Textual Data Mining Applications for Industrial Knowledge Management Solutions

By
Nadeem Ur-Rahman

A Doctoral Thesis

Submitted in partial fulfilment of the requirements for the award of

Doctor of Philosophy of Loughborough University

June 2010

Abstract

In recent years knowledge has become an important resource to enhance the business and many activities are required to manage these knowledge resources well and help companies to remain competitive within industrial environments. The data available in most industrial setups is complex in nature and multiple different data formats may be generated to track the progress of different projects either related to developing new products or providing better services to the customers. Knowledge Discovery from different databases requires considerable efforts and energies and data mining techniques serve the purpose through handling structured data formats. If however the data is semi-structured or unstructured the combined efforts of data and text mining technologies may be needed to bring fruitful results. This thesis focuses on issues related to discovery of knowledge from semi-structured or unstructured data formats through the applications of textual data mining techniques to automate the classification of textual information into two different categories or classes which can then be used to help manage the knowledge available in multiple data formats.

Applications of different data mining techniques to discover valuable information and knowledge from manufacturing or construction industries have been explored as part of a literature review. The application of text mining techniques to handle semi-structured or unstructured data has been discussed in detail. A novel integration of different data and text mining tools has been proposed in the form of a framework in which knowledge discovery and its refinement processes are performed through the application of Clustering and Apriori Association Rule of Mining algorithms. Finally the hypothesis of acquiring better classification accuracies has been detailed through the application of the methodology on case study data available in the form of Post Project Reviews (PPRs) reports. The process of discovering useful knowledge, its interpretation and utilisation has been automated to classify the textual data into two classes.

Keywords:

Knowledge Discovery, Knowledge Management, Data Mining, Text Mining, Clustering, MKTPKS Termset Mining, Decision Trees, K-nearest Neighbouring (K-NN), Naïve Bayes, Support Vector Machines (SVMs), Post Project Reviews (PPRs)

Dedication

TO MY PARENTS

Who always stand with me and endeavoured to support me in attainment of ultimate goal of life

AND FAMILY

Who loved and cared me to collect sweet memories and had enjoyable moments in life

Acknowledgements

First of all thanks to Allah Almighty Who helped me to finish this thesis.

I thank to Dr. Jenny Harding whose generosity, kindness, patience, continuous support and professional guidance helped me to complete Ph.D. studies. I might not be able to find a personality like her as a supervisor who is fully touched with feelings, cooperative in work, constructive in thoughts and logical in her approach towards research. I really enjoyed working with Dr. Jenny Harding and hope that this would not be an end of relationship and cooperation but a beginning of new relationship where research work will be harnessed by her valuable suggestions and guidance in future.

I want to thank my parents and all my family members whose love provided me strength to stand against any hardship and helped me to attain my objective. I am also thankful to my wife who always stood beside me in every hour of need. I would also like to share the great moments of happiness that I had ever enjoyed as being a father of baby girl during writing up my thesis.

I am also thankful to Wolfson School of Mechanical and Manufacturing Engineering, Loughborough University, UK for giving me an opportunity as a Ph.D. student and provided financial support for pursuing research studies.

I would also like to pay special thank to Muhammad Shahbaz for his moral support and feel that there might be a few such friends in life.

I am also thankful to all my friends and colleagues specially George Guninderan, B.P. Das, Srinivas, A.K. Choudhary, Rahul Swarnkar, Natesh Khilwani and Claire Palmer for making lab a place where we can share our feelings, research ideas and have friendly environment.

I also thank to the industrial collaborators for providing case study data to perform experimental work during this research studies.

Glossary of Terms

AI	Artificial Intelligence
ANN	Artificial Neural Networks
APN	Automated Perceptions Network
AR-rules	Attribute Relationship Rules
BN	Bayesian Network
CAD	Computer Aided Design
CAM	Computer Aided Manufacturing
CMM	Coordinate Measuring Machine
CNC	Computer Numerical Controlled
CRISP-DM	Cross-Industry Standard Process for Data Mining
DM	Data Mining
Document	Set of textual data or information available in the form of case study data
DTE	Decision Tree Expert
EBC	Environment-Based Constraints
EDA	Exploratory Data Analysis
EDM	Evidential Data Mining
ELT	Extraction-Transformation-Load
FTS	Frequent Termset Sequences
GBOM	Generic Bill of Material
HG-Trees	Hierarchical Generalization Trees
HTML	Hypertext Markup Language
IMS	Intelligent Manufacturing Systems
IR	Information Retrieval
IT	Information Technology
KDD	Knowledge Discovery from Databases
KDT	Knowledge Discovery from Textual Data
KM	Knowledge Management
K-NN	K-Nearest Neighbouring
KREFS	Knowledge Refinement System,
LS-SVM	Least Square Support Vector Machines
MKS	Mining Kernel System

MKTPKS	Multiple Key Term Phrasal Knowledge Sequences
ML	Machine Learning
NLP	Natural Language Processing
NNs	Neural Networks
OLAP	Online Analytical Processing
PDD	Product Design and Development
PDM	Product Data Management
PLM	Product Lifecycle Management
PPRs	Post Project Reviews
SGML	Standardized Markup Langauge
SVMs	Support Vector Machines
TDM	Textual Data Mining
TM	Text Mining
VSM	Vector Space Model
XML	Extensible Markup Language

Table of Contents

Chapter 1 Introduction

1.1 Research Context and Introduction

In recent years Knowledge Management (KM) has become an important contributor to the success of enterprises. Increasing product complexity, globalization, virtual organizations and customer oriented demand require a more thorough and systematic management of knowledge both within individual enterprises and between several cooperating enterprises. Information Technology (IT) supported KM solutions are often built around an organizational memory that integrates informal, semi-formal and formal knowledge to facilitate the access, sharing and reuse of knowledge by members of the organization(s) to solve their individual or collective tasks. In such a context, knowledge has to be modelled, appropriately structured and interlinked to support flexible integration and its personalized presentation to the customer.

In this era of information technology a vast amount of data is collected, stored and reused to improve the product or service quality and for customer needs identification within industrial setups. Manual handling of such a large amount of data is expensive in terms of time and money spent, and it is particularly costly and difficult to fully exploit information that is available in different or mixed data formats (i.e. structured, semi-structured or unstructured).

Machine Learning (ML) and advanced Artificial Intelligence (AI) tools are gaining more importance in this data rich and information poor world of knowledge. Different data analysis tools are available to provide both online and off the shelf solutions to the problems associated with handling large amounts of data and information which is available in different data formats. These tools are used to discover useful information through exploiting relationships between different attributes of data and help to discover patterns which can ultimately be transformed into a source of knowledge. The most recent use of these tools can be categorized as forms of Data Mining technologies. The technological efforts available under this heading have potential to uncover useful patterns in data and discover valuable knowledge in any industrial setups (as seen in Chapter 3) but use of these technological efforts is still in its infancy. To handle multiple different formats (i.e. semi-structured or un-structured

data) existing tools may need to be combined with other technological efforts i.e. Text Mining (TM) which might be quite useful in discovering valuable information and converting it into a form of knowledge to help in providing knowledge based solutions.

(Nonaka and Takeuchi 1995) proposed that knowledge creation was achieved by the interaction of tacit and explicit knowledge which further helped to generate organizational knowledge. (Ler 1999) pointed out that knowledge management involves collecting information and transferring it to the demanders. These activities include knowledge obtaining (or creation), knowledge refinement, knowledge storing and knowledge sharing. Therefore knowledge management has become a major manufacturing tool and pre-requisite for success in production environments and competitive advantages can be obtained by taking into consideration and carefully selecting and applying appropriate knowledge management techniques.

This research will therefore consider the path defined by (Ler 1999) to manage knowledge but will approach this by using Data and Text Mining techniques. The aim of this research will be to examine how knowledge may be more accurately identified and classified automatically from textual reports, so that more reliable knowledge will available to be automatically identified, refined and transferred to the demanders as required by (Ler 1999).

Different Data Mining tools including Clustering, Association Rules, Decision Trees, K-NN, Naïve Bayes, Support Vector Machines (SVMs) and Text Mining tools in particular based on Information Retrieval (IR) techniques for data structuring have been considered and well investigated for handling textual databases. Both individual and hybrid approaches have been used to harness Knowledge Discovery and Management Solutions for which a conceptual background is provided by the theoretical development of the methodology detailed in Chapter 4. Different knowledge discovery functions (i.e. Clustering and Frequent Termset Mining Methods) are discussed in Chapter 5 while an architecture has been detailed in Chapter 6 based on the hybridized efforts of both Data and Text Mining technologies.

The proposed methodology in Chapter 6 has been applied and tested on the semi-structured textual data available in the form of Post Project Reviews (PPRs) reports from real construction industry environments. The results, in the form of discovered knowledge, have been compared with the results of domain experts (detailed in Chapter 7) in the form of multiple key term phrases identified by the system. The limitations in terms of classifying textual data through unsupervised Clustering and Apriori Association Rule of Mining techniques have been addressed and consequently knowledge of how to apply these tools and methods have been advanced by this research. This advancement is done through use of Supervised methods of classifying textual data into two categories or classes (as discussed and implemented in Chapter 8 & Chapter 9).

1.2 Thesis Structure

The overall thesis has been divided into four sections, detailed in the following paragraphs and shown in figure 1.1.

1.2.1 Background, Context and Scope of the Research

This section consists of three chapters. Chapter 1 discusses the introductory material containing research context and main aim of the research. Chapter 2 discusses the research problem, scope and issues. The aim and objectives of the research and research approach adopted are also detailed in this chapter. In Chapter 3 the literature is reviewed to form the basis of knowledge discovery process from multiple data formats through applications of different data mining techniques. The knowledge and information needs for intelligent decision making are also discussed in this chapter. It also provides support to identify the gap in terms of applications of different data mining techniques to exploit knowledge resources in new industrial and business setups.

1.2.2 Concept Realization and Methodological Development

This section is composed of three Chapters where Chapter 4 is focused on defining conceptual relationships among different parts of knowledge discovery and management technologies i.e. Data Mining, Text Mining and Business Intelligence to

develop a framework which is detailed in Chapter 6. Chapter 5 discusses knowledge discovery functions and their implementation issues to help in discovery of useful information in terms of finding term based relationships by forming Multiple Key Term Phrasal Knowledge Sequences (MKTPKS).

1.2.3 Implementation and Analysis

This section further consists of three chapters. Chapter 7 discusses the implementation of different functions required within the methodology for finding the knowledge relationships among different terms in the textual data available in the form of Post Project Reviews (PPRs). Chapter 8 discusses the text classification methods and their implementation on the PPRs data to classify it into two different categories of Good or Bad information documents. The accuracy of different classifiers is tested on the experimental data and better classification accuracies are reported by adopting the proposed methodology. Chapter 9 discusses the semantic analysis methods based on Information Retrieval (IR) techniques to classify the PPRs data into two different classes. The main hypothesis investigated in this research is that better classification accuracies will be achieved using the proposed methodology.

1.2.4 Conclusions and Future Work

This section contains chapter 10 which discusses the results obtained and their reliability and mentions some of the research limitations which could not be explored during this research. The further improvements to the proposed system are also discussed and future dimensions for improving the product or service quality based on the proposed methodology are also articulated at the end.

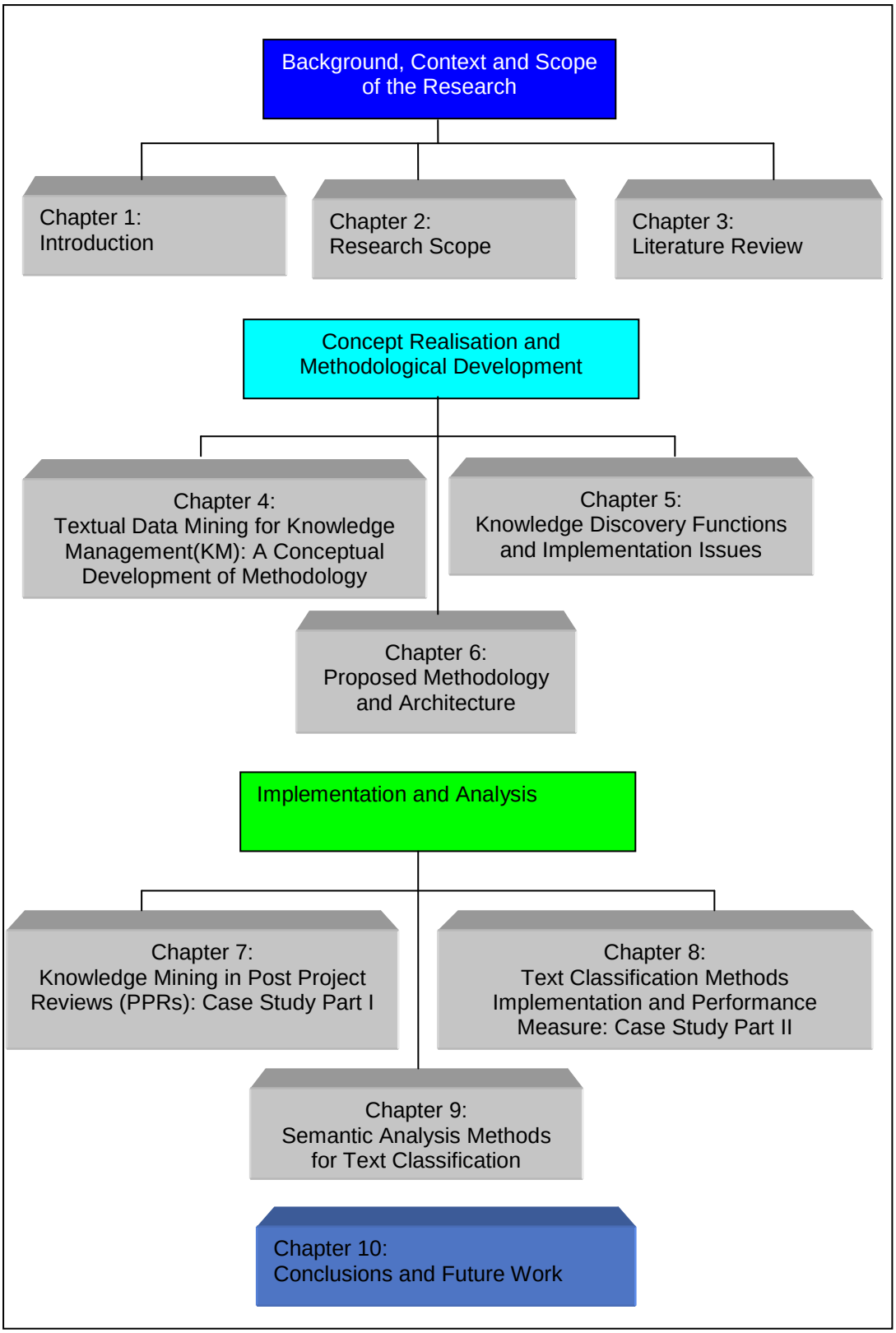

Figure 1.1: Overview of thesis structure

Chapter 2 Research Scope

2.1 Introduction

This chapter provides the research background and explains the research problem scope and issues. It also details the aims and objectives of the research and briefly explains the adopted research approach.

2.2 Research Issues

The research reported in this thesis examines ways of discovering useful knowledge through the classification of textual data into one of two categories of document i.e. good or bad information classes. The system proposed is based on Hybrid Applications of Data and Text Mining Techniques to provide knowledge discovery and management solutions for industrial or business environments. The term "***document***" used in this research work refers to the set of information available in the form of textual data under different headings (i.e. time, cost, planning) in Post Project Reviews (PPRs).

2.2.1 Research Background & Motivations

Knowledge discovery and knowledge management solutions can provide great benefits and better services in many industrial contexts and these will be discussed further in Chapter 3. Business Intelligence Systems may be used to help to uncover useful information and convert it into a form of knowledge by utilising Data and Text Mining techniques. There is a wide range of these techniques available which can be used independently or in combinations with other tools, to discover valid, novel, potentially useful and ultimately understandable knowledge. These techniques will be discussed further in Chapter 3 and Chapter 4.

In current industrial business setups a large amount of information is transferred through the use of internet technologies. Most of this data is available in the form of online documents, customer feedback etc. which is used to update information about a product or service to help to improve quality and competitive performance. These data types include:

- Structured data files, information stored in database management systems or specific applications, such as data warehousing, enterprise resource planning, cost estimating, scheduling, payroll, finance and accounting packages;
- Semi-structured data files such as Hypertext Markup Language (HTML), Extensible Markup Language (XML), or Standardized Markup Langauge (SGML) files;
- Unstructured text data files, such as product catalogues, quality reports, internal documents, memos, Failure Mode and Effect analysis (FMEA) documents, contracts, specifications, catalogues, change orders , requests for information, field reports and meeting minutes;
- Unstructured graphic files stored in a binary format such as 2D and 3D drawings; and finally
- Unstructured multimedia files such as pictures, audio and video files.

These textual sources of information are read and understood by people but manual handling is very slow, tedious and prone to errors. Therefore automatic text classification or text categorization methods can be used to overcome these difficulties. The handling of these sources of information can provide benefits in various ways to engineers (Kasravi 2004) i.e.

- Improving the engineering processes through having access to the information and downstream analysis,
- Patent analysis, handling product warranty claims and quality issues of the product,
- Failure mode and effect analysis (FMEA) and finally
- gaining the competitive assessments by reviewing the product announcements information available in the form of web pages which are almost impossible to handle manually

In addition to the above some other specific advantages related to the manufacturing industry can be obtained by;

- Improving manufacturing product design (Dong and Agogino 1997) ,
- Analyzing customer service database for machine fault diagnosis (Fong and Hui 2001)

- In automotive industry mapping auto problems to some diagnostic categories such as engine, transmission, electrical, brakes etc. (Huang and Murphey 2006)
- Integrating supply chain into total process, capturing and re-using best practices and specifying key characteristics at early stages of product development. These are identified as needs for next generation product development (Rezyat 2000)

In the construction industry automated text classification methods can provide benefits in organizing and finding the desired information from project documents according to the project components (Cladas and Soibelman 2003). This information can further be used to extract new and interesting information like accessing facility condition assessments provided at different campuses (Ng et al. 2006). There are other areas of research which also benefit from data analysis techniques like DNA sequence analysis, pattern discovery, image recognition and speech recognition.

The above discussion shows the benefits that can be achieved through textual data analysis methods. Therefore analyzing this information provides new sources of information and knowledge to the knowledge workers and an enterprise can thereby gain some competitive advantages.

2.2.2 Gap Analysis and Research Scope

Data mining techniques have been used to discover knowledge from manufacturing or construction data to support decision makers to improve product design, integrate shop floor activities and manage the customer knowledge. However the literature reviewed in Chapter 3 and in (Harding et al. 2006) showed that these techniques are mainly used to exploit information hidden in the numerical or more structured forms of data whilst there are fewer reported applications in handling the semi-structured or unstructured data types. Key points which have motivated this research and determined the scope of the studies are as follow:-

1. Data Mining techniques are mainly used to discover knowledge from structured databases. Hence there may be opportunities to apply knowledge discovery techniques on semi-structured or unstructured databases.
2. The applications of Data Mining techniques for knowledge discovery processes are mostly by using individual tools which help to identify relationships among different attributes defined in the data but are limited in their ability to discover useful knowledge in an understandable pattern.
3. The hybrid applications of data mining techniques appear to have potential to handle data and information, but currently there are less reported applications of these approaches.
4. Discovering knowledge from semi-structured or unstructured databases requires some additional efforts but they have potential to help to uncover useful patterns of information and knowledge in textual data formats. Data mining techniques can be used on their own to find potentially useful knowledge but handling semi-structured or unstructured data requires additional efforts, possibly by applying Information Retrieval (IR) and Natural Language Processing (NLP) techniques.
5. The combined efforts of Data and Text Mining technologies could be used to manage multiple data and information resources and enhance knowledge management activities. The combined efforts of data and text mining techniques can be fruitful as shown in Chapter 3 and Chapter 4.
6. The rapid growth of information in electronic and digital formats in globally distributed industrial environments could capture useful information about customers needs. The identification of these needs and classification of information available in the form of free formatted textual data into two different classes (i.e. good or bad information) might help in retaining the customers to the industry. But little attention has been paid to this area of information and knowledge research.
7. Domain experts play an effective role in the process of discovering useful knowledge through applications of data mining techniques where the data mining expert and domain expert sometimes work interactively. However if the knowledge discovery process can be automated fully this would help to reduce the effort required by the knowledge workers in terms of time and money spent.

Therefore the scope of this research takes into account the above points related to Data and Text Mining techniques to manage the data and information sources for knowledge discovery and knowledge management purposes.

2.2.2.1 Data Mining

Data mining techniques have influenced various sectors of life ranging from finances to life sciences but these techniques have only been used for exploiting manufacturing data for discovery of useful knowledge in the last decade or two. Handling these data and information sources usually requires additional knowledge to be provided by domain experts. An overview of a range of data mining techniques is given in Chapter 3 and an introduction to new techniques or hybridized methods which gave useful results has been undertaken within this research. The reported tools such as clustering and association rules analysis have been tested in manufacturing related data analysis, so the implementation of these tools has been tested along with their scope for discovering useful knowledge by exploiting term based relationships and the refinement of this process has also been considered. Finally, some supervised classification techniques are used to classify data into two classes in terms of good or bad information documents .

2.2.2.2 Text Mining

Text mining techniques are examined in Chapter 4 and it is clear that to discover knowledge from semi-structured or unstructured text data requires some additional efforts which have been used from the area of Information Retrieval (IR).These methods are mainly based on term frequency, term frequency inverse document frequency, and binary representation methods. These methods provide solutions to structure data and make it ready for the application of different data mining techniques and also help to overcome the difficulty of losing key information in the textual databases. Classifying textual data into two classes is highly dependant on the choice of these structuring techniques where different data mining algorithms use various functions to measure the degree of closeness of one document from another i.e. to classify as either good or bad information documents in the current research context.

2.3 Aims and Objectives of Research

The main aim of this research was defined in Section 1.1 and it will be achieved by satisfying the following objectives:-

1. Understanding different applications of DM/ TM techniques which have been well reported in the literature, and testing them on the exemplary data relevant to the current research context.
2. Establish which current applications provide potentially useful knowledge in this context to establish a datum for comparison with new or hybridized applications of different data mining techniques.
3. Identify new or less commonly used DM / TM techniques in handling the real industrial data and test their reliability in managing the information and knowledge resources.
4. Determine a framework and methodology to analyse textual data and transfer identified knowledge to demanders as required. Thus further shifting the paradigm from knowledge discovery to knowledge management.

2.4 Research Overview

The proposed framework should be able to accommodate semi-structured or unstructured text data. The architecture proposed is entirely based on knowledge management methods defined by (Ler 1999) , but will utilize the combined efforts of data and text mining techniques. The efforts made under the cover of these technological efforts of data and text mining should play an effective role in integrating the KM activities by identifying valid, novel, potentially useful and ultimately understandable patterns in the data. The purpose of these combined efforts will therefore be to provide valuable understanding of the new ways of automatically identifying knowledge in textual reports. The research is based on the hypothesis that knowledge management methods and in particular knowledge generation methods have strong impact on the effective use and exploitation of valuable knowledge assets in many industrial contexts. The second hypothesis is that knowledge generation can be done effectively through the combined use of Data and Text Mining Techniques. The generated knowledge may require refinement and this will be achieved by the use of better predictive and classification techniques. These techniques will utilise either

single data mining techniques or a hybridization of several techniques available under the heading of textual data mining technologies.

2.4.1 Research Approach

The research approach used in the thesis is divided into two stages i.e. the conceptual development stage and the experimental stage to verify the research hypothesis. Conceptual development of the methodology is based on the knowledge discovery and management methods detailed in Section 4.3 which provide strength to handle textual databases. The detailed process of discovering knowledge from semi-structured or unstructured databases is followed by using the path defined within this section. Thus the following activities have been carried out to develop the theoretical grounds for developing methodologies:

- Identification of different Data and Text Mining techniques used in industrial environments.
- Examination of different tools for discovering useful knowledge from semi-structured or unstructured databases, defined in the Data Mining and Text Mining technologies.
- Revisiting of Data structuring methods used for discovering useful information available in the form of textual data.

