found that the number of topics can influence the quality of macrostructure and microstructure. As shown in Figure 4.9, including too few or too many topics would both decrease the performance of MMI evaluation. The best performance was achieved when 50 topics were chosen. This finding suggests that the number of topics should be appropriately chosen in order to achieve the optimal performance of macrostructure and microstructure in summarization modeling and evaluation.

4.5 Conclusion of the Chapter

A well known challenge for MDS is that there does not exist a single best or “gold standard” summary, i.e. there is often little consensus among reference summaries written by different authors for a same document set. Through analysis of DUC corpus, it was found that although different manual summaries varied a lot in terms of words or sentences, they might still follow a similar structure.

Based on the analysis, the notions of macrostructure and microstructure were proposed. Macrostructure is defined as the significant topics shared among different input documents, while microstructure is defined as sentences or clauses that act as elaborating or complementary information for macrostructure.

Two experiments were conducted to examine the influence of macrostructure and microstructure on summarization performance. The first experiment demonstrated that human summarizers heavily relied on the macrostructure in writing their summaries. The more significant topics from the input documents are more likely to appear in the manual summaries and more likely to be agreed by different human summarizers. The second experiment suggested that microstructure offered complementary information for macrostructure and the two structures constitute the important information in summarization modeling and evaluation.

This thesis focuses on summarization in the domain of technical papers. However, the DUC corpus, which is composed of news articles, is applied in this chapter, because it is a standard corpus widely used in existing summarization research (Lin, 2004; Moens et al., 2005; Nenkova and Passonneau, 2004). Therefore, the macrostructure and microstructure proposed in this chapter can be applied to a generic domain of textual documents.

Macrostructure and microstructure can better represent the actual relationship among multiple documents than clusters discussed in the previous chapter. The discussion of macrostructure and microstructure in this chapter is based on the general corpus of DUC due to its popularity and credibility. In the next chapter, the domain of technical papers will be focused on and the issues of applying macrostructure & microstructure in multi-paper summarization will be examined.

Chapter 5

Multi-Paper Summarization Based on Macrostructure and Microstructure

Compared to clustering structure, macrostructure and microstructure proposed in the previous chapter can better represent the actual relationship among multiple documents. Moreover, the information from macrostructure- and microstructure-level has been demonstrated to have great influence on the MDS performance. The discussion and experimentation of macrostructure and microstructure in the previous chapter were based on the general corpus of DUC. This chapter focuses on summarization in the domain of technical papers and the issues of applying macrostructure & microstructure in multi-paper summarization are examined.

5.1 Summarization Based on Structure Analysis

There exist a few studies which implemented summarization based on the structure analysis of input documents. Most of these studies focused on summarization of a single document and the three major methods used (discourse structure, lexical chains and text segmentation) are reviewed in this section. In terms of multi-document summarization, unfortunately, structure analysis has been largely ignored, mainly due to the diversified topics and structures across multiple documents.

5.1.1 Structure Analysis in Single-Document Summarization

Some existing studies of single-document summarization have applied structure analysis in summarization process. Three typical methods, discourse structure (Hobbs, 1993; Marcu, 1999; Polanyi, 1993), lexical chains (Barzilay and Elhadad, 1997) and text segmentation (Hearst, 1997), are briefly reviewed here.

5.1.1.1 Discourse Structure

Discourse structure analysis was driven mostly by research in natural language processing and generation (Mann and Thompson, 1988; Marcu, 1997). Central to the theory is the notion of discourse relation, which is a relation that holds between two non-overlapping text spans called nucleus and satellite. Some relations between nucleus and satellite are listed in Table 5.1. The distinction between nucleus and satellite is that the nucleus expresses what is more essential to the writer's purpose than the satellite; and that the nucleus of a discourse relation is comprehensible independent of the satellite, but not vice versa. Satellite can be viewed as the complementary information for nucleus.

Table 5.1 Relations between nucleus and satellite

RELATION	NUCLEUS	SATELLITE
Background	Text whose understanding is being facilitated	Text for facilitating understanding
Elaboration	Basic information	Additional information
Evidence	A claim	Information intended to increase the reader's belief in the claim
Interpretation	A situation	An interpretation of the situation

Discourse relations can also hold between equally important text spans, i.e. relations between multi-nuclei, as shown in Table 5.2. The full list of discourse relations can be found at http://www.sfu.ca/rst/index.html.

Table 5.2 Relations between multi-nuclei

RELATION	NUCLEUS	NUCLEUS
Contrast	One alternate	The other alternate
List	An item	A next item

For an article, discourse relations can be extracted and assembled into discourse structure trees by recursively applying individual relations to spans that range in size from one clause-like unit to the whole text. For example, the text shown in Figure 5.1 can be first broken into ten elementary units which are surrounded by square brackets. The cue phrases shown in italics in Figure 5.1 can help discourse parsing algorithm (Marcu, 1997) to hypothesize the discourse relations among these elementary units, e.g. the word *because* in unit 6 indicate that this unit may act as a satellite unit to provide the cause for its nucleus unit 5. By recursively integrating relations among the text spans, the discourse structure tree of this text can be generated as Figure 5.2.

[1 *With* its distant orbit 50 percent farther from the sun than Earth and slim atmospheric blanket,] [2 Mars experiences frigid weather conditions.] [3 Surface temperatures typically average about -60 degrees Celsius (-76 degrees Fahrenheit) at the equator and can dip to -123 degrees C near the poles.] [4 Only the midday sun at tropical latitudes is warm enough to thaw ice on occasion,] [5 *but* any liquid water formed in this way would evaporate almost instantly] [6 *because* of the low atmospheric pressure.]

Figure 5.1 Example text for discourse structure analysis

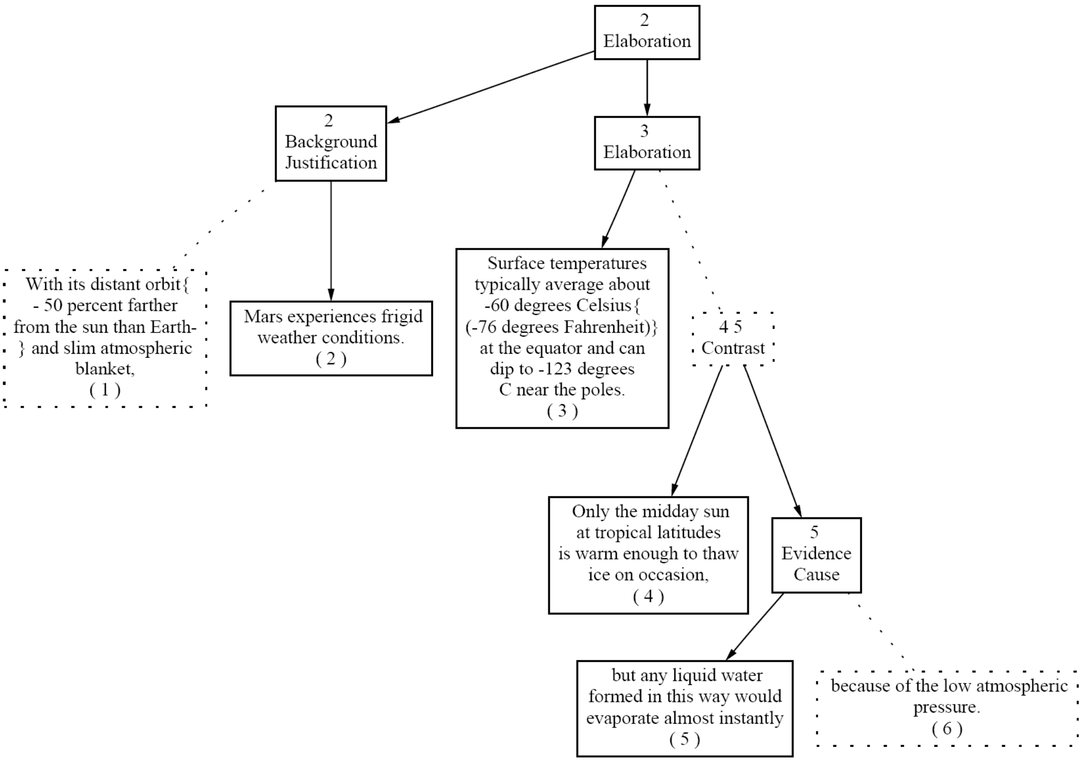

2
Elaboration
2
Background
Justification
3
Elaboration
With its distant orbit{ - 50 percent farther from the sun than Earth- } and slim atmospheric blanket,
(1)
Mars experiences frigid weather conditions.
(2)
Surface temperatures typically average about -60 degrees Celsius{ (-76 degrees Fahrenheit)} at the equator and can dip to -123 degrees C near the poles.
(3)
4 5
Contrast
Only the midday sun at tropical latitudes is warm enough to thaw ice on occasion,
(4)
5
Evidence
Cause
but any liquid water formed in this way would evaporate almost instantly
(5)
because of the low atmospheric pressure.
(6)

Figure 5.2 Discourse structure tree for text in Figure 5.1

The importance of the text units can be estimated based on the discourse structure tree. For example, as shown in Figure 5.2, unit 2 is the most important text unit because it is the only unit associated with the root node. Similarly, it can be determined that unit 3 is the most important unit of the span [3-6] and that units 4 and 5 are the most important units of span [4-6]. Therefore, the text units in the discourse structure tree can be ranked based on their importance and the summarization can be performed by selecting the top ranked text units.

5.1.1.2 Lexical Chains

Lexical chain is a sequence of related words spanning a topical unit of the text (Morris and Hirst, 1991). The words are grouped together by relationships of repetition, synonym, hypernymy (the semantic relation of being superordinate or generic, e.g. *plant* is a hypernym of *flower* and *tree*), antonymy and holonymy (the semantic relation that holds between a whole and its parts, e.g. *body* is a holonym of *arm* and *leg*), etc. These semantic relations can be derived from the WordNet thesaurus (Miller, 1995). Morris and Hirst (1991) argued that lexical chains may be useful in identifying topical segments in text.

Barzilay and Elhadad (1997) applied lexical chains in automatic text summarization. Due to the high degree of polysemy of English words, there are usually many candidate chains for one text. In the method of Barzilay and Elhadad, the best chain among all candidate chains is chosen based on the number and weight of different relations in the chain. Sentences are then extracted from chains based on a variety of heuristics, such as the frequency in the document of members of the chain.

5.1.1.3 Text Segmentation

Segmenting a text into topical regions is an important way to discover the structure of text and has been widely applied into single-document summarization (Choi, 2000; Hearst, 1997; Moens and De Busser, 2001; Ponte and Croft, 1997). The typical work of text segmentation was done by Hearst (1997). She compared blocks of text based on vocabulary overlap to identify topic boundaries. Her TextTiling algorithm divides a document into fixed-length text segments, e.g. 20 words. Adjacent blocks of segments (each block being, say, six segments long) are compared for similarity based on a vocabulary overlap measure. The system then assigns topic boundaries to the gaps between blocks with the sharp similarity change.

5.1.2 Structure Analysis in Multi-Document Summarization

Unlike single-document summarization, in MDS the structure of documents has been largely ignored in the previous studies, although there have been some initial investigations into the structures across multiple documents.

Radev (2000) proposed the CST which was a taxonomy of the information relationships among related documents. CST relations can be at document-level, passage-level, phrase-level and word-level. Typical relations include identity, subsumption, contradiction, elaboration, etc.

CST has not yet been widely applied in MDS, mainly due to the difficulty for automatic identification of the CST relations and the difficulty for linking CST relations with sentence selection. As compared to CST, macrostructure and microstructure described in the previous chapter are generated by statistical methods and therefore are easier to implement. Moreover, they have shown great influence on MDS performance. Therefore, in this chapter, a summarization framework based on macrostructure and microstructure is proposed. This summarization framework focuses on multi-paper summarization. Therefore, the special characteristics of technical paper domain will be addressed.

5.2 Multi-Paper Summarization Based on Textual Structures

As has been demonstrated in the previous chapter, macrostructure and microstructure constitute important information across multiple documents and greatly influence the MDS performance. Therefore, a multi-paper summarization framework is proposed based on macrostructure and microstructure, as shown in Figure 5.3.

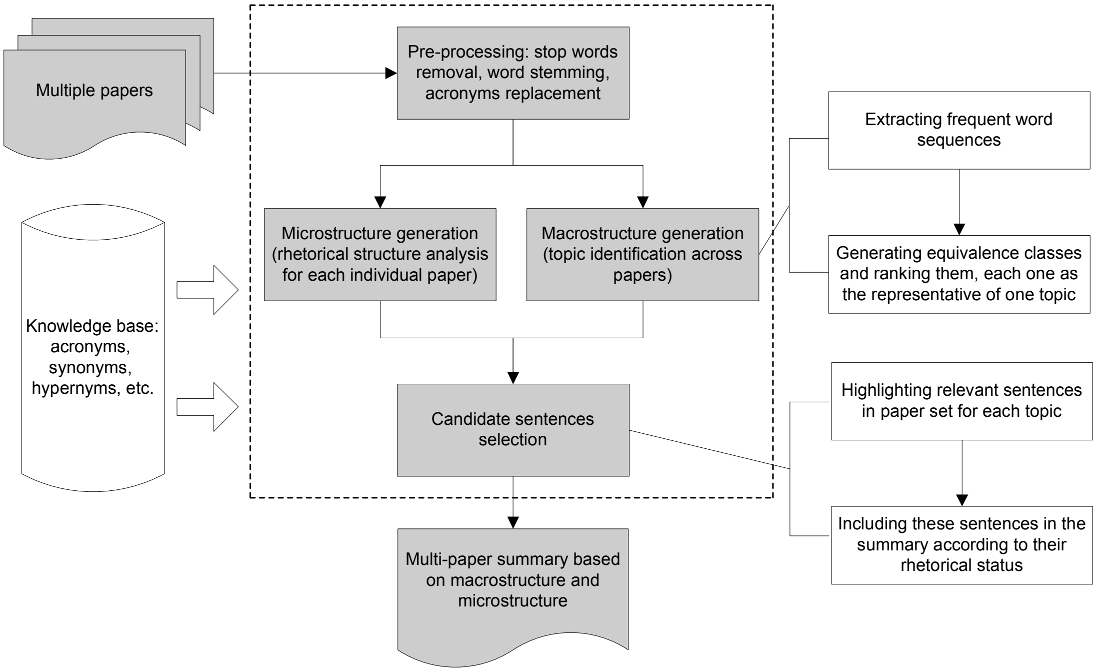

Multiple papers
Pre-processing: stop words removal, word stemming, acronyms replacement
Extracting frequent word sequences
Microstructure generation (rhetorical structure analysis for each individual paper)
Macrostructure generation (topic identification across papers)
Generating equivalence classes and ranking them, each one as the representative of one topic
Knowledge base: acronyms, synonyms, hypernyms, etc.
Highlighting relevant sentences in paper set for each topic
Candidate sentences selection
Including these sentences in the summary according to their rhetorical status
Multi-paper summary based on macrostructure and microstructure

Figure 5.3 Multi-paper summarization based on macrostructure and microstructure

The multi-paper summarization process starts with a set of technical papers as the input, followed by the pre-processing steps including stop words removal, word stemming, acronyms identification and replacement. Macrostructure and microstructure are then generated based on the pre-processed documents. In the next, candidate sentences are selected based on macrostructure and microstructure, and composed into summary. The detailed steps and methods are given as follows.

5.3 Macrostructure within Multiple Papers

As has been defined, macrostructure consists of significant topics across multiple documents. Our method of topic identification is based on Frequent word Sequences (FSs), as shown in Algorithm 4.1. In the rest of this chapter, an example document set *computer integrated manufacturing* is used to illustrate the algorithms and outputs in the summarization system. This document set contains 29 articles about the topic *computer integrated manufacturing*.

5.3.1 Topic Identification: FSs and Equivalence Classes

A comparison to this study is the work done by Yap et al. (2006) in which the authors utilized Maximal Frequent word Sequences (MFSs) (Ahonen, 1999) in topic identification. A MFS is defined as a FS which is not contained in any other longer FS. In our system, all FSs are considered as candidates for topical phrases instead of only using MFSs, since we intend to cover concepts at different levels. For example, for a given threshold of two supporting documents, *complex computer integrated manufacturing* is a MFS, which is not contained in any other longer sequences, as shown in Figure 5.4. Thus its subsequence *computer integrated manufacturing* will be removed from MFSs list. *Complex computer integrated manufacturing* only occurs twice in the document set while *computer integrated manufacturing* represents a more general concept which occurs in many more articles. If only MFSs are considered, some more general but still important concepts, like *computer integrated* *manufacturing*, will be discarded. Therefore, in our method, phrases at different concept levels are all included for consideration, i.e. topics are extracted based on all FSs.

1. In **computer integrated manufacturing** environments, dependability is a crucial attribute for the production management and control information system, which should be carefully assessed during system design.
2. The configuration design of **complex computer integrated manufacturing** systems such as semiconductor wafer fabricaton plants is a multi-objective, multi-criterion design problem.
3. A key component of **computer integrated manufacturing** is computer aided process planning.
4. The recent developments in **computer integrated manufacturing** systems have made the traditional dimensional inspections bottlenecks in the production line.
5. The fundamental concept of **computer integrated manufacturing** is to integrate the information flow, material flow and control flow of the whole manufacturing enterprise.
6. Specific emphasis is given to the company's most **complex computer integrated manufacturing** venture, and the difficulties, lessons, limitations and benefits gained from AMT.