The second phase consists of developing an architecture and testing different functions defined in the integrated framework to improve the knowledge discovery, its refinement and then utilizing the knowledge for categorizing of textual data into two different categories (i.e. as good or bad information documents within this research context). The data used was from Post Project Reviews (PPRs) taken from the construction industry.

2.5 Summary of the Chapter and Conclusion

In this chapter the research background and its scope has been discussed in detail where the focus was on two major area of knowledge discovery and management i.e. Data Mining and Text Mining technologies. The aim and objective of research are detailed as an outcome of the research gap found through detailed survey of the literature relevant to the data mining techniques applications in manufacturing and construction domains.

Chapter 3 Literature Review

3.1 Introduction

The purpose of this chapter is to provide a context for this research by giving an overview of the importance of knowledge in modern business environments, and a discussion on the types of knowledge and knowledge management needs for manufacturing companies. As the focus of this research is on Textual Data Mining (TDM) solutions to these challenges, a review has been carried out of the literature about the different DM techniques and their applications in Manufacturing or Construction Environments. It will serve the purpose of identifying the capabilities of these techniques to handle large amounts of information available in different databases and to discover valuable knowledge to support the activities carried out in manufacturing or construction industrial environments. This chapter will further explore the possible applications of data mining to analyse structured or semi-structured databases to manage the knowledge resources to improve the product or service quality within industrial environments.

Section 3.2 discusses the importance of business intelligence solutions in manufacturing. One of the roles of data mining techniques is to provide support in finding the optimal solutions to the existing problems. This is done through discovering useful knowledge in terms of rules which can be stored in knowledge bases. Knowledge can then be further updated through various applications of intelligent data analysis tools. Therefore knowledge and information needs for intelligent manufacturing systems and methods to handle these knowledge resources are discussed in sections 3.3 and 3.4. The role of data mining techniques for providing intelligent solutions is explored in section 3.5. The role of Information retrieval (IR) for handling textual data and data mining to support these activities are considered in sections 3.6. Finally section 3.7 reports textual data mining solutions in manufacturing business improvement environments.

3.2 Business Intelligence in Manufacturing

To survive in competitive business environments, lower cost, higher quality and rapid response are the important issues mentioned by most enterprises. The improvement of

manufacturing processes and controlling the product variables highly influence product quality. The manufacturing system needs to be robust to perform operational activities and should be designed to improve the product or operational process in such a manner that the company can attain the desired target with minimum variations (Peace 1993).

Business Intelligence (BI) can , therefore, bridge the gap between different parts of information and knowledge stored in databases and also can be viewed as an important tool for e-business (Grigori et al. 2001). Since the focus of BI and Enterprise Resource Planning (ERP) are to control the operation and execute activities in an organisation their main purpose is to deal with enterprise analysis and discovery of knowledge for decision making to help enterprise managers (Powell 2000; Reimer et al. 2000; Brezocnik et al. 2003). In any competitive business environment companies have to rely heavily on both design and process automation technologies and use professional manufacturing knowledge to enhance their intelligence solutions. But if these technologies and professional competencies are used together to advance product design and production capabilities then product development time can be shortened and the quality and competitive capabilities can be enhanced (Tong and Hsieh 2000; Hsieh and Tong 2001; Hsieh et al. 2005).

Business Intelligence can be interpreted as a term for decision support and is sometimes referred to as Enterprise Business Intelligence (EBI). Business Intelligence tools can be categorised into three different categories on the basis of producing answers to questions “what”, “where” and “why” i.e. what is required, where it is needed and why it is needed. The answers to the first two of these questions can be provided by utilising Data Warehouse and On-line Analytical Processing (OLAP) tools which apply different queries to the databases. But the third question i.e. “why” all this is happening and how this could be addressed is more difficult to answer. This leads to the use of different tools to answer “why” this is happening on the basis of existing data, and these tools include forecasting, Classification, Statistical Analysis, Data Mining and Text Mining techniques. These tools can be used to provide feedback information to support important decision making in any business environment. With the inception of knowledge centred product and service quality

improvement solutions, most enterprises have to pay particular attention to the issue of knowledge management to fully exploit their valuable knowledge resources.

3.3 Knowledge and Information Needs for Next Generation Intelligent Manufacturing Systems

Throughout modern history, the global concept of Manufacturing Systems has been closely related with the principles and findings of science, philosophy, and arts. The manufacturing concepts can be seen as reflecting those principles, criteria, and values which are generally accepted by society and considered as the most important. For example, scientific facts mainly exposed the concepts of exchangeability, determinism, rationality, exactness, and causality in the , , and the first half of the century. That period of history can be considered as an era when the society was predominated by the concept of production. The inception of information technology and its advancement in the second half of the century assured the formal conditions necessary for the expansion of various organizational forms. The second half of the century can thus be regarded as an era when organizational aspects were prevailing (Brezocnik et al. 2003).

Today the service life of products is reducing, the number of product versions is increasing and the time for product conception to their manufacture is reducing. The novelties are being introduced in many areas at a greater speed and changes in one area have a strong interdisciplinary character and often affect other areas which seem to be independent at first glance. Although there is a great advancement in the field of science and technology, the global purpose of human activities and subsequently the requirement for manufacturing concepts needs to be well defined in future. This is essential because human creativity and the manufacture of goods are obviously the basic needs of human beings. The central question is not only the rational manufacture of goods but defining the global meaning and purpose of goods will also be important.

The deterministic approaches are being used in particular for synchronization of material, energy, and information flows in present manufacturing systems and methods based on exact mathematical findings and the rules of logic are used in

practice for modelling, optimization, and functioning of systems. Since production is a very dynamic process where many unexpected events often occur and cause new requirements, conventional methods are insufficient for exact description of a system. Mathematical models are often derived by making simplifying assumptions and consequently may not be in accordance with the functioning of the real system. To develop new products or meet the needs of service oriented industrial environments, the systems based on Mathematical modelling are not suitable and flexible enough to respond efficiently to the requirements. In recent years the paradigm has shifted in other areas of science and technology where Intelligent Manufacturing Systems (IMS) have been introduced which are capable of learning and responding efficiently. Machine learning as an area of Artificial Intelligence (AI) has gained much importance in the last one and half decades as successful intelligent systems have been conceived by the methods of Machine Learning (ML) (Gen and Cheng 1997; Mitchell 1997; Brezocnik et al. 2002).

In this era of information technology, the information technology has great impact on every aspect of society and also influences the traditional manufacturing systems. Due to increased competition in the global market, companies have to respond to the needs of their customers by making the best use of their core technological competencies. Manufacturing companies are driven by a knowledge based economy where the acquisition, management and utilization of knowledge give them a leading edge. Use of both internal and external sources of information is of great importance while manufacturing a product. At the stage of a new product design eight different types of information needs have been defined in (Zahay et al. 2003). This information extends across both internal and external sources and includes the following points;

- Strategic
- Financial
- Project management
- Customer
- Customer needs
- Technical
- Competitor

- Regulatory

There are many knowledge management problems related to the integration and reuse of different knowledge sources as they are often generated and stored in different ways. Strategic information is taken in by the governing board or from the corporate business unit. Financial information is usually imported from the finance department. Project Management information is generated and controlled to a very large extent by the design teams. Customer information and customer needs have been defined separately in (Zahay et al. 2003). The customer information may be stored as a database about the potential customers while customer needs are separately identified as the needs, desires and preferences of customers. Technical information is available in various forms and sources such as Product Data Management (PDM) and Product Lifecycle Management (PLM) systems and external patent databases, to be used by design teams for product innovation, creation and trouble shooting. Finally the information collected from competitors and government regulatory agencies are also crucial since new products must be compatible with the newly issued government regulations and should either be superior to what competitors can offer or be distinct in some way in the market. For the purpose of utilisation of these information sources they are further categorised into two main categories i.e. internal and external sources (Liu et al. 2006).

From a technical perspective the internal information is divided into two major parts in which the internal portion refers to the technical experience and knowledge generated through a firm's endeavours in Product Design and Development (PDD). This experience and knowledge should be owned and stored within the company and be accessible company-wide. The external source refers to the synthesis of broad information and knowledge out of the company, such as patent information managed by the government agency, breakthroughs and nascent technologies incubated at the public research institutes, customer information and needs investigated by marketing firms, competitors' latest product developments and government's recent law enforcements relevant to the company's products. These sets of information and knowledge exist both as numeric data and textual documents and potentially these sources can hold useful but hidden or implicit knowledge about the organisation

operation or products and therefore they need to be used to discover patterns in terms of rules.

In a typical industrial supply chain, the information passes through many companies before reaching the final user of the product. Processing of this information manually requires a lot of efforts and causes errors (Toyryla 1999), and therefore printed information has commonly been replaced by digital data transfer as 70% of companies convey product information to their clientele, digitally. Many companies in the supply chain may not need the information for their own activities but they still have to be able both to receive and transmit the information to all their partners. To operate effectively, all companies have to be able to communicate with each other. If one of the companies in the supply chain is unable to receive and transmit the information then the information flow is interrupted. The product information that is sent and stored at various downstream companies is difficult to keep up-to-date. The producer of new information may not know what parties to inform about updates and thus companies with outdated information risk making decisions based on wrong information (Karkkainen et al. 2003), and transmission of product information may cause overflow of information at the downstream level of a supply chain (Beulens et al. 1999). It is therefore an established fact and has been widely accepted that the future of manufacturing organisations is information theoretic and knowledge driven. This information can be exploited to improve manufacturing operations and build global manufacturing environments (Rahman et al. 1999).

3.3.1 Intelligent Systems and Learning Capabilities in Manufacturing

Manufacturing systems are complex with many variables involved and there are complex hierarchical problems associated with these systems. Due to the large range of problems associated with each of these systems the synchronization of different material, energy and information flows becomes difficult. The learning capabilities of current intelligent systems can be divided into three groups where their learning is based on conventional knowledge bases, learning through interaction with the environment in which they exist and also from other environments as well.

- Learning through Knowledge Bases

A large number of applications of the intelligent systems are based on the knowledge bases which have capabilities to maintain information and knowledge in the form of rules (i.e. if-then rules, decision trees etc). If information about different scenarios is stored in these systems, it can provide a basis for taking any suitable action in many unexpected circumstances. These systems perform their functions based on the environmental properties and transform them into relevant actions in accordance with the instructions in the knowledge base. However if new situations occur which have not been previously defined in the system then these systems fail to respond intelligently (Brezocnik et al. 2003).

An example of an inference of rule in the knowledge base is given in (Kusiak 1990). Consider a semantic network space partially representing the concept of machine shown in figure 3.1 where application of inference rule helps to prove or disprove the conclusions or goals. Using the inference rule method of Modus Ponendo Ponens or modus ponens rule (i.e. A=> B, A ⊢ B stated as If A is TRUE Then B is TRUE, A is TRUE Therefore B is TRUE) it is possible to infer the new fact "L-001 has a motor" from the fact "L-001 is a machine" and the rule "If **X** is a machine, THEN **X** has a motor."

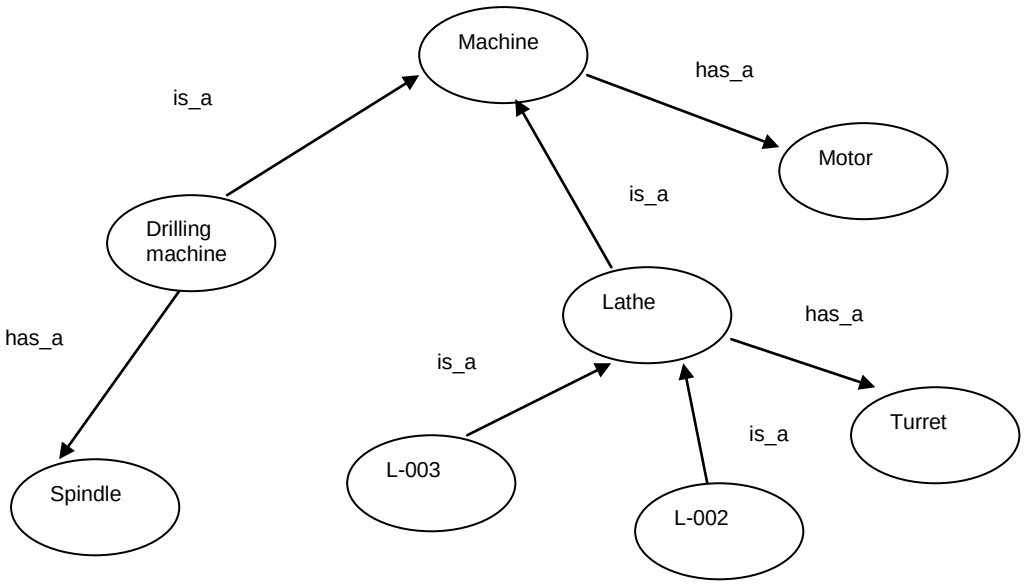

Figure 3.1: Simple Semantic network partially representing the concept of a machine[adapted from (Kusiak 1990)]

- Learning with Interactions with Environment

The systems which learn through interactions with their environment are able to induce new knowledge on the basis of learning examples (Brezocnik et al. 2003). These environments may be static or dynamic. Their main characteristic is to show intelligent behaviour while working in a narrow space of the static business environment, with a large number of optimization parameters which are unpredictable and therefore cause difficulties in learning to these intelligent systems. These systems are meant to solve problems where the interpolation barriers cause an explosion of combinatorial factors. It therefore brings the need that the learning capabilities, in terms of knowledge upgradation, should be done within these systems to work in any static or dynamic manufacturing environment. The examples of such systems are software systems, assembly systems, manufacturing systems and information systems. For example dynamic scheduling in a flexible manufacturing environment a scheme was proposed in (Lee et al. 1997) which uses the decision tree C4.5 at first stage to select the best rule to control the input flow of jobs to the system. Then a genetic algorithm was used at second level to select the most appropriate dispatching rules for each of the system's machines.

- Learning through Interactions with the Internal Environment and External Environments

These systems not only learn through interactions with the environment in which they exist but also learn through interactions with other environments and can develop methods and techniques to deal with or handle the hardest problems (Brezocnik et al. 2003). This type of intelligent system works in any environment and interacts with other environments working like living organisms. These systems are therefore expected to behave like human beings and overcome the hardest problems of dialectic barriers. Such systems exist only in human beings where the complex hierarchical structures exist. For example living cells are associated with more complex structures of tissues which are then associated with organs and these organs are related with individuals which ultimately forms communities of most different shapes. The example of such an intelligent system for self organising assembly of parts into final products is presented in (Brezocnik 2001). The simulation of a self organising assembly of shaft is introduced in this research which imitates a general principle of evolution of organisms from basic to higher level of hierarchical units. Under the influence of a production environment and genetic contents of basic components, the basic parts of a shaft grow into a final shaft.

3.4 Knowledge and Information Management Methods

Knowledge is a valuable resource in a company's business and is sometimes said to pass through the different phases of the 'data-information-knowledge' sequence. This can be defined as "a framework of different experiences, values, contextual information and experts insight that provides a framework for evaluating and incorporating new experiences and information" (Davenport and Prusak 1998). Knowledge can be divided mainly into two parts i.e. explicit and tacit knowledge (Polanyi 1967; Nonaka and Takeuchi 1995) where the explicit knowledge can be codified and stored in computer based systems while the tacit knowledge is hidden in a person's mind and is difficult to codify, hard to formalise and communicate. In an organisational point of view the knowledge can be defined as a set of routines, processes, products, rules and culture that enable actions to be taken in any industrial environment (Beckman 1999). Nowadays, the focus of researchers is to consider the

emergent nature of knowledge processes (Markus et al. 2002). Three major schools of thought are at work in the Knowledge Management (KM) communities i.e. technocratic, commercial and behavioural (Earl 2001). Technocratic schools believe that it is possible to capture specialist knowledge and codify it so that the knowledge base can help to make it transferable and reusable. This school of thought can see the advantages that can be gained from the fast growing capabilities of information technology. The greatest flaw of this approach is that it implies that knowledge is static and hence discards the possibilities of embedding the knowledge gained through practice. Commercial schools of thought believe that KM is an economical perspective i.e. knowledge as an asset. The focus of the behavioural schools of thought is on the socialising aspects of knowledge and enhancing its productivity through exchanges within a social network of motivated people. Information Technology (IT) can play an effective role in the knowledge management capabilities of an organisation through the use of groupware and knowledge representation tools (Earl 2001).

3.4.1 Knowledge Based Systems

Knowledge based systems are defined to have four components i.e. a knowledge base, an inference engine, a knowledge engineering tool, and a specific user interface (Dhaliwal and Benbasat 1996). These are also supposed to include all those organizational information technology applications that may prove helpful for managing the knowledge assets of an organization such as Expert Systems, rule-based systems, groupware, and database management systems (Laudon and Laudon 2002).

3.4.2 Agent Based Systems

In agent based approaches, the challenging task of making information available about all the aspects of the product without having risk of overflow at the downstream level is controlled by using software agents. The use of these agents makes it possible to answer the challenges involved in information management systems. Agent technology was a big paradigm shift in computer programming during the 1980s, in which the old procedural programming paradigm changed and shifted into an object oriented paradigm. The main reason lying behind this shift was because it is easier to

manage data and functionality of a program by dividing it into objects. By having a reference of an object one can access information about the object through methods declared in the object's public interface while hiding the object's implementation. Object oriented programming has therefore become a dominant paradigm in software engineering. A number of characteristics available under the cover of object oriented programming can be used to create agent based product information management systems. In distributed intelligent manufacturing systems agents are used to represent the manufacturing resources such as workers, cells, machines, tools, products, parts and operations which are used to facilitate the manufacturing resource planning, scheduling and execution control (Shen and Norrie 1999).

3.4.3 CAD/CAM and (PDM) Systems

The product information that is created and required throughout the product lifecycle can be captured through the use of Computer Aided Design and Manufacturing (CAD/CAM) and Product Data Management (PDM) systems (Ameri and Dutta 2005). CAD systems emerged in the early 1980s and enabled designers to create geometric models of the product more easily than on paper. These digital models can easily be manipulated and reused. The Initial Graphic Exchange Specification (IGES) was designed as a neutral format to exchange CAD data and used as a standard for geometric information transfer by many CAD/ CAM systems. It does not however fulfil the complete requirements of representing product data (Bloor and Owen 1991). But due to limitations of these systems the inception of product data management systems appeared in 1980s (Dutta and Wolowicz 2005). A PDM system provides a framework that enables manufacturers to manage and control engineering information, specifically, data surrounding new product designs and engineering processes (Gascoigne 1995). PDM systems provide quick and secure access to data created during product design. The early PDM systems manage information about geometric models, bill of materials and finite element analysis and can provide any required engineering knowledge. However these systems have limitations as they do not support information and knowledge related to sales, marketing and supply chain and other external entities that play an important role like customers and suppliers. With the advent of the Internet, Web-based PDM Systems have became more accessible to suppliers and other parties outside of the enterprise. But still these

systems are confined to engineering information without considering other aspects of the product's lifecycle (Ameri and Dutta 2005).

3.4.4 Product Lifecycle Management (PLM) Systems

PLM systems are generally defined as "a strategic business approach for the effective management and use of corporate intellectual capital" (Amann 2002). Product Lifecycle Management (PLM) appeared late in the 1990s with the aim of moving beyond mere engineering aspects of an enterprise. The aim of these systems was to manage information beyond the engineering domain in an enterprise by managing information throughout the stages of a product lifecycle such as design, manufacturing, marketing, sales and after-sales service (Ameri and Dutta 2004). Thus PLM systems can provide a number of benefits in terms of delivering more innovative products and services, shorter time-to-market and comprehensive and collaborative relationships among customers, suppliers, and business partners (Amann 2002; Ameri and Dutta 2004; Dutta and Wolowicz 2005). Several vendors are in the market like SAP, IBM, Dassault Systems and UGS which offer PLM solutions but still no comprehensive PLM system exists today as the application of these systems is still five years behind the state of the art solutions (Abramovici and Siege 2002) available.

In an extended enterprise the components of the product are fabricated by external suppliers with close relationships to the company designing the product. However, this means that the manufacturing knowledge no longer lives within the walls of the company therefore when a new product is conceptualized the manufacturing knowledge cannot be used easily to address issues such as manufacturability and cost during design (Rezyat 2000). There is a need to integrate the supply chain, its subsidiaries and affiliated partners with the lifecycle of a particular family of products while PLM systems are still limited to product design (Abramovici and Siege 2002).

3.4.5 Web-based Systems

Internet and extranet have facilitated the proliferation of information all over the world enabling some or all enterprise activities to be moved into virtual spaces. The success of an enterprise environment is based on successful or effective exchange of information. The heterogeneity of information resources poses some challenges to the

enterprise and proper knowledge is needed to handle appropriate information processes. There are two types of data or information which an enterprise needs to manage i.e. business data which includes accounting and personnel data etc. and product data e.g. CAD and CAM data. An integrated web based system using Java solution and CORBA-ORG broking technologies for design and manufacturing has been proposed in (Cheng et al. 2001).

The previous sections give details of methods and techniques that are prevalent in manufacturing industrial environments for using and managing valuable knowledge. The focus of this research is to look into methods that can be used as a source of generating or discovering knowledge from different databases or other file sources and transmitting it through the use of an IT infrastructure. This type of managing knowledge is related to the technocratic school of thought.

3.5 Need for Data Mining Techniques and Intelligent Decision Making in Manufacturing

One of the concepts for identifying useful knowledge from the codified information is Data Mining which can be defined as "the process of discovering interesting knowledge from large amounts of data stored either in databases, data warehouses or other information repositories." (Han and Kamber 2001). The word interestingness is important to interpret in this definition as it refers to the fact that rules or patterns or relationships existed in databases, but not in a form that was easily understandable to human beings. The extraction of rules and making these rules interpretable is always a focus of the data mining process which goes through different stages in an interactive sequence of data cleaning, data integration, data selection, data transformation and knowledge representation (Han and Kamber 2001).

Current applications of data mining in Manufacturing generally explore large volume databases to discover valuable information and transform this into a valuable source of knowledge or patterns. Data Mining has been applied in various domains and has been used to explore knowledge sources ranging from finance to life sciences. Data Mining has been less commonly applied in manufacturing than in other domains and the reason for this may be because of the required effort in terms of time and expertise

from both Data Miner experts and Domain Experts (Shahbaz et al. 2006). Further reasons are explored in (Wang 2007) as follow:

- Researchers are lacking knowledge of how to apply data mining techniques in manufacturing
- The complex manufacturing processes are not understandable for theoretical based researchers in the field of Data Mining Application
- Manufacturing data are less accessible due to propriety and sensitivity issues to the data mining researchers
- The difficulty of evaluating the benefits of results during applications of Data Mining techniques in Manufacturing

Therefore, the value of using these techniques has not been fully ascertained in manufacturing environments but applications which have been made so far have provided better "Knowledge Management" solutions in various aspects of manufacturing. A detailed survey of applications of data mining techniques in manufacturing has been done in (Harding et al. 2006) . Therefore the term data mining tool which is used to extract useful knowledge from data can be taken as a tool of Business Intelligence. The applications of combined efforts of Data Mining, Business Intelligence and Knowledge Management (KM) can change the current competitive status of companies. It can help an enterprise to collect data or information, extract patterns from it and transform it into useful knowledge and deliver the best manufacturing information and knowledge to the remain competitive in a business environment (Wang 2007). The concurrent applications of data mining techniques could give useful results when compared to the applications of these techniques independently due to the increased complexities of the manufacturing systems (Wang 2005). That is the systematic application of hybridized applications of data mining approaches can give better results rather the application of a single technique.

3.5.1 Rules Extraction and Updating domain Knowledge

Understanding Domain knowledge is the first step in the data mining process and is used to guide the whole knowledge discovery process. It helps to evaluate the interestingness of the resulting patterns. Knowledge of user beliefs is used to assess

the pattern's interestingness based on its unexpectedness (i.e. if the pattern was not anticipated by the domain expert or obvious from the source database). The measure of interestingness is very important in the knowledge (rules) extraction. Since a data mining system has the potential to generate thousands or even millions of patterns, or rules the natural question is whether all these patterns are interesting. The answer of this question is typically not.

Only a small fraction of patterns generated would be of interest to any given user. An interesting pattern represents knowledge. Both user-driven(or "subjective") and data-driven (or "objective") approaches are used to measure the degree of interestingness (Freitas 2006). A user-driven approach is based on using the domain knowledge, beliefs or preferences of the user, while the data-driven approach is based on statistical properties of the patterns. The Data driven approach is more generic, independent of the application domain. This makes it easier to use this approach, avoiding difficult issues associated with manual acquisition of the user's background knowledge and its transformation into a computational form suitable for a data mining algorithm. On the other hand, the user-driven approach tends to be more effective at discovering truly novel or surprising knowledge for the user, since it explicitly takes into account the user's background knowledge. Thus the study of rules in the process of data mining is very important for the discovery of knowledge which is really understandable, valid on new or test data with some degree of certainty, potentially useful and novel in the end as well. Some of instances of rules generation, its interestingness which help in updating domain knowledge have been reported as follows:

(Annand et al. 1995) discussed the use of domain knowledge within Data Mining and they defined three classes of domain knowledge: Hierarchical Generalization Trees (HG-Trees), Attribute Relationship Rules (AR-rules) and Environment-Based Constraints (EBC) and also discussed that one of these types of domain knowledge is incorporated into the discovery process within the EDM (Evidential Data Mining). (Djoko et al. 1997) presented a method for guiding the discovery process with domain specific knowledge. They used the SUBDUE discovery system to evaluate the benefits of using domain knowledge to guide the discovery process. (Pohle 2003) proposed that formalized domain knowledge be employed for accessing the

interestingness of mining results and also recommended that a next-generation data mining environment should actively support a user to both incorporate his domain knowledge into the mining process and update this domain knowledge with the mining results.

(Park et al. 2001) presented and evaluated a knowledge refinement system, KREFS, which refined knowledge by intelligently self-guiding the generation of new training examples. This system used induced decision tree to extract patterns from data which are further utilized to refine the knowledge. (Yoon et al. 1999) introduced a method to utilize three types of domain knowledge i.e. interfield , category and correlation in reducing the cost of finding a potentially interesting and relevant portion of the data while improving the quality of discovered knowledge. They proposed that relevant domain knowledge should be selected by defining clusters of attributes which avoid un-necessary searches on a large body of irrelevant domain knowledge. (Padmanabhan and Tuzhilin 1998) proposed a new method of discovering unexpected patterns that takes into consideration prior background knowledge of decision makers. (Nguyen and Skowron 2004) presented a method to incorporate domain knowledge into the design and development of a classification system by using a rough approximation framework. They demonstrated that an approximate reasoning scheme can be used in the process of knowledge transfer from a human expert's ontology, often expressed in natural language, into computable pattern features. (Bykowski and Rigotti 2001) presented the idea to extract a condensed representation of the frequent patterns, called disjunction-free sets, instead of extracting the whole frequent pattern collection.

(Liu et al. 2000) developed a new approach to find interesting rules (in particular unexpected rules) from a set of discovered association rules. (Chia et al. 2006) developed a novel technique for neural logic networks (or neulonets) learning by composing net rules using genetic programming. (Lin and Tseng 2006) proposed an automatic support specification for efficiently mining high-confidence and positive-lift associations without consulting the users.

(Last and Kandel 2004) presented a novel, perception-based method, called Automated Perceptions Network (APN), for automated construction of compact and

interpretable models from highly noisy data sets. (McErlean et al. 1999) introduced a new evidential approach for the updating of causal networks which is to be added to an existing general data mining system prototype-the mining Kernel System (MKS). They presented a data mining tool which addresses both the discovery and update of causal networks hidden in database systems and contributes towards the discovery of knowledge which links rules (knowledge) and which is normally considered domain knowledge. (Zhou et al. 2001) presented an intelligent data mining system named decision tree expert (DTE). The rule induction method in DTE is based on the C4.5 algorithm. Concise and accurate conceptual design rules were generated from drop test data after the incorporation of domain knowledge from human experts. (Cooper and Giuffrida 2000) developed and illustrated a new knowledge discovery algorithm tailored to the action requirements of management science applications. Information is extracted from continuous variables by using traditional market-response model and data mining techniques are used to extract information from the many-valued nominal variables, such as the manufacturer or merchandise category.

3.5.2 Decision Trees Analysis

Decision trees play an important role in making decisions based on distribution of information in terms of binary classification trees. These are used to present the decision rules in terms of a binary tree where each node is subdivided into sub nodes. In computer integrated manufacturing (CIM), (Kwak and Yih 2004) presented a decision tree based methodology for testing and rework purposes and the rules generated showed the effect in decision making process. The earliest version of Decision Trees is ID3 which was used for solving a variety of problems in different application fields. A generalised ID3 was proposed by (Irani 1993) as a part of an expert system for diagnosis and process modelling for semi-conductor manufacturing. A decision tree based model was also used for the accurate assessment of probabilistic failure of avionics by using the historical data relating to environment and operation condition (Skormin 2002). A decision tree based analysis was made for mining the job scheduling in order to support date assignment in a dynamic job shop environment in (Sha and Liu 2005) and rules are presented in terms of IF-THEN rules. (Zhou et al. 2001) applied C4.5 algorithm for drop test analysis of electronic goods where the focus was to predict the integrity of solder points for large components on the printed

wiring boards and also could be used to other parts. (Kusiak 2006) proposed a decision tree based algorithm for learning and prediction of incoming faults of watery chemistry.

3.5.3 Clustering Techniques

Clustering is a data mining technique which is used to analyse data and classify it into different classes. (Chien 2007) developed a framework on the basis of k-means clustering to increase the yield of semi-conductor manufacturing. (Liao et al. 1999) used fuzzy based clustering techniques to detect the welding flaws and also presented the comparative study between fuzzy k-nearest neighbour and fuzzy C clustering. (Liao 2006) used a genetic clustering algorithm for exploratory mining of feature vectors and time series data. The approach used showed good results which are comparable to the k means clustering algorithm. Various clustering algorithms are used in manufacturing environments where the main focus is to use k-means, fuzzy k-means, fuzzy C means and artificial neural networks approaches to enhance the quality of product or services oriented jobs. Artificial neural networks have also been used to solve multiple problems in manufacturing. These techniques have long been used to learn from historical databases and perform both supervised and unsupervised learning from databases.

3.5.4 Association Rule Analysis

In (Shahbaz et al. 2006) association rule mining was applied on product (Fan Blade) data to extract information about process limitations and knowledge about relationships among product dimensions. The information and knowledge extracted could then be used as a feedback for design and quality improvement. Association rule mining was also used for the subassembly based analysis of prior orders received from the customers (Agard and Kusiak 2004a). The extracted knowledge can be used for the selection of subassemblies in order to timely deliver the product from the suppliers to the contractors. (Cunha 2006) used association rule analysis for detecting the faults in an assembly line to improve the quality of assembly operations. The semi conductor manufacturing industry has been highly influenced by the use of these techniques where (Chen et al. 2005) extracted association rules to detect the defects in semiconductor manufacturing and finally used these rules and their relationships to

identify the defective machines. (Chen 2005) used association rule mining techniques to mine the information about customer demands from the order databases which contain information in terms of frequently ordered product item sets. (Jiao and Zhang 2005) applied association rule of mining to extract rules among customer needs, marketing folks and designers to develop a decision support system about product portfolio. (Shao et al. 2006) proposed an architecture to discover associations between clusters of product specifications and configurations alternatives.

3.5.5 Support Vector Machines (SVMs)

(Samanta et al. 2003) compared the performance of bearing fault detection, by selecting both with and without automatic selection of features and classifying parameters, using two different classifiers, namely, artificial neural networks (ANNs) and support vector machines (SVMs). Genetic Algorithm (GA) is used to select the optimised input features and classifier parameters (e.g. mean, root mean square (rms), variance, skewness, higher order normalised moments) to distinguish between the normal and defective bearings. (Vong et al. 2006) used least squares support vector machines and Bayesian inference for prediction of automotive engine power and torque. Least square support vector machines (LS-SVM) is used to determine the approximated power and torque of an engine. (Cho et al. 2005) proposed an intelligent tool breakage detection system which used Support Vector Regression (SVR) analysis to detect the process abnormalities and suggest the corrective actions to be taken during the manufacturing process especially the milling process. The results are compared with a multiple variable regression approach. (Kwon et al. 2006) presented a comparative study on Coordinate Measuring Machine (CMM) and probe readings to investigate closed-loop measurement error in Computer Numerical Controlled (CNC) milling relating to two different inspection techniques. Adaptive support vector regression analysis was used to measure closed-loop inspection accuracy where different material types and parameter settings(e.g. cutting force, spindle vibration and tool wear) to simulate the results. (Ramesh et al. 2003) presented a hybrid Support Vector Machines (SVM) and a Bayesian Network (BN) model to predict machine tool thermal error which depends considerably upon the structure of error model. The experimental data is first classified using a BN model

with a rule-based system. Once the classification has been achieved, the error is predicted using a SVM model.

3.5.6 Hybridized Approaches for Analysis

The combinations of different data mining approaches are also gaining popularity for solving manufacturing problems. A couple of instances have also been reported in the literature for the analysis of the data. A hybrid kernel based clustering approach and outlier detection methods were used for customer segmentation and outlier detection in (Wang 2008). The methods were tested on two real domain data sets i.e. iris and automobile maintenance data sets. A hybrid system was proposed for statistical process control based on decision tree and neural network approaches in (Guh 2005). These approaches were used to solve the problem of false recognition and increase the control chart pattern classification in different situations and produced promising results.

3.6 Information Retrieval (IR) for Textual Data Handling

Information retrieval(IR), in a broad sense, includes representation, storage, organization, and access to information. In practice, many aspects of work produce documents, i.e. items that carry information. Thus, it is common to use information retrieval as synonymous with document retrieval, with an understanding that the notion of a document is used very flexibly. Most information retrieval research has been focused on identifying documents or portions of documents that may satisfy the user's information need. For example, in response to the user's query 'electric cars' it is reasonable to assume that any document that provides information about 'electric cars' satisfies the user's information need. However, in other situations, it may be necessary to interpret the query based on a wider context. For example, in order to process a query for recent news about faster electric cars the system would need to disambiguate under-specified query terms 'recent' and 'faster' that have a meaning relative to the user's experience and perception of the world.

IR covers various types of information access: search, browsing, pro-active information gathering and filtering. In the case of browsing, the user's information need may be less well defined and highly affected by the user's interaction with

documents through viewing, skimming, and reading. In search, the need is sufficiently defined so that the user can express it verbally, in a form of a query. It is expected that the query terms carry sufficient semantic characterization of the need to enable search over the data set. However, the user needs to refine or reformulate the query. This can be accomplished by using a general purpose thesaurus or by extracting relevant vocabulary directly from the content of the database being searched. The advancement in information technology and computation techniques has greatly influenced the field of IR. The general approach of IR is to create a suitable representation for the query and the document, then to apply a retrieval technique that derives from the model adopted for representing the documents (Srinivasan et al. 2001). To implement the query, the search engine plays a crucial role in IR. A search engine operates by indexing the content of documents and allowing users to search the indexes. When a user poses a query to the system, the query is indexed by its terms and the weights are associated with the query term (Singhal et al. 1996). After that, a numerical similarity is computed between the user query and all the documents in the collection (Salton 1989). Such a similarity supposedly measures the potential usefulness of a document for the user query (Singhal et al. 1996). The documents in the document collection are ranked by their decreasing similarity to the query and are presented to the user in this order. Using the term-based (e.g. keyword) approach to represent documents is a mainstream method in document retrieval. IR technology has matured to the point where there are now reasonably sophisticated operational and research systems to support IR (Srinivasan et al. 2001). However, despite the recent advances of IR or search technologies, studies show that the performance of search engines is not quite up to the expectations of the end users (Gordon and Pathak 1999).

There are various reasons contributing to the dissatisfaction of the end users, among them are imprecise query formulation, poor document representations and an unfamiliarity with system usage. It has been found that there are retrieved documents whose contexts are not consistent to the query (Kang and Choi 1997). Users often have to waste time sifting through ‘hit lists’ that are full of irrelevant results (Weiguo et al. 2004), thus reducing their satisfaction of search result. Therefore, increasing the effectiveness of retrieval algorithms remains an important goal (Srinivasan et al. 2001). To achieve this goal both new retrieval models and extensions of existing models, in particular the Vector Space Model (VSM), have been used, mainly with a

two fold aim (1) to make the query language more expressive and natural; and (2) to incorporate a relevance feedback mechanism to control the production of relevant retrieval results (Salton 1989). (Bordogna and Pasi 1995) suggested that providing the IR system with a powerful query language or a sophisticated retrieval mechanism is not sufficient to achieve effective results if the representation strongly simplifies their information content. So there is a need to develop some new models to refine and improve the retrieval task of the documents.

3.6.1 Data Mining to Support IR Based Solutions

Data mining involves the exploration and analysis of large quantities of data to discover meaningful patterns and rules using automatic and semiautomatic methods. However, applications to handle textual data or documents are less reported in literature. Some instances of these are as reported below :

(Lin et al. 2005) presented a study on the integration of information retrieval and data mining techniques to discover project team coordination patterns from project documents written in Chinese. (Lin and Hsueh 2002) proposed knowledge map creation and maintenance approaches by utilizing information retrieval and data mining techniques to facilitate knowledge management in virtual communities of practice.

(Tan 2005) proposed the neighbour-weighted K-nearest neighbour algorithm, i.e. NWKNN to deal with uneven text sets. (Tan 2006) proposed a new refinement strategy, which is called DragPushing, for the k-Nearest Neighbours (KNN) Classifier, which is widely used in the text categorization community but suffers some model misfits due to its presumption that training data are widely distributed among all categories. (Huang et al. 2006) proposed a rough-set-based approach to enrich document representation where the classification rules are generated and the premise terms are provided by the rough-set approach. (Saravanan et al. 2003) proposed a text-mining framework in which classification and summarization systems are treated as constituents of a knowledge discovery process for text corpora. (Spertus 1997) discussed the varieties of link information: not only the hyperlinks on the web but also how the web differs from conventional hypertext, and how the links can be exploited

to build useful applications. (Ngu and Wu 1997) proposed an alternative way in assisting all the web servers and further proposed that each server should do its own housekeeping. (Ngu and Wu 1997) showed that a large annotated general-English corpus is not sufficient for building a part-of-speech tagged model adequate for tagging documents from the medical domain.

3.7 Textual Data Mining Solutions in Manufacturing

The literature reviewed during the research shows that data mining can serve the purpose of managing the knowledge and information from the product design to through product life cycle activities. It is an interdisciplinary field that combines Artificial Intelligence (AI), Computer Science, Machine Learning, Database Management, Data Visualization, Mathematics Algorithms, and Statistics. Data Mining is defined as a technology for discovering knowledge from databases (KDD). It provides different methodologies for decision making, problem solving, analysis, planning etc. Exploratory Data Analysis (EDA) provides a general framework for data mining based on evidence theory. This provides a method for representing knowledge and allows prior knowledge from the user or knowledge discovered by another discovery process to be incorporated into the knowledge discovery process (Apte et al. 1994) . Knowledge discovery from textual data (KDT) or textual data mining and Text Mining (TM) can be defined as the special fields of knowledge discovery from databases (KDD). Text mining techniques combined with the data mining tools can be used effectively to discover hidden patterns in the textual databases. The next section focuses on the application of these efforts to handle the information and knowledge sources.

3.7.1 Manufacturing Product/ Service Quality Improvement

In this section applications of Text/data mining techniques have been reported which further help to identify the needs of discovering valuable knowledge from textual databases.

The design stage in product development plays a key role in the product lifecycle. A great deal of time is consumed in the product design stage as many different technological efforts have to be used. The role of data /text mining is quite effective to

support variant design activities as a generic bill of material (GBOM) approach was proposed in (Romanowski and Nagi 2004). (Romanowski and Nagi 2005) then found structural similarity of BOMs using tree matching techniques. This approach helped to advance the definition and design of product families using text mining techniques and association rule mining (Agard and Kusiak 2004a) .The focus of this research was to identify the relationships between functional requirements and design solutions.

Data Mining techniques have also been used to find generic routings for large amounts of production information and process data available in a firm's legacy systems (Jiao et al. 2007). The generic routing identification goes through three consecutive stages, including routing similarity measures, routing clustering and routing unification. Text mining and tree matching techniques were used to handle information hidden within textual and structural types of data underlying generic routings.

To identify a 'shared understanding' in design by analysing the design documentation, a formal methodology was described in (Hill et al. 2001). The premise of the paper was the topical similarity and voice similarity are identifiers of the shared frame of reference of the design. The voice of the designer operating in a team was defined more as the ability of a designer to borrow the shared vision of a design team. Using the computational linguistic tool of latent semantic analysis, engineering courseware of documents (.needs.org) written by various authors were analysed to reveal highly correlated group of topics. This study also showed that there were characteristics within documents that allow the author of a document to be identified.

(Yang et al. 1998) discuss how to make 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. They used informal information found in electronic notebooks since it is captured as it is generated, thereby capturing the design process. They investigated schemes for improving access to such informal design information using hierarchical thesauri overlaid on generic information retrieval (IR) tools. They made

use of the Singular Value Decomposition (SVD) technique to aid in the automated thesauri generation.

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

Data mining techniques to generate relationships among design concepts were used in (Dong and Agogino 1997). In the first stage the syntactic relationships within the design documents are analysed to determine content carrying phrases which serve as the representation of the documents. In the second stage, these phrases are clustered to discover inter-term dependencies which are then used in the building of a Bayesian belief network which describes a conceptual hierarchy specific to the domain of the design.

A data mining technique to mine unstructured, textual data from the customer service database for online machine fault diagnosis, was developed in (Fong and Hui 2001). The data mining techniques integrated neural networks (NNs) and rule based reasoning (RBR) with case based reasoning (CBR). In particular , NNs were used within the CBR framework for indexing and retrieval of the most appropriate service records based on a user's fault description.

An unsupervised information system OPINE was introduced in (Popescu and Etzioni 2005) to mine reviews in order to build a model of important product features, their evaluation by reviewers and their relative quality across product. They decomposed the problem of review mining into the four main subtasks of identifying product features, opinions regarding product features, determining the polarity of opinions and ranking those opinions on the basis of their strengths.

Textual data mining techniques i.e. clustering and classification based upon decision tree and neural networks were used to analyse pump station maintenance logs stored in the form of free text in spread sheet (Edwards et al. 2008).

In the product development process textual data mining techniques were used for automatic classification of textual data to facilitate the fast feedback in the product development process (Menon et al. 2003). Different document representation techniques were tested in this case study.

A methodology was proposed based on text mining techniques of morphological analysis to develop a technological road map through identification of key words and their relationships for new product development and technological advancement in (Yoon et al. 2008). The proposed methodology is based upon the three major steps of data collection, product and technological analysis and mapping the analysed set of information from product and technological related key words to develop a road map for the new technology.

Text mining based solutions have been proposed in (Huang and Murphey 2006) to diagnose engineering problems through textual data classification. The automotive industry problems are often descriptive in nature and it becomes difficult to map the problems to their diagnostic categories such as engine, transmission, electrical, brake etc. In this paper the text mining methods, in particular text classification, has been used to mine the automotive problems and map these information to their correct diagnostic categories.

(Kasravi 2004) discussed various application domains in the engineering sector i.e. predictive warranty analysis, quality improvements, patent analysis, competitive assessments, FMEA, and product searches where text mining methods can help to untap the vast amount of information available in textual data. This specifically improves processes through tracking of information from top to downstream levels and through complex data analysis.

3.7.2 Business Process Improvement through Customer Knowledge Management

The business process quality improvement issue has been handled in (Grigori et al. 2001) through analysing, predicting and preventing the occurrences of exceptions with application of data mining techniques. A complete tool suite was presented in (Grigori et al. 2004) through applications of data warehouse and data mining technologies to support IT users to manage the business process execution quality.

In today's business environments, business analysts have to predict their customer behaviours in various ways including using their past histories and communicating effectively with them through face to face interaction, using calls to the customers through service centre calling, recording their comments of the web sites and recording their views on e-mail systems. The nature of information available is in the form of data that is unstructured and any useful knowledge within it is implicit in nature. The attributes and their corresponding fields in a database are mostly structured and data mining techniques are generally only able to handle the structured form of data. If the unstructured data is not taken into consideration this may cause some valuable information to be lost and only remain available in the form of unstructured data bases (Grigori et al. 2004). The data warehouse and data mining techniques were used to analyse the customer behaviour to build customer profiles and provide methods to help companies to retain their customers. This was adapted by developing marketing strategies through discovery of hidden knowledge within the databases of a company (Chang et al. 2009). The decision tree C4.5 and content analysis were used to segment customers into different categories by identifying their needs.

Text Mining techniques were applied to categorise the customers feedback made on phone calls surveys in (Grievel 2005). Documents were assigned predefined classes and divided into dynamical categories respectively.

3.8 Summary of the Chapter and Conclusions

The literature reviewed in this chapter shows that intelligent decision making is of great importance in many contexts within manufacturing, construction and business generally. Business intelligence tools, which can be interpreted as decision support tools, are of increasing importance to companies for their success within competitive

global markets. However, these tools are dependent on the relevancy, accuracy and overall quality of the knowledge on which they are based and which they use.

Potential knowledge sources are very varied, and include data warehouses, various databases, files and web sources and can contain numerical and/ or textual data in structured and / or unstructured forms. The reviewed literature shows that data mining has been used successfully in many design, operational and manufacturing contexts. However, to date, most of these applications have used numerical and / or structured knowledge sources. There is great potential still for data mining to be used further to better manage the needs of next generation business and in particular to exploit the implicit or tacit knowledge which is likely to be available in unstructured textual knowledge sources. However, this will require greater research and adoption of textual data mining (TDM) handling techniques or Text Mining methods. The literature reviewed also showed that there are not many instances reported in exploiting textual sources of information in manufacturing or construction industrial environments.

Chapter 4 Textual Data Mining (TDM) for Knowledge Management (KM): A Conceptual Development of Methodology

4.1 Introduction

In the previous chapter the literature was reviewed about various applications of data mining and textual data mining techniques to enhance business intelligence solutions. This chapter focuses on Text Mining and its application technologies to generate valuable knowledge to improve the business integrated solutions. The nature of technologies used for this purpose will be discussed in detail to solve the issues of handling semi-structured or unstructured data formats. The power of Text Mining combined with other technological efforts has been explored during this research work . The focus of applications examined in particular is Information Retrieval (IR) and Natural Language Processing (NLP).

4.2 Text Mining Needs

The continuous growth of digital based information in various sectors has led to the identification of various issues associated with the handling of multiple sets of information. In all areas of life the quantity of information continues to grow that it becomes a cumbersome job to handle the information manually (Sholom et al. 2005). A large part of information within any corporate business environment is available in the form of unstructured data i.e. documents and Web pages, business information in data repositories on Intranets and the Internet. 80% of companies estimated information are in the form of textual information such as emails, memos, customer correspondence, and reports (Tan 1999; Spinakis 2001; El Wakil 2002; Karanikas and Theodoulidis 2002). The analysis and careful handling of these sources of information could give competitive advantages to a company and help it to be successful in the era of the knowledge-based economy (El Wakil 2002). Therefore use of automatic methods, algorithms, and tools for dealing with such a large amounts of textual data have become necessary (Lagus 2000). In order to solve these issues a relatively new field of Text Mining (TM) for management of multiple data formats has evolved to address the potential issues of mining large numbers of textual databases automatically (Spinakis and Peristera 2003; Fan et al. 2006). Text Mining can be defined as a sub field of data mining if data mining techniques are used to discover

patterns or information from textual data. It also inherently requires techniques from other fields of Information Retrieval, data mining and Computational linguistics (Bolasco et al. 2002) as shown in figure 4.1. Text Mining techniques are also aimed at finding the Business Intelligence solution to help companies to remain competitive in the market (Bolasco et al. 2005).