Figure 5.4 FS and MFS: both “computer integrated manufacturing” and “complex computer integrated manufacturing” are FSs, but only “complex computer integrated manufacturing” is MFS

After all FSs are extracted, they will be grouped into equivalence classes (Ahonen, 1999; Yap et al., 2006) according to their co-occurrences with each other. The purpose is to reduce the redundancy within macrostructure. Some FSs which are not informative, e.g. *paper prove*, *in summary* and *for example*, will be removed first since they do not clearly indicate their relation with a particular topic. All candidate FSs which appear in the same set of papers will be grouped into one equivalence class. Figure 5.5 demonstrates four equivalence classes extracted from the paper set c*omputer integrated manufacturing* and the supporting papers for them.

EQV CLASS 1	EQV CLASS 2	EQV CLASS 3	EQV CLASS 4
(D_5, D_{10}, D_{15}, D_{20}):	(D_1, D_2, D_{25}):	(D_5, D_{16}, D_{23}):	(D_7, D_{12}, D_{13}):
small medium	autom guid vehicl	aid process plan	materi requir plan
small medium size	guid vehicl	aid process	materi requir
medium size	autom guid	comput aid process	requir plan
		comput aid process plan	
		process plan system	

Figure 5.5 Top four equivalence classes extracted from the paper set "computer integrated manufacturing" (The words shown here are after stemming.)

5.3.2 Ranking of Topics

An equivalence class is considered as the representative of one topic. All equivalence classes are ranked based on the average scores of their FSs. The score of a FS is a combination of three parts: frequency, length, penalty of query terms.

As has been demonstrated in the previous chapter, the frequency is a very important factor for the significance of a FS. Through experiment, we found that *tf* ranking scheme achieved better performance than *tf.idf*, because *tf.idf* is good at extracting those topics which differentiate one document from others rather than extracting those topics which share across documents (Therefore, it is often applied in the task of document classification). Therefore, *tf* scheme will be applied in our system, since one important purpose of multi-paper summarization is to extract the commonalities among documents.

In our system, the length of a FS is another factor which affects the significance of the FS. The assumption is that given the same frequency, the longer FSs will be more significant than those with shorter lengths.

Multiple papers for summarization are usually retrieved from search engines on a query or grouped under one general topic (which can also be treated as a query). The purpose of a query penalty is to let important subtopics surface more easily. For example, given the query *computer integrated manufacturing*, word sequences such as *computer integrated manufacturing*, *computer integrated* and *integrated manufacturing* must be very frequent in the retrieved papers. Without a penalty to such phrases, they would probably dominate highly ranked equivalence classes and prevent other significant topics from emerging. From the user's perspective, however, many more subtopics are also expected, e.g. *automated guided vehicle*, *computed aided process planning*, rather than *computer integrated manufacturing* and its subsequences only. Therefore, with the query *computer integrated manufacturing*, we give penalty to the FSs in which any of the word sequences in Table 5.3 shows up. The penalty score is the maximal overlap between candidate FS and the query. For example, for FS *computer integrated system*, the penalty is 2 since *computer integrated* appears; for FS *computer integrated manufacturing system*, the penalty is 3 since *computer integrated manufacturing* appears.

Table 5.3 Penalty to word sequences in query

Word sequences	Penalty score
computer integrated manufacturing	3
computer integrated	2
integrated manufacturing	2
computer	1
integrated	1
manufacturing	1

Therefore, the combined score of a FS is:

$$score = tf \cdot \log_2 l \cdot \log_2 \frac{Q+2}{penalty+1} \tag{5.1}$$

where *tf* is the appearing frequency of the FS in the document set, *l* is the length of sequence, *Q* is the number of query words excluding stop words (*Q*=3 for the query *computer integrated manufacturing*), *penalty* is the maximal overlap between the FS and the query (0 for FSs which have no overlap with the query). It is apparent that the value of *penalty* ranges from 0 to *Q*. Therefore, 1 is added to denominator to prevent it from equaling 0, and 2 is added to numerator to assure that the numerator is always greater than denominator (so that the combined score is always greater than 0).

5.3.3 Macrostructure: Topical Structure

The four equivalence classes of the paper set c*omputer integrated manufacturing* in Figure 5.5 are the top four topics based on the ranking scheme of Equation 5.1. Figure 5.6 demonstrates the macrostructure, i.e. topical structure, of the paper set. The key part of this macrostructure is the list of ranked topics. The relations between topics and papers are also indicated in this macrostructure. As can be seen, such form of macrostructure can better represent the actual relationship among multiple papers: one topic can appear in different papers and one paper can be associated with different topics (the paper D_5 and D_{15} are both associated with multiple topics).

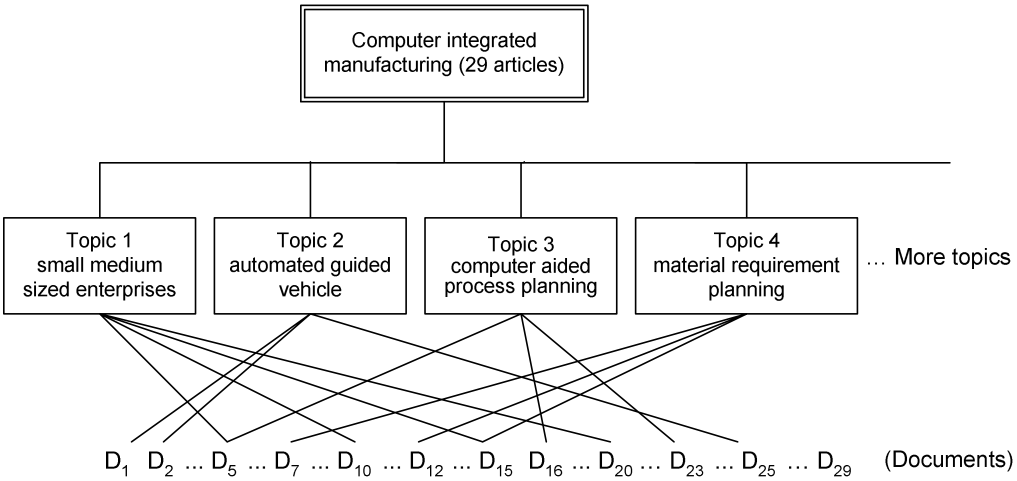

Computer integrated manufacturing (29 articles)
Topic 1 small medium sized enterprises
Topic 2 automated guided vehicle
Topic 3 computer aided process planning
Topic 4 material requirement planning
... More topics ...
D1 D2 ... D5 ... D7 ... D10 ... D12 ... D15 D16 ... D20 ... D23 ... D25 ... D29
(Documents)

Figure 5.6 Macrostructure for the paper set of "computer integrated manufacturing": topical structure

5.4 Microstructure within Multiple Papers

Microstructure is defined as the structure within each single paper in the paper set. In multi-paper summarization, microstructure should provide information like how the papers develop based on the topics and how they differentiate among each other in terms of the topics.

5.4.1 Problem-Solving Structure

Since this summarization system is focused on the domain of technical papers, the microstructure should address the special characteristics of technical paper summarization. Technical papers and paper abstracts are often presented in a problem-solving structure: problem introduction and definition (reviewing other researchers' work), solutions, testing, results, etc. (Trawiński, 1989; Zappen, 1983), as shown in Figure 5.7. Identification of this problem-solving structure is useful to identify the role of every passage in the input papers. In this system, microstructure of each single paper is defined as the problem-solving structure.

Background	Computer aided process planning (CAPP) is generally acknowledged as a significant activity to achieve computer-integrated manufacturing (CIM). In coping with the dynamic changes in the modern manufacturing environment, the awareness of developing intelligent CAPP systems has to be raised, in an attempt to generate more successful implementations of intelligent manufacturing systems.
Contribution	In this paper, the architecture of a hybrid intelligent inference model for implementing the intelligent CAPP system is developed. The detailed structure for such a model is also constructed.
Results & conclusion	The establishment of the hybrid intelligent inference model will enable the CAPP system to adapt automatically to the dynamic manufacturing environment, with a view to the ultimate realization of full implementation of intelligent manufacturing systems in enterprises.

Figure 5.7 Microstructure for a paper abstract in the paper set "computer integrated manufacturing": problem-solving structure

As has been discussed in Chapter 3, in the domain of technical papers, a summarization system should be able to present the general information and background knowledge of common topics across papers, and indicate the different contributions of each paper as well. With the clearly identified microstructure in Figure 5.7, we can distinguish between common knowledge and author's unique contribution so that our summarization system is able to summarize the similarities and difference among papers.

5.4.2 Rhetorical Analysis

The method used to identify microstructure is rhetorical analysis (Teufel and Moens, 2002). Rhetorical analysis is to identify rhetorical zones (like Figure 5.7) by assigning rhetorical categories to every sentence or clause in the paper article. The annotation scheme of rhetorical categories used in this study is given in Table 5.4. This annotation scheme is designed for paper abstracts.

Table 5.4 Annotation scheme of rhetorical categories for paper abstracts

RHETORICAL CATEGORY	DESCRIPTION
R_1. BACKGROUND & OTHER WORK	Background knowledge, common sense, or work from other researchers
R_2. CONTRIBUTION	Statements to summarize the major contribution of the paper, usually one or two sentences
R_3. METHODOLOGY & EXPERIMENTS	Description of researcher's own work: methodology, experiment process, etc.
R_4. RESULTS & CONCLUSION	Description of researcher's own work: experimental results, conclusion, discussion, etc.

There are four major types of rhetorical categories. The annotation scheme is non-overlapping and non-hierarchical, and each sentence or clause in the paper article must be assigned with exactly one rhetorical category. The adjacent sentences of the same category can be considered to form a zone of the same rhetorical category, which is called rhetorical zone.

5.4.3 Experiment of Rhetorical Classification

The rhetorical analysis is actually formed as a problem of automatic text classification: to automatically classify sentences to the four rhetorical categories based on the features of sentences.

5.4.3.1 Experimental Data Sets

1425 manually categorized sentences and clauses from 246 paper abstracts were used as the training samples in building the classification model of rhetorical categories. Since the four rhetorical categories were clearly defined and paper abstracts were usually highly structured, the inter-agreement among different human subjects was very high (at 94%). Within the training samples, R_3 was the most popular category which possessed a half of all sentences. The other three categories were nearly equally distributed within the rest of sentences.

The features of sentences and clauses used in the categorization included:

- **Absolute location**: equals *i* for the *i*th sentence in the document (ranges from 1 to *N*, *N* is the total number of sentences in the document)

  In the domain of news articles, sentence location is the most important feature for sentence selection (Brandow et al., 1995). In the domain of technical papers, sentence location, although less dominant, can still give a useful indication.

- **Relative location**: equals *i*/*N* for the *i*th sentence (ranges from 1/*N* to 1)

  Rhetorical zones appear in typical positions in the article, as problem-solving structure follows certain patterns (Swales, 1990). For example, the paper abstract often starts with background knowledge and introduction of previous studies, while the authors' own contribution can usually be found in the middle and the end of the abstract. Therefore, we intend to model this by adding the feature of relative location.

- **Voice**: active, passive or no verb

  Previous studies found that linguistic features like voice and tense often correlated with rhetorical zones (Biber, 1995; Riley, 1991).

- **Tense**: nine tenses (simple present, present continuous, present perfect, simple past, past continuous, past perfect, simple future, future continuous, future perfect) or no verb

- **Modal**: modal or no modal verb

  The presence and absence of a modal auxiliary might be relevant for detecting the statements in which the author signals low certainty, e.g. *these results might prove that* ... (Hyland, 1998).

- **Category of preceding sentence**: R_0 (The current sentence is the leading one.), R_1, R_2, R_3 or R_4

  Because rhetorical structure follows a certain pattern of scientific argumentation, there is definitely correlation between adjacent sentences. Therefore, the rhetorical category of preceding sentence is also considered as an important feature of a sentence.

- **Action verbs**

  Previous studies demonstrated that action verb is an important indicator for rhetorical category (Myers, 1992; Thompson and Yiyun, 1991). Teufel and Moens (2002) have grouped 365 verbs into 20 classes based on semantic concepts like presentation, contrast and argumentation. The 20 classes were further pruned and some of them were removed because they did not show a high association with any category. Table 5.5 lists some of the classes of action verbs and their examples.

Table 5.5 Action verbs

Type	Example
SOLUTION	We solve this problem by …
USE	We employ CITE's method …
SIMILAR	Our approach resembles that of …
INTEREST	We are concerned with …
CONTRAST	Our approach differs from …

- **Formulaic expressions**

Formulaic expressions are semantic indicators that are expected to be helpful for rhetorical classification, e.g. *in general*, *in this paper*, etc. In this study, 73 formulaic expressions were extracted from 150 paper abstracts through manual analysis. In order to minimize the bias, these 150 paper abstracts were different from those samples that would be later applied to build classification models. These 73 formulaic expressions are divided into seven semantic classes, as shown in Table 5.6. The reason that formulaic expressions are clustered is because of the data sparseness. Teufel and Moens (2002) have shown that clustered list performs much better than the unclustered list. The guidelines for clustering formulaic expressions were: some expressions only appear frequently in one rhetorical category and thus they are grouped into one cluster, e.g. *attract … attention*, *in the past*, *generally* only appearing frequently in category R_1. BACKGROUND & OTHER WORK; those formulaic expressions which cannot be grouped in the above way are clustered according to their natural similarity, e.g. *however*, *on the other hand*, *in addition* all belonging to connecting adverbs.

Table 5.6 Formulaic expressions

Type	Examples
BACKGROUND_FORMULA	attract … attention, in the past, generally, recently
METHODOLOGY_FORMULA	there are … steps, specifically
RESULT_FORMULA	as the result, consequently, through experimental analysis
COMPARATIVE	10 times lower than, give better performance than
CONNECT_ADVERBIAL	however, on the other hand, in addition
THIS_WORK	in this paper/study/project
PREVIOUS_WORK	unlike previous research, contrast to the early work

5.4.3.2 Classification Algorithm

Two popular classification algorithms, Naïve Bayes (Lewis and Gale, 1994) and SVMs (Vapnik, 1995) were applied to these samples. The implementations used were Weka (Witten and Frank, 2005) and SVM^{light} (Joachims, 1998 and 1999) respectively.

The basic idea in Naïve Bayes (NB) algorithm is to use the joint probabilities of features to estimate the probabilities of categories (McCallum and Nigam, 1998; Mitchell, 1996), which is in the form as:

$$P(c_i \mid F_1, F_2, ...F_k) = \frac{P(F_1, F_2, ...F_k \mid c_i)P(c_i)}{P(F_1, F_2, ...F_k)} \tag{5.2}$$

where

$P(c_i \mid F_1, F_2, ...F_k)$ is the posterior probability of observing category c_i given the feature set (F_1, F_2,…F_k);

$P(F_1, F_2, ...F_k \mid c_i)$ is the prior probability of observing feature set (F_1, F_2,…F_k) given occurrence of class c_i;

$P(c_i)$ is the probability that a randomly picked sample belongs to class c_i;

$P(F_1, F_2,...F_k)$ is the probability that a randomly picked sample has the feature set (F_1, F_2,…F_k).

The naïve part of NB method is the assumption of feature independence. This assumption makes the computation of the NB classifiers far more efficient than non-naïve Bayes approaches. The Equation 5.2 can be converted into the following form which is more easily to computed based on the training samples:

$$P(c_i \mid F_1, F_2,...F_k) = \frac{\prod_{j=1}^{k} P(F_j \mid c_i)P(c_i)}{\prod_{j=1}^{k} P(F_j)} \tag{5.3}$$

Support Vector Machines (SVMs) is a learning approach introduced by Vapnik (1995) for solving two-class pattern recognition problems. It is based on the structural risk minimization principle for which error-bound analysis has been theoretically motivated (Cortes and Vapnik, 1995). The method is defined over a vector space where the problem is to find a hyperplane that "best" separates the data points in two classes, as shown in Figure 5.8.

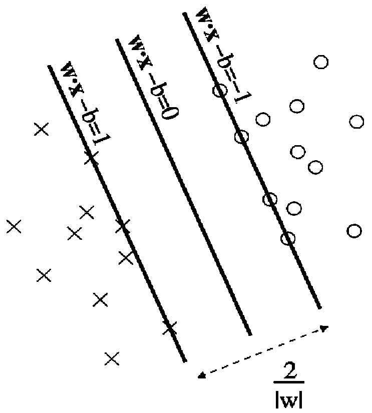


Figure 5.8 Hyperplanes for SVMs trained with samples from two classes (Samples along the hyperplanes are called the support vectors.)

More precisely, the hyperplane of SVMs can be written as

$$\vec{w} \cdot \vec{x} - b = 0 \tag{5.3}$$

where $\vec{x}$ is a data point to be classified; vector $\vec{w}$ and constant b are learned from a training data set $\{(\vec{x}_i, y_i)\}$, $y_i \in \{\pm 1\}$. SVMs is to find $\vec{w}$ and b that satisfy the following constrains:

$$\vec{w} \cdot \vec{x}_i - b \geq +1 \quad \text{for} \quad y_i = +1 \tag{5.4}$$

$$\vec{w} \cdot \vec{x}_i - b \geq +1 \quad \text{for} \quad y_i = -1 \tag{5.5}$$

An important property of SVMs is that the hyperplane is determined only by the data points which have exactly the distance $\frac{1}{|\vec{w}|}$ from the hyperplane. Those points are called the support vectors, which are the only effective elements in the training set.

5.4.3.3 Experimental Results

The algorithms of Naïve Bayes and SVMs were applied in the samples of 1425 sentences to build the classification models. The performance between these two methods was compared based on five-fold cross validation. In five-fold cross validation, the initial samples are randomly partitioned into five mutually exclusive subsets or "folds", S_1, S_2, S_3, S_4, S_5, each of approximately equal size. Training and testing is performed five times. In each iteration, one of the subsets is reserved as the test set and the remaining subsets are combined to train the classification model.

The performance of the classification is measured by recall (R), precision (P) and F-score, as defined in Section 3.3.3. The calculation of these three measures is based on the confusion matrix shown in Table 5.7. This matrix gives the overlap between machine classification and ideal classification. For example, there are a+b samples that have been classified by machine as belonging to the category, and among them only a samples are correctly classified.

Table 5.7 Confusion matrix for classification

		Actual	
		Yes	No
Predicted	Yes	a	b
	No	c	d

Thus, recall (R), precision (P) and F-score are computed as:

$$R = \frac{a}{a+c}, \quad P = \frac{a}{a+b}, \quad F = \frac{2PR}{P+R} \tag{5.6}$$

Recall, precision and F-score were recorded for each of the four rhetorical categories. All the measures and averages across the four categories are given in Table 5.8.

Table 5.8 Comparison of Naïve Bayes and SVMs in rhetorical classification (five-fold cross validation)

Category	**Naïve Bayes**			**SVMs**		
	Precision	Recall	*F*-measure	Precision	Recall	*F*-measure
R_1	0.807	0.816	0.811	0.740	0.851	0.791
R_2	0.650	0.703	0.675	0.721	0.662	0.690
R_3	0.833	0.830	0.832	0.881	0.899	0.890
R_4	0.705	0.642	0.672	0.836	0.687	0.754
Average	0.749	0.748	**0.748**	0.795	0.775	**0.781**

The results showed that SVMs outperformed Naïve Bayes in terms of *F*-measure. The possible reason was that Naïve Bayes assumed that the features were statistically independent of each other. However, statistical analysis showed that in our classification model, some features were highly correlated with each other, e.g. the correlation between features “absolute location” and “relative location” was 0.774. “Action verbs” were also highly correlated with “formulaic expressions”. Therefore, SVMs model was used in our system to decide rhetorical zones of paper abstracts.

5.5 Generation and Presentation of Summary

The output summary is developed based on macrostructure and microstructure from multiple papers. For each topic in the macrostructure, all relevant sentences in which the topic appears are extracted from input papers and added into a pool as candidate segments for summary. Each sentence is accompanied by a label including its source article ID and rhetorical category.

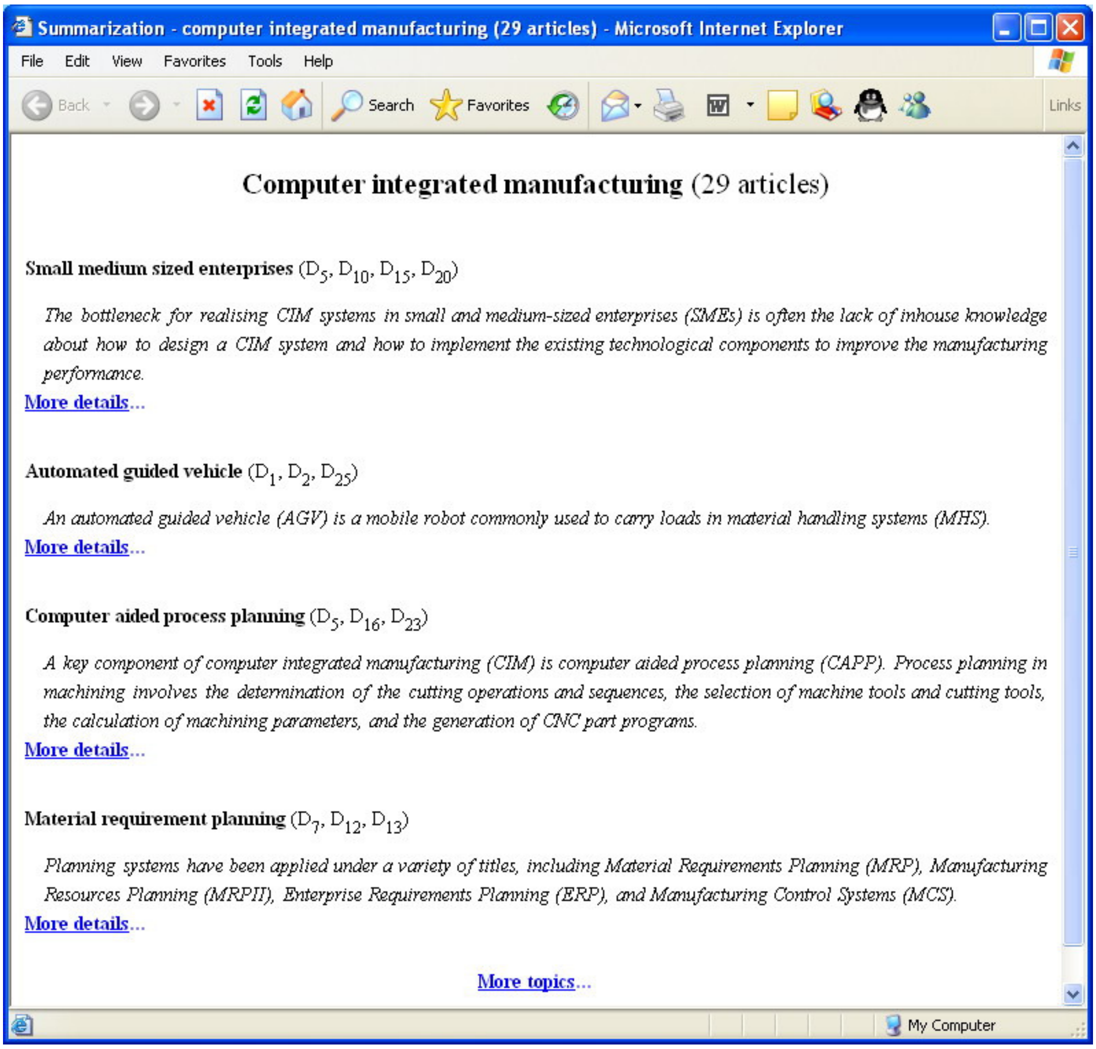
Table 5.9 All candidate passages relevant to the topic “computer aided process planning” in the paper set “computer integrated manufacturing”

Sentence	Source article	Rhetorical status
A key component of computer integrated manufacturing (CIM) is computer aided process planning (CAPP).	D_5	R_1
Process planning in machining involves the determination of the cutting operations and sequences, the selection of machine tools and cutting tools, the calculation of machining parameters, and the generation of CNC part programs.	D_5	R_1
A prototype system, the integrated intelligent process planning system (IIPPS), is described for machining; it was developed on the basis of an IIS and constructed using three levels of effort: (1) AutoCAD, (2) dBASE III and (3) KnowledgePro.	D_5	R_2
The system may be utilized not only by a process planning engineer in a company, but also by students of mechanical or industrial engineering.	D_5	R_4
Computer aided process planning (CAPP) is generally acknowledged as a significant activity to achieve computer-integrated manufacturing (CIM).	D_{16}	R_1
In coping with the dynamic changes in the modern manufacturing environment, the awareness of developing intelligent CAPP systems has to be raised, in an attempt to generate more successful implementations of intelligent manufacturing systems.	D_{16}	R_1
In this paper, the architecture of a hybrid intelligent inference model for implementing the intelligent CAPP system is developed.	D_{16}	R_2
The establishment of the hybrid intelligent inference model will enable the CAPP system to adapt automatically to the dynamic manufacturing environment, with a view to the ultimate realization of full implementation of intelligent manufacturing systems in enterprises.	D_{16}	R_4
A well-constructed generative computer-aided process planning (CAPP) system is suitable for computer-integrated manufacturing systems and intelligent manufacturing systems.	D_{23}	R_1
Most generative CAPP systems developed in the last decade employed a linear and batch approach as their underlying methodology, but because of low efficiency and quality, many such systems cannot be applied effectively to industrial enterprises.	D_{23}	R_1
To overcome these weaknesses, a novel methodology, called prototypebased incremental process planning (PIPP), is presented for CAPP in this paper which offers a new approach for increasing the efficiency and quality of process planning, and for fully supporting concurrent engineering.	D_{23}	R_2
Based on this methodology, an experimental CAD/CAPP concurrent design system (HFCAD/CAPP) has been built, and a case study is presented to illustrate the characteristics of this system.	D_{23}	R_3

Table 5.9 lists the sentences extracted for the topic *computer aided process planning* in the paper set *computer integrated manufacturing*.

The output summary is composed of two parts based on topics. The first part is the description of the topic's general information and its background knowledge. Sentences with rhetorical status R_1 constitute a pool of candidate segments for this part. The second part of the summary includes the uniqueness of each paper with regard to this topic. Sentences with rhetorical status R_2, R_3 and R_4 are considered as candidate segments for this part. The method of Maximal Marginal Relevance (MMR) (Carbonell and Goldstein, 1998) introduced in Section 4.3.2 is implemented to reduce the redundancy in the first part of the summary, since the same background information is often repeated by different authors. Maximal phrases, which are not the subsequences of any other phrases in the equivalence classes, are chosen to represent the topics, e.g. in Figure 5.5 *autom guid vehicl* is the maximal phrase of equivalence class 2.

The purpose of separating summary into two parts is to present the summarization result in a better organized format and to fulfill the reader's information requirement in a more efficient way. Figure 5.9 shows the first page of summary presented to readers. Topics are ranked according to their significance in the paper set. Several sentences are included to briefly describe the background or common knowledge of each topic. The articles relevant to each topic have been identified and hyperlinked. If readers want to browse more topics, they can click on the hyperlink "More topics…" located at the page bottom.



Summarization - computer integrated manufacturing (29 articles) - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Search
Favorites
Links
My Computer

Computer integrated manufacturing (29 articles)

Small medium sized enterprises (D_5, D_{10}, D_{15}, D_{20})

The bottleneck for realising CIM systems in small and medium-sized enterprises (SMEs) is often the lack of inhouse knowledge about how to design a CIM system and how to implement the existing technological components to improve the manufacturing performance.

More details...

Automated guided vehicle (D_1, D_2, D_{25})

An automated guided vehicle (AGV) is a mobile robot commonly used to carry loads in material handling systems (MHS).

More details...

Computer aided process planning (D_5, D_{16}, D_{23})

A key component of computer integrated manufacturing (CIM) is computer aided process planning (CAPP). Process planning in machining involves the determination of the cutting operations and sequences, the selection of machine tools and cutting tools, the calculation of machining parameters, and the generation of CNC part programs.

More details...

Material requirement planning (D_7, D_{12}, D_{13})

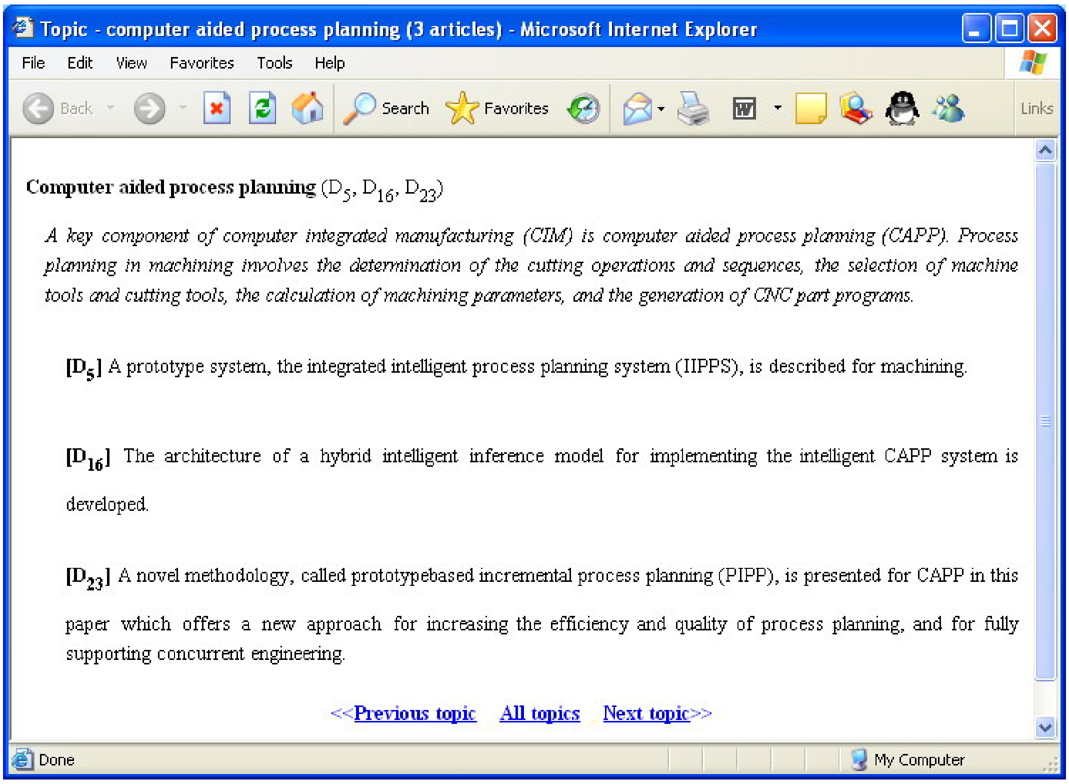
Planning systems have been applied under a variety of titles, including Material Requirements Planning (MRP), Manufacturing Resources Planning (MRPII), Enterprise Requirements Planning (ERP), and Manufacturing Control Systems (MCS).

More details...

More topics...

Figure 5.9 Summarization output for paper set "computer integrated manufacturing": ranked topics and their general information

If users are interested in a particular topic after reading the general topic information, they can choose "More details…" at the bottom of the topic to view the detailed summary, as shown in Figure 5.10. The unique information of articles in this topic is then presented so that readers are able to find out the difference among articles and proceed to choose interesting ones for further readings.


Topic - computer aided process planning (3 articles) - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Links
Done
My Computer

Computer aided process planning (D_5, D_{16}, D_{23})

A key component of computer integrated manufacturing (CIM) is computer aided process planning (CAPP). Process planning in machining involves the determination of the cutting operations and sequences, the selection of machine tools and cutting tools, the calculation of machining parameters, and the generation of CNC part programs.

[D_5] A prototype system, the integrated intelligent process planning system (IIPPS), is described for machining.

[D_{16}] The architecture of a hybrid intelligent inference model for implementing the intelligent CAPP system is developed.

[D_{23}] A novel methodology, called prototypebased incremental process planning (PIPP), is presented for CAPP in this paper which offers a new approach for increasing the efficiency and quality of process planning, and for fully supporting concurrent engineering.

<<Previous topic All topics Next topic>>

Figure 5.10 Summarization output: difference of the papers with respect to one topic

5.6 Conclusion of the Chapter

In this chapter, a multi-paper summarization system based on macrostructure and microstructure has been proposed. Macrostructure consists of significant topics which are generated by grouping FSs into equivalence classes, while microstructure is the rhetorical structure within each individual paper. Candidate sentences for summarization are extracted based on macrostructure and microstructure. The output summary is composed of two parts: general information of ranked topics and differences among papers in terms of topics. In the next chapter, the proposed summarization system will be evaluated based on the existing summarization benchmarks.

Chapter 6

Evaluation of Summarization Performance

It is useful to evaluate the performance of a summarization system, because summarization evaluation can contribute in two ways: to optimize a summarization system and to compare it with other peer systems. There are a few factors that may affect the performance of a summarization system and evaluation experiment can help to find the optimal parameterizations for it. On the other hand, evaluation experiment can validate the effectiveness and efficiency of a summarization system by comparing it with other peer systems. In this chapter, the proposed multi-paper summarization system is investigated under different parameterizations and compared with two other peer summarization systems.

6.1 Methods of Summarization Evaluation

Evaluation has become an independent discipline in automatic text summarization, although there is not a widely held agreement upon set of methods for carrying out summarization evaluation (Jing et al., 1998). In general, methods for evaluating text summarization can be classified into two categories, intrinsic and extrinsic methods. Intrinsic methods evaluate the summarization based on the analysis of summary itself, either performed manually or automatically (Jing et al., 1998; Lin, 2004). Extrinsic methods measure the summarization performance based on the influence on some other tasks, such as reading comprehension and information retrieval (Mani et al., 1998; Morris et al., 1992; Tombros and Sanderson, 1998).

6.1.1 Intrinsic Methods

Most evaluations of summarization systems use intrinsic methods (Edmundson, 1969; Kupiec et al., 1995; Paice, 1990). Instead of evaluating summaries' linguistic quality like fluency, grammaticality and readability (Mani, 2001; Minel et al., 1997), this study intends to evaluate summaries' informativeness, which is also the major focus of existing summarization evaluation researches. The informativeness of a system-generated summary is usually evaluated based on the comparison with one or more "ideal" reference summaries. The reference summaries are generated by human summarizers. There are two popularly used intrinsic evaluation methods: ROUGE and Pyramid.

6.1.1.1 ROUGE

ROUGE (Recall-Oriented Understudy for Gisting Evaluation), the most frequently used automated summary evaluation package (Lin, 2004), is closely modeled after BLEU for machine translation evaluation (Papineni et al., 2001). As an intrinsic evaluation method, ROUGE automatically measures the quality of a candidate summary by comparing it to human-generated reference summaries. The measures count the number of overlapping content units such as *n*-grams between system-generated summary and human-generated reference summaries (Lin and Hovy, 2003; Saggion et al., 2002). An *n*-gram is a subsequence of *n* words from a given word sequence (Manning and Schütze, 1999).

ROUGE-*n* is an *n*-gram recall between a candidate summary and a few reference summaries, which is computed as follows:

$$\text{ROUGE-}n = \frac{\sum_{S \in \{reference\_summaries\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{reference\_summaries\}} \sum_{gram_n \in S} Count(gram_n)} \tag{6.1}$$

where *n* stands for the length of the *n*-grams ($gram_n$), $Count_{match}(gram_n)$ counts the total number of *n*-grams co-occuring in the candidate summary and all reference summaries, and $Count(gram_n)$ counts the total number of *n*-grams in all reference summaries. It is clear that ROUGE-*n* is a recall-related measure because the denominator of Equation 6.1 is the sum of the number of *n*-grams occurring at the reference summary side.