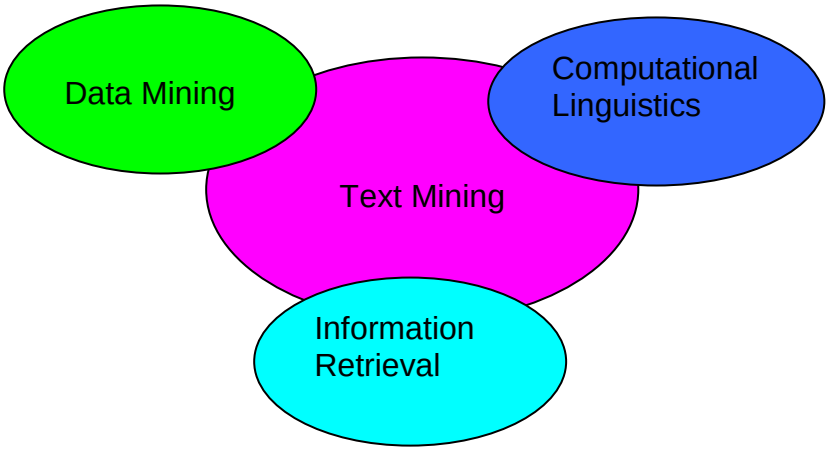

Figure 4.1: Text Mining as an Interdisciplinary Field

The objectives of data mining techniques can be characterised by the different functions that they perform through the process of discovering knowledge from databases e.g. clustering, association, classification etc. Thus text mining processes can be defined in different aspects of information and knowledge management. In terms of handling information text mining can be defined as “the process of extracting useful information from textual databases through the application of computer based methods and techniques” (Fan et al. 2006).

In terms of discovering knowledge from textual databases text mining can also be defined as, “the non trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in unstructured data ” (Karanikas and Theodoulidis 2002). Thus it can be defined as an extension of the field of Data Mining to explore

the possible solutions of discovering knowledge and its management in any business oriented environments.

Text Mining performs nine different tasks for handling rich sources of information and making it suitable for discovering useful knowledge from them. These tasks range from extracting useful information to its visualization and are categorised as information extraction, text-based navigation, search and retrieval, clustering, categorization, summarization, trends analysis, associations, and visualizations (Fan et al. 2006; Singh et al. 2007; Gupta and Lehal 2009). These tasks are briefly described as follows:

Information extraction deals with finding useful information from a text. It identifies information in terms of key phrases and their relationship within text. The various information objects are defined in the textual data to characterise the information about person, place, organization etc. It also deals with finding information through counting the number of times a term occurred in the text and their corresponding term sequences and knowledge is presented in terms of relationships between terms and their sequences.

Text-base navigation systems are concerned with finding the related terms discussed in some specific context in the text and finding the relationships between them.

Searching and retrieval strategies are meant to answer the questions of specific needs of their users and help to bring the relevant information to the desk. This technique helps the user to ask a question from the computer and be provided with the related answer. In a company or industrial setup, the employees are enabled to search internal databases to find the answers of their common questions.

Clustering, *Categorisation* and *Summarisation* tasks are most widely used for drawing key information and knowledge from the text. The process of Clustering or grouping the documents is done to find information hidden in the text by finding the similarities between the documents. This method groups similar documents on the basis of strong similarities within each cluster and dissimilarities to the documents

outside the cluster. This technique is useful to organise thousands of documents in an industrial or organisational information management systems.

Categorization is the process of identifying the similarities in the documents based on content mining technologies and putting these documents into pre-defined sets of categories or classes or topics for analysis. The process of categorization of documents relies on methods of taking the whole document as a set of words or "bag of words" where the information is extracted on the basis of words counts, the relationships are identified by looking terms in broader and narrowing aspects of these terms, and their synonyms.

Summarization process helps to reduce the content of documents and makes it readable to others whilst still retaining the sense of the topic discussed in it. In practice humans read through the text and understand the meaning of this and mention or highlight the main topic or point discussed in the text. Computers lack this capability of understanding the text therefore certain methods or techniques (e.g. sentence extraction) are used to find the useful information by using statistical weighting methods. These methods are used to find the key information in terms of phrases to define main theme of the text.

Trend Analysis and Association Analysis are used to find trends or predict future patterns based on time dependant data and associate these patterns to the other extracted patterns. Visualization is defined as representing the extracted features with respect to the key terms and helps identifying main topics or concepts by the degree of their importance on the representation. It is further used to easily discover the location of the documents in graphical representation.

4.3 Data and Text Mining For Discovering Knowledge

Data mining tasks can be characterised as different from other technologies due to its handling of multiple data formats or databases such as relational databases, data warehouses, transactional databases, etc. Among these databases, text databases are "databases that contain word descriptions for objects" (Han and Kamber 2001), such as papers, reports, messages, notes, Web pages, or others. Text databases may be

unstructured, semi- structured or structured (Han and Kamber 2001). They therefore, support the concept of defining text mining as Textual Data Mining or Knowledge Discovery from Textual Databases. Though text mining processes rely heavily on applications of data mining techniques for knowledge discovery from textual databases, there are inherently more challenges due to the properties of handling very complex databases which are unstructured or fuzzy in nature (Tan 1999) when compared to numerical data formats.

4.3.1 Cross- Industry Standard Process for Data Mining (CRISP-DM)

A standard data mining process modelling approach was developed in 1996 by the group of analysts which was termed as the Cross-Industry Standard Process for Data Mining (CRISP-DM) (Chapman et al. 2000). The group who developed this approach included DaimlerChrysler, SPSS and NCR. The intent behind defining this standard was to make the process of data mining as industry-neutral, tool-neutral, and application-neutral. According to this standard process there are six different stages of Data Mining defined as follows;

4.3.1.1 Understanding and Defining the Business Problem

In this stage of the CRISP-DM process the initial understanding of the problem or the research need is defined. That is the process of data mining needs to be started by clearly defining the business/ project needs and objectives in terms of the whole business or research unit. When translating the goals and formulating the data mining problem definition one should design a plan to achieve these goals and objectives. Thus this is a very essential step in the process of data mining which needs to be performed carefully.

4.3.1.2 Understanding Data or Information

At this stage of analysis some raw data or information is collected and some exploratory data analysis techniques are used to get an initial insight of the data. Dealing with data quality issues and selection of some interesting subsets of information may serve the purpose of discovering useful patterns.

4.3.1.3 Preparing Data for Analysis

At this stage of handling data or information the data is made ready for analysis and use in the subsequent stage of application of data mining techniques. This is a quite labour intensive task which includes selection of variables which will be effective in the analysis, transformation of different variables (if needed) and removal of un-necessary information from the data which are less effective in the knowledge discovery process.

4.3.1.4 Modelling

At this stage of analysis an appropriate data mining tool or technique is selected which is used to optimise the results. Different data mining techniques may be used for finding the solution of the problem. An iterative procedure will be adopted based on the initial data preparation phase to meet the specific needs of applying the appropriate data mining algorithms.

4.3.1.5 Model Evaluation

This is a decision making stage where the quality and effectiveness of the models used at the modelling stage is tested before deploying the model in a real industrial or business problem.

4.3.1.6 Model Deployment

Creating a model by following the different stages of CRISP-DM, the last stage is to deploy the designed model on the industrial context and generate a report.

Data Mining and Text mining are similar in many ways as they are both used for discovering knowledge from multiple databases but they differ in a number of points as below (Spinakis and Chatzimakri 2005):-

1. Data formats which are handled by the text mining techniques are usually more difficult to decipher than data handled by data mining technologies.

2. Text Mining techniques consider the syntactic and semantic relationships of textual data in depth and decode the relationships among them whilst the focus

of data mining tends to be on distance measures for measuring the similarities of terms or context defined in textual data.

4.3.2 Text Mining Process

The text mining process mainly consists of three different stages i.e. text preparation, text processing and text analysis(Natarajan 2005).

4.3.2.1 Text preparation

At this stage of data handling the initial process of selecting the suitable variables, cleansing and pre-processing of textual data is done. This process of handling text should be done under the guidance of human experts who can help to identify which terms or phrases are more suitable in the analysis of the data. Some pre-processing techniques are applied at this stage of analysis e.g. stop words removal, stemming etc.

4.3.2.2 Text processing

Once pre-processing is done then the next stage is to store the multiple sets of information in a homogenous format on which some data mining or natural language techniques can be used. These techniques help to process the information by finding the relationships among terms in the textual data and help to explore relationships existing in terms of people, companies, organizations, etc. Added to this some conceptual relationships between entities can be explored to find information related to particular aspects of interest. These relationships among entities help to extract meaningful features through the applications of techniques such as decision trees, neural networks, case-based learning, association rules or genetic algorithms.

4.3.2.3 Text analysis

The most important requirement of textual data handling is that the knowledge discovered must be understandable and useful for meeting the business needs. The discovered knowledge is often presented using visualisation tools to help the analysis in identifying the corresponding relationships existing in terms to their links with other terms of the same category or different set of terms.

The figure 4.2 shows the interactive and iterative procedures which are adopted during the process of discovering knowledge from textual databases. The information

available in the form of textual information sources is used as an input to the text preparation and text processing procedures. Both the text preparation and the text processing stages should work interactively to form the basis of finding useful and understandable patterns in data which are then visualised in the text analysis stage. Finally the results are published in the form of graphs or tables.

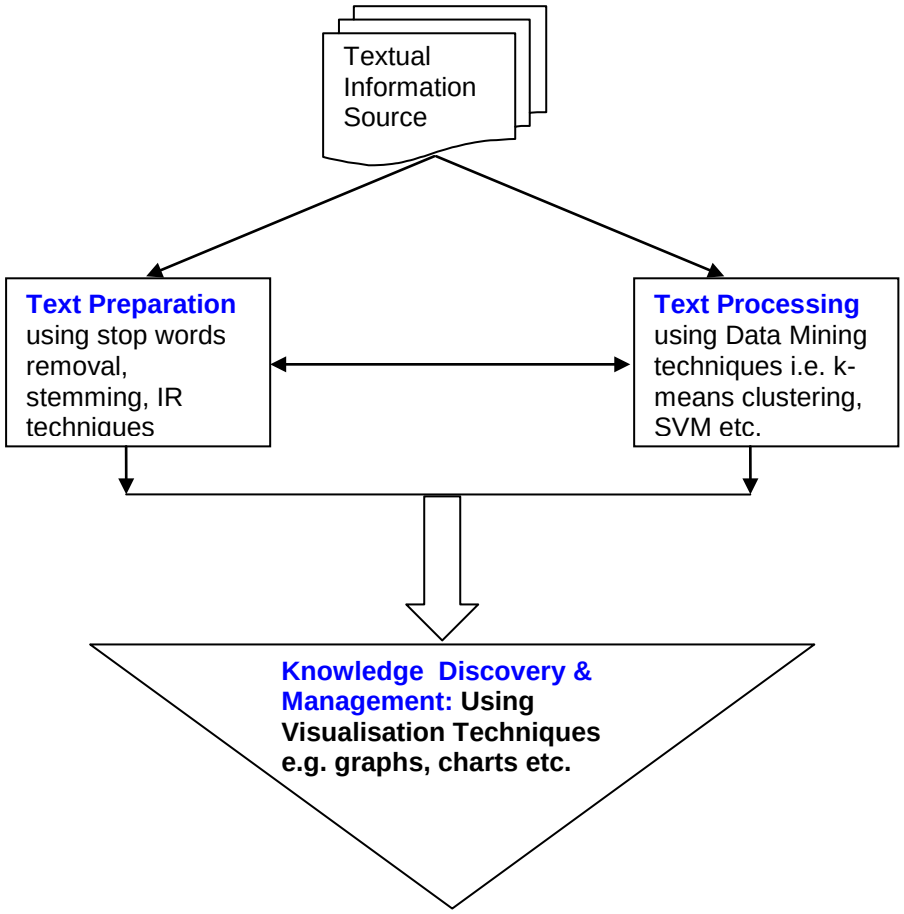

Figure: 4.2 Text Mining process as Interactive and Iterative Procedures

4.3.3 Text Mining and Core Technologies for Information Processing

This section describes the efficiencies of text mining techniques compared with other information handling technologies. Text mining shares techniques from other areas of information processing and knowledge handling techniques i.e. information retrieval, natural language processing and pattern recognition. These tools or techniques are

used to solve the problems of extracting business intelligence solutions from the text (Baeza-Yates and Ribeiro-Neto 1999).

4.3.3.1 Information Retrieval

Information retrieval methods are taken as a first step in handling the source of information which may be available in textual data formats.

Information Retrieval is defined as the methods used for representation, storage and accessing of information items (Joachims 1998) where the information handled is mostly in the form of textual documents, newspapers, research papers and books which are retrieved from databases according to the user request or queries. Information Retrieval Systems (IRS) are meant to find information which matches their customers or users needs. A Text Mining process differs from information retrieval in the sense it identifies the "knowledge" as a consequence of applications of data mining techniques which is new, potentially useful and ultimately understandable. Thus the focus of text mining is more generic in comparison to the IRS since the information is not already known as it is in the case of IRS (Hearst 1999).

In a typical IRS the query is constructed by the user which is analysed and compared to the documents available within databases and the required information is brought to satisfy the user's needs. These systems were first adopted in libraries to satisfy their user's needs (originally using card catalogues, but now using digital resource and information management systems). World Wide Web pages have attracted the attention of users by providing capabilities for accessing information and knowledge across hundreds or thousands of web pages. Measures of performance are common in IRS and these include measures of precision, recall and F-measures, which is most difficult to decipher and relates to how close two documents are to each other (Joachims 1998). Information retrieval methods share techniques from a couple of other areas which have helped to develop models and techniques to present large collections of text. A big research problem is how to present and identify the documents about particular topics and subtopics. Potentially, the greatest benefit of text mining techniques lies in generating multiple clusters.

4.3.3.2 Computational Linguistics

Text mining techniques share methods from natural language processing to deal with textual information hidden in natural language text based databases. The ability to handle information of this type and make it understandable to the computer lies at the core of text mining technological efforts. Computational efforts are being made to make the computer understand the human natural language but efficient methods are not yet achievable for processing these types of information and extracting useful knowledge patterns. Therefore text mining techniques can offer benefits for processing human natural language information with speed and accuracy (Gao et al. 2005).

The area of processing information hidden within textual databases falls under Natural Language Processing techniques which have started to fill the gap between natural language and the computer's ability to process information. These methods and techniques generate patterns and teach computers to analyse, understand and generate information which can further be processed by applying further data mining algorithms. These algorithms can help to discover useful patterns for part of speech tagging, word sense disambiguation or creation of bilingual dictionaries.

Information Retrieval (IR) can select relevant documents of interest for a user enabling him to use his time more productively on these particularly in cases where he could not read through all the available documents and therefore would miss some useful information. Natural language processing therefore targets appropriate information by digging deep into the structures of textual data.

4.3.3.3 Pattern Recognition

Pattern Recognition is the process of searching for predefined sequences in text. In a text mining scenario it is taken as a process of matching the patterns using words as well as morphological and syntactic properties. Two different methods for pattern recognition are terms or word matching and relevancy signatures. Word and term matching methods are easier to implement but need manual efforts as well whereas relevancy signatures are based upon methods of morphological and syntactic information processing techniques.

4.4 Text Mining Role for Advancement of KM and BI Solutions

Business Intelligence Systems are used as a term for integrated sets of tools, technologies and programme products which are used to collect, integrate, analyse data and making it suitable for particular business decision making (Reinschmidt and Francoise 2000). These systems are used to optimise the business process and resources, increase profit and improve the decision making processes. The key components of knowledge management and Business Intelligence are articulated in figure 4.3.

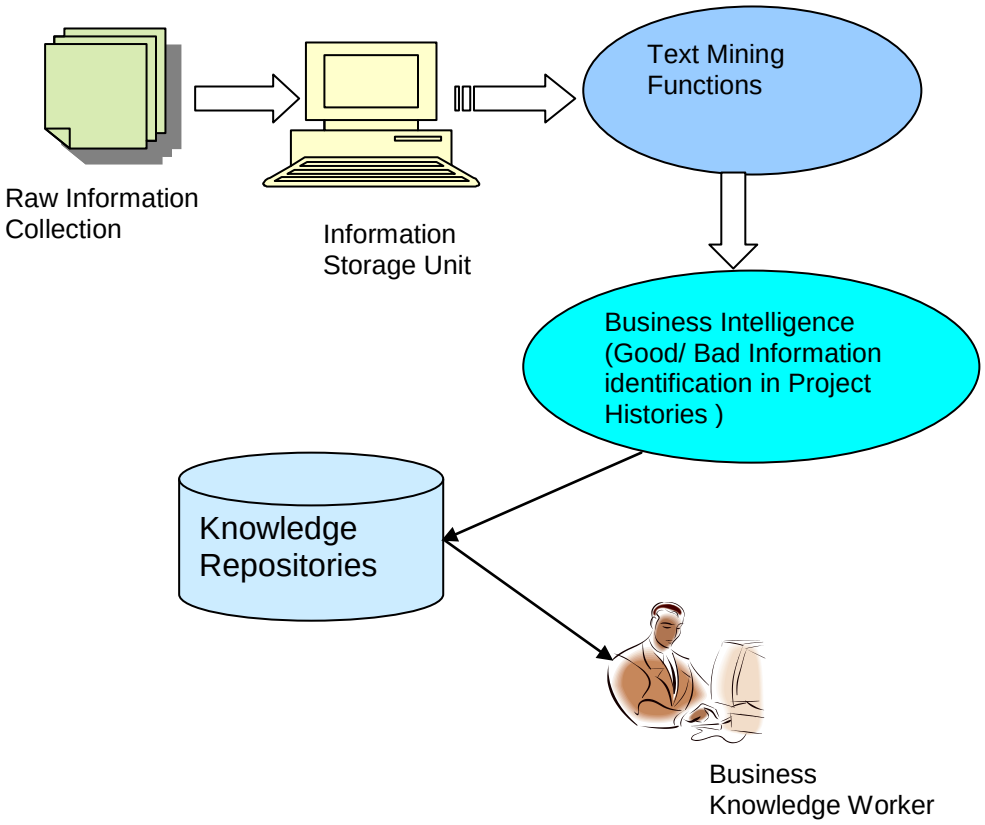

Figure 4.3: Text Mining for Knowledge Management and Business Intelligence

The key business intelligence components described in (Cloza and Ziemba 2006) are defined as;

- Extraction-Transformation-Load (ELT) and Data Warehouse are taken as Key Information Technological tools.

- Potential Information Technologies are taken as data analysis and its presentation techniques mainly rely on Online Analytical Processing (OLAP) and Data Mining techniques and finally.
- Business Intelligence applications support effective decision making on production, sales, competition, monitoring, finance etc. (Kalakota and Robinson 1999).

Different areas of applications in which business intelligences solutions are useful range from trading companies, banking and finance, telecommunication and manufacturing. Some of the key roles played by Business Intelligence Solutions in Manufacturing are as follow (Reinschmidt and Francoise 2000):-

- Sales. Analysing customer transactional databases
- Forecasting. To forecast customer demands and define inventory requirements
- Ordering and replenishment. Order optimum quantities of items
- Purchasing. Provide help to the distribution centres to manage requirements for increased volumes.
- Distribution and logistics. Utilising the advance shipment information in order to schedule and consolidate inbound and outbound freight.
- Transportation management. Developing optimal plans for load consolidation and routing schedules.
- Inventory planning. Identifying the needs at inventory level and ensure a given grade of services.

The reported applications of data mining techniques in Chapter 3, which provide business intelligence solutions in manufacturing, are sufficient to show the power of these techniques to solve problems related to product or service quality improvement. Since the powerful characteristic of knowledge management is to provide a systematic approach to manage organisational knowledge in a beneficial manner (Davenport and Prusak 1997), information technological efforts should be effectively used to transform tacit knowledge into an explicit knowledge (Marwick 2001). However data mining techniques are less efficient at handling traditional databases where the information sources are more unstructured. Solutions can also be provided

by other techniques such as data warehouse, multidimensional models and ad hoc reports but these techniques are unable to cover the full scope of business intelligence solutions (Berry and Linoff 2004). Text Mining methods can give additional advantages by better management of the knowledge resources and knowledge management activities (Hafeez et al. 2002).

The important factor or component of Knowledge Management methods is knowledge discovery which is purposefully used to derive useful information in terms of knowledge from available data. For example the knowledge in terms of useful information may be about (Spinakis 2001):-

- finding what new markets are there for the existing products
- what information is available on internet or intranet
- keeping track of what the industrial competitors are doing at
- the customers needs and what they think about a particular product and services
- new developments made in the market or market trends in industrial environments

Text mining as a term of discovering useful information in terms of knowledge can help to process the information and improve the productivity of knowledge workers and consequently add value to the corporate information by facilitating the process of decision making at less cost than other text processing techniques (Spinakis 2001). To gain more competitive advantages in newly developing industrial business environments there are pressing needs to utilise multiple information sources and consider knowledge discovery techniques. So more attention should be paid towards text mining techniques in business intelligence solutions (Nasukawa and Nagano 2001; Gao et al. 2005). The knowledge discovery and management process to gain competitive business advantages is shown in figure 4.4.

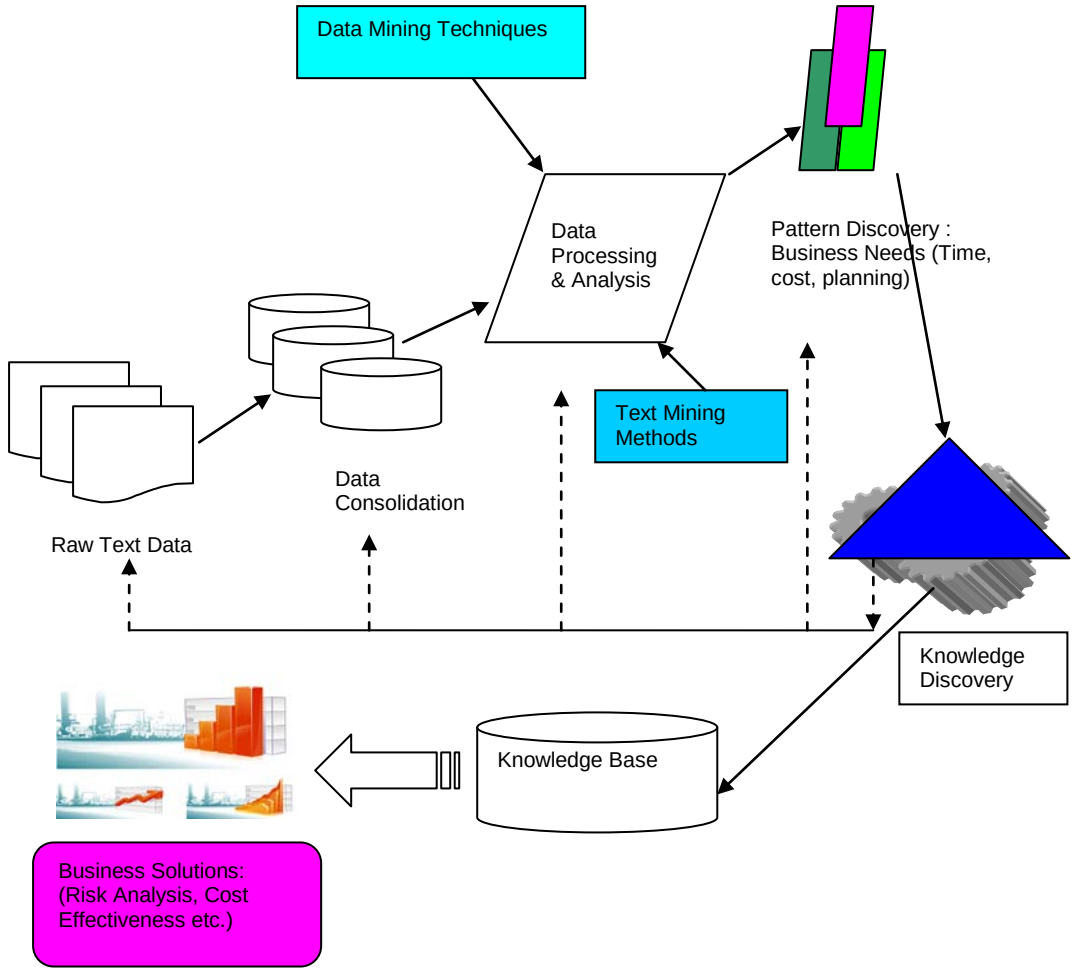

Figure 4.4: Textual Data Mining for Downstream Knowledge Discovery and Management Solutions

4.4.1 Applications of Text Mining in Real Domains

Various applications of text mining in real domains show the power of these technologies to advance the business intelligence solutions in competitive business environments.

- **E-mail Monitoring and Management Systems**

The increased uptake of electronic communication systems (e.g. e-mail) in a business environments has increased the volume of information to billions of text documents

(Spinakis 2001; Weiss et al. 2005). Many companies make sincere efforts to provide their employees with a good working environment but fail to deal with unnecessary information like viruses, chain letters and non-business letters. Capturing such traffic from the systems will protect both the employer and employees. In addition to this some special rules are applied to the contents of the emails within certain industries. For example many financial companies are bound by law to communicate with their customers by abiding to the securities law and regulations. To prevent improper or illegal communications many manual review procedures have been put in place, yet text mining systems may provide better solutions than these procedures.

- **Document Management Systems**

Large volumes of textual data exists in terms of archived documents in company databases. These documents can be in the form of research reports, sales information and product documentation or could also be external sources of information such as the competitors backgrounds and news releases. Many companies might take greater advantage from these sources of information by indexing, cataloguing and extracting useful information. Text Mining technology can potentially offer better solutions in these cases by tracking and finding the key relationships among persons names, location and their specific domains (Spinakis 2001). For example if “Mr Smith became CEO of XYZ corp.” and "XYZ Corp is opening a new branch office in Some City, USA" then relationships between these two texts might be found to provide knowledge of what key players are doing . So text mining techniques offer potential benefits in these type of analysis or solutions.

- **Market Research and Automated Help Desk**

Text Mining techniques can help to find the occurrence of words, phrases, or themes that are useful for finding market needs and trends in the market research areas. They can also provide solutions to automatically analyse customer email systems and route or categorise them (Spinakis 2001; Gupta and Lehal 2009). For example mails from a customer regarding some complaints are sent to customer service department for handling. Similarly a technical department has to produce the answers to the queries from their customers about some specific product. Analysis of complaints and suitable responses to questions sometimes can be handled automatically. Text mining can be

used to provide solutions to them by categorising questions and providing answers to them automatically.

- **Business Intelligence Solutions**

Due to the ever increasing volume of information and technological advancements there is a need to interact with these resources and related information in a timely and efficient manner. Text mining techniques may help company business intelligence officials or analysts to perform their duties efficiently by collecting relevant information about markets and competitors (Spinakis 2001; Gupta and Lehal 2009).

4.4.2 Manufacturing Knowledge Management

It is said that the amount of information doubles every twenty months and therefore the size of databases increases faster than before (Loh et al. 2002). With this massive increase in the size of databases statistical methods can be used to produce nuggets of knowledge and improve quality issues and suggest design improvements (Loh et al. 2002). In manufacturing or construction industrial environments a large amount of information is about product failures or in maintenance reports. However, the knowledge in these is difficult to identify, capture and manage (Loh et al. 2002; Harding et al. 2006). Data mining techniques are well established to handle numerical databases whereas most corporate knowledge is available in the form of textual documents (Tan 1999; Spinakis 2001; El Wakil 2002; Karanikas and Theodoulidis 2002). Handling textual documents or text in databases and analysing it to deal with product or service quality improvement issues could give potentially useful results . Since it is expected that 80% of corporate information is available in the form of textual databases better methods for handling this information in general and in manufacturing in particular could provide significant advantages (Menon et al. 2004; Harding et al. 2006).

In a complex manufacturing environment such as the Xerox copy centre which has more than 2000 product parts, its development process involves generating 12000 engineering problems which are solved by taking 1,000,000 decision making steps (Hauser 2002). Thus increase in product yield and dealing with quality issues is a very

challenging task, but text mining can offer solutions to these problems (Gardner and Bieker 2000).

Applications of text mining techniques in manufacturing industry have shown their potential to change business performance (Braha 2002). For example, text mining technology was used in the Ford Company for early detection of warranty defects which helped to reduce the expenses of the industry and improved customer satisfaction (Duverge et al. 2005). They used text mining efforts to extract useful information by searching through data files, such as vehicle mileage, part codes, and labour operation codes, which ultimately helped to save tens of millions of dollars (Duverge et al. 2005).

In the National Highway Traffic Safety Agency in the United States, customer complaints are collected and gathered to investigate the underlying product problems regarding safety related vehicle defects and crashes. Companies may be directed to give customers services free of charge and even, if needed, the agency is empowered to ask the manufacturer to conduct a recall, if warranted by the problem. The data or information related to automobile information was collected in the form of descriptive text information on damages and accidents. Manufacturers are therefore potentially required to explore the factors lying behind these problems and automate the problem solving methods by applying data and text mining techniques (Drewes 2005). To handle these multiple tasks several qualitative and quantitative tools have been devised (Spinakis and Peristera 2003). These tools are based on the synergetic efforts of different areas described in section 4.3.

4.5 Summary of the Chapter and Conclusion

This chapter discusses the needs of text mining in the context of Knowledge and Information seeking communities. This provides a background review to begin and to address the key question in the thesis i.e. “How to shift the paradigm of knowledge discovery to Knowledge Management to support Business Intelligence Solutions in manufacturing and construction environments?”. The basic tools for knowledge discovery and knowledge management include many different data and text mining tools to handle multiple data formats within company data which are estimated to

represent 80% of corporate information. ***This chapter is however intended to describe a conceptual relationship among different parts of knowledge discovery and management technologies i.e. Data Mining, Text Mining and Business Intelligence and thereby to form a basis for developing an integrated knowledge discovery and text classification framework to analyse the semi-structured or unstructured databases proposed in Chapter 6***. It provides background knowledge of enterprise knowledge discovery solutions which are capable of handling textual data or information to discover useful knowledge relationships.

Chapter 5 Knowledge Discovery Functions and Implementation Issues

5.1 Introduction

In this chapter the algorithms used in the proposed framework are discussed. There are three levels of knowledge considered in the Knowledge Generation and Classification Module these are Level Knowledge Processing and Storing Unit, Level Knowledge Refinement Unit and Level Knowledge Utilisation and Text Classification Unit. The algorithms used in the first level knowledge processing and storing unit and second level of knowledge refinement unit are discussed in detail in this chapter. The potential advantages and weakness of these algorithms are also discussed in terms of experiments performed for discovering useful knowledge from free formatted textual databases. Some information structuring methods are also discussed which play a very important role in handling textual databases for discovering useful knowledge.

5.2 Expected Benefits Associated with Term Based Analysis

The benefits associated with term based analysis for discovering useful knowledge are defined as follows;

- The analysis of textual databases using a term based analysis method is useful in finding sensible relationships by exploiting the co-occurrence of terms in the text.
- Clustering algorithm applications at the document level i.e. clustering documents, may find those documents within a cluster which share no common terms or concepts but analysing textual data on the basis of term based relationships helps to overcome this difficulty.
- Clustering algorithm implementations at the terms level help to generate disjoint clusters of terms which share some meaning defined in the text. So key information which is defined in the textual data is transformed into a useful source of knowledge.

There are various algorithms proposed in the literature for finding the solutions to different problems associated with industrial or other knowledge domains. The following sections focus on defining two well known algorithms which will play a

key role in de-codifying the information and will therefore form a sound basis for defining a new hybridised approach in the Chapter 6 to handle the information and discover useful knowledge.

5.3 Clustering and Apriori Association Rule of Mining Techniques

5.3.1 K-means Clustering Algorithm

The K-means clustering algorithm (MacQueen 1967) is a well accepted algorithm for uncovering the information hidden in the data and this algorithm is therefore considered to play a key role in the field of knowledge engineering. The algorithm is very straight forward and effective in terms of converting the information space of words (or terms in the current research scenario) defined in text documents into a space of K clusters. It works on the principle of taking input in the form of document (or term) vectors and assigning these vectors to clusters by finding the similarities between different terms represented in a multi dimensional vector space model. The simple form of k means clustering takes input and assigns clusters by choosing their respective centroid and measuring the distance between each centroid to the term vector. The process of choosing a centroid and assigning terms to respective clusters is repeated until the cluster membership no longer changes.

The formal structure and flow of the algorithm is given in the following steps which are defined in (Larose 2005);

Step1*: *Selecting the number of clusters

The functionality of the k-means clustering algorithm can be defined as **(*k*, T)** where ***k*** is the number of desired clusters i.e. k=2,3,4 or more. Consider the set of documents as $D = \{d_1, d_2, ..., d_n\}$ and **T** = $\{t_1, t_2, ..., t_m\}$ is taken as the set of terms defined within these documents.

Step2: Initialising the Cluster Centroids

Initialise *k* centroids $(m_1, ..., m_k)$. Then each cluster C_j is associated with a centroid m_j, $j \varepsilon \{1, ..., k\}$

Step3: Classifying Terms to Nearest Cluster Centroids

Repeat

for each input vector t_l , where $l \varepsilon \{1,...,m\}$,

do

Assign t_l to the cluster C_j with nearest centroid m_j

Step4: Updating Clusters Centroids

for each cluster C_j , where $j \varepsilon \{1,...,k\}$

do

Update the centroid to be the centroid of all samples currently in C_j ,

so that $m_j = \sum_{t_l \varepsilon C_j} \frac{t_l}{|C_j|}$

Step5: Terminating Condition

Compute the error function Sum of Squared Error (SSE) where

$$SSE = \sum_{j=1}^{k} \sum_{t_l \varepsilon C_j} d(t_l, m_j)^2$$

Until *SSE* does not change significantly or cluster membership no longer changes.

The flow of information in the above algorithm is shown in the figure 5.1 below:

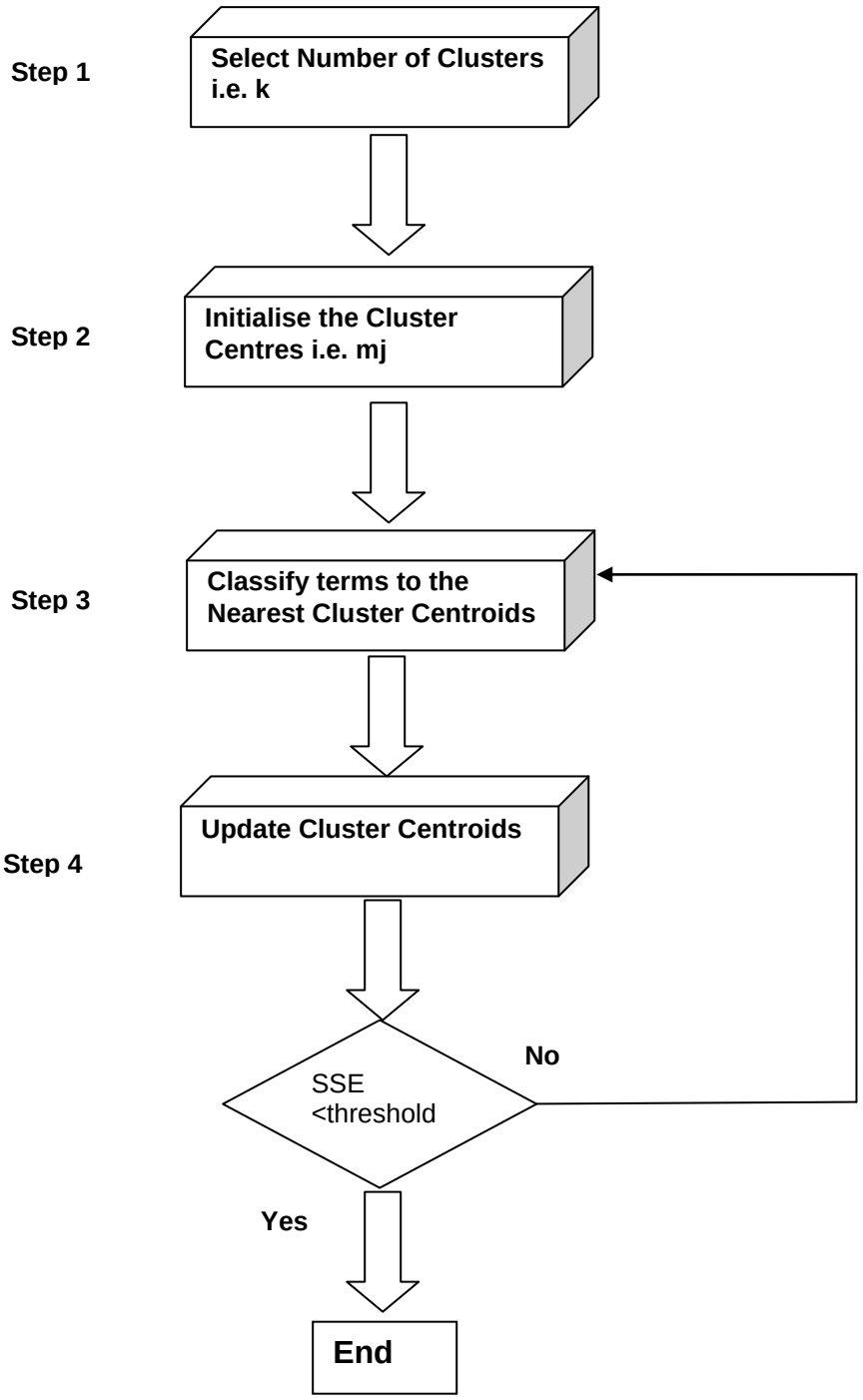

Figure 5.1: Flow of Information in Clustering Algorithm

The terms and the related clusters are shown in the Figure 5.2 where the distance between two terms is measured using some distance measure and then terms closer to

one centroid are put together in a cluster with centroid as m1, m2 or m3. The process is repeated until disjoint clusters of terms are formed i.e. each cluster consists of member terms which are different from other cluster's member terms. Three different clusters with centroids m1, m2, and m3 are shown in figure 5.2 below.

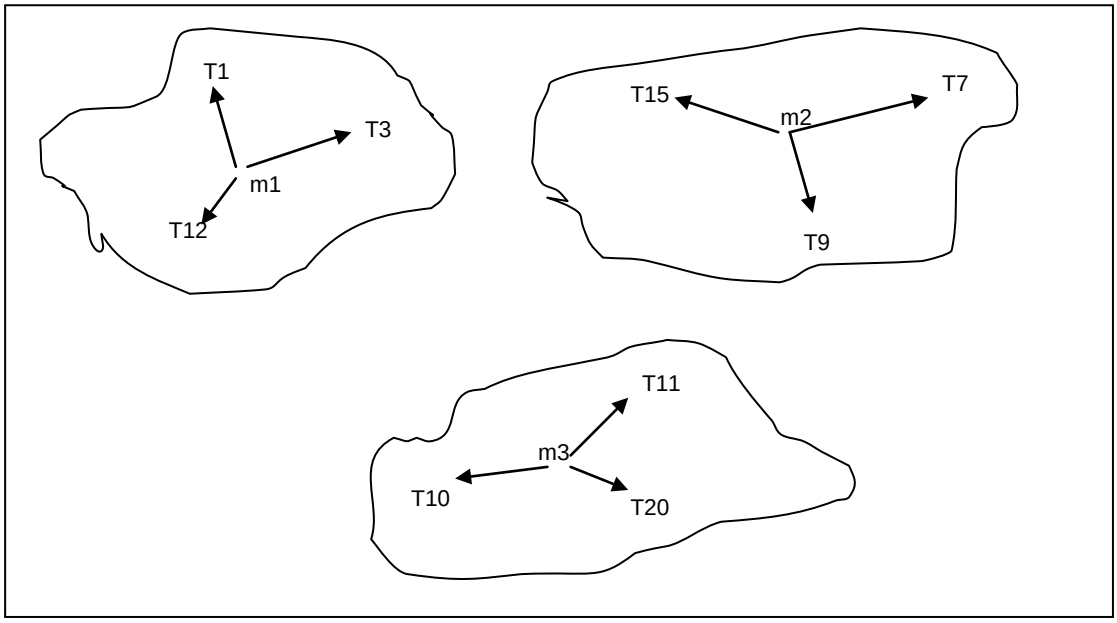

Figure 5.2: Clustering on terms based vector model

5.3.2 Apriori Algorithm for MKTPKS Generation

5.3.2.1 Background

The term Apriori Assocation Rule of Mining originated from the customer behaviour of buying products in a supermarket and it aims to help the market place to arrange their items on the shelves in order to improve customer services. In the current research scenario the aim is to find the relationships among terms in the textual data by finding patterns which identify good or bad information documents. The automatic discovery of frequent termsets or MKTPKS (in the current research scenario) is a very promising area of research in the field of data mining. It is also emerging as a sub-field within statistics and pattern recognition, concerned with finding patterns and connections between elements in large databases (Bodon 2003).

A frequent termset is defined as a set of terms that occurs within the text document set more than the chosen minimum support times. A common example is described in

(Han and Kamber 2001) which is used to demonstrate customer buying behaviours based on super market's transactional databases. From a statistical analysis perspective, if the analysis of the transactional databases show that a man often buys some diapers and beer after work then a beer refrigerator and diaper aisle should be put close together to facilitate their customer's shopping.

The problem of finding frequent termsets starts with a database of transactions T:

$T := \{t_1, t_2, ..., t_n\}$

With each transaction t_i being a termset $(t_i \subseteq I)$ and the support of a termset $l (\varepsilon I)$ in T is defined as the number of transactions that contain l as a subset and it is denoted as follows:

$\text{suppT}(l) = \{t_j \varepsilon T : l \subseteq t_j\}$

Any termset l is frequent if its support is greater than a given minimum support i.e. $(|\text{suppT}(l)| \geq \text{min supp})$. Any termset with k elements that is frequent is called a frequent *k*-itemset or termset (Bodon 2003). The problem is thus to find all frequent termsets in a given database of transactions T . One of the most important contributions for solving this problem in an efficient way is the Apriori Algorithm, which was proposed by (Agrawal et al. 1993). Apriori has quickly become the "gold standard" that all other frequent itemset or termset algorithms are measured against and today the notion "Apriori" covers a whole family of algorithms based on the same basic ideas (Bodon 2003).

However, the original Apriori Algorithm as defined in (Agrawal et al. 1993; Han and Kamber 2001) will suffice for the purpose of the current research work presented in this thesis.

5.3.2.2 Apriori Algorithm

In terms of explaining the Apriori Algorithm the notations used are illustrated below;

• Lk is the set of frequent *k*-termsets (i.e. those with at least the required minimum support), each member is represented by the termset and the support count.

• Ck is the set of candidate *k*-termsets (i.e. potentially MKTPKS termsets), each member is also represented by the termset and the support count.

It is also assumed in the following paragraphs that all termsets are represented as ordered sets.

The Apriori Algorithm is actually rather simple. Apriori exploits the basic fact that all subsets of a frequent termset are also frequent. The whole algorithm can be divided into the following steps;

Step1: Initial Data Scanning for MKTPKS 1-termset

The initial pass first scans through the data to find the minimum support of each item or term. These are then used to find the MKTPKS 1-termsets by counting terms and finding the ones that are frequent (having min support or higher support levels). The term support is defined as the percentage of data containing both the terms T1 and T2 together in a transactional database.

The tasks in this step are summarised as follow:

- Scanning the whole data to find the support S of each item or term
- Comparing the S against minimum support
- Finding the MKTPKS 1-termset

Step2: Candidate Termset Generation & Prunning

This is the step where Lk-1 is joined with itself to generate the candidates termsets Ck. It is the stage at which each frequent (k-1)-termset Lk-1 is used to generate possible candidates for the frequent k-termset Lk. The name of Apriori is derived from the property of building the candidate termset by pairing termsets which have their first k-1 terms in common and that no superset of an infrequent termset will be frequent (Larose 2005). The information derived from step 1 is used to determine the candidates that need to be examined in this step. Then candidate termsets are pruned using the Apriori property i.e. no superset of an infrequent termset will be frequent alternatively any subset of the frequent termset is frequent. That is all (k- 1) subsets

of the candidates in Ck are checked as being frequent, otherwise the candidate is removed from Ck.

The tasks in this step are summarised as follow:

- Lk-1 is joined with Lk i.e. (Lk-1►◄Lk) to generate candidate k-itemsets or termsets
- The infrequent termsets are pruned using the Apriori property

Step 3: MKTPKS Termset Scanning and Refinement

This step determines the support of the candidates in the transactional databases denoted as T= $\{t_1, t_2, t_3, ..., t_n\}$. All candidates with less than the required minimum support are removed and not saved in the frequent termset Lk.

- Scanning further the frequent termsets and finding the support of each termset
- Pruning the infrequent termsets using the Apriori property

Step 4: MKTPKS Termset Formation

This step checks whether the candidate frequent termset has been formed satisfying the minimum support value otherwise the process is repeated. This is the most data intensive part of the algorithm since it is necessary to iterate through all transactions in the database. Then finally the MKTPKS termsets Lk are generated.

The detailed flow of information in forming the MKTPKS termset is shown in the Figure 5.3 below;

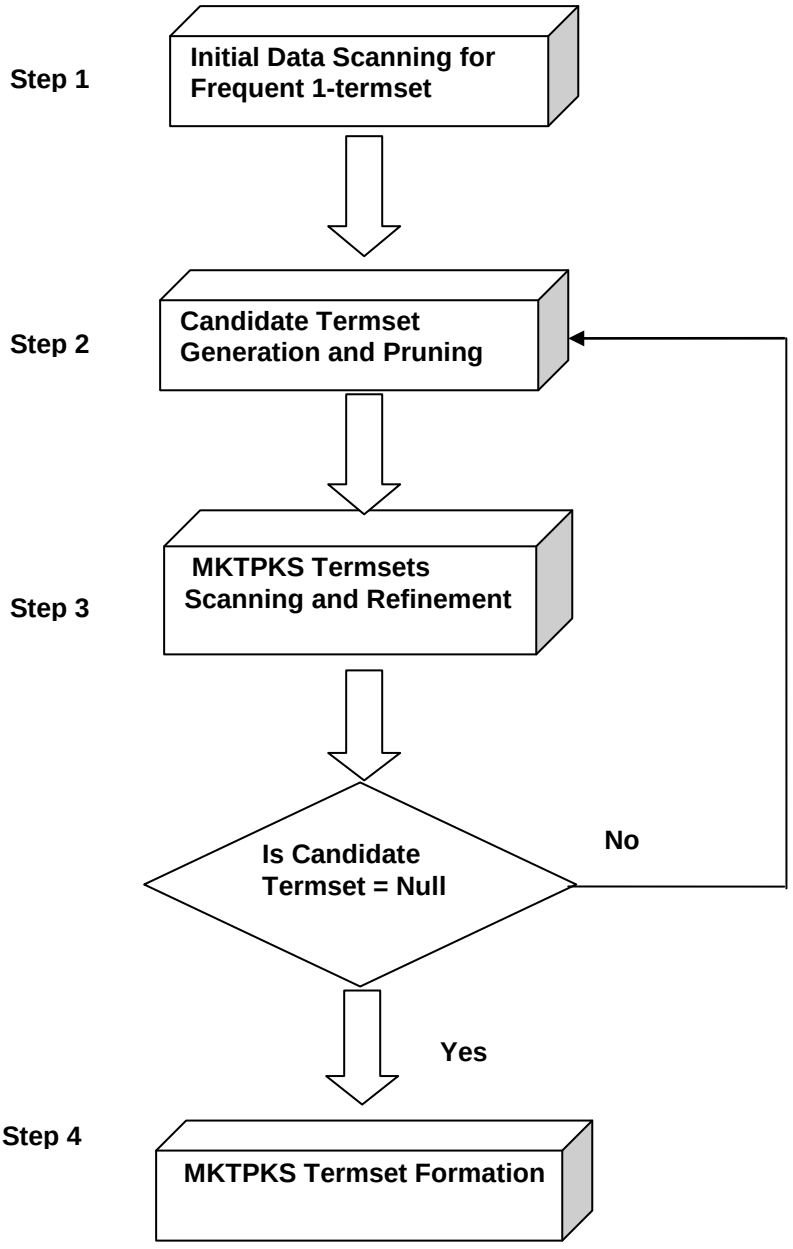

Figure 5.3: Flow Diagram for Apriori Algorithm for MKTPKS Formation

In the coming paragraphs the formation of frequent 3-termset generation process is illustrated with the help of a representative dataset taken from the textual data. The terms are represented in the form of representative terms i.e. T1, T3, T4, T5, T7, T9, T11 in which each has a minimum level of support 2. Table 5.1 shows the

representative candidate 1-termsets formed through scanning the whole database. The Table 5.2 shows the MKTPKS 1-termset formed from the candidate 1-termset. Table 5.3 shows the candidate 2-termset formation and pruning (i.e. the candidate 2-termsets with minimum support less than the required minimum level of 2 are eliminated) which is used to generate the corresponding MKTPKS 2-termsets shown in the Table 5.4. The final candidate 3-termset and corresponding MKTPKS 3-termset generation is shown in Table 5.5-5.6. Thus overall process from Step 1 to Step 4 is shown with the help of Tables 5.1- 5.6.