Lin's experiments showed that *n*, the length of *n*-grams, would greatly affect the performance of ROUGE-*n* evaluation (Lin, 2004). For MDS, ROUGE-2 and ROUGE-3 achieved better performance than other parameterizations.

6.1.1.2 Pyramid

Similar to ROUGE, Pyramid is an intrinsic evaluation method in which system-generated candidate summary is compared with one or more human-generated reference summaries. The difference between Pyramid and ROUGE lies in how summaries are fragmented into content units so that they can be aligned and compared.

In ROUGE, the fragmentation is performed in an automatic process in which summaries are fragmented into *n*-grams of fixed length. Unlike ROUGE, Pyramid has a manual fragmentation process in which humans define what constitutes content fragments (Halteren and Teufel, 2003; Nenkova and Passonneau, 2004).

Pyramid method is based on Summarization Content Units (SCUs). Figure 6.1 & 6.2 demonstrate the manual extraction of two SCUs from a set of reference summaries. Figure 6.1 gives four similar sentences identified by human subjects from different reference summaries. In the next, two SCUs from the underlined portions of the sentences are obtained. Each SCU has a weight corresponding to the number of summaries it appears in, as shown in Figure 6.2: SCU$_1$ has weight of 4 and SCU$_2$ has weight of 3.

A. In 1998 two Libyans indicted in 1991 for the Lockerbie bombing were still in Libya.
B. Two Libyans were indicted in 1991 for blowing up a Pan Am jumbo jet over Lockerbie, Scotland in 1988.
C. Two Libyans, accused by the United States and Britain of bombing a New York bound Pan Am jet over Lockerbie, Scotland in 1988, killing 270 people, for 10 years were harbored by Libya who claimed the suspects could not get a fair trail in America or Britain.
D. Two Libyan suspects were indicted in 1991.

Figure 6.1 Example text for SCU extraction

SCU$_1$ (weight=4): two Libyans were officially accused of the Lockerbie bombing A, B, C, D	SCU$_2$ (weight=3): the indictment of the two Lockerbie suspects was in 1991 A, B, D

Figure 6.2 Summarization Content Units (SCUs)

The remaining parts of the four sentences in Figure 6.1 end up as contributors to nine other SCUs of different weights. After all SCUs are extracted from the reference summaries, they are ranked by weights and used to evaluate the candidate summary with the assumption that an ideal summary should only contain top ranked SCUs.

Although pyramid can avoid some low-informative fragments, such as “of the”, it will definitely cost a lot of manual work and introduce human bias. On the contrary, ROUGE is more efficient and robust, and is thus more widely applied in summarization evaluation work.

6.1.2 Extrinsic Methods

Unlike intrinsic evaluation, in extrinsic evaluation, the quality of a summary is judged on how it affects the completion of some other tasks. These possible tasks include human subjects answering questions as well as determining the relevance of documents to topics based upon reading summaries (Jing et al., 1998; Mani et al., 1998; Morris et al., 1992; Tombros et al., 1998).

Morris et al. (1992) reported on an extrinsic summarization evaluation in a task of question-answering. The authors picked four Graduate Management Admission Test (GMAT) reading comprehension exercises. The exercises were multiple-choice, with a single answer to be selected from answers shown alongside each question. There were eight questions for each exercise, with five possible answers shown for each question. The authors intended to examine whether summary could substitute original text in such reading comprehension task, by measuring how many questions were correctly answered given original texts and their summaries.

Mani et al. (1998) reported another extrinsic task-based evaluation, measuring the impact of summarization on time cost and accuracy in assessing document relevance. They evaluated the summarization performance by examining whether use of summaries instead of full documents could save time in document relevance assessment, without impacting accuracy.

It was found that automatic text summarization was very effective in these question-answering and relevance assessment tasks. For example, the work of Mani et al. (1998) showed that summaries as short as 17% of full text length could speed up decision making by almost a factor of two with no statistically significant degradation in accuracy.

6.2 Experimental Design of Summarization Evaluation

One goal of the experiments was to evaluate the proposed multi-paper summarization system based on macro- and micro-structure under different parameterizations and to investigate the influence of different factors on summarization performance. The factors that were of interest in the experiments are listed in Table 6.1. Another goal of the experiments was to evaluate whether the summarization system, compared with other peer systems, could better identify the major topics in a set of papers and identify the similarities and difference among papers.

Table 6.1 Factors in experimental design of summarization evaluation

Factor	Levels
Threshold for FS extraction (σ)	2, 3, 4
Topic ranking scheme	With/without query penalty
Compression ratio	10%, 30%

6.2.1 Factors in Experimental Design

The first two factors in Table 6.1 are designed for macrostructure generation. The first factor, threshold for FS extraction (σ), is used in Algorithm 4.1 as the threshold for supporting documents to extract FSs. In this experiment, the levels of 2, 3 and 4 would be examined. The second factor, with/without query penalty, is applied in the topic ranking of macrostructure generation process, as introduced in Section 5.4.2.

The last factor is the compression ratio of summarization, i.e. the ratio between the length of summary and the length of original document (Mani et al., 2002). Too short a summary discards a lot of useful information, while too long a summary costs more reading time. Therefore, a summarization system should find an optimal compression ratio so that important information is kept and the reading time is reduced to a minimum.

The first two factors in Table 6.1 would affect the content composition of output summaries, and therefore could be examined by intrinsic evaluation method, say ROUGE, which compares the content overlap between candidate summaries and reference summaries. On the other hand, the factor of compression ratio was not suitable to be examined by intrinsic evaluation since it did not affect the method to choose content units in summarization process. Therefore, the factor of compression ratio would be examined by extrinsic evaluation.

6.2.2 Peer Summarization Systems

Two peer summarization systems were applied as the baseline in the evaluation experiments: Copernic summarizer (http://www.copernic.com) and clustering-summarization method. Copernic summarizer is a commercial software which can pinpoint the key concepts from source documents and extract the most relevant sentences using undisclosed statistical or linguistic algorithms. Sentences from different documents are treated as the same in the pool of candidate segments for summarization.

The method of clustering-summarization is a popular method for multi-document summarization, especially in the context of information retrieval (Maña-López, 2004; Radev et al., 2004; Roussinov and Chen, 2001). In this method, a document set is separated into several clusters with the assumption that each cluster discusses one topic and summarization is then performed in each cluster. For clustering process, *K*-means clustering implemented in CLUTO (Karypis, 2002) was used in our experiments. In each cluster of papers, salient sentences were extracted using MMR in the form of Equation 4.1.

6.2.3 Experimental Data Sets

The data sets used in the experiments included 15 sets of technical papers from Manufacturing Corpus Version 1 (MCV1) (Liu, 2005), with 20 to 120 papers in each set sharing a common topic. For each paper set, two manual summaries were generated by different human summarizers in order to reduce the subjectiveness. Since all the papers are from technical domain, the human summarizers with engineering background were chosen in this manual summarization task. In their summarization process, the human summarizers were instructed to develop a summary at the length 10-20% of the original texts, by extracting the most important information across the paper set and highlighting similarities and difference among papers. The manual summaries were used as reference summaries in the intrinsic evaluation process.

As has been discussed, our study focused on indicative summarization which aimed to help readers to decide whether it is worth reading the full papers, instead of aimed to generate summaries that could substitute original papers. Therefore, paper abstracts, which are usually enough for readers to get the gist of papers, were used as experimental data instead of full papers. Compared to full papers, abstracts were more concise, less redundant and usually well structured. This characterized abstracts a suitable choice for our purpose.

6.2.4 Factor Analysis: ROUGE Evaluation

The first two factors in Table 6.1 were examined using ROUGE evaluation. In order to find out whether these two factors affected summarization performance, a two-factor factorial experiment was designed as in Table 6.2.

Table 6.2 Two-factor factorial experiment

		Topic ranking scheme	
		With query penalty	**Without query penalty**
Threshold for FS extraction	**2**	$y_{1,1,1}, y_{1,1,2}, \dots, y_{1,1,15}$	$y_{1,2,1}, y_{1,2,2}, \dots, y_{1,2,15}$
	3	$y_{2,1,1}, y_{2,1,2}, \dots, y_{2,1,15}$	$y_{2,2,1}, y_{2,2,2}, \dots, y_{2,2,15}$
	4	$y_{3,1,1}, y_{3,1,2}, \dots, y_{3,1,15}$	$y_{3,2,1}, y_{3,2,2}, \dots, y_{3,2,15}$

There were two factors in this experiment:

1. Threshold for FS extraction, with three levels: 2, 3, 4
2. Topic ranking scheme, with two levels: with/without query penalty

For each of the six combinations of parameters, the summarization system generated summaries for 15 paper sets. For each paper set, topics were extracted and ranked after the pre-processing of the paper set. Top ten topics were chosen and relevant sentences were selected to compose the summary. System-generated summaries were evaluated by ROUGE-2 and ROUGE-3 based on two manual summaries. ROUGE-2 and ROUGE-3 were used since they have demonstrated better performance than other ROUGE parameterizations for multi-document summarization (Lin, 2004). Summarization score was calculated as the average of ROUGE-2 and ROUGE-3 scores. Therefore, for each combination of parameters in Table 6.2, there were 15 replicas of scores.

The linear statistical model of this factorial experiment was:

$$y_{i,j,k} = \mu + \tau_i + \beta_j + (\tau\beta)_{i,j} + \varepsilon_{i,j,k} \tag{6.2}$$

where

$y_{i,j,k}$ denotes the score of the *k*th replica (*k*th document set), under the *i*th level of factor "threshold for FS extraction" and the *j*th level of factor "topic ranking scheme";

μ denotes the overall mean score;

τ_i denotes the effect of the *i*th level of factor "threshold for FS extraction";

β_j denotes the effect of the *j*th level of factor "topic ranking scheme";

$(\tau\beta)_{i,j}$ denotes the effect of interaction between the two factors;

$\varepsilon_{i,j,k}$ denotes the random error component.

Since there were two factors involved in the analysis, a two-way ANOVA (ANalysis Of VAriance) would be applied to test the following hypotheses (Montgomery and Runger, 2006):

1. H$_0$: $\tau_1 = \tau_2 = \tau_3 = 0$ (no effect from factor "threshold for FS extraction"),

   H$_1$: at least one $\tau_i \neq 0$

2. H$_0$: $\beta_1 = \beta_2 = 0$ (no effect from factor "topic ranking scheme"),

   H$_1$: at least one $\beta_j \neq 0$

3. H$_0$: $(\tau\beta)_{1,1} = (\tau\beta)_{1,2} = ... = (\tau\beta)_{3,2} = 0$ (no effect from the factors' interaction),

   H$_1$: at least one $(\tau\beta)_{i,j} \neq 0$

6.2.5 Comparison with Peer Systems: Extrinsic Evaluation

Extrinsic evaluation was designed to compare our summarization system with the peer systems of Copernic summarizer and clustering-summarization. Moreover, it would examine the effect of compression ratio on summarization performance.

The extrinsic evaluation included two tasks. The first task focused on readers' responsiveness. Human assessors were required to give scores for system-generated summaries according to the following questions:

Is the summary helpful for you to

- *get an initial understanding of the paper set?*
- *identify different topics in the paper collection?*
- *identify the similarities and difference among the papers?*

The score was an integer between 1 and 5 (with 1 and 5 inclusive), where 1 stood for “not at all”, 3 for “somewhat” and 5 for “greatly”.

The second task was to evaluate how summaries helped in the manual categorization of technical papers. Evaluation was based on the expectation that through summaries readers could correctly identify as many papers as possible for each category in a short period of time.

MCV1 was a corpus with hierarchical classification scheme and each paper set used in our experiments was organized in a hierarchical way. For example, Figure 6.3

shows a paper set used in the experiments. All the papers in this set discussed a general topic C0719 *machining specific materials* and could be further classified into more specific categories, i.e. C071901, C071902, C071903 and C071904. In the second task of extrinsic evaluation, human subjects were required to read the summaries for a paper set (e.g. C0719) and to assign as many papers as possible for all sub-categories (e.g. C071901, C071902, C071903 and C071904) in the paper set according to the information they acquired from the summaries. However, in order not to affect summarization process, this hierarchical information of paper sets was made blind to summarization system.

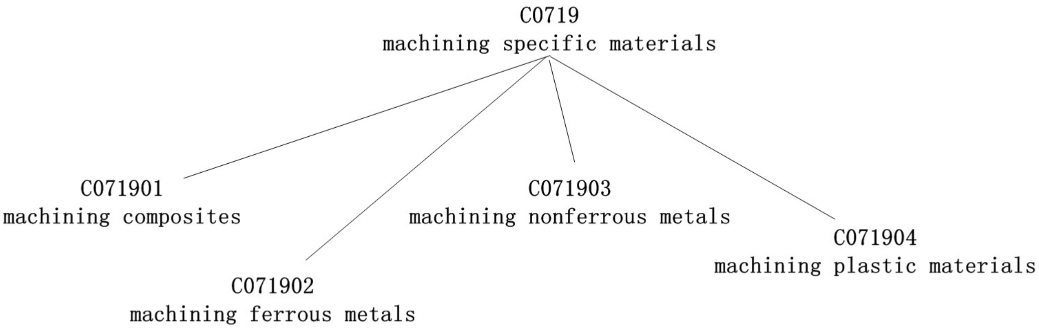

C0719
machining specific materials
C071901
machining composites
C071903
machining nonferrous metals
C071904
machining plastic materials
C071902
machining ferrous metals

Figure 6.3 Hierarchical classification scheme of paper set “machining specific materials”

6.3 Experimental Results

All the 15 paper sets were applied in ROUGE evaluation for ANOVA test. In extrinsic evaluation, ten sets were used in task 1 and five sets were used in task 2. Before summarization, pre-processing steps were conducted on these paper sets, including acronyms identification, stop words removal and word stemming.

6.3.1 Factor Analysis: ROUGE Evaluation

The experimental results for the factorial experiment are presented in Table 6.3. For each combination of parameterizations in Table 6.3, there show 15 scores as well as their mean and standard error. ANOVA table of this factorial experiment is given in Table 6.4. The total Sum of Squares (SS) is partitioned into components related to the effects in the model of Equation 6.2. The Degree of Freedom (DF) and Mean of Squares (MS) for each effect are also given in the table. The F-test statistic is computed as the ratio between MS of the effect and MS of the error with according DFs. The p-values of the F-tests are given in the last column of Table 6.4.

Table 6.3 Results of two-factor factorial experiment

		Topic ranking scheme	
		With query penalty	**Without query penalty**
Threshold for FS extraction	**2**	0.073, 0.117, 0.115, 0.087, 0.080, 0.095, 0.020, 0.107, 0.113, 0.193, 0.134, 0.063, 0.231, 0.135, 0.090	0.044, 0.028, 0.047, 0.110, 0.070, 0.045, 0.048, 0.052, 0.073, 0.027, 0.045, 0.070, 0.078, 0.084, 0.090
		Mean: 0.110 Standard error: 0.013	Mean: 0.061 Standard error: 0.006
	3	0.058, 0.082, 0.080, 0.035, 0.056, 0.095, 0.013, 0.112, 0.049, 0.081, 0.051, 0.139, 0.037, 0.078, 0.054	0.113, 0.061, 0.105, 0.063, 0.045, 0.032, 0.022, 0.091, 0.041, 0.029, 0.045, 0.030, 0.030, 0.058, 0.106
		Mean: 0.068 Standard error: 0.008	Mean: 0.058 Standard error: 0.008
	4	0.054, 0.070, 0.063, 0.027, 0.039, 0.072, 0.023, 0.074, 0.030, 0.059, 0.041, 0.120, 0.032, 0.069, 0.112	0.049, 0.035, 0.047, 0.057, 0.036, 0.040, 0.031, 0.056, 0.029, 0.021, 0.044, 0.022, 0.043, 0.054, 0.097
		Mean: 0.059 Standard error: 0.007	Mean: 0.044 Standard error: 0.005

Table 6.4 ANOVA table

Source of Variation	SS	DF	MS	F	P-value
Threshold for FS extraction	0.0179	2	0.0089	8.4154	0.00047
Topic ranking scheme	0.0138	1	0.0138	13.0105	0.00052
Interaction	0.0070	2	0.0035	3.2735	0.04276
Error	0.0892	84	0.0011		
Total	0.1278	89			

As can be seen in Table 6.4, given α=0.05 as the test's level of significance, the following hypotheses from the model of Equation 6.2 had to be rejected:

H$_0$: $\tau_1 = \tau_2 = \tau_3 = 0$ (no effect from factor "threshold for FS extraction");

H$_0$: $\beta_1 = \beta_2 = 0$ (no effect from factor "topic ranking scheme").

Therefore, the two factors tested here could both significantly affect the summarization performance. Table 6.3 demonstrates that choosing 2 as the threshold of supporting documents to extract FSs could generally improve the summarization performance than choosing higher thresholds. This was probably because the document sets used in this experiment were moderate-sized and low threshold was thus helpful to let more topics to surface. Moreover, the experimental results also proved our assumption that incorporating query penalty in the topic ranking scheme could achieve better summarization performance, as can been seen in Table 6.3.

Given α=0.05 as the test's level of significance, the following hypothesis was also rejected:

H$_0$: $(\tau\beta)_{1,1} = (\tau\beta)_{1,2} = ... = (\tau\beta)_{3,2} = 0$ (no effect from the factors' interaction).

However, the p-value of this test was 0.04276, which was close to the α value and much higher than the other two p-values (Table 6.4). This probably means that the interaction between the two factors in this experiment might somewhat affect the summarization performance, but the effect was not that significantly like the effects of individual factors.

The optimal parameterizations (threshold for FS extraction=2, ranking topic with query penalty) acquired in this experiment were used in the further experiments, e.g. extrinsic evaluation.

6.3.2 Comparison with Peer Systems: Extrinsic Evaluation

In this experiment, the summaries generated by our system, under different compression ratios (10% and 30%), were presented to readers for evaluation. Two other peer summarization systems, Copernic summarizer and clustering-summarization, were applied for comparison. The compression ratio of the peer summarization systems was set to 10% so that they could be compared with our system. The summary generated by Copernic summarizer was a set of ranked sentences and the summary generated by clustering-summarization was divided into clusters. In these summaries, all sentences were accompanied by their source paper IDs for readers to make decisions in the evaluation tasks.

6.3.2.1 Evaluation Task 1: Responsiveness

In task 1, three human assessors with different backgrounds joined the scoring process. For one paper set, all the summaries generated by Copernic, clustering-summarization and our system were evaluated by one of the three human assessors so that the hypothesis testing (paired t-test) could be performed.

Table 6.5 shows the average responsiveness scores of Copernic summarizer, clustering-summarization method and our system based on all the ten paper sets. Table 6.6 presents the results of paired t-test between our system and other two methods. Moreover, the paired t-test of our system under 30% and 10% compression ratios is also presented.

The results demonstrated that our system and peer systems were almost equally helpful in readers' initial understanding of a paper set. The possible reason was that most summarization systems could extract sentences pinpointing the significant ideas of a paper set and these sentences helped readers to get an initial understanding of the paper set. However, for identification of various topics and similarities and difference among papers, our approach scored the highest and was significant better than other two methods in the paired t-test with confidence α=0.05 (Table 6.6).

It can also be found from the last row of Table 6.6 that increasing the compression ratio from 10% to 30%, i.e. increasing the summary length, did not result in significant improvement of summarization performance.

Table 6.5 Subjects' responsiveness scores to questionnaire (average scores based on ten paper sets)

	To get an initial understanding of the paper set	To cover different topics	To identify similarities and difference among papers
Copernic summarizer (10%)	3.1	1.9	1.7
Clustering-summarization (10%)	2.9	2.9	2.7
Our system (10%)	3.3	4.0	4.0
Our system (30%)	3.3	4.2	4.3

Table 6.6 P-values for hypothesis testing (paired t-test, α=0.05)

Null hypothesis (H_0): There is no difference between the two methods.
Alternative hypothesis (H_1): The first method outperforms the second one.

	To get an initial understanding of the paper set	To cover different topics	To identify similarities and difference among papers
Our system vs. Copernic summarizer (10%)	0.39	4.3×10^{-6}	2.2×10^{-5}
Our system vs. clustering-summarization (10%)	0.25	8.7×10^{-5}	4.8×10^{-4}
Our system 30% vs. 10%	0.50	0.08	0.17

Overall, Copernic summarizer performed worse than the other two summarization methods. The possible reason was that Copernic summarizer did not take into account the case of MDS and treated all sentences from a paper set as the same in the pool of candidate segments for summarization.

6.3.2.2 Evaluation Task 2: Manual Categorization

In task 2, the subjects were five students from Mechanical Engineering. Three summaries were generated for each of the five paper sets using Copernic summarizer, clustering-summarization and our system under the compression ratio of 10%. Therefore, we had a combination of 5×3=15 evaluation experiments. The allocation of five subjects is given in Table 6.7. The measures used here for classification performance were recall and precision. Recall was the ratio between the number of documents correctly identified and the total number of documents for one category. Precision measured how many documents were correctly identified out of all identified documents for one category. All the recall and precision scores shown in Table 6.8 were the average values across all categories in the collection. The time for the subject to complete the task was also recorded in Table 6.8.

Table 6.7 The allocation of five human subjects (a, b, c, d, e) in the 15 experiments in the evaluation task 2

Paper set	C1	C2	C3	C4	C5
Copernic summarizer	a	b	c	d	e
Clustering-summarization	e	a	b	c	d
Our system	d	e	a	b	c

Table 6.8 Comparison of the three approaches in the evaluation task 2 (categorization task)

Paper set		C1	C2	C3	C4	C5
Number of papers (categories)		26 (4)	118 (5)	33 (5)	71 (7)	21 (3)
Copernic summarizer	Recall	0.306	0.194	0.512	0.435	0.561
	Precision	0.861	0.809	0.797	0.510	0.905
	Time	13’20”	62’30”	10’20”	26’30”	9’40”
Clustering-summarization	Recall	0.328	0.257	0.537	0.549	0.604
	Precision	0.847	0.849	0.804	0.537	0.890
	Time	10’10”	50’30”	8’00”	22’50”	5’50”
Our system	Recall	0.366	0.290	0.538	0.576	0.654
	Precision	0.921	0.923	0.940	0.607	0.947
	Time	11’00”	45’30”	8’10”	21’00”	5’40”

From the results it was noted that clustering-summarization and our system could both save time for readers to make decision compared to Copernic summarizer, which was consistent with previous findings that grouping documents into clusters could be an effective way to substitute the traditional interface of rank list (Zamir and Etzioni, 1999; Maña-López, 2004).

As shown in Table 6.8, the recall score of our system was much higher than Copernic summarizer. In Copernic summarizer, the relationship and structure among articles were ignored. Hence, only a few papers were given attention while others were ignored in summarization process, which probably resulted in the low recall score of Copernic summarizer. Our system was also slightly better than clustering-summarization in terms of recall. In real-world technical paper corpora like MCV1, each article could have multiple category labels. This characteristic was better considered by macrostructure of our approach, in which one article could belong to various topics. Therefore, our approach could achieve higher recall score than clustering-summarization which assigned only one category to each article.

In terms of precision, clustering-summarization was sometimes better than Copernic summarizer and sometimes worse. However, our system demonstrated significant improvement in precision compared to both baseline systems. As shown in Figure 5.10, our approach gave a description and the background knowledge for each topic, which could help users to gain a better understanding about the topics and probably caused the improvement in precision.

6.4 Conclusion of the Chapter

This chapter has evaluated the proposed multi-paper summarization system under different parameterizations. By ROUGE evaluation using ANOVA test, it has been found that the threshold for supporting documents in topic extraction could significantly affect the summarization performance. The optimal threshold value could vary for different document sets. In our experiment, choosing two as the threshold was better than higher threshold values. This was probably because the document sets used in the experiment were moderate-sized with tens of documents and high threshold could probably prevent some important topics to surface. For a larger-sized document set in real world, a higher threshold value may be required to screen out unimportant topics. Moreover, it was found that including query penalty in the topic ranking scheme could significantly improve the summarization performance.

Extrinsic evaluation has been adopted to compare the performance of our proposed system with the peer systems of Copernic summarizer and clustering-summarization. The results showed that our summarization approach could better present the topical relationship among various papers and better recognize their similarities and as well as difference. The evaluation, when benchmarked with the peer systems, also demonstrated the effectiveness of our approach in terms of precision and recall in categorizing real-world technical papers.

Chapter 7

Case Studies: Applications of Summarization in Engineering Information Management and Text Mining

Summarization is a process to transform unstructured textual documents to structured or semi-structured documents by distilling the most important information and as well reducing irrelevant and redundant information. Therefore, it may help to facilitate other tasks within engineering information management and text mining. This chapter introduces two case studies to examine these issues.

The first case study was to apply the summarization system proposed in Chapter 5 in the domain of online customer reviews. Since there already existed some studies on processing customer reviews like opinion mining, this case study intended to examine the feasibility of summarizing multiple customer reviews and to compare the performance between summarization and opinion mining.

The second case study was to investigate whether substituting documents with their summaries could improve the performance of text classification, since summary was a condensed version of original document and would reduce the redundancy of dimensionality in text classification.

7.1 Case Study 1: Summarization of Customer Reviews

Online customer reviews offer valuable information for product designers, merchants and potential shoppers in e-Commerce and e-Business. However, even for a single product, the number of reviews often amounts to hundreds or thousands. This case study aimed to apply our proposed summarization system in the domain of customer reviews and to extract the important issues from multiple customer reviews that designers, merchants and customers were concerned about.

7.1.1 Motivation

Nowadays, with the rapid development of e-Commerce and e-Business, it is common that products are sold on the websites such as *Amazon.com*. Customers are invited to write reviews to share their experiences, comments and recommendations with respect to different products. Also, in modern enterprises, a lot of emails are received from customers every day regarding products and services. These product reviews are valuable for designers and manufacturers to keep track of customers' feedback and make improvements on their products or services. Moreover, the reviews posted on WWW offer recommendations to potential buyers for their decision making. However, the number of reviews can grow very quickly and it is time-consuming to read through all of them manually. For example, there are hundreds of reviews posted on the web for some popular products in *Amazon.com*; and thousands of customer emails may be received by the manufacturer regarding one particular product.

Some work has been reported dealing with the vast amount of customer reviews (Hu and Liu, 2004; Popescu and Etzioni, 2005; Turney, 2001). All these work focused on opinion mining which was to discover the reviewers' orientations, whether positive or negative, regarding various features of a product, e.g. weight of a laptop and picture quality of a digital camera. However, we noticed that although some comments regarding product features could not be labelled as positive or negative, they were still valuable. For example, the following two sentences are extracted from the customer reviews of mobile phone Nokia 6610 in Hu's corpus (Hu and Liu, 2004):

#1: *The phone's* ***sound quality*** *is great.*

#2: *The most important thing for me is* ***sound quality***.

Both sentences discuss the product feature *sound quality*. Unlike the first sentence, the second one does not offer any orientation, neither positive nor negative, regarding the specific phone Nokia 6610, yet it does provide valuable information for designers and manufacturers about what mobile phone consumers are really concerned about. Such neutral comments and suggestions are currently not considered in the method of opinion mining.

Moreover, opinion mining focuses mainly on product features, but product features cannot cover all significant issues in customer reviews. Figure 7.1 shows some sentences extracted from the customer reviews of Nokia 6610. These sentences all discuss *flip phone* and they reveal the different perspectives from customers about *flip phone*. Some customers also elaborate on the reasons for their choices. This information is believed to be valuable for designers and manufacturers. However, in the method of opinion mining, such important issues are not pointed out because *flip phone* is not an explicit product feature of Nokia 6610.

- *As much as I like Nokia phones the **flip phones** are much better because a) you won't scratch your screens/keys b) you don't need to lock your phone all the time to prevent accidentally hitting the keys.*
- *Personally I like the Samsung phones better because I found myself liking the **flip phones** so much more.*
- *My past two phones were all **flip phones**, and I was beginning to tire of them.*
- *Nokia was my first non-**flip phone**, and I'm glad I decided to go with them.*
- *This is probably your best bet if you are looking for a phone in this price range, or like me, do not have the patience to deal with annoying **flip phones**.*

Figure 7.1 Sentences discussing "flip phone" from customer reviews of Nokia 6610

Therefore, opinion mining is not enough to extract all the important information from customer reviews and there is a desire to apply summarization technique to identify the significant topics from multiple customer reviews.

7.1.2 Summarization Approach

When applied to the domain of customer reviews, the approach of clustering-summarization (Boros et al., 2001; Radev et al., 2004) may still have the two limitations that has been discussed in Section 2.3.1, i.e. the number of clusters is difficult to determine without prior knowledge regarding the review set and topics are not perfectly distributed in the non-overlapping clusters of reviews in a real-world document set.

Based on the analysis of Hu's corpus (Hu and Liu, 2004), it was observed that in a set of customer reviews, topics often overlapped with each other and were not perfectly distributed in the non-overlapping clusters. As shown in Figure 7.2 which lists some topics in the review set of Nokia 6610 and review IDs relevant to these topics, review 18 has comments regarding all the topics and some other reviews are also associated with multiple topics. The approach of clustering-summarization is not suitable in this situation since clustering this collection into non-overlapping groups will cut off the relationship among reviews.

- *Sound quality 8,13,**18**,20,27,33,34,40*
- *Battery life 2,5,10,13,17,**18**,26,28,29,30,37*
- *Flip phone 4,**18**,26,33*
- *Nokia phone 1,2,16,17,**18**,31,37*
- *Samsung phone **18**,40*
- …

Figure 7.2 Some topics from the review set of Nokia 6610

Therefore, the summarization framework proposed in Chapter 5 was applied to summarize multiple customer reviews because of its capability to handle such kind of macrostructure. Some modifications were made with the system in order to cater for the domain of customer reviews, as shown in Figure 7.3.

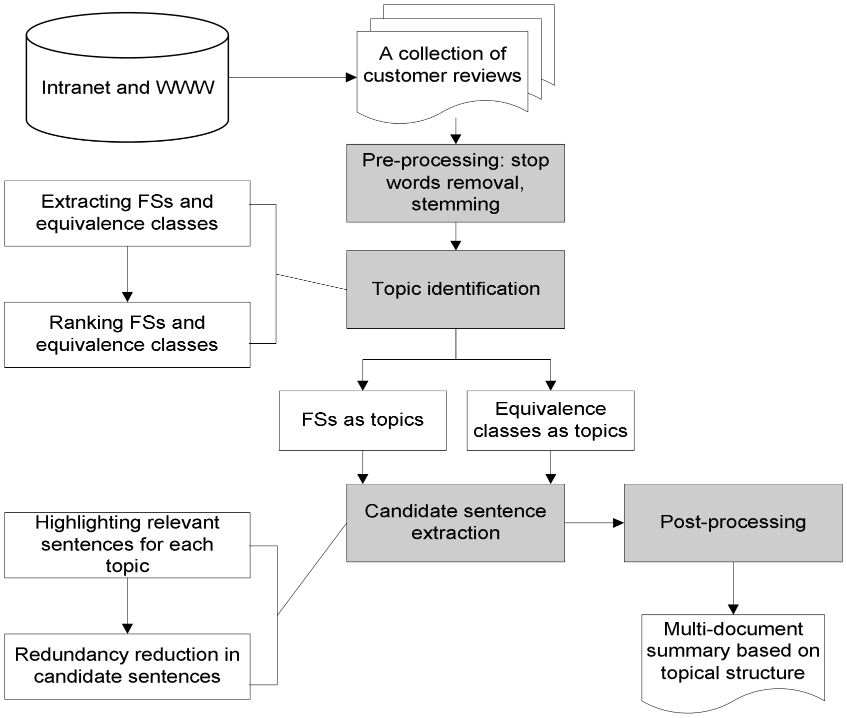
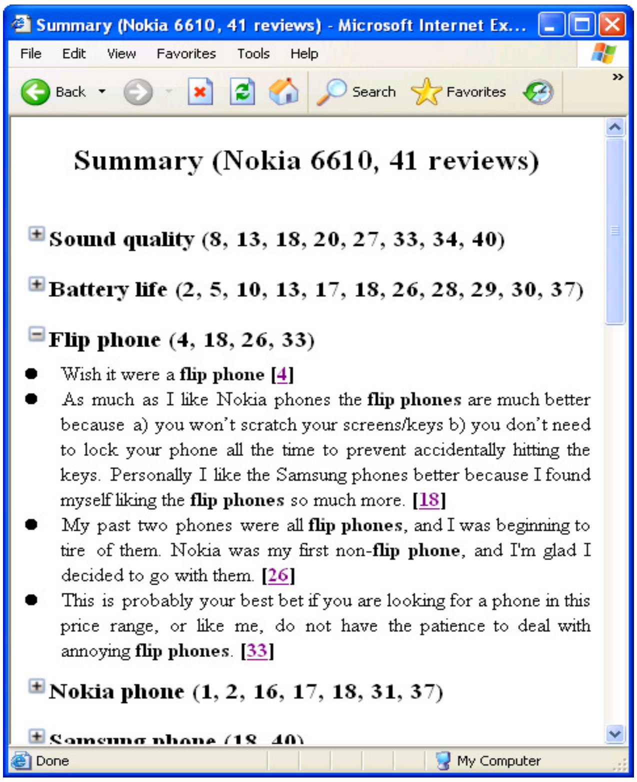

Intranet and WWW
A collection of customer reviews
Pre-processing: stop words removal, stemming
Extracting FSs and equivalence classes
Topic identification
Ranking FSs and equivalence classes
FSs as topics
Equivalence classes as topics
Candidate sentence extraction
Post-processing
Highlighting relevant sentences for each topic
Redundancy reduction in candidate sentences
Multi-document summary based on topical structure

Figure 7.3 Summarization of customer reviews based on macrostructure

The summarization process starts with a set of customer reviews as the input. These reviews are collected from WWW or retrieved from Intranet, e.g. all customer emails regarding a product. After pre-processing steps like stop words removal and word stemming, FSs are extracted and grouped into equivalence classes. A FS or an equivalence class is considered as the representative of one topic in a review set. In the following experiments, the performance between FSs and equivalence classes as topics would be compared. The topics are ranked using the ranking scheme given in Section 5.4.2.

For each topic in a review set, all relevant sentences are extracted and added into a pool as candidate segments of final summary until the expected summary length is reached. Each sentence will be accompanied by a label including its source review ID. The method of MMR is implemented to reduce the redundancy in the sentence selection process.

Figure 7.4 shows an example of the summary presented to readers. Topics are ranked according to their significance in the review set. Reviews relevant to each topic have been identified and hyperlinked, with their IDs included in the parenthesis following the topical phrase, to make it easy for users to browse the details of each review article. If users are interested in a particular topic, they can click the unfolding button prior to the topical phrase to expand this topic and the detailed information will then be presented. In Figure 7.4, the topic *flip phone* is unfolded and all the relevant sentences to this topic are displayed along with reviews' IDs.