Table 5.1: Representative Candidate 1-termset possible combinations

Candidate 1-termset	{T1}	{T3}	{T4}	{T5}	{T7}	{T9}	{T11}
Count	2	3	2	3	3	2	3

Table 5.2: Representative MKTPKS 1-termset

MKTPKS 1-termset L1	{T1}	{T3}	{T4}	{T5}	{T7}	{T9}	{T11}
Count	2	3	2	3	3	2	3

Table 5.3: Representative Candidate 2-termset and pruning

Candidate 2-termset C2	~~{T1, T3}~~	~~{T1, T4}~~	{T1, T5}	{T1, T9}	{T3, T4}	~~{T3, T5}~~	{T5, T7}	{T5, T9}	~~{T5, T11}~~	~~{T7, T9}~~
Count	~~1~~	~~0~~	2	3	2	~~1~~	2	3	~~1~~	~~1~~

Table 5.4: Representative MKTPKS 2-termset

MKTPKS 2-termset L2	{T1, T5}	{T1, T9}	{T3, T4}	{T5, T7}	{T5, T9}
Count	2	3	2	2	2

Table 5.5 : Representative candidate 3-termset

Candidate 3-termset (C3)	{T1, T5, T9}	{T5, T7, T9}
Count	2	2

Table 5.6 : Representative MKTPKS 3-termset

MKTPKS 3-termset (L3)	{T1, T5, T9}	{T5, T7, T9}
Count	2	2

5.4 Potential Strengths and Limitations of Clustering and Apriori Association Rule of Mining Techniques

5.4.1 Clustering

5.4.1.1 Strengths

The following strengths can be associated with using the above clustering functions in the process of discovering useful knowledge (in terms of finding key term phrasal knowledge sequences using case study data). The observations made are as follows;

- Firstly the number of clusters, k for which the cluster centroids are selected affects the length of processing time required. The steps (i.e. step1 to step 5 defined in the section 5.3.1) are performed for the number of clusters which were initially selected as k. The computer time and memory space used are therefore of order O(k), where k is the number of desired clusters.
- Secondly the clustering task process is repeated till the cluster does not change it membership i.e. terms within each cluster remain the same. So the process is repeated for some threshold value (η) in terms of SSE.
- Thirdly the distances of each term from the centroids are computed which gives an order of computation as O(kn), where n is the total number of terms in the document vector space and k is the number of desired clusters.

So the overall algorithm requires memory and space requirements in terms of assigning terms to their clusters as $O(\eta kn)$. This shows that K-Means clustering

algorithm therefore has , linear memory requirements, in order to store the documents, determine cluster membership of the terms and determine the cluster centroids.

The main advantage of K-Means for assigning terms to their respective clusters is that its very simple in terms of time and memory used for running the algorithm and it produces good results which was confirmed in the case of its implementation on the case study data set.

5.4.1.2 Limitations/ Weaknesses

There are some of disadvantages or weakness which can be identified in terms of the application of the k-means clustering algorithm in spite of its popularity. The issues that exist also limit the advantages of using the k-means clustering algorithm.

- First of all, in terms of selecting the number of clusters, it is not at clear how many clusters would be sufficient to de-codify information into different clusters i.e. determining the optimal number of clusters is the most apparent issue in any information and knowledge handling space. Choosing a low number of clusters might cause the information to be gathered in a single cluster while using too many clusters could disperse the key information carrying terms into different clusters (and this has been experienced in handling the case study data). In terms of implementation it was very difficult to find the exact number of clusters that best de-codify information in discovering first level of knowledge. So in this research the number of clusters retained was six as will be discussed in Chapter 7.

- Secondly, the direct consequence of implementation of k-means algorithm is to partition the data into clusters with useful information. The nature of its implementation is to partition the information which adds to the problem of how to generate the correct number of clusters and uncertainty over the problem of cluster-number increased. In any industrial or manufacturing domain the information needs to be clustered so that it could be transformed into a sensible key information or knowledge to but the partitioning nature of this algorithm also limits the potential advantages. The problem was faced

during its implementation on the case study data and it was addressed by paying special consideration to clustering those terms which form some meaningful structures as defined in the case study data.

- Thirdly the key information captured within each cluster also relied on the selection of the centroids, the initial cluster centres. The choice of these cluster centroids also generates problems of uncertainty. So there is no way to know whether two centres are placed in well-separated clusters, or if they end up in the same cluster. This results in generating poor quality clusters and either some clusters carrying good information are split or some clusters are joined together and contain poor quality information as observed in the process of implementation of this algorithm on the case study data.

5.4.2 Apriori Association Rule of Mining

5.4.2.1 Strengths

The application of Apriori Association Rule of mining can give the following potential benefits while mining the textual databases.

- The extraction of different relationships among terms defined within textual data are easy to understand
- It provides clear understanding of relationships among terms in order to characterise the knowledge discovered through application of frequent termset mining techniques.
- The information given as input is taken as a whole to discover the knowledge and there would therefore be less chance of losing key information in terms of processing knowledge from the textual data.
- The knowledge discovery process is simple as the minimum support measure is needed to find the refined knowledge in terms of MKTPKS .

5.4.2.2 Limitations

The applications of Apriori Association Rule of Mining methods have some limitations as follows;

- The determination of MKTPKS is a sequential process where first MKTPKS 1-termset are found and are then used to generate the MKTPKS 2-termset

and so on. This causes the size of each termset to grow and large number of MKTPKS are generated so a change in the minimum support level is necessary.

- The MKTPKS are determined on the basis of frequency of occurrence of terms in the text documents so a difficulty arises in finding the right number of terms used for forming the MKTPKS by choosing the minimum support value.

5.5 Structural Data Representation Methods

The representation of textual data into suitable formats plays an important role in handling information for data analysis and classification. Different studies have shown that the following representation techniques are useful for performing the clustering and textual data classification tasks (Leopold and Kindermann 2002). Three different data representation techniques have been considered during this study given as follows;

1. Binary representation
2. Term frequency
3. Term frequency inverse document frequency length normalised (tf-idf) representation

- **Binary Representation**

Let a_{ij} be the weight of term i in document j, f_{ij} be the frequency of term i in document j, df_i be the number of documents in which term i occurs, N be the number of documents and M be the number of distinct terms in the document collection.

The binary representation is one of the simplest but effective forms of representation. A word present in the document would be given a feature value of 1 whilst a zero is used if the word is absent. The formula below shows the weight for each word under the binary representation scheme.

$$a_{ik} = \begin{cases} 1 \ if \ \ f_{ik} > 0 \\ 0 \ otherwise \end{cases} \qquad (5.1)$$

- **Term Frequency Representation**

In this type of representation each text document is represented in the matrix form where the frequency of each term occurring in the document is represented. It is a similar representation to the binary representation except that the actual occurrence of terms are recorded in the matrix form.

- **Term Frequency Inverse Document Frequency Representation**

The previous schemes do not take into account of the frequency of occurrence of the word throughout all the documents in the collection (Salton and Buckley 1988). A well known approach for computing word weights is the TF*IDF weighting, which assigns the weight to word *i*, in proportion to the number of occurrences of the word in the document, and in inverse proportion to the number of documents in the collection for which the word occurs at least once. This weighting scheme does not depend on the distribution of terms in documents but only on the number of different documents in which a term occurs. The formula for the data representation using term frequency inverse document frequency method is given below;

$$a_{ik} = f_{ik} * \log\left(\frac{N}{df_i}\right) \qquad (5.2)$$

5.6 Summary of the Chapter and Conclusion

The main purpose of this chapter is to provide details of the two techniques and algorithms which will be included in the proposed architecture and which were used in the initial case study experimentation. This chapter therefore provides a technical and computational background for the discussion of the proposed methodology which is given in the next chapter 6.

Chapter 6 Proposed Methodology and Architecture

6.1 Introduction

In this chapter the two major functions of knowledge discovery defined in the Chapter 5 are incorporated to define a new architecture and methodology for discovering useful knowledge in terms of Multiple Key Term Phrasal Knowledge Sequences (MKTPKS) . The discovery of key information in terms of MKTPKS is used to categorise the documents into predefined categories or classes. The first level of knowledge is discovered in terms of single key term phrases using the k-means clustering algorithm and are then used to generate MKTPKS through the application of the well known method of Association Rule Mining. A hybrid application of these methods is used to form a relational database of knowledge based on MKTPKS as this can be used to analyse and better manage the textual data. The ultimate benefit of the proposed methodology is to automate the process of categorising the textual data or documents into two different classes (e.g. good and bad information) based on key knowledge in terms of MKTPKS .

6.2 Proposed Architecture or Framework

In this section a framework is outlined to analyse textual databases. This framework consists of two main parts (as shown in Figure 6.1);

1) A data handling section called “Text Mining Module” and
2) A knowledge Discovery and Text Classification Section for the discovery of Multiple Key Term Phrasal Knowledge Sequences (MKTPKS). This is called the “Knowledge Generation and Classification Module”.

The Text Mining Module works on free formatted text documents. In the experiments carried out during this research a collection of textual data has been used relating to time, cost and planning information from the Post Project Reviews. Information pre-processing and structuring techniques are then applied on these free formatted text documents with the help of the Information Pre-processing Unit and Information Structuring Unit.

When the information has been structured with the help of the Text Mining Module, it is then passed to the Knowledge Generation and Classification Module for further processing and to discover the Multiple Key Term Phrasal Knowledge Sequences (MKTPKS). This is a new "***Term***" used in this thesis for the sequence of words or terms which define some key information or refer to some issue in the textual documents. **Multiple Key Term Phrasal Knowledge Sequences (MKTPKS)** has been identified through application of **Level Knowledge Refinement Unit** using the Association Rule of Mining Algorithm. In the current research context every document is composed of different words or terms to define some concept or issue in the document, and mining through these documents generates the MKTPKS . The discovery of MKTPKS enables key issues to be summarised within each particular document. The MKTPKS based matrix model is then used to classify documents into two different categories. The figure 6.1 illustrates the proposed framework showing the flow of information, its processing modules and how the processed information is transformed into a set of MKTPKS carrying useful knowledge to further classify the textual data into two different classes.

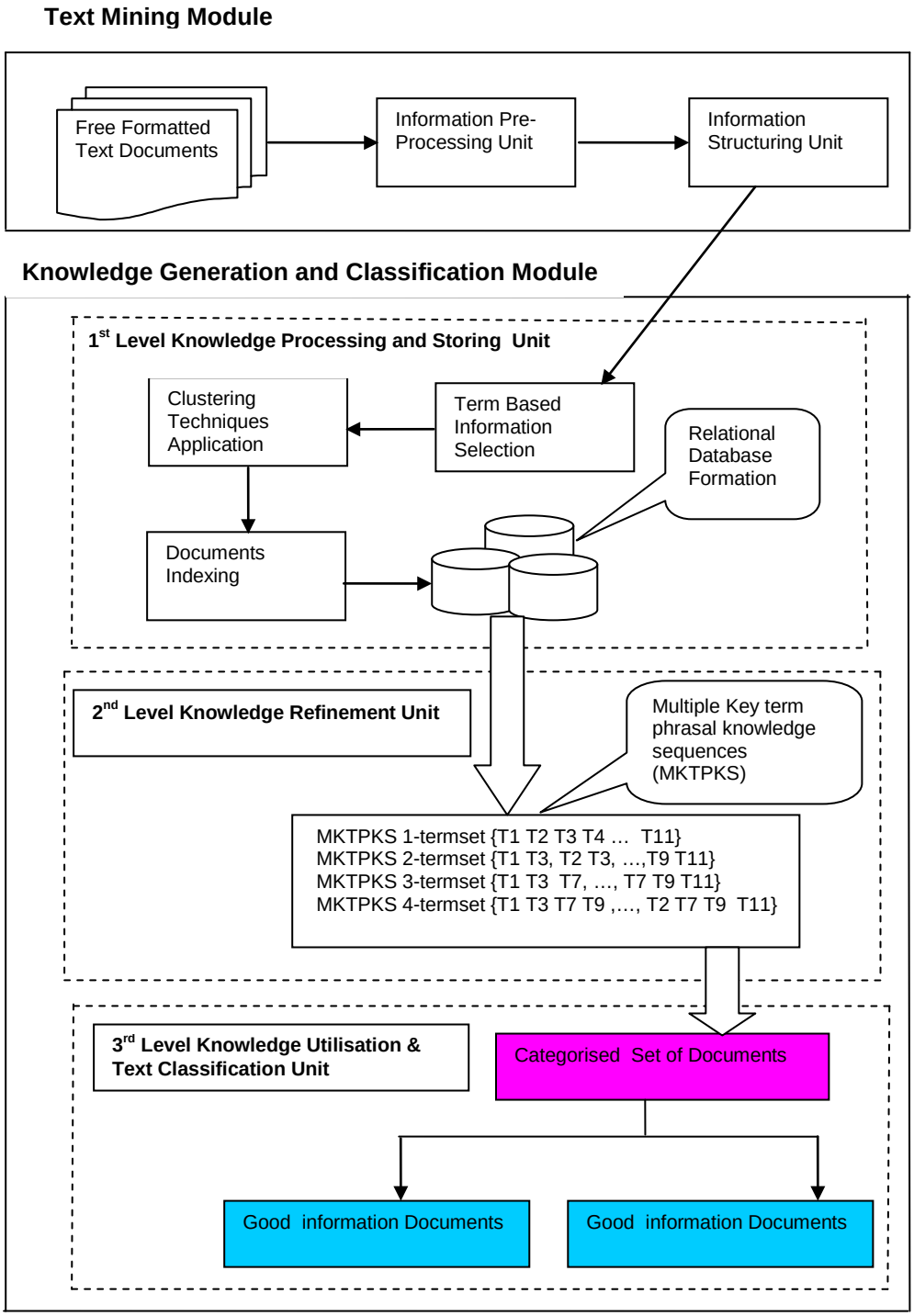

Fig 6.1: Multiple Key term Phrasal Knowledge Sequences based Text Classification System

The framework applies the combined efforts of Data Mining and Text Mining techniques which are referred to as Textual Data Mining (TDM), to analyse free formatted textual data and discover useful knowledge in terms of multiple key term phrasal knowledge sequences .

6.3 Text Mining Module

This module performs two different tasks using the Information Pre-processing unit and Information Structuring Unit. These units are used to remove un-necessary information from the textual data available in the form of free formatted text documents and structure it for the application of different data mining algorithms for subsequent analysis of the text. Therefore the first step towards performing the analysis is to process the information and this is done through the information pre-processing unit.

6.3.1 Information Pre-processing Unit

The process of handling textual data or information in any industrial setup starts by initially considering the opinion of domain experts who might help the data mining expert to define the business needs. Both data mining and domain experts work interactively to identify the input variables which help to start the process of analysing the textual data. Since, the information available in the form of data in manufacturing or construction industry is crucial in decision making, decisions made at this early stage of analysis highly affect the success of Knowledge Discovery (KD) process. The task of selecting input variables, therefore needs to be performed carefully by the data mining and domain experts. The input information must then be codified in a format suitable for the TDM tasks.

The next step is to remove the un-necessary words which are less effective in textual data analysis. These words include some verbs (e.g. is, am, are etc.), pronouns, conjunctions, disjunctions e.g. a, an, the, of , and, ***I***, etc. which are termed as stop words and need to be removed from the list. The assumption behind removal of these words is that text can be assumed and interpreted more easily this way. Removal of these less informative words increases the efficiency and accuracy of results in processing the text and is a common technique in text analysis (van Rijsbergen 1979).

Word stemming is also important and this is the process of shortening derived words to their actual stem, base or root form. For example in English the words like *design*, *designing*, and *designed* are stemmed to their root word *design*. So in this study a simple suffix stripping technique was used to reduce the loss of meaning or context with the help of terms defined in the textual data. This method is therefore used to capture more explicit relationships among terms defined in the text. Thus within this unit three different steps were taken to improve the overall process of text analysis i.e.

- Selection of decision variables or attributes
- Stop words removal
- Stemming

6.3.2 Information Structuring Unit

After the initial pre-processing stage, the next essential part of the proposed framework is to structure the information. Therefore, the pre-processed information is then considered so that it can be structured for further analysis. To perform this task some structuring techniques available in the literature have been applied. The whole words representation methods commonly known as bag of words (BoW) approaches have been used. The reason for choosing these techniques is that the whole information space is taken into account so that there is no information loss. These methods are independent of the structures of text and are represented in the vector form where each term is taken as a word vector. These methods are commonly reported in literature and have been adopted in many studies due to their simplicity and effectiveness (Salton 1989).

6.4 Knowledge Generation and Classification Module

This module works with the help of three main parts i.e.

1. Level Knowledge Processing and Storing Unit
2. Level Knowledge Refinement Unit
3. Level Knowledge Utilisation and Text Classification Unit

A short description of each of these main parts and their sub-parts is given in the following paragraphs.

6.4.1 1st Level Knowledge Processing and Storing Unit

6.4.1.1 Term Based Information Selection

At this stage of analysis the input information is available in different structural representations obtained through the application of the Information Structuring Unit. These structures are available in a word or term frequency (TF) matrix consisting of the vectors describing the information in a document. Each word or term can additionally be weighted in the document collection using Inverse document frequency (IDF) and term frequency inverse document frequency (TF*IDF) matrices. It is therefore necessary to select a suitable representation for performing further analysis on the text data. To avoid losing some key information, a simple term based representation model has been considered where the terms and their corresponding frequencies are counted. The information matrix so formed is taken as an input for the application of clustering techniques.

6.4.1.2 Clustering Techniques Application

Clustering is defined as a process of grouping data or information into groups of similar types of information using some physical or quantitative measures (see section 5.3). These quantitative measures are based upon a distance function, which is measured from some centre point i.e. termed as the centroid of the cluster. The Euclidean distance measure is commonly used as a natural distance function but this may be replaced with other similarity measures to find the similarities between documents. Within the current research context the information is already available as an input matrix based on term frequencies so the similarities are found between terms. Thus clustering techniques are then implemented on these information matrix and output is generated in the form of correlated terms based on their natural relationships existing within each document.

The process of clustering helps to capture information in the form of different clusters formed on the basis of their natural relationships found among terms defined in each document. The information captured within each cluster is carefully observed during the clustering techniques application stage to reduce the risk of loosing key information. This stage will determine the number of clusters to be made and used in further knowledge processing task.

6.4.1.3 Documents Indexing

After performing the clustering task the information in terms of single key term phrases is obtained and this is used to index the documents. The documents corresponding to these single key term phrases are marked with their identification codes. The task of indexing the documents at this stage helps to store information in a useful format for classifying documents based on the information possessed within a cluster.

6.4.1.4 Relational Database Formation

The first level knowledge captured in the previous stages must be stored in the form of relational tables (in a database) for further use in discovering useful relationships among terms by generating multiple key term phrasal knowledge sequences. The tables are based on cluster labels, indexed documents identification (IDs) and their respective key single term phrases. The documents are considered as transactions and the terms captured as a result of clustering are considered as items (a commonly used term in market basket analysis). This helps to form an input space of information which is used in the next stage of **2nd Level Knowledge Refinement Unit** working with the help of Apriori Association Rule of Mining .

6.4.2 2nd Level Knowledge Refinement Unit

In this part of the analysis the input matrix is in the form of relational tables where information are represented in the form of transactional databases. The Apriori Association Rule of Mining (Agrawal et al. 1993) methods are applied on the information matrix formed where documents are the transactions and terms are taken as items. These methods are used to form multiple key term phrasal knowledge sequences based on the single term phrases previously identified by the clustering techniques. The key to implementing these techniques is the process of finding co-occurring terms by searching through all information or document space. The most useful knowledge carrying terms can be found by varying the level of support (the term used for finding the co-occurrence of terms with some defined percentage) see section 5.3.2. This unit is focused on finding the MKTPKS rather than on determining association rules since it is likely that a large number of association rules

will occur. Finding the MKTPKS will ultimately help to overcome the difficulty of populating the knowledge base with too many association rules as this affects the process of discovering useful knowledge from these knowledge bases.

6.4.3 3rd Level Knowledge Utilisation and Text Classification Unit

Data is mainly stored as semi-structured rather than fully structured or unstructured forms. To classify textual data into some predefined categories or classes it needs to be partitioned manually into different classes to test the accuracies of the classifiers. This task is performed through manual inspection of data with the help of domain experts. A categorical attribute is set as a class variable or target variable. The given data was therefore first divided into two different classes of good or bad information documents (in the current research case study) using the information available in the form of interpreted single or multiple term phrasal knowledge sequences identified by the knowledge workers of the TrackStore project (http://www.lboro.ac.uk/departments/cv/projects/trackstore/index.htm).

During the research into the specification and implementation of this unit different classifiers were used to study their effect on classification of data into their respective categories. In the current research focus has been placed on application of Decision Trees (C4.5), K- Nearest Neighbouring (K-NN), Naïve Bayes and Support Vector Machines (SVMs) algorithms discussed in detail in Section 8.4. The reason behind the application of these different classifiers is their variability of information selection criteria for classification of documents. In the case of Decision tree (C4.5 or J48) the information is selected using entropy measure, K-NN uses simple distance measure termed as Euclidean Distance Measure is defined in equation 8.4 (page No. 112), probabilistic information selection is considered in Naïve Bayes Algorithm while the SVMs applies kernel based methods for selection of information. The purpose of the Knowledge Utilisation and Text Classification Unit in the framework is to validate the hypothesis that the proposed MKTPKS based classification method improves the classification accuracies of the classifier when compared to simple term based matrix models for classification of textual data.

Thus in this part of the proposed methodology emphasis is placed on the classification of documents into their predefined categories or classes where different classification techniques are tested. The study of applications of these classifiers ranges from simple distance measures to probabilistic based distance measure to find the co-occurrences of terms within textual databases. The differences in the range of classifier applications considered will help to study the effects of classifying textual databases into two different classes possessing some good or bad information documents. An introduction to these classifiers with their applications to analyse the textual data is given in Chapter8.

6.4.3.1 Categorised Set of Documents: Good or Bad Information Documents

The overall output of the framework is to categorise free formatted textual documents into two different categories or classes in general and into good or bad information documents in the current research context. The text classification methods Decision Trees (J48 or C4.5), K-Nearest Neighbouring (K-NN), Naïve Bayes and Support Vector Machines (SVMs) methods used in this research were defined in the Weka 3.4 software, and this has been used to classify the documents into their predefined categories. The methods were tested on predefined categories of documents in order to test the accuracy of the classifiers. The results were then compared with the proposed MKTPKS based method of classifying documents to validate the hypothesis of acquiring better accuracies in terms of classification of documents using the proposed architecture or framework.

6.5 Summary of the Chapter and Conclusion

In this chapter a new architecture is proposed for discovering useful knowledge and then utilising it for the classification of textual data. The complete implementation of the methodology is divided into two parts and discussed in detail in the following Chapter 7 and Chapter 8.

Chapter 7 Knowledge Mining in Post Project Reviews (PPRs): Case Study Part 1

7.1 Introduction

In this chapter the implementation of the proposed framework for discovering useful knowledge in terms of multiple key term phrasal knowledge sequences (MKTPKS) is discussed in the scenario of Post Project Reviews (PPRs). The benefit of discovering FTS is that they can be used to identify some key issues discussed in the PPRs and for classifying these as good or bad information documents defined in the free formatted textual data. The results obtained in the form of MKTPKS are compared with the domain experts key term phrases to determine the effectiveness of the knowledge processing units of the proposed system.

7.2 Benefits Associated with Framework

The expected benefits of implementation of this framework are as follows;

- Pre-processing the information and then clustering will reduce the efforts required to decipher the results of text analysis.
- A new document or set of information can be handled easily by assigning it to the corresponding cluster.
- Application of Association Rule Mining will form the multiple key term phrasal knowledge sequences which are used to map the discovered knowledge to their appropriate category of textual documents.