Summary (Nokia 6610, 41 reviews)

⊞ **Sound quality (8, 13, 18, 20, 27, 33, 34, 40)**

⊞ **Battery life (2, 5, 10, 13, 17, 18, 26, 28, 29, 30, 37)**

⊟ **Flip phone (4, 18, 26, 33)**

- Wish it were a **flip phone** [4]
- As much as I like Nokia phones the **flip phones** are much better because a) you won't scratch your screens/keys b) you don't need to lock your phone all the time to prevent accidentally hitting the keys. Personally I like the Samsung phones better because I found myself liking the **flip phones** so much more. [18]
- My past two phones were all **flip phones**, and I was beginning to tire of them. Nokia was my first non-**flip phone**, and I'm glad I decided to go with them. [26]
- This is probably your best bet if you are looking for a phone in this price range, or like me, do not have the patience to deal with annoying **flip phones**. [33]

⊞ **Nokia phone (1, 2, 16, 17, 18, 31, 37)**

⊞ **Samsung phone (18, 40)**

Figure 7.4 Summarization output for the review collection of Nokia 6610

7.1.3 Experiment and Results

The summarization performance was compared with the output generated by opinion mining and the method of clustering-summarization. The data sets used in the experiment included five sets from Hu's corpus (Hu and Liu, 2004) and three sets from *Amazon.com*. These document sets were moderate-sized with 40 to 100 documents per set. Therefore, FSs were extracted with at least two supporting documents. The compression ratio of summarization was set to 10%, i.e. the length ratio of summary to original text was 10%. The summary generated by clustering-summarization was divided into clusters, as shown in Figure 7.5 (only three clusters are shown here).

Cluster 1 (4 reviews)

Sound - excellent polyphonic ringing tones are very nice (check cons) it also doubles as a radio, which is a nice feature when you are bored.

Cons: ring tones only come with crazy songs and annoying rings, there is only one ring that sounds close to a regular ring.

…

Cluster 2 (3 reviews)

Nice and small and excellent when it comes to downloading games, graphics and ringtones from www.crazycellphone.com I thought this was the ultimate phone when it comes to basic features, but I was dissapointed when I saw that it was only a gsm comaptible phone.

…

Cluster 3 (17 reviews)

I've had an assortment of cell phones over the years (motorola, sony ericsson, nokia etc.) and in my opinion, nokia has the best menus and promps hands down.

No other color phone has the combination of features that the 6610 offers.

From the speakerphone that can be used up to 15 feet away with clarity, to the downloadable poly-graphic megatones that adds a personal touch to this nifty phone.

…

Figure 7.5 Summary generated by the method of clustering-summarization for the review collection of Nokia 6610 (Only three clusters are shown here.)

Summarization performance was evaluated according to users' responsiveness. Human assessors were required to give a score for each summary based on its structure and coverage of important topics in the review collection. The score was an integer between 1 and 5, with 1 being the least responsive and 5 being the most responsive. In order to reduce bias in the evaluation, three human assessors from different background joined the scoring process. For one set, all the peer summaries were evaluated by the same human assessor so that the hypothesis testing (paired t-test) could be performed to compare the peer summaries.

Table 7.1 shows the average responsiveness scores of opinion mining, clustering-summarization and our approach (using FSs and equivalence classes as topics) based on all the review sets. Table 7.2 presents the results of paired t-test between our approach (using FSs as topics) and other methods. The comparison between FSs and equivalence classes as topics is also presented in Table 7.2

It could be found that our approach based on macrostructure performed significantly better than other peer methods (Table 7.1 & 7.2). The clustering effectiveness of customer reviews was also analyzed in this experiment. Table 7.3 shows the intra-cluster similarity and inter-cluster similarity for the review set of Nokia 6610. As can be seen, there was not much difference between intra-cluster similarity and inter-cluster similarity, especially for cluster 4 and 5 which were the two major clusters in the set. This implied that the review sets were difficult to be clustered into non-overlapping clusters.

As shown in Table 7.1 & 7.2, it was also found that using FSs as topics was significantly better than equivalence classes with the p-value of 0.0008 in paired t-test. Unlike technical paper authors, review writers usually write in an arbitrary style and cover different topics in a review (these topics may have little sensible relationship among each other). Therefore, using equivalence classes might introduce much noisy information, since equivalence classes group topics based on their co-occurrences.

Table 7.1 Average responsiveness scores

		Responsiveness score
Opinion mining		2.9
Clustering-summarization		2.3
Our approach	FSs	4.3
	Equivalence classes	2.6

Table 7.2 Hypothesis testing (paired t-test)

Null hypothesis (H_0): There is no difference between the two methods.
Alternative hypothesis (H_1): The first method outperforms the second one.

	P-value
Our approach (FSs) vs. opinion mining	1.91×10^{-3}
Our approach (FSs) vs. clustering-summarization	2.43×10^{-4}
Our approach FSs vs. equivalence classes	7.68×10^{-4}

Table 7.3 Intra-cluster similarity and inter-cluster similarity of the review set Nokia 6610 (41 reviews, 5 clusters)

Cluster ID	Size	Intra-cluster similarity	Inter-cluster similarity
1	2	0.684	0.343
2	4	0.592	0.431
3	3	0.606	0.454
4	17	0.692	0.546
5	15	0.645	0.553

7.1.4 Conclusion of Case Study 1

Summarization of online customer reviews is a process to transfer reviews from unstructured free texts to a structured or semi-structured summary which can reveal the commonalities and links among reviews. The automation of this process, in the context of e-Commerce and e-Business, should be able to assist potential consumers in seeking information and to facilitate knowledge management in enterprises as well.

The application of our proposed summarization approach on the domain of customer reviews has demonstrated better performance than the method of opinion mining in terms of readers' satisfaction. Compared to opinion mining, this approach is more capable of addressing different concerns from potential consumers, product designers and merchants. Potential consumers usually concentrate on the positive or negative comments given by other consumers. Designers and manufacturers, on the other hand, may be more concerned about the overall important issues and the reasons why customers are favoring or criticizing their products.

Compared to technical paper, customer review is a type of documents with relatively loose structure and review writers may cover different topics which have little sensible relationship in a review. This characteristic of customer reviews might result in the low performance of equivalence classes as topic candidates. Experimental results have shown that FSs achieved better performance than equivalence classes as topic candidates in the domain of customer reviews.

7.2 Case Study 2: Applying Summarization in Text Classification

Automatic text classification, or text categorization, is an important component in information management tasks, defined as assigning pre-defined category labels to new documents based on the likelihood suggested by a training set of labeled documents (Lee et al., 2002; Yang and Liu, 1999). Since summary is a distilled version of original document, it may substitute original document in text classification task to reduce redundancy and improve accuracy. This case study investigates the issues of applying summarization in text classification.

7.2.1 Motivation

A lot of supervised machine learning techniques have been applied in text classification, including Naïve Bayes (McCallum and Nigam, 1998), Rocchio (Joachims, 1997), K-Nearest Neighbor (Rahal and Perrizo, 2004), C4.5 (Gabrilovich and Markovitch, 2004), SVMs (Vapnik, 1995). Previous research showed that SVMs was the most robust algorithm for text classification problem, outperforming other methods substantially and significantly (Joachims, 1998; Yang and Liu, 1999).

For most classification algorithms, including SVMs, the “bag of words” representation is employed, where each document is transformed into a vector counting the number of occurrences of different words as features. One of the major problems of this representation scheme is the high dimensionality of the feature space (even tens or hundreds of thousands), which not only considerably increases the running time of the classification learning algorithm, but also results in the high level of feature redundancy and may thus reduce the classification accuracy. Feature selection is one technique to deal with such problem. Prior studies found that the accuracy for some classifiers could be improved by selecting an optimal subset from the feature space (Lewis and Ringuette, 1994; Rogati and Yang, 2002; Yang and Pedersen, 1997). However, SVMs, the best learning algorithm for text classification, has proven to be much less sensitive to feature selection. Reduction of feature space has no improvement or even small degradation on the performance of SVMs (Joachims, 1998 and 2001; Rogati and Yang, 2002).

Unlike feature selection techniques which only rank all the features and select top ones, summarization can be viewed as a process to select an optimal feature subset and re-weights all the features in the subset. Some studies have been reported by previous researchers to apply summarization in text classification (Ko et al., 2002; Kolcz et al., 2001; Shen et al., 2004). Kolcz et al. (2001) used summarization technique to select features and built classifier based on the reduced feature space. The result was competitive with state-of-the-art feature selection techniques. Ko et al. (2002) used summarization technique to calculate the importance of sentences and re-weighted all the features. Improvement was achieved on several classifiers including SVMs. Shen et al. (2004) employed summarization technique to increase the performance of web page classification. However, none of prior researches investigated the effect on classification from redundant information in summaries. In this study, we reduced redundancy in summaries to different levels in order to investigate its effect on SVMs performance. The particular method to reduce redundant information in summary is MMR, i.e. to scale down the scores of all the sentences not yet included in the summary by an amount proportional to their similarity to the summary generated so far (Carbonell and Goldstein, 1998).

7.2.2 Experimental Design

The experimental data were based on the corpus Reuters-21578 which was a standard corpus widely used in text classification research (Joachims, 1998, Yang and Liu, 1999). There are 21578 documents and 135 categories in this corpus. Each document may belong to one or more categories. To simplify the experiment, only the documents with single category were considered. After removing those articles with multiple labels, the remaining corpus had 9494 documents and 66 categories, in which many categories contained only one or two documents. Therefore, ten most populous categories were selected (see Table 7.4).

Table 7.4 Ten most populous categories in Reuters-21578

Category	Number of documents
earn	3945
acq	2362
crude	408
trade	361
money-fx	307
interest	285
ship	158
sugar	143
coffee	116
gnp	83

In the next, we removed very short documents for which summarization does not make sense, finally obtaining a corpus of 2130 documents (see Table 7.5). Further experiments were based on this corpus (denoted as Reuters-2130 in the rest of the chapter).

Table 7.5 The corpus Reuters-2130 used in the experiments

Category	Number of documents
acq	670
earn	499
trade	255
crude	234
money-fx	125
interest	82
ship	74
sugar	72
coffee	67
gnp	52

The purpose of this study was to investigate the redundant information in summarization and its effect on text classification performance. The method of MMR was used to reduce the redundancy in summaries. The initial goal of MMR was to reduce redundancy while preserving query relevance in re-ordering retrieved documents in information retrieval system (Carbonell and Goldstein, 1998). Since summarization has a similar process as information retrieval in terms of ranking and selection, MMR can also be used in summarization to reduce redundancy. The MMR-based summarization process in this experiment was a little complex than that in section 4.3.2. A parameter λ is added in order to tune the redundancy in the summary. The detailed summarization steps are described as follows:

- Calculate Marginal Relevance for each sentence in *D*-*S* (*D* is the whole document, *S* is the summary, *D-S* is the set of sentences in *D* which have not been included into *S*). The definition of Marginal Relevance for sentence s_i is:

$$MR(s_i) = \lambda Sim(s_i, D) - (1-\lambda)\max_{s_j \in S} Sim(s_i, s_j) \tag{7.1}$$

where $0 \le \lambda \le 1$.

In the first part of Equation 7.1, $Sim(s_i,D)$ is the similarity between s_i and the whole document *D*, which indicates the relevance of this sentence to the main topic of this document. In the second part, $Sim(s_i,s_j)$ is the similarity between s_i and s_j (s_j is any sentence from summary *S*). Therefore, the second part measures the redundant information carried by sentence s_i with respect to the summary *S*. The higher this value, the more redundant information is contained in this sentence.

- Pick up the sentence with maximal value of Marginal Relevance and add it into *S*.
- Repeat the first two steps until expected summary length is reached.

It can be seen from Equation 7.1, the redundancy contained in the summary was tuned by the value λ, which is ranged from 0 to 1. When λ is 1, the second part of Marginal Relevance equals 0 and there is no redundancy reduction in summarization process. When λ decreases to 0, the first part of Marginal Relevance is 0 and redundancy in the summary is reduced to minimal. Therefore, λ value actually indicates the level of redundancy contained in the final summary. In this experiment, summaries were generated for all documents in Reuters-2130 based on λ values of 0, 0.3, 0.7, 1 and were used for further classification tasks. Two examples of summaries based on λ=0 and λ=1 are listed here:

Summaries λ=0:

> *Dome Petroleum Ltd's proposal to restructure debt of more than 6.10 billion Canadian dlrs includes provisions that may force the company to sell its 42 pct stake in <Encor Energy Corp Inc>, Dome said in a U.S. Securities and Exchange Commission filing. "However, the final outcome of the negotiations cannot be predicted at this time," it said.*

Summaries λ=1:

> *Dome Petroleum Ltd's proposal to restructure debt of more than 6.10 billion Canadian dlrs includes provisions that may force the company to sell its 42 pct stake in <Encor Energy Corp Inc>, Dome said in a U.S. Securities and Exchange Commission filing. Dome said in the filing that its debt plan proposes making payments under a five year income debenture to the lender whose debt is secured by Dome's Encor shares.*

The classification performance was evaluated by average *F*-scores across all categories, under five-fold cross validation, as introduced in Section 5.5.3.3.

7.2.3 Experimental Results

Table 7.6 shows classification result of iteration 1 in the five-fold cross validation. Classification results based on original full documents are recorded in the column “Original”. The last four columns record the classification results based on summaries when λ value ranges from 0 to 1. Average *F* values across all categories are given at the end of this table.

From Table 7.6 it can be found that classification performance for some categories was greatly improved through summarization, such as categories “coffee” and “sugar”. Minor improvement or degradation was obtained for other categories. Summaries with λ=0 offered the best classification performance, which achieved better performance than full articles for all categories except the category "trade". On average, summaries with λ=0 achieved 77.09% of *F* value and about 5% more than original documents, and was also better than the performance of other λ values.

Table 7.6 SVMs classification results for iteration 1 in five-fold cross validation (percent)

Category	Original	Summaries			
		λ=0	λ=0.3	λ=0.7	λ=1
acq	90.63	93.24	93.75	93.24	93.78
coffee	88.37	97.87	97.87	97.87	95.65
crude	87.77	89.05	89.21	89.36	89.21
earn	84.21	87.13	83.79	86.38	81.91
gnp	75.00	78.26	69.56	63.64	69.56
interest	60.47	61.91	61.91	54.00	52.38
money-fx	62.07	71.88	58.18	62.07	55.56
ship	0	11.11	11.11	5.71	11.11
sugar	85.71	97.87	97.87	97.87	97.87
trade	87.50	82.63	84.88	84.03	84.21
Average *F*	**72.17**	**77.09**	**74.81**	**73.52**	**73.12**

Classification results for all iterations in five-fold cross validation are presented in Table 7.7 and Figure 7.6. The average *F* values for all iterations are given at the end of Table 7.7. From the results it was found that for all iterations, summaries made better classification performance than full articles and the best accuracy was achieved when λ=0, i.e. redundancy in summaries was reduced to minimal. On average, summaries with λ=0 could improve SVMs performance with more than 6% increase on *F* measure. The results also showed that lower λ value tended to generate higher classification performance, i.e. redundancy reduction in summaries was helpful for improving text classification accuracy.

Table 7.7 SVMs classification results for all iterations in five-fold cross validation (percent)

Iterations	Original	Summaries			
		λ=0	λ=0.3	λ=0.7	λ=1
1	72.17	77.09	74.81	73.52	73.12
2	70.82	76.96	75.02	75.63	75.10
3	66.39	74.27	74.19	73.57	72.18
4	69.84	79.02	77.92	77.14	75.10
5	73.68	78.05	76.11	76.75	74.53
Average	**70.58**	**77.08**	**75.61**	**75.32**	**74.01**

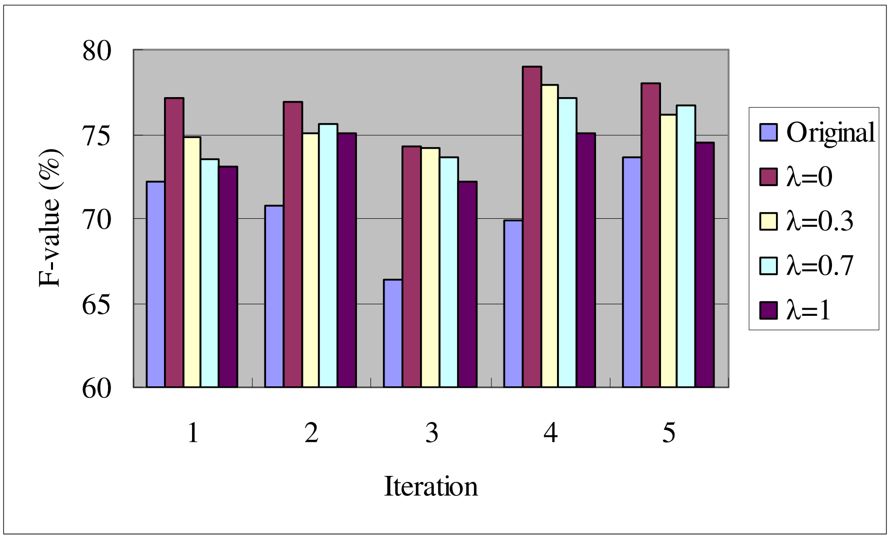


Figure 7.6 SVMs classification results for all iterations in five-fold cross validation

7.2.4 Further Discussion

As shown in the experimental results, summarization could improve the classification performance and redundancy reduction in summaries was helpful for SVMs algorithm. These results seem to contradict with previous studies which reported that feature selection was not effective with SVMs and might even degrade the classification performance on Reuters corpus (Joachims, 1998 and 2001; Rogati and Yang, 2002). In fact, summarization process not only selects an optimal feature subset, but also re-weights all the features in the subset. This process is different from traditional feature selection techniques which only rank all the features according to their importance and select top ones. This is clarified as follows.

After stop words removal and word stemming, Reuters-2130 had a set of 7889 features (noted as FS-7889). After summarization with λ=0 and compression ratio of 10%, the dimension of feature space reduced to 3871 (noted as FS-3871). The difference between FS-7889 and FS-3871 was a set of 4018 features, which was noted as FS-4018. We conducted classification of original full articles based on FS-7889, FS-3871 and FS-4018. The training and testing set were the same with iteration 1 in the previous five-fold cross validation procedure. The results are presented in Table 7.8.

Table 7.8 shows that classifier trained on FS-3871 with original documents achieved 71.68% of average *F* score, which was slightly lower than that of FS-7889 (72.17%) and much better than that of FS-4018, which was only 37.19%. This was probably because summarization actually selected an optimal subset (FS-3871) of features from the whole corpus (FS-7889) and FS-4018 contained most of the noisy features. This result was consistent with previous researches which showed that SVMs was not sensitive and even had minor degraded performance with feature selection.

The classification result of summaries (λ=0 and 10% compression ratio) based on feature space FS-3871 is also listed in Table 7.8. It was found that based on the same feature space of FS-3871, summaries achieved better classification performance (77.09%) than original documents (71.68%). The reason was probably that although based on the same feature space, summaries and full documents had different weights for each feature. The results showed that summaries offered a better weighting scheme for text classification.

Table 7.8: Comparison between summarization and feature selection

Category	Original FS-7889	Original FS-3871 (Optimal)	Original FS-4018 (Noisy)	Summaries FS-3871 (Optimal)
acq	90.63	90.14	73.48	93.24
coffee	88.37	90.91	28.58	97.87
crude	87.77	87.77	62.99	89.05
earn	84.21	85.62	60.00	87.13
gnp	75.00	75.00	0	78.26
interest	60.47	50.00	0	61.91
money-fx	62.07	56.60	40.00	71.88
ship	0	5.71	5.71	11.11
sugar	85.71	85.71	50.00	97.87
trade	87.50	89.38	51.13	82.63
Average *F*	**72.17**	**71.68**	**37.19**	**77.09**

7.2.5 Conclusion of Case Study 2

This case study focused on the application of summarization to improve text classification. Redundancy in summaries was reduced to different levels and its effect on classification performance was examined. The classification algorithm used here was SVMs which has proven to be very effective and robust for text classification. Experimental results showed that redundancy reduction was helpful to improve classification accuracy and summaries with lowest redundancy could improve the classification performance of Reuters corpus with more than 6% increase on average *F* measure.

In order to explain why the results showed that SVMs performance was improved by using summarization while previous studies reported that SVMs was not sensitive with feature selection, a further experiment was conducted to demonstrate the difference between summarization and traditional feature selection techniques. We trained classifiers using original documents based on the native feature space FS-7889 and optimal feature space FS-3871. The results showed that FS-3871 generated a slightly lower *F* score than FS-7889. This was consistent with previous studies which reported that SVMs was not sensitive with feature selection. However, it was also found that classifier trained using summaries was much better than classifier trained using original full documents based on the same feature space FS-3871, which probably means that summarization can re-weight the features and this re-weighting process is helpful for SVMs classification.

7.3 Conclusion of the Chapter

This chapter reports two case studies regarding text summarization. One case study was to apply summarization in processing online customer reviews to help product designers, merchants and potential shoppers for their information seeking. The other case study was to utilize summarization to improve the performance of automatic text classification. These two case studies have consolidated this research in the sense of applying summarization in another domain of documents and in downstream text mining tasks like text classification.

Chapter 8

Conclusions and Future Work

This chapter describes the conclusions that have been drawn from this study. Since this is a pioneering study regarding automatic text summarization within the engineering domain, there exist a few directions for future research. Therefore, the recommendations for possible future work in this area are also presented in this chapter.

8.1 Conclusions of the Study

This research investigated the significant issues of automatic text summarization within the engineering domain, with the focus on technical papers, in order to facilitate engineering information management.

This work might bridge the gap between engineering information management and automatic text summarization. Existing tools of engineering information management mainly focused on the domain of numerical data. Textual data, such as technical papers, patent documents, e-mails and customer reviews, which constitute a significant part of engineering information, have been somewhat ignored. On the other hand, the technique of automatic text summarization has been increasingly applied in information management in the past few years. However, the main focus of summarization, especially for multi-document summarization, was on the domain of news articles, and little interest has been taken in summarization of technical papers. The existing abstracts in digital libraries are all based on single papers and are unable to fulfill the diverse information needs in current knowledge-intensive engineering information management. A typical information need is to integrate information from different sources, i.e. multiple technical papers.

Compared to news articles, summarization of technical papers demands different approaches, mainly due to the differences in terms of readers' information requirements and the special characteristics of this document genre. In the summarization of multiple news articles, the summary aims to be an informative one which can substitute for the source articles to some extent. On the other hand, readers of multi-paper summary are more interested in the similarity and dissimilarity relationships among papers, so that they can further choose the papers most interesting to them. In this sense, the summary is an indicative one which provides a clue for further readings of the full papers. **The sum-up of differences between the domain of news articles and technical papers in Chapter 3 could provide a basis for further researches.**

In the preliminary investigation of multi-paper summarization, it was found that **the clustering-summarization method, one of the most popularly used multi-document summarization methods, did not address the specialties of the technical paper domain and could not reveal the internal structures of multiple**

papers, e.g. the topics within a set of documents are not perfectly distributed into non-overlapping clusters of documents. Therefore, it motivated the detailed investigation into the structures within multiple real-world documents and how these structures could help in multi-document summarization.

Based on the qualitative analysis of the DUC corpus of manual summaries, **the notions of macrostructure and microstructure were proposed and these two structures were believed to cover the most important information in the process of multi-document summarization.** Macrostructure was defined as the significant topics shared among different input documents, while microstructure was defined as sentences or clauses that act as elaborating or complementary information for macrostructure.

Two experiments were conducted to examine the influence of macrostructure and microstructure on summarization performance based on the general corpus of DUC. The first experiment demonstrated that human summarizers heavily relied on the macrostructure, i.e. topical structure, in writing their summaries. The more significant topics from the input documents were more likely to appear in the manual summaries and more likely to be agreed by different human summarizers. The topics was ranked by the ranking schemes of *tf*, *tf.df* and *tf.idf* in which *tf* and *tf.df* were found to achieve better performance than *tf.idf*, possibly because the macrostructure aimed to cover the common topics that appeared frequently across documents. The second experiment suggested that microstructure offered complementary information for macrostructure and the two structures constitute the important information in summarization modeling and evaluation.

The experiments proved the assumption that summary authors greatly relied on macrostructure in summarization process and they might include different details because of authors' different backgrounds, composition skills and understanding of the documents. **This finding might somewhat find a solution to the well-known challenge in multi-document summarization research that there does not exist a single best or "gold standard" summary.** Some previous studies reported this challenge because they found that there was often little consensus among reference summaries written by different authors for a same document set (Halteren and Teufel, 2003; Nenkova and Passonneau, 2004). By introducing the concept of macrostructure, different manual summaries might share a consensus in a macrostructure-level although they varied a lot in terms of word overlap.

Next, **a multi-paper summarization framework based on macrostructure and microstructure was proposed.** The following significant findings were acquired through experiments and evaluation of the proposed system:

- In the domain of technical papers, the microstructure was defined as rhetorical structure within each single paper, e.g. the paper starting with background, following with experiments and results, finally conclusion. The identification of such rhetorical structure has been transformed into a problem of automatically assigning rhetorical categories to every sentence or clause in the paper article. The algorithms of Naïve Bayes and SVMs were applied to build the classification models. The results showed that SVMs outperformed Naïve Bayes in terms of F-measure. The possible reason was that Naïve Bayes assumed that the features of the model were statistically independent of each other, whereas statistical analysis showed that in the rhetorical classification model, some features were highly correlated with each other, like the features “absolute location” and “relative location”, “action verbs” and “formulaic expressions”.

- Macrostructure was generated by grouping FSs into equivalence classes and each equivalence class is a representation for a topic. The factors in macrostructure generation were examined by ANOVA test using ROUGE measure. It was found that the threshold for supporting documents in topic extraction could significantly affect the summarization performance, and choosing 2 as the threshold was better than higher threshold values. This was probably because the document sets used in the experiments were moderate-sized with tens of documents and high threshold could probably prevent some important topics to surface. Moreover, it was found that including query penalty in the topic ranking scheme could significantly improve the summarization performance.

- Extrinsic evaluation has been adopted to compare the performance of the proposed summarization system with the peer systems of Copernic summarizer and clustering-summarization. The results showed that the summarization approach based on macrostructure and microstructure could better present the topical relationship among various papers and better recognize their similarities and difference. The evaluation, when benchmarked with the peer systems, also demonstrated the effectiveness of our approach in terms of precision and recall in assisting manual categorization of real-world technical papers.

Finally, **two case studies were introduced to consolidate and extend this research in the sense of applying summarization within engineering information management and text mining:**

- One case study was to apply summarization in processing online customer reviews to help product designers, merchants and potential shoppers for their information seeking. The application of our proposed summarization approach on the domain of customer reviews has demonstrated better performance than the method of opinion mining in terms of readers' satisfaction. Unlike technical paper, customer review is a type of documents with relatively loose structure and review writers may cover different topics which have little sensible relationship in a same review. This characteristic of customer reviews might result in the low performance of equivalence classes as topic candidates. Experimental results have shown that FSs achieved better performance than equivalence classes as topic candidates in the domain of customer reviews.

- The other case study examined the application of summarization to improve text classification and the effect of redundancy on classification performance. Experimental results showed that redundancy reduction was helpful to improve SVMs classification accuracy and summaries with lowest redundancy could improve the classification performance of Reuters corpus with more than 6% increase on average *F* measure. Moreover, this case study explained why SVMs performance was improved by using summarization while previous studies reported that SVMs was not sensitive with feature selection. Unlike normal feature selection techniques, summarization is a process to re-weight the selected features and this re-weighting process may be helpful for SVMs classification.

8.2 Recommendations for Future Work

This research is an initial study regarding automatic text summarization within the engineering domain. Therefore, it leaves a few directions for future work, which are listed as follows:

- In the proposed multi-document summarization approach, macrostructure was a list of topics which were generated by extracting FSs and grouping them into equivalence classes according to their co-occurrences. The topics in the macrostructure were organized in a parallel form rather than in a hierarchical form, which was helpful to simplify the experiment and was powerful enough to deal with the moderate-sized document sets in the experiments. However, when extending the proposed summarization approach to much larger document sets, the macrostructure topics may need to be handled in a hierarchical way, because of the topics' complexity and inherent hierarchy.

- This study has discussed some problems of summarization's linguistic quality, e.g. acronyms identification. The full aspects of linguistic quality, i.e. coherence and grammar, may be addressed in future work. One significant issue is regarding anaphoric reference, such as *this method*, *those experiments* used to avoid repetition. Anaphoric reference is an inevitable problem in the domain of technical articles which has not yet been solved effectively (Paice, 1990). The focus of future studies may be automatic detection of anaphoric references and linking them with their candidate substitutes in the source articles.

- In the experiments of this study, paper abstracts were utilized. Compared to full article which contains much more detailed information, abstract is a concise, non-redundant version. The purpose of paper abstract is to let readers know the main idea and decide whether it is worthwhile to read the full article. Also, readers can gain some idea about which parts of the full article are interesting to them. Therefore, abstracts were applied in the current experiments since indicative summarization was focused on. However, paper abstracts usually concentrate on authors' own contributions without much emphasis on other researchers' work. In the future studies, other parts of technical papers may be included in the experiments, such as *introduction* and *literature review*, because these parts may contain valuable information of background knowledge and review of existing research.

- In the case study of processing online customer reviews, the proposed summarization framework has been applied in the domain of customer reviews. The emergence of Blogs and e-Opinion portals has offered customers novel platforms to exchange their experiences, comments and recommendations. Reviews for a particular product may be obtained from various sources in very different writing styles. How to integrate information from such diverse sources can be another focus in the future research.

References

Ahonen, H. Finding all maximal frequent sequences in text. In Proceedings of the ICML'99 Workshop on Machine Learning in Text Data Analysis, Bled, Slovenia. 1999.

Ake, K., J. Clemons and M. Cubine. Information Technology for Manufacturing. St. Lucie Press. 2004.

Anderson, K. and C. Kerr. Customer Relationship Management. NY: McGraw-Hill Trade. 2001.

Barzilay, R. and M. Elhadad. Using lexical chains for text summarization. In Proceedings of the Intelligent Scalable Text Summarization Workshop (ISTS'97), ACL, Madrid, Spain, 1997.

Baxendale, P.B. Man-made index for technical literature - an experiment. IBM Journal of Research and Development, *2*(4), pp. 354-361. 1958.

Beil, F., M. Ester and X. Xu. Frequent term-based text clustering. In Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, Canada, 2002.

Biber, D. Dimensions of Register Variation: A Cross-Linguistic Comparison. Cambridge University Press, Cambridge, England. 1995.

Bishop, C.M. Neural Networks for Pattern Recognition. Oxford, England: Oxford University Press. 1995.

Blumberg, R. and S. Atre. The problem with unstructured data. DM Review. 2003.

Boros, E., P.B. Kantor and D.J. Neu. A clustering based approach to creating multi-document summaries. In Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New Orleans, LA, 2001.

Brandow, R., K. Mitze and L.F. Rau. Automatic condensation of electronic publications by sentence selection. Information Processing and Management, *31*(5), pp. 675-685. 1995.

Carbonell, J. and J. Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In Proceedings of the 21th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Melbourne, Australia, pp. 335-336, 1998.

Chaffey, D. and S. Wood. Business Information Management, Improving Performance Using Information Systems. Reading, MA: Addison-Wesley. 2004.

Choi, F.Y.Y. Advances in domain independent linear text segmentation. In Proceedings of the 1st North American Chapter of the Association for Computational Linguistics, Seattle, WA, pp. 26-33, 2000.

Clifton, C., R. Cooley and J. Rennie. TopCat: data mining for topic identification in a text corpus. IEEE Transactions on Knowledge and Data Engineering, *16*(8), pp. 949-964. 2004.

Cortes, C. and V. Vapnik. Support vector networks. Machine Learning, *20*, pp.

273-297. 1995.

Curtis, G. and D. Cobham. Business Information Systems: Analysis, Design and Practice (4th ed.). Reading, MA: Addison-Wesley. 2000.

Deerwester, S., S.T. Dumais, G.W. Furnas, T.K. Landauer and R. Harshman. Indexing by latent semantic analysis. Journal of the American Society of Information Science, *41*(6), pp. 391-407. 1990.

Earl, L.L. Experiments in automatic extracting and indexing. Information Storage and Retrieval, *6*(6), pp. 313-334. 1970.

Edmundson, H.P. New methods in automatic extracting. Journal of the Association for Computing Machinery, *16*(2), pp. 264-285. 1969.

Edwards, A.L. An Introduction to Linear Regression and Correlation. San Francisco, CA: W. H. Freeman. 1976.

Fayyad, U.M. Data mining and knowledge discovery: making sense out of data. IEEE Expert: Intelligent Systems and Their Applications, *11*(5), pp. 20-25. 1996.

Fong, A.C.M. and S.C. Hui. An intelligent online machine fault diagnosis system. Computing and Control Engineering Journal, *12*(5), pp. 217-223. 2001.

Gabrilovich, E. and S. Markovitch. Text categorization with many redundant features: using aggressive feature selection to make SVMs competitive with C4.5. In Proceedings of the 21st International Conference on Machine Learning, Banff, Canada, 2004.

Ganapathy, S., C. Ranganathan and B. Sankaranarayanan. Visualization strategies and tools for enhancing customer relationship management. Communications of the ACM, *47*(11), pp. 92-99. 2004.

Gardner, M. and J. Bieker. Data mining solves tough semiconductor manufacturing problems. In Proceedings of the 6th International ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Boston, MA, pp. 376-383, 2000.

Gentle, J.E. Singular value factorization. Numerical Linear Algebra for Applications in Statistics (pp. 102-103). Berlin: Springer-Verlag. 1998.

Goldstein, J., M. Kantrowitz, V. Mittal and J. Carbonell. Summarizing text documents: sentence selection and evaluation metrics. In Proceedings of the 22nd International ACM SIGIR Conference on Research and Development in Information Retrieval, Berkeley, CA, pp. 121-128, 1999.

Gong, Y. and X. Liu. Generic text summarization using relevance measure and latent semantic analysis. In Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New Orleans, LA, pp. 19-25, 2001.

Halteren, H. and S. Teufel. Examining the consensus between human summaries: initial experiments with factoid analysis. In Proceedings of the HLT-NAACL03 on Text Summarization Workshop, pp. 57-64, 2003.

Han, J. and M. Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann. 2001.

Hearst, M.A. TextTiling: segmenting text into multi-paragraph subtopic passages. Computational Linguistics, *23*(1), pp. 33-64. 1997.

Hicks, B.J., S.J. Culley and C.A. McMahon. A study of issues relating to information management across engineering SMEs. International Journal of Information Management, *26*(4), pp. 267-289. 2006.

Hobbs, J. Summaries from structure. In Working Notes of the Dagstuhl Seminar on Summarizing Text for Intelligent Communication. 1993.

Hovy, E. and C.-Y. Lin. Automated text summarization in SUMMARIST. In Advances in Automatic Text Summarization, I. Mani and M. Maybury (editors), pp. 81-94. MIT Press. 1999.

Hu, M. and B. Liu. Mining and summarizing customer reviews. In Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, WA, pp. 168-177, 2004.

Hutchins, J. Summarization: some problems and methods. In K. P. Jones (Ed), Meaning: the Frontier of Informatics, Informatics 9 (pp. 151-173). 1987.

Hyland, K. Persuasion and context: the pragmatics of academic metadiscourse. Journal of Pragmatics, *30*(4), pp. 437-455. 1998.

Ishino, Y. and Y. Jin (Eds.) Data Mining for Knowledge Acquisitions in Engineering Design. Netherlands: Kluwer Academic Publishers. 2001.