7.3 Implementation of Different Functionalities of Methodology

The framework proposed in Section 6.2 has been implemented and is now demonstrated using a case study example of PPRs taken from the construction industry. Initially the decision variables are selected to start the process of analysing the textual data in the PPRs and there are single or multiple term sequences that were located with the help of domain experts. The decision variables consist of key words or phrases which are considered to represent the important areas of knowledge which might be covered during PPRs e.g. cost, time, planning and financial Issues etc.

Whilst it is believed that the framework and methodology reported in this thesis is applicable to many (or possible all) industrial uses of semi-structured data files, the experimentation and testing of the framework and methodology was mainly carried out using PPRs which were collected and analysed as part of the TrackStore Project (http://www.lboro.ac.uk/departments/cv/projects/trackstore/index.htm) as these contained appropriate raw data and domain expert analysis had been obtained during the TrackStore Project.

PPRs are routinely carried out at the end of construction projects. The business processes of the construction industry require efficient use of resources in terms of time, cost and planning and must achieve good customer satisfaction levels. Identifying information related to these issues and tracing the causes and effects of problems in previous projects can reduce the repetition of these issues and improve the chances of success in current and future projects (Choudhary et al. 2008).

7.3.1 PPRs for Business Intelligence and Information Description

The previous knowledge and experience of a project manager can affect the success of a project and satisfaction of the customer. However if a project manager's decisions are also based on the past practices or lessons learnt from reports produced by previous projects even better results may be achieved. PPRs are a useful form of information available in a construction industry environment, and have huge potential as sources of knowledge for workers on subsequent or similar projects. PPRs can also be considered as a necessary tool for knowledge management and a valuable source of shared knowledge across the boundaries of an enterprise (Tan 2006). These reviews help in learning collective lessons (Carrillo 2005) and the lessons learnt might then be used to prevent similar mistakes being made in the future (Pitman 1991). Thus discovery of useful information or valuable knowledge from these reviews will provide solutions to improve future business processes of an industry. PPRs are a collection of information coded with key phrases that are represented by either single or multiple terms therefore, these types of information need to be handled with special care to structure them. There are various Text Mining methods that can be used to decode this information. The combined use of data and text mining methods will

provide opportunities for useful lessons to be learnt from these reports, which can be beneficially used in the future.

The example data used in this research was available in the form of PPRs which were already divided into sixteen different headings as given below;

- General outcomes
- Estimating
- Planning
- Method of work
- Material procurement
- Subcontract procurement
- Mistakes or errors
- Innovations
- Quality assurance
- Waste/ Environmental Issues
- Health and Safety
- Interaction with Design Teams
- Interaction with Client
- H&S / O&M Manuals
- Snagging/ Defects
- General

These headings were further divided into sub-headings e.g. cost, time, prelims, subcontractors etc. The topic discussed in these reviews ranged from general outcomes in terms of cost and time, to general levels of satisfaction acquired during the whole project. The knowledge these reviews contain therefore covers different stages of interaction with design teams, clients, errors and mistakes and health & safety status etc. observed during the project. The topics or issues discussed in these reviews were identified with the help of domain experts. The defined key words or phrases or sentences refer to some particular topics discussed in the PPRs. Some examples of the topics and useful knowledge phrases identified in the sample PPRs are shown in the Table 7.1;

Table 7.1: Key topics or knowledge areas identified by domain experts in PPRs

Main/ Sub-headings	Key Phrases/ Knowledge Areas
Time	"work completed on time", "no issues with the programme", "causing additional delay", "time consuming" etc.
Safety	"No accidents", "reportable accidents", "problems with programme and safety" etc.
Financial Issues	"financial account is slightly less", "business unit target", "good margin", "considerably less than the estimated figure" etc.
Quality	"scope of works", "carried out necessary remedial work", "any specific problem, leaking, faults, errors, mistakes" etc.
Communication	"a good relationship", "would not communicate", "knowledge gained from the earlier work was not passed on" etc.

The purpose of this Case Study is to determine whether the proposed framework can do as well, or better than manual inspection, in identifying the key knowledge areas within a PPRs. If these key knowledge terms can be identified automatically or semi-automatically then knowledge captured within these terms can be passed more quickly and easily to other projects, so that workers may avoid the practices that may lead to bad results or can benefit by using good information based practices identified in previous projects. The identification of these key phrases can be done manually, but it is a very time consuming job. Also it becomes impractical to search manually through very large databases of potentially useful PPRs as the number of reports increases over time. This research aims to overcome this difficulty by fully automating or semi-automating the process of extracting information and converting it into a useful source of knowledge.

This study will focus on issues associated with "***time***" in the PPRs e.g. identification of key information that a project has been completed on time, or before time or late and the words (or terms) associated with time that could help to trace the reasons for delay and this is likely to provide useful knowledge for future projects.

A phrase like “causing an additional delay” is a time related example of a knowledge phrase defined in the text of the PPRs. It gives information about some issue that is delaying the project’s handing over to the client. Identification of such key phrases will help to uncover the hidden information in the text and their corresponding causes.

7.3.2 Implementation of Text Mining Module on PPRs

The application of the Text Mining Module as shown in figure 6.1 and described in Chapter 6 ensures that the data is made available in the form of free formatted text documents. This task is done by removing the headings or sub-headings from the PPRs data and storing them in a text file with different document IDs. This file is then passed to the information pre-processing unit where different activities are performed e.g. setting the aim or objectives of data analysis by selecting decision attributes or variables, removing redundant information in the form of stop words e.g. a, an , the, and also performing the process of simple stemming where the words are conflated to their original stem by removing the suffix ‘ing’, ‘ed’ and ‘ly’.

The next step towards analysis of the data by the text mining module is to structure the information into different representations using the information structuring unit. The text is represented using the bag of words (see section 6.3.2) approach in which documents are represented in the form of a term frequency (TF) matrix. This task was performed by writing java code to count the words and their corresponding frequencies. The output is then saved in a comma separated (csv) file to be used for further clustering techniques applications.

7.3.3 Implementation of 1st Level Knowledge Processing and Storing Unit on PPRs

In the current PPR example the analysis of data is made using the term frequency matrix which contains terms and their respective frequency counts within each document. Since the intention is to find the similarities between terms the representation is done showing terms along rows and documents along columns. The similarities are determined by calculating the Euclidean Distance using formula given in equation 7.1;

$$d(\mathbf{X}, \mathbf{Y}) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \qquad (7.1)$$

Where X= $x_1, x_2, ..., x_n$ and Y= $y_1, y_2, ..., y_n$ represent the n attribute values (i.e. terms with their corresponding frequencies) within each document whose dimension is equal to the number of documents taken at a time for textual data analysis e.g. $x_1 = T_1$ (1,0,0,0,…,1) and $x_2 = T_2$ (0,1,0,…,0). A two dimensional view of the distance calculated between two terms is shown in the figure 7.1;

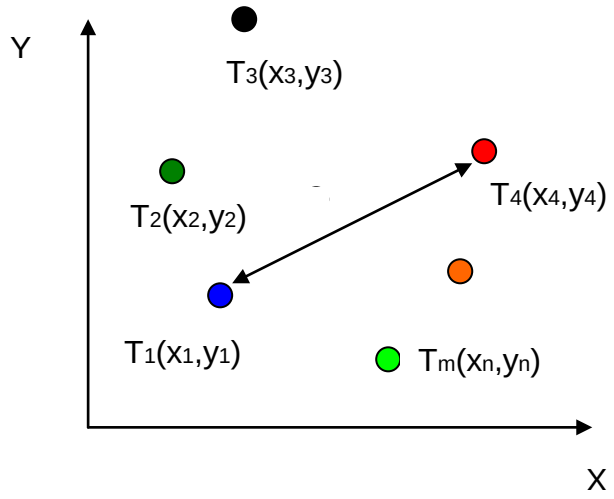

Figure 7.1: Two dimensional view of distance measure between terms

After selecting a suitable matrix representation for the term based information selection the data is ready for the application of some clustering techniques.

Weka 3.4 software was used to support the activities defined within the 1st level knowledge processing and storing unit before the application of 2nd Level Knowledge Refinement Unit which is used to form the MKTPKS. Weka 3.4 software is based upon a java environment which is open source and allows the user to customise or add new features to it. It offers different functionalities ranging from pre-processing to data classification and analysis. For further details and handling data or information see the reference (Witten and Frank 2000). There are variants of clustering algorithms available in the Weka 3.4 software but in the present work the k-means clustering

algorithm has been used due to its linearity (i.e. it has linear memory requirements for storing documents, cluster membership of the documents plus the cluster centroids). Clustering is also an essential part of the proposed methodology because it helps to divide the information space into multiple sub-spaces, each carrying useful information based upon single key term phrases. Thus Weka 3.4 can be used in both the information pre-processing unit of the Text Mining Module and in capturing first level of knowledge by identifying the relationships among key terms. In this work as a part of the 1st Level Knowledge Processing and Storing Unit, Weka 3.4 is used for the application of k-means clustering algorithm application. The algorithm was applied on the comma separated value (csv) data file which was obtained as a result from the Information Structuring Unit of the Text Mining Module. A large number of experiments were made to find the suitable number of clusters to capture information in terms of single key term phrases. These experiments showed that using a large number of clusters increases the risk of loosing key information. The number of clusters therefore for the current task was taken to be six. The selection of the number of clusters varies with the data size and in this research work for ten (10) different documents (or sets of information) with the processed number of terms ranging from (100-300) after applications of the different units defined in the text mining modules six clusters were found to be appropriate (which were selected as a result of extensive experimentation).

The set of information (or documents) taken from the sub-headings of 'time' , 'cost' and 'planning' were used to provide the sequence of terms to identify the good or bad information documents. So the process of generating clusters should be carefully handled and observed to overcome the difficulty of loosing useful information in terms of capturing single key term phrases within each cluster. The experiments were done by selecting the number of clusters to be between 2-10 and the K-means Clustering algorithm application proved to be most useful in retaining the useful information structures defined in the text. Thus K-means was used as an effective tool for capturing useful single term phrases within each cluster (i.e. CL1, CL2,…, CL6) and the number of clusters was retained as six to overcome the difficulty of losing useful information defined with the help of single term phrases within each cluster. Secondly since the data under consideration was also very sparse with low frequencies of occurrence it was difficult to capture useful information with the help

of single term phrases. Thus special care was taken while handling the type of data available in the PPRs, as in cases where data is sparse in nature and the intention is to discover useful knowledge sequences the selection of the number of clusters to be used needs to be handled carefully to overcome the difficulty of loosing useful information. The application of the K-means clustering algorithm helped to identify single term phrases within each cluster are shown for three clusters in the Table 7.2.

Table 7.2: Single Key Term Phrases Identified by Clustering Technique

Clusters ID's	Number of Instances clustered	Single Key Term Phrases Identified
CL1	11	"agreed", "complete", "customer", "job", "period", "suggest", "time", "twentyone", "week", "within", "work"
CL2	07	"actual", "contract", "eight", "extension", "fortyeight", "forty", "including"
CL3	10	"behind", "just", "handed", "one", "programme", "over", "ran", "take", "two", "under"
CL4	05	"certificate", "defect", "each", "end", "locate"
CL5	18	"allowed", "build", "fifty", "few", "instructions", "good", "issued", "KPI", "***A***", "prior", "project", "noted", "simply", "start", "variations", "year", "very", "give"
CL6	74	"additional", "all", "alternative", "arrange", "because", "before", "books", "both", "***B***", "cable", "cause", "claim", "concession", "cost", "crossing", "damage", "deliver", "demand", "diversion", "***C***", "due", "event", "existing", "extra", "fiftytwo", "finish", "five", "fortyfive", "forynine", "fortyseven", "four", "framework", "***D***", "get", "granted", "***E***", "head", "inclement", "large", "late", "liquidated", "made", "manager", "months", "morturay", "most", "need", "obtain", "office", "own", "paid", "paper", "***F***", "plan", "poor", "possible", "practical", "problem", "re-roofing", "responsibility", "seven", "significant", "site", "small", "still", "***G***", "thirteen", "three", "twelve", "twentysix",

		"understood", "unlikely", "weather"

The letters (i.e. ***A***-***G***) are used to represent some company names or information to be kept hidden.

The application of clustering is an essential part of the proposed framework and results shown in Table 7.2 give useful information. The information contained within each cluster is based on single key term phrases which refer to some key issue discussed in the PPRs. A difficulty arises however, when these results are presented to the user, because only a few meaningful structures can be presented. For example the information captured in CL1 comes from multiple documents and the human observer may try to *interpret* a cluster description like that of a single key term phrase "*time*" which may refer to the issue of "*delivered on time*" or "*completed on time*" or "*extension of time*" where these key multiple key terms phrases refer to three different issues discussed in PPRs documents. So it is difficult to map these key single term phrase based information or knowledge to identify some good or bad information documents. Similarly the key term phrase "*job*" may be used to define the concept of "*job finished late*", "*job took just under one year*" or "*job should be done within twenty one weeks*". So defining these structures within each clustered single term phrases is not an easy task. Similarly the terms defining the concepts in the documents and identified in the clusters CL2 and CL3 are not their own giving specific information from within a single document, and this can be confusing particularly when the data size is very large. Thus a difficulty arises in defining the concept on the basis of these single key term phrases and in identifying the exact document matching to these information. So if this model is used on its own, only as far as this stage the terms captured within each cluster may be of some importance in capturing first level of knowledge but interpretation of this would depend upon the concept being defined by the user. So some further technique is needed to help to restrict the information domain (i.e. good or bad information) identified by these single key term phrases.

In case of clustering each cluster containing with single term base information can be referred to different concepts at the same time so further information pruning techniques are required for which 2nd Level Knowledge Refinement Unit with the applications of Apriori Association Rule of Mining will perform this task. Before

passing information to the next stage of analysis, the identified key information is used to index the documents and store it in the form of a relational tables using a Relational Database Formation function. Then this key information is passed to other stages of the analysis to generate Multiple Key Term Phrasal Knowledge Sequences MKTPKS. These multiple key term phrasal knowledge sequences (MKTPKS) will ultimately serve the purpose of representing key information captured within each cluster. The Apriori Association Rule of Mining is used to search through each document and prune the key information. The process of refining the key information is discussed in the next section.

7.3.4 2nd Level Knowledge Refinement Unit

The term frequent itemset or termset mining comes from supermarket transactional databases where each frequent itemset is regarded as products which are most likely to be purchased together in one transaction. For example if a person buys "*milk*" then he might be interested in also buying "*egg*". Thus finding frequent itemsets in a transactional database serves the purpose of finding the items which appear most often together. So in this research context this algorithm is applied to find the terms which occur together in documents.

The simple Apriori Algorithm is explained in Section 5.3.2. The MKTPKS, corresponding to each cluster as shown in Table 7.2, formed through application of Apriori Association Rule of Mining technique on the case study data are shown in the Table 7.3.

Table 7.3 : Identification of Key Term Phrasal Knowledge Sequences

Clusters ID's	Multiple Key term phrasal Knowledge Sequences (MKTPKS)
CL1	{agree complete customer time week within work}
CL2	{contract eight extension forty fortyeight including}
CL3	{handed one over programme take}
CL4	{certificate defect each end locate}
CL5	{***A*** project start}
CL6	{all finish ***C*** few}

Knowledge discovered with the help of MKTPKS shown in Table 7.3 are used to map key knowledge discovered to the good or bad information documents about project. Since these multiple key term phrases are considered as a single unit of knowledge it is easy to find the document which is related to good or bad information documents within this research context. For example the MKTPKS discovered in cluster CL1 refer to a unique good information document about a project that a customer suggested that job should be done in some weeks time and it was completed on time. The frequent terms uniquely represent this set of information in cluster CL1 "{agree, complete, customer, time, week, within, work}" and these enable the relevant document to be found. Similarly the MKTPKS representing cluster CL2 uniquely represent the key information available as good information document where terms sequences showing that both the client and the company were agreed on the time required to finish some job. The preliminary results obtained from the experiments carried out at this stage of the framework are highly useful to identify the documents carrying key information using multiple key term phrasal knowledge sequences .

The discovery of MKTPKS is useful in two different ways for analysing textual data available in the form of PPRs;

- Firstly forming the sequences of terms with useful knowledge to compare with those identified by the domain experts in the PPRs.
- Secondly to map these knowledge sequences to identify the good or bad information documents and thereby classify the documents into two classes.

The results obtained in terms of MKTPKS are compared with those of domain experts and this evaluation will be given in the Section 7.3.5. However a secondary benefit can also be obtained beyond identification of these key term phrases and this is if these results are used to map information to some unique document which is then marked as good or bad information documents. In cases when the data size is small this benefit was also obtained. But there is a natural question of how to handle large databases since as the data size grows large then mapping key term knowledge sequences to unique documents and categorising these documents into two different classes would be an increasingly challenging task. Thus an unsupervised learning method of clustering would suffer problems. This challenge further motivated an

examination of supervised learning techniques of classifying the documents into two different classes of good or bad information documents about project. The details of these classification methods and implementation are discussed in Chapter8.

The discovery of MKTPKS is based on using varying levels of support (i.e. 10 - 20%), defined as, for any two terms A and B it is the percentage of transactions in the database that contain both terms. Different levels of forming multiple key term phrasal knowledge sequences are done through the process of forming a lattice of concepts as shown in figure 7.2 below;

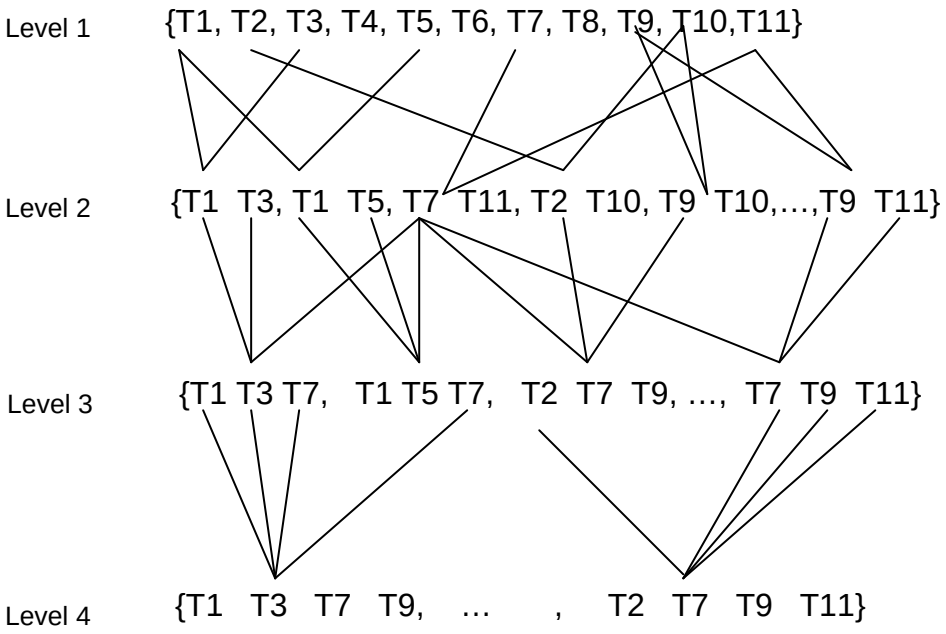

Fig 7.2: Levels of forming multiple key term phrasal knowledge sequences

The figure 7.2 shows that at the first level (or Level 1) MKTPKS 1-termsets are identified by searching through whole document space and then at second level (or level 2), the data is scanned again to form frequent 2-termsets from the discovered MKTPKS 1-termsets. This process continues until MKTPKS of fourth level (or level 4) are not obtained and the process stops. Consequently MKTPKS are formed as a lattice of concepts by searching through the whole data or information space.

7.3.5 Evaluation of the Proposed Systems

Evaluation of such methods for generating the summaries of the textual data and extraction of useful key term sequences of knowledge can generally be discussed in two ways (Jing et al. 1998). The first method is task based and called an extrinsic method while the second is an intrinsic method which is based on user judgements. In terms of evaluating the proposed system an intrinsic method of evaluation has been adopted because the objective of this work is to try to identify automatic text analysis methods which are as good or better at generating knowledge sequences than the results identified by the domain experts. The F-measure was used to measure the performance of the system for discovering knowledge in terms of multiple key term phrasal knowledge sequences. It is defined as the Harmonic Mean of Precision and Recall ratios. A recall measure is defined as the proportion of all the relevant multiple key term phrases that have been identified from the collection of multiple key term phrases whereas the Precision is the ratio of relevant multiple key term phrases to the multiple key term phrases identified. In Mathematical notations these terms are defined as;

$$\text{Recall (R)} = \frac{TP}{TP + FP} \tag{7.2}$$

$$\text{Precision (P)} = \frac{TP}{TP + FN} \tag{7.3}$$

$$\text{F-measure} = \frac{2 \times R \times P}{R + P} \tag{7.4}$$

Where TP (True Positive), TN (True Negative), FP (False Positive), and FN (False Negative) are defined as correctly identified, correctly rejected, incorrectly rejected and incorrectly identified MKTPKS respectively .

For example in the multiple key term phrasal knowledge sequences {the work was completed on time} identified by the domain expert, the stop words ‘the’ , ‘was’ and ‘on’ were removed during the pre-processing stage of handling textual data. So the whole sequence was taken as {work complete time} referring to key information

defined in the text. The multiple key term phrasal knowledge sequence identified by the proposed system is {agree complete customer time week within work}, which contains the key term phrases identified by the domain expert so the Precision and Recall values are calculated as defined above.

The results obtained with these efforts are recall and precision values as shown in the Table 7.4.

Table 7.4: Performance measure of the system

Measure	Accuracy of the Method
Recall	50%
Precision	29%
F-Measure	37%

Although the results obtained had low accuracy, they still indicate that the approach can be useful in finding the key good or bad information within the textual data. The use of clustering at first stage and then application of Apriori Association Rule of Mining at the second stage therefore proved to be useful in identification of useful information about project.

7.4 Results and Discussion

The MKTPKS identified in Table 7.3 refer to the key issue that both customer and contractors were agreed to finish some job within a time of twentyone weeks and it was done within a time. So this marks the identification of good information document on the basis of multiple key terms phrasal knowledge sequences with the application of 2nd Level Knowledge Refinement unit. Thus a document carrying key information can be retrieved on the basis of these multiple key term phrasal knowledge sequences. Secondly since the process of forming MKTPKS goes through the stages of finding information within each document set by generating MKTPKS 1-termset and then MTPKS 2-termset and so on, also helps to form a ***lattice of concepts*** by exploiting the property of monotonicity i.e. ***lattice structure*** of all 1-subsets of MKTPKS 2-sets are also frequent, all 2-subsets of MKTPKS 3-sets are also frequent etc shown in figure 7.2. Thirdly the system helped to identify the key

term phrasal knowledge sequences which had been marked as useful by the domain experts and it further helped even to mark key issues which were initially identified by the domain experts. However couple of merits and demerits of the proposed framework on the basis of experiments performed at 1st Level Knowledge Processing and Storing Unit and 2nd Level of Knowledge Refinement Unit are enumerated as follows;

1. Information available in the form of textual data is processed and useful knowledge in terms of single term phrases has been identified through applications of Clustering Technique which helped to identify the topics or issues discussed in the case study data.
2. The knowledge discovered at first level is further processed to refine it without involving user or human to interpret the key single term phrase based knowledge and map the discovered knowledge to unique to mark the good or bad information documents discussed in the case study data.
3. The refinement of knowledge through application of Apriori Association Rule of Mining technique at one end helped to refine the knowledge and identify key information sequence of knowledge but some loss of information has also been witnessed as in case of Cluster CL6. The reason behind this is the sparse nature of data under consideration available in the form of case study. However using minimum support value helped to overcome this difficulty and useful sequences of knowledge were discovered.
4. The strong benefit associated with the experimental work and knowledge discovered was that it greatly helped to reduce the human efforts to interpret knowledge available in the form of single key term phrases obtained through applications of Clustering technique.