Jing, H., R. Barzilay, K. McKeown and M. Elhadad. Summarization evaluation methods: experiments and analysis. In Proceedings of the AAAI'98 Workshop on Intelligent Text Summarization, Stanford, CA, pp. 60-68, 1998.

Joachims, T. A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization. In Proceedings of the 14th International Conference on Machine Learning, Nashville, TN, 1997.

Joachims, T. Text categorization with support vector machines: learning with many relevant features. In Proceedings of the 10th European Conference on Machine Learning, Chemnitz, Germany, pp. 137-142, 1998.

Joachims, T. Making large-scale support vector machine learning practical. In B. Schölkopf, C. Burges and A. Smola (Eds.), Advances in Kernel Methods: Support Vector Learning (pp. 169-184). Cambridge, MA: The MIT Press. 1999.

Joachims, T. A statistical learning model of text classification for support vector machines. In Proceedings of the 24th ACM SIGIR International Conference on Research and Development in Information Retrieval, 2001.

Kan, M.-L., K.R. McKeown and J.L. Klavans. Domain-specific informative and indicative summarization for information retrieval. In Proceedings of DUC 2001 Workshop on Text Summarization, New Orleans, LA. 2001.

Karypis, G. Cluto: A software package for clustering high dimensional datasets. Release 1.5. Department of Computer Science, University of Minnesota. 2002.

Knight, K. and D. Marcu. Summarization beyond sentence extraction: a probabilistic approach to sentence compression. Artificial Intelligence, *139*(1), pp. 91-107. 2002.

Ko, Y., J. Park and J. Seo. Automatic text categorization using the importance of sentences. In Proceedings of COLING, 2002.

Kohonen, T. Self-Organizing Maps. New York: Springer-Verlag. 1997.

Kolcz, A., V. Prabakarmurthi and J. Kalita. Summarization as feature selection for text categorization. In Proceedings of the 10th International Conference on Information and Knowledge Management, Atlanta, GA, 2001.

Kupiec, J., J. Pedersen and F. Chen. A trainable document summarizer. In Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, WA, pp. 68-73, 1995.

Laudon, K.C. and J.P. Laudon. Management Information Systems: New Approaches to Organization and Technology. NJ: Prentice-Hall. 1996.

Lee, K.H., J. Kay, B.H. Kang and U. Rosebrock. A comparative study on statistical machine learning algorithms and thresholding strategies for automatic text categorization. In Proceedings of the 7th Pacific Rim International Conference on Artificial Intelligence, pp. 444-453, 2002.

Lee, J.-H. and S.-H. Park. Data mining for high quality and quick response manufacturing. In D. Braha (Ed.), Data Mining for Design and Manufacturing: Methods and Applications (pp. 179-206). Netherlands: Kluwer Academic Publisher. 2001.

Leong, K.K., K.M. Yu and W.B. Lee. Product data allocation for distributed product data management system. Computers in Industry, *47*(3), pp. 289-298. 2002.

Lewis, D.D. and W.A. Gale. A sequential algorithm for training text classifiers. In Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Dublin, Ireland, pp. 3-12, 1994

Lewis, D.D. and M. Ringuette. A comparison of two learning algorithms for text categorization. In Proceedings of the 3rd Annual Symposium on Document Analysis and Information Retrieval, 1994.

Li, H. and K. Yamanishi. Mining from open answers in questionnaire data. In Proceedings of Knowledge Discovery and Data Mining, San Francisco, CA, pp. 443-449, 2001.

Lin, C.-Y. ROUGE: a package for automatic evaluation of summaries. In Proceedings of Workshop on Text Summarization Branches Out, Post-Conference Workshop of ACL 2004, Barcelona, Spain, 2004.

Lin, C.-Y. and E. Hovy. Identifying topics by position. In Proceedings of the Applied Natural Language Processing Conference (ANLP-97), Washington D.C., pp. 283-290, 1997.

Lin, C.-Y. and E. Hovy. Automatic evaluation of summaries using n-gram co-occurrence statistics. In proceedings of Language Technology Conference (HLT-NAACL 2003), Edmonton, Canada, 2003.

Liu, D.T. and X.W. Xu. A review of web-based product data management systems. Computers in Industry, *44*(3), pp. 251-262. 2001.

Liu, X., H. Bo, Y. Ma and Q. Meng. A new approach for planning and scheduling problems in hybrid distributed manufacturing execution system. In Proceedings of the 6th World Congress on Intelligent Control and Automation, Dalian, China, 2006.

Liu, Y. A concept-based text classification system for manufacturing information retrieval. Ph.D. Thesis, National University of Singapore. 2005.

Loh, H.T., C. He and L. Shen. Automatic classification of patent documents for TRIZ users. World Patent Information, *28*(1), pp. 6-13. 2006.

Lovins, J. Development of a stemming algorithm. Mechanical Translation and

Computational Linguistics, *11*, pp. 22-31. 1968.

Luhn, H.P. The automatic creation of literature abstracts. IBM Journal of Research and Development, *2*(2), pp. 159-165. 1958.

Maña-López, M.J. Multidocument summarization: an added value to clustering in interactive retrieval. ACM Transaction on Information Systems, *22*(2), pp. 215-241. 2004.

Mani, I. Automatic Summarization. John Benjamins Publishing Company. 2001.

Mani, I. and E. Bloedorn. Summarizing similarities and differences among related documents. Information Retrieval, *1*(1-2), pp. 35-67. 1999.

Mani, I., T. Firmin, D. House, M. Chrzanowski, G. Klein, L. Hirschman, B. Sundheim and L. Obrst. The TIPSTER SUMMAC text summarization evaluation: final report. MITRE Technical Report MTR 98W0000138. Mclean, VA: The MITRE Corporation. 1998.

Mani, I., B. Gates and E. Bloedorn. Improving summaries by revising them. In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, College Park, MD, pp. 558–565, 1999.

Mani, I., G. Klein, D. House, L. Hirschman, T. Firmin and B. Sundheim. SUMMAC: a text summarization evaluation. Natural Language Engineering, *8*, pp. 43-68. 2002.

Mani, I. and M.T. Maybury (Eds.). Advances in Automatic Text Summarization. MIT Press. 1999.

Mann, W. and S. Thompson. Rhetorical structure theory: toward a functional theory of text organization. Text *8*(3), pp. 243-281. 1988.

Manning, C.D. and H. Schütze. Foundations of Statistical Natural Language Processing. MIT Press. 1999.

Marcu, D. The rhetorical parsing of natural language texts. In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics, pp. 96-103, 1997.

Marcu, D. Discourse trees are good indicators of importance in text. In I. Mani and M. Maybury (Eds.), Advances in Automatic Text Summarization (pp. 123-136). Cambridge, MA: The MIT Press. 1999.

McCallum, A. and K. Nigam. A comparison of event models for Naïve Bayes text classification. In AAAI-98 Workshop on Learning for Text Categorization, 1998.

McKeown, K. and D.R. Radev. Generating summaries of multiple news articles. In Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, WA, pp. 74-82, 1995.

Menon, R., H.T. Loh, S.S. Keerthi, A.C. Brombacher and C. Leong. The needs and benefits of applying textual data mining within the product development process. Quality and Reliability Engineering International, *20*(1), pp. 1-15. 2004.

Miller, G. WordNet: a lexical database for english. Communications of the ACM, *38*(1), pp. 39-41. 1995.

Minel, J., S. Nugier and G. Piat. How to appreciate the quality of automatic text summarization. In Proceedings of the ACL/EACL’97 Workshop on Intelligent Scalable Text Summarization, Madrid, Spain, 1997.

Mitchell, T. Machine Learning. McGraw Hill. 1996.

Moens, M.-F., R. Angheluta and J. Dumortier. Generic technologies for single- and multi-document summarization. Information Processing and Management, *41*(3), pp. 569-586. 2005.

Moens, M.-F. and R. De Busser. Generic topic segmentation of document texts. In Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New Orleans, LA, pp. 418-419, 2001.

Montgomery, D.C. and G.C. Runger. Applied Statistics and Probability for Engineers. John Wiley & Sons Inc. 2006.

Morris, J. and G. Hirst. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. Computational Linguistics, *17*(1), pp. 21-43. 1991.

Morris, A., G. Kasper and D. Adams. The effects and limitations of automatic text condensing on reading comprehension performance. Information Systems Research, *3*(1), pp. 17-35. 1992.

Nanba, H. and M. Okumura. Towards multi-paper summarization using reference information. In Proceedings of the 16th International Joint Conference on Artificial Intelligence, Stockholm, Sweden, pp. 926-931, 1999.

Nenkova, A. and R.J. Passonneau. Evaluating content selection in summarization: the pyramid method. In Proceedings of the NAACL 2004, Boston, MA, 2004.

Paice, C.D. Constructing literature abstracts by computer: techniques and prospects. Information Processing and Management, *26*(1), pp. 171-186. 1990.

Paice, C.D. and P.A. Jones. The identification of important concepts in highly structured technical papers. In Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Pittsburgh, PA, pp. 69-78, 1993.

Papineni, K., S. Roukos, T. Ward and W. Zhu. Bleu: a method for automatic evaluation of machine translation. In Proceedings of the Conference of the Association for Computational Linguistics, Philadelphia, PA, pp. 311-318, 2001.

Polanyi, L. Linguistic dimensions of text summarization. In Working Notes of the Dagstuhl Seminar on Summarizing Text for Intelligent Communication. 1993.

Ponte, J.M., and W.B. Croft. Text segmentation by topic. In Proceedings of the 1st European Conference on Research on Advanced Technology for Digital Libraries, Pisa, Italy, pp. 113-125, 1997.

Popescu, A.-M. and O. Etzioni. Extracting product features and opinions from reviews. In Proceedings of Joint Conference on Human Language Technology / Empirical Methods in Natural Language Processing (HLT/EMNLP’05), Vancouver, Canada, pp. 339-346, 2005.

Porter, M.F. An algorithm for suffix stripping. Program, *14*(3), pp. 130-137. 1980.

Radev, D.R. A common theory of information fusion from multiple text sources, step one: cross-document structure. In Proceedings of the 1st ACL SIGDIAL Workshop on Discourse and Dialogue, HK S.A.R., China. 2000.

Radev, D.R., H. Jing, M. Styś and D. Tam. Centroid-based summarization of multiple documents. Information Processing and Management, *40*(6), pp. 919-938. 2004.

Rahal, I. and W. Perrizo. An optimized approach for KNN text categorization using

P-trees. In Proceedings of ACM Symposium on Applied Computing, 2004.

Riley, K. Passive voice and rhetorical role in scientific writing. Journal of Technical Writing and Communication, *21*(3), pp. 239-257. 1991.

Rogati, M. and Y. Yang. High-performing feature selection for text classification. In Proceedings of the 11th International Conference on Information and Knowledge Management, McLean, VA, pp. 659-661, 2002.

Roussinov, D.G. and H. Chen. Information navigation on the web by clustering and summarizing query results. Information Processing and Management, *37*(6), pp. 789-816. 2001.

Saggion H., D. Radev, S. Teufel and W. Lam. Meta-evaluation of summaries in a cross-lingual environment using content-based metrics. In Proceedings of COLING, Taipei, Taiwan R.O.C., 2002.

Salton, G. and C. Buckley. Term-weighting approaches in automatic text retrieval. Information Processing and Management, *24*(5), pp. 513-523. 1988.

Salton, G., A. Singhal, M. Mitra and C. Buckley. Automatic text structuring and summarization. Information Processing and Management, *33*(2), pp. 193-207. 1997.

Salton G., A. Wong and C.S. Yang. A vector space model for automatic indexing. Communications of the ACM, *18*(11), pp. 613-620. 1975.

Schlesinger, J.D., J.M. Conroy, M.E. Okurowski and D.P. O'Leary. Machine and human performance for single and multidocument summarization. IEEE Intelligent Systems (special issue on Natural Language Processing), *18*(1), pp. 46-54. 2003.

Schutze, H. and C. Silverstein. Projections for efficient document clustering. In Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.74-81. 1997.

Schwabacher, M., T. Ellman and H. Hirsh. Learning to set up numerical optimizations of engineering designs. In D. Braha (Ed.), Data Mining for Design and Manufacturing: Methods and Applications (pp. 87-127). Netherlands: Kluwer Academic Publisher. 2001.

Scott, S. and S. Matwin. Feature engineering for text classification. In Proceedings of the 16th International Conference on Machine Learning, pp. 379-388, 1999.

Shafiei, F. and D. Sundaram. Multi-enterprise collaborative enterprise resource planning and decision support systems. In Proceedings of the 37th Hawaii International Conference on System Sciences, 2004.

Shen, D., Z. Chen, H.-J. Zeng, B. Zhang, Q. Yang, W.-Y. Ma and Y. Lu. Web-page classification through summarization. In Proceedings of the 27th Annual International ACM SIGIR on Research and Development in Information Retrieval, Sheffield, United Kingdom, 2004.

Stark, J. Engineering Information Management Systems: Beyond CAD/CAM, to Concurrent Engineering Support. NY: Van Nostrand Reinhold. 1992.

Stark, J. Product Lifecycle Management: 21st Century Paradigm for Product Realization. London, UK: Springer-Verlag. 2005.

Steinbach, M., G. Karypis and V. Kumar. A comparison of document clustering techniques. In Proceedings of KDD Workshop on Text Mining. 2000.

Sullivan, D. Document Warehousing and Text Mining. John Wiley & Sons. 2001.

Swales, J. Research articles in English. In Genre Analysis: English in Academic and Research Settings. Cambridge University Press, Cambridge, chapter 7, pp. 110-176. 1990.

Tan, P.-N., H. Blau, S. Harp and R. Goldman. Textual data mining of service centre call records. In Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Boston, MA, pp. 417-423, 2000.

Tanaka, F. and T. Kishinami. STEP-based quality diagnosis of shape data of product models for collaborative e-engineering, Computers in Industry, *57*(3), pp. 245-260. 2006.

Teufel, S. and M. Moens. Summarizing scientific articles: experiments with relevance and rhetorical status. Computational Linguistics, *28*(4), pp. 409-445. 2002.

Tirpack, T.M. Design-to-manufacturing information management for electronics assembly. International Journal of Flexible Manufacturing Systems, *12*(2), pp. 189-205. 2000.

Tkach, D. (Ed.). Text Mining Technology, Turning Information into Knowledge: A White Paper from IBM. IBM Corporation. 1997.

Tombros, A. and M. Sanderson. Advantages of query biased summaries in information retrieval. In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Melbourne, Australia, pp. 2-10. 1998.

Trawiński, B. A methodology for writing problem-structured abstracts. Information Processing and Management, *25*(6), pp. 693-702. 1989.

Turney, P.D. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Philadelphia, PA, pp. 417-424, 2001.

Van Rijsbergen, C. Information Retrieval. Butter Worths. 1979.

Vapnik, V. The Nature of Statistical Learning Theory. Springer, New York. 1995.

Visa, A. Technology of text mining. In Proceedings of Machine Learning and Data Mining in Pattern Recognition, Second International Workshop, Leipzig, Germany, pp. 1-11, 2001.

Voorhees, E.M. Implementing agglomerative hierarchical clustering algorithms for use in document retrieval. Information Processing and Management, *22*(6), pp. 465-476. 1986.

Willcocks, L.P. and R. Sykes. Enterprise resource planning: the role of the CIO and its function in ERP. Communications of the ACM, *43*(4), pp. 32-38. 2000.

Witten, I.H. and E. Frank. Data Mining: Practical Machine Learning Tools and Techniques (2nd Edition). San Francisco: Morgan Kaufmann. 2005.

Wood, W.H., M.C. Yang and M.R. Cutkosky. Design information retrieval: improving access to the informal side of design. In Proceedings of ASME DETC Design Theory and Methodology Conference, 1998.

Yan, E., C.H. Chen and L.P. Khoo. A radial basis function neural network multicultural factors evaluation engine for product concept development. Expert Systems, *18*(5), pp. 219-232. 2001.

Yang, M.C., W.H. Wood and M.R. Cutkosky. Data mining for thesaurus generation in informal design information retrieval. In Proceedings of Congress on Computing in Civil Engineering, Boston, MA, pp. 189-200, 1998.

Yang, Y. and C.G. Chute. An example-based mapping method for text categorization and retrieval. ACM Transactions on Information Systems, *12*(3), pp. 252-277. 1994.

Yang, Y. and X. Liu. A re-examination of text categorization methods. In Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 42-49, 1999.

Yang, Y. and J.O. Pedersen. A comparative study on feature selection in text categorization. In Proceedings of the 14th International Conference on Machine Learning, 1997.

Yap, I., H.T. Loh, L. Shen and Y. Liu. Topic detection using MFSs. In Proceedings of the 19th International Conference on Industrial & Engineering Applications of Artificial Intelligence & Expert Systems, Annecy, France. 2006.

Yeh, J.-Y., H.-R. Ke, W.-P. Yang and I.-H. Meng. Text summarization using a trainable summarizer and latent semantic analysis. Information Processing and Management, *41*(1), pp. 75-95. 2005

Yen, P.H., B. Tseng and C.C. Huang. Rough set based approach to feature selection in customer relationship management. In Proceedings of the 15th International Conference on Information Management, Taiwan, 2004.

Zamir, O. and O. Etzioni. Grouper: a dynamic clustering interface to Web search results. Computer Networks, *31*(11-16), pp. 1361-1374. 1999.

Zappen, J.P. A rhetoric for research in sciences and technologies. In P.V. Anderson, R.J. Brockman and C.R. Miller (Eds.), New Essays in Technical and Scientific Communication Research Theory Practice (pp. 123-138). Baywood, Farmingdale, NY. 1983.

Zhan, J., H.T. Loh and Y. Liu. Automatic summarization of online customer reviews. In Proceedings of the 3rd International Conference on Web Information Systems and Technologies (WEBIST), Barcelona, Spain, 2007.