7.5 Novelty in the Research Work

This research contributes to the hybridised application of textual data mining techniques to find the multiple key term phrasal knowledge sequences. The example presented in this chapter showed how the proposed framework helps to identify potentially useful multiple key term phrases to mark some key issues discussed in the PPRs and their causes so that this knowledge can be used to help in the decision making process for future projects. ***The originality of work is based on the original***

integration of textual data mining techniques for generating first level of knowledge in terms of single key term phrases through the application of clustering rather than on the application of these algorithms individually. Additionally these single key term phrases are then processed by the applications of 2nd Level Knowledge Refinement unit (Apriori Association Rule of Mining Applications) for generating multiple key term phrasal knowledge sequences (MKTPKS). This integration serves the purpose of achieving benefits in three main dimensions. Firstly it helps to identify the key issues discussed in the PPRs by finding multiple key term phrasal knowledge sequences in the real dataset collected from an industrial setup. Secondly it divides the whole information space into a subspace where clustering techniques are applied to group and discriminate larger text into different subsets of information e.g. either good or bad information documents in this research context. Thirdly restricting a large document space into subspaces is meant to isolate the one cluster from another on the basis of key terms uniquely identified within each cluster for further analysis of text documents.

7.6 Summary of the Chapter and Conclusion

In this chapter the detailed application of the proposed methodology in terms of pre-processing the free formatted textual data and discovering the useful knowledge in terms of multiple key term phrasal knowledge sequences has been discussed in detail. The applications of different functions described in the methodology helped to discover useful knowledge and provide a basis of this knowledge for future activities of classifying textual databases. Discovery of the MKTPKS forms the basis of classifying textual data into two different categories whose implementation is further discussed in the following Chapter8 and Chapter 9.

Chapter8 Text Classification Methods Implementation and Performance Measure: Case Study Part II

8.1 Introduction

In the previous chapter the main parts of the knowledge discovery process from textual data have been described. The important part of textual data handling, which is the pre-processing of data i.e. to make it suitable for analysis, has also been discussed in detail. This chapter examines ways of making the identified knowledge easier to use by providing methods and techniques that can be used for classifying the textual data into different classes. The major work in terms of information classification considered here is to divide the information into two main classes which identify good or bad information documents that have been recorded within the PPRs. Different data mining techniques for performing the classification task are considered in this chapter and the main focus is on the applications of Decision Trees (C4.5), K- nearest neighbouring (K-NN) and Naïve Bayes Algorithms. The performance of these classifiers are tested on data sets collected from construction industrial environments available in the form of Post Project Reviews (PPRs). The accuracies of the classifiers are compared using a simple term based representation and a frequent term based representation method where the dimension of the feature space is reduced. In this chapter the proposed model is implemented as defined in Section 6.4.

8.2 Text Classification

8.2.1 Background Knowledge

Text classification methods were first proposed in 1950 where the word frequency was used to classify documents automatically. In 1960 the first paper was published on automatic text classification and until early 90's it was a major sub-field of information systems. In past the text data was available in the form of paper tape which was read with the help of expensive computers with limited memory which limited the use of this technology. Currently in the age of information technology the amount of information or text data available in digital form which gave real impetus to the applications of text mining technology. The availability of enormous amount of

data in digital format has renewed the interest in the area of automated text classification and data mining techniques applications. Applications of machine learning techniques have reduced the amount of manual effort required and have improved the accuracy of the results. There are many text miner software products on the market which can be used to perform different tasks of handling textual databases and classifying the text to discover useful information (Tan 1999). Considerable research work has been done in defining new algorithms for handling textual based information and performing the task of text classification such as K-nearest neighbouring (KNN) algorithm, Bayesian classifier based algorithms, Neural Networks, Support Vector Machines (SVMs), Decision Trees Algorithms etc.(Yong et al. 2009).

8.2.2 Problem Description and Objective of Research

Text analysis and classification systems help to identify the key issues discussed in textual databases and play an effective role in future decision making processes. The information specific to some product or service quality issues is normally available in the form of numerical or textual data formats including perhaps colour coding, abbreviations and special text features. Common requirements for manufacturing organisations include better project management, reducing product lead time to the market and improving customer satisfaction levels or service quality. Decisions must be cost effective and efficient to meet the current and future requirements, and it is important therefore to identify t good or bad information documents which may exist in company reports and other documents.

This research therefore addresses the challenges of identifying such knowledge automatically as shown in Chapter 7 and continued in this Chapter, which addresses the problem of automatically classifying the identified knowledge (in its related documents) as good or bad information documents. Data is commonly classified into two different categories, and this is termed as binary classification problem. Some manual efforts are employed to perform this task, by creating a training set to use with different data mining algorithms or classifiers.

The hypothesis made for this experimental work is that , ***"The generation and use of multiple key term phrasal knowledge sequences in classifying the documents will provide better classification accuracy than single term based classification methods."***

Therefore the objectives of this Chapter are to show:

1. Application of textual data mining techniques for capturing first level of knowledge in terms of single key term phrases (as previously described in Chapter 6/ 7).
2. Generation of multiple key term phrasal knowledge sequences termed as MKTPKS to represent key knowledge discovered through clustering applications (as previously described in Chapter 6/7).
3. Studying the effects of single term based representations and multiple key term phrasal knowledge sequences in classification of textual documents. This will validate the hypothesis that multiple key term phrasal knowledge sequences based classification gives better accuracy than the single term based method.

8.3 Textual Data Mining for Information processing and Knowledge Discovery

The first step towards handling textual databases or classifying their text into classes starts by going through the whole text based information available in free formatted text documents (i.e. the PPRs reports in this case). The information is first read through by domain experts who manually identify the good or bad information documents. The results from this manual process will eventually tested and compared with the results of the automatic classifiers (to validate the experiments). The textual data is then structured (e.g. using stop word removal and stemming) into a format suitable for the application of different data mining algorithms as detailed in section 6.3.1.

8.4 Text Classification Methods

In this Chapter different classifiers are compared , in particular Decision Trees (C4.5), K-NN , Naïve Bayes and Support Vector Machines (SVMs) algorithms. The reason

for examining these different types of classifiers is that they use different selection criteria on the information variables as these range from simple distance measures to probabilistic based distance measures to find the similarities between documents and classify them into their categories. The purpose of these experiments is to validate the hypothesis that the proposed method based on multiple key term phrasal knowledge sequences provides better accuracy than the simple term based data classification method. Decision Trees, Naïve Bayes, K-NN and SVMs algorithms are tested and their accuracies are measured in terms of classifying data into two different categories or classes. A brief introduction to these classifiers is given in the following paragraphs.

8.4.1 Decision Trees (C4.5) Algorithm

Decision trees analysis algorithms are most useful for classification problems and the process of building a decision tree starts with the selection of a decision node and splitting it into its sub nodes or leafs. A decision tree algorithm C 4.5 is an extension of Quinlan's algorithm ID3 which generates decision trees (Quinlan 1992) by splitting each decision node to select an optimal split and continues its search until no further split is possible. It uses the concept of *information gain or entropy reduction* to select the optimal split. Different steps of the algorithm for forming a decision tree are defined below;

Step1:

Suppose a variable ***X*** for which k possible values have probabilities p1, p2, p3, …,pk. Then *entropy* of ***X*** is defined as;

$$H(\boldsymbol{X}) = -\sum_{j} p_j \log_2\left(p_j\right) \qquad (8.1)$$

Step2:

The means information requirement can then be calculated as the weighted sum of the entropies for the individual subsets, as follows:

$$H_S(T) = \sum_{i=1}^{k} P_i H_S(T_i) \qquad (8.2)$$

Where P_i represents the proportion of records in subset i.

Step3:

Information gain is defined as, the increase in information produced by partitioning the training data *T* according to this candidate split *S*, given by;

Information gain IG(S) = $H(T) - H_S(T)$ (8.3)

The selection of optimal split at each decision node is based on the greatest information gain, **IG(S)**.

Some of the advantages and disadvantages associated with the applications of decision tree algorithm are enumerated as follows;

Advantages

- The rules generated by the application of a decision trees algorithm are easily interpretable and understandable as they are presented in the form of if-then rules.
- The series of these algorithms (i.e. ID3, C4.5) can handle large databases and the formation of the decision tree's root and branch nodes are independent of the size of the database.
- The decision node is formed by considering each attribute defined in the database and the corresponding tree is formed on the basis of this information selection criteria where the time taken is proportional to the height of the tree.

Disadvantages

- Firstly the branches formed using a decision tree algorithm can be so large that it becomes difficult to interpret the rules.
- Secondly it is not easy to handle continuous data using a decision tree as the data needs to be divided into categorical attributes.

8.4.2 K-NN Algorithm

K nearest neighbouring (KNN) algorithm is a technique that can be used to classify data using distance measures. The K nearest neighbouring algorithm works by learning through the training samples where the entire training set includes not only the data in the set but also the desired classification for each item. In effect the training data becomes the model. The K- nearest neighbouring algorithm works on the principle of finding the minimum distance from the new or incoming instance to the training samples (Han and Kamber 2001). On the basis of finding the minimum distance only the K closest entries in the training set are considered and the new item is placed into the class which contains K closest items. The distance between the new or incoming item to the existing one is calculated by using some distance measure. The most common distance function is the *Euclidean distance* given in ***equation 8.4***.

$$D(X,Y) = \sqrt{\sum_{i=1}^{m}(x_i - y_i)^2} \qquad (8.4)$$

Where $X = (x_1, x_2, ..., x_m)$ and $Y = (y_1, y_2, ..., y_m)$. In case of classifying documents into their respective categories if term based representation is used then corresponding attributes are given as $x_1 = D1$ (1,2,0,0,…,1) and $x_2 = D2$ (0,1,1,0,…,0) where the numeric values shows the frequency of occurrence of each term in documents.

A two dimensional view of classifying a new document is shown in the figure 8.1 where the distance between the entries or terms of one document to the other is calculated using the distance measure given in equation 8.4. For K=2, the new document Dμ is placed in the class for which the distance is minimum as compared to the other class of document. Since the distance between Dμ and D3 is less than the D1 so the new document Dμ will be placed in the class for which D3 lies.

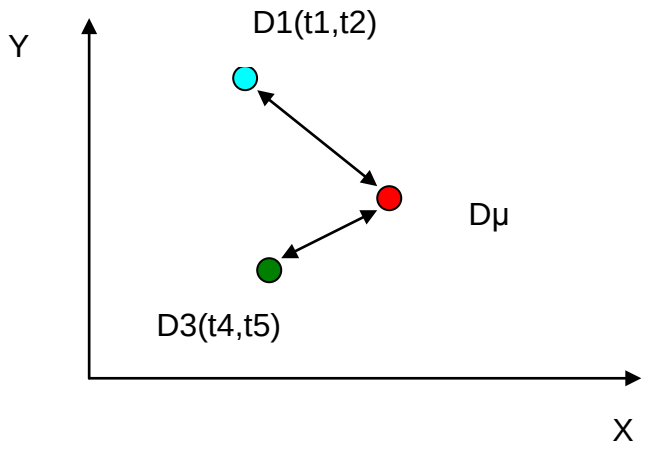

Figure: 8.1 Two dimensional view of document classification using K-NN

So the overall algorithm has two main steps of finding the nearest neighbouring class for a document with unknown class variable.

Step1:

Selecting or deciding the number of nearest neighbours i.e. the value of K

Step 2:

Measuring the distance between the new or incoming instance to the training sample.

The main dis-advantage associated with implementation of the algorithm is to find the best number of values of nearest neighbours to use i.e. to find the appropriate value of K to make the best decision for classifying the documents. Changing the value of K will affect the classification accuracy of the classifier. In the case of application of this algorithm on the case study data the better classification accuracies were found for selecting K =10, but this would be different for other data. The best value of 'K' to use varies with the size of data under consideration and is highly dependant on the quality of the data.

8.4.3 Naive Bayes Algorithm

A Naïve Bayes algorithm is simple and a well known classifier which is used in solving practical domain problems. The Naïve Bayes classifiers are used to find the joint probabilities of words and classes with a given set of records (Witten and Frank 2000). This approach is based on the Naïve Bayes Theorem. In the context of text classification the different steps of the Algorithm are defined as follows;

Step1:

For a binary classification problem (i.e. two class variables of A and B) the probability of a class c, is given by $p(c)$, known as the prior probabilities.

Step2

Using the prior probabilistic information the probabilities of a new incoming instance or document d_j i.e. $p(d_j / c)$ are calculated.

Step3:

Sum of the probabilities or likelihood of new document i.e. $p(d_j)$ are then calculated.

Step4:

Finally the actual probabilities or posterior of the new document i.e. $p(c / d_j)$ for a given document d_j is calculated by using the Bayes Theorem shown below;

$$P(c / d_j) = \frac{p(d_j / c)\, p(c)}{p(d_j)} \qquad (8.5)$$

As P(c) is constant for all classes, only $P(c / d_j)\, p(d_j)$, where j=1,2,3,…, m, need to be maximized. If the class prior probabilities are not known, then it is commonly assumed that the classes are equally likely, that is:

$$P(d_1) = P(d_2) = ... = P(d_m) \qquad (8.6)$$

The prior probabilities of the class may be estimated by

$$P(d_j) = T_j / T \qquad (8.7)$$

where T_j is the number of training samples of class c and T is the total number of training samples.

It is assumed that classes are independent of each other which is called the Naïve assumption of class conditional independence and it is made while evaluating the classifier. Naïve Bayes classifier performs well on data where the problem is to categorise the incoming object into its appropriate class and classification task is carried out on the basis of prior information and likelihood of the incoming information to form a posterior probability model of classification.

Some of the advantages and disadvantages associated with the application of the Naïve Bayes algorithm are given as follows;

Advantages

- The algorithms are easy to use and scan through the whole database once to classify it into the respective categories.
- The probabilities calculated at each attribute provide a means to overcome the difficulty of handling missing values in the data as the probabilities calculated are omitted during analysis.

Disadvantages

- The algorithm does not handle continuous data as the data is divided into their ranges to solve the problem. Dividing continuous data into different ranges is difficult and will ultimately affect the results.

8.4.4 Support Vector Machines (SVMs)

The Support Vector Machine was first developed in Russia in the 1960s by (Vapnik and Lerner 1963; Vapnik and Chernonenkis 1964). This is a non linear classification algorithm which uses kernel methods to map data from an input space or parametric space into a higher dimensional feature space. The non linear boundaries in

parametric space may become linear in the feature space. A Support Vector Machine can be used as a tool for text classification and the design of SVM classifiers is now discussed in the following sections.

8.4.4.1 Benefits Associated with SVMs Applications

There are many existing data mining techniques i.e. Decision Trees, K-means clustering, Association Rules of Analysis, Neural Nets etc, but each of these has its merits and demerits in terms of their application. It is therefore useful to also consider the application of some other new techniques or combinations of these techniques. SVMs show promising results according to reports of their application, they have been compared to neural networks and Nonlinear Regression. Their benefits include (Cho et al. 2005);

- Good Generalization Performance: Given a training set Support Vector Machines are able to learn rules to often correctly classify a new object.
- Computational Efficiency: SVMs are efficient in terms of speed and complexity involved in real world data.
- Robust in High Dimension: It is difficult for learning algorithms to deal with high dimensional data because of the over-fitting problem, SVMs are more robust to over-fitting than other algorithms.

Therefore SVMs methods have potential to give good results. These methods potentially could be very useful in manufacturing or construction industry where data is more complex and diverse than in other applications areas such as finance, customer call centre, retails etc.

8.4.4.2 Constructing SVM

Support Vector Machines (SVMs) were developed for two types of classification problems one is binary classification and the other is multi class classification. Since the focus of study in the current research is to consider the binary classification problem this section only discusses the binary classification problem.

Consider a binary classification problem where ω1 and ω2 are two classes of a training data set given as X= $\{x_1, x_2, ..., x_n\}$ having class labels y_i = {-1,+1}.That is the data set is labelled by the rule that if $x_i \varepsilon w_1$ then y_i =+1 otherwise y_i =-1 for $x_i \varepsilon w_2$.The basic idea of SVM estimation is to project input observation vectors non linearly into high dimensional feature space and then compute a non linear function in that space. The hyperplane which is separated by two classes is described by;

$$f(x) = \text{sgn}(\langle w.x \rangle + b) \qquad (8.8)$$

where w is the coefficient vector and b is the bias of the hyper plane and sgn[.] stands for the bipolar sign function. The objective of the SVM algorithm is to choose the optimal separating hyperplane that maximizes the margin between two classes (Vapnik 1995). The figure 8.2 below shows the classification of documents into two different classes based on SVMs ;

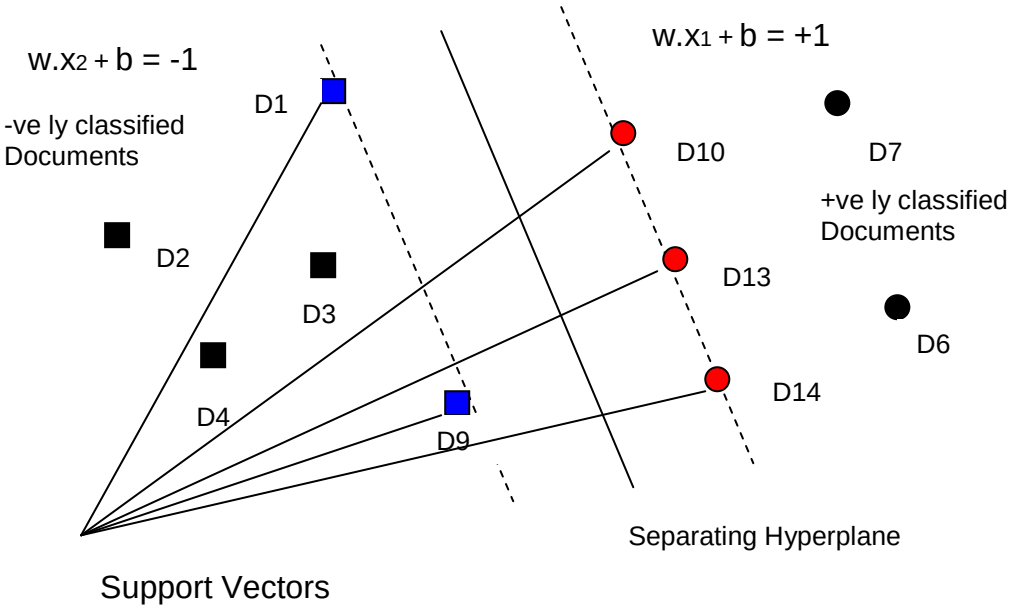

Figure 8.2: SVMs based binary Classification of Documents

The hyperplane that has the maximum distance to the closest point is called the optimal separating hyper plane. The distance from the hyperplane to the closest point is 1/ ω and 2/ ω is called the margin between the hyperplanes. This margin provides the measure of the generalization ability of the hyperplanes to separate the data into corresponding classes. The larger the margin the better the generalization abilities are expected to be (Christiniani and Shawe-Taylor 2000). The optimization problem that yields the hyperplane which can be written as ;

$$\text{minimize}_{w,b} \frac{1}{2}\|w\|^2 \qquad (8.9)$$

Subject to

$$y_i(\langle w.x_i \rangle + b) \geq 1 \; ; \quad \text{for i=1,2,…,N} \qquad (8.10)$$

8.4.4.3 Kernel Induced Feature Space

In the context of machine learning the kernel trick was first introduced by (Aizermann et al. 1964). Kernel trick involves changing the representation of the data, i.e.

$$x = (x_1,...,x_n) \rightarrow \phi(x) = (\phi_1(x),...,\phi_N(x)) \qquad (8.11)$$

where $\phi(.)$ is a non-linear operator mapping from input space ***X*** to feature space ***F*** i.e. $\phi : X \rightarrow F$.

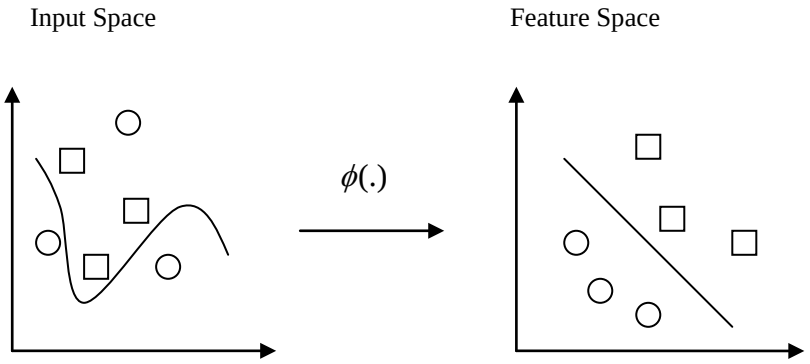

Figure 8.3: Mapping from Input Space to Feature Space

The mapping can greatly simplify the learning task (Scholkopf and Smola 2001). The above Figure 8.3 shows the mapping of data from an input space to a higher dimensional feature space by a non-linear operator $\phi(.)$, in order to classify the data by a linear boundary. However, a problem associated with high dimensional feature spaces is that as the number of features grow, the generalization performance can degrade and the solution can become computationally expensive. This phenomenon is known as the curse of dimensionality (Christiniani and Shawe-Taylor 2000). Although, dimensionality reduction can be performed by removing the features corresponding to low variance in the data, there is no guarantee that these features are

not essential for learning. Support vector machines are inherently equipped with a linear combination of the dot product between the data points that turn out to be support vectors and hence can defy the curse of dimensionality by using the dot product kernel functions.

The dot product kernel function is defined as;

$$k(x, x') = \langle \phi(x), \phi(x') \rangle \quad (8.12)$$

The kernel matrix or Gram Matrix (Gunn 1997; Christiniani and Shawe-Taylor 2000) in the form of an inner product space between documents is shown in Table 8.1 given below;

Table 8.1: Gram Matrix for building Support Vector Machine Classifier

Gram Matrix (K) =

K(D1, D1)	K(D1,D2)	K(D1,D3)	…	K(D1,Dm)
K(D2,D1)	K(D2, D2)	K(D2, D3)	…	K(D2, Dm)
K(D3, D1)	K(D3, D2)	K(D3, D3)	…	K(D3, Dm)
…	…	…	…	…
K(Dm, D1)	K(Dm, D2)	K(Dm, D3)	…	K(Dm, Dm)

The matrix should satisfy the condition given in equation (8.13) known as Mercer Condition;

$$\int_{X^2} k(x, x') f(x) f(x') dx dx' \geq 0; \forall f \in L_2(x) \quad (8.13)$$

According to Mercer theorem (Mercer 1909), the kernel $k(x, x')$ is any continuous and symmetric function that satisfies the condition of positive semi-definiteness given by (8.13).Such a function defines a dot product in the feature space given by .

$$f(x) = \sum_{i=1}^{N} (\alpha_i^* - \alpha_i)\langle x_i . x\rangle + b \qquad (8.14)$$

Linear SVM can be readily extended to non-linear SVM by using (8.12) and (8.14) can be written as;

$$f(x) = \sum_{i=1}^{N} (\alpha_i^* - \alpha_i)\langle \phi(x_i).\phi(x)\rangle + b$$

$$= \sum_{i=1}^{N} (\alpha_i^* - \alpha_i) k(x_i, x) + b \qquad (8.15)$$

The equation (8.15) shows a basic formulation of the problem for finding the support vectors through optimising the function $f(x)$ where the input information space is transformed into higher dimensional feature space by introducing kernel function. So there is no need to feature map its properties explicitly. However, the knowledge associated with this feature map and its properties can provide some additional insight about the support vector kernels and might be helpful in answering the question why this mapping usually provides good results (Vapnik 1998).

8.4.5 Illustrative Example for Information Handling

Several different techniques are considered in this chapter, as each uses a different measure (i.e. Euclidean Distance, entropy measure, kernel functions and probabilistic methods) to classify textual data into two different classes. The different criteria of information selection and how they are used to classify documents are demonstrated with the help of an illustrative example. The Table 8.2 shows part of the experimental set of information (i.e. six documents) which will be classified as either positive or negative and for each document the corresponding terms frequencies are given with a

class attribute shown as A or B representing positive and negative (good or bad) information. The data set shown in Table 8.2 only gives details of five terms (T1,…, T5) whereas these documents actually contains approximately hundreds of terms.

Table 8.2: Table showing the documents and corresponding term frequencies

DocsId	**T_1**	**T_2**	**T_3**	**T_4**	**T_5**	**Class**
D1	1	0	2	3	1	**A**
D2	0	1	1	0	2	**B**
D3	1	0	0	1	1	**B**
D4	0	0	1	2	1	**A**
D5	1	1	0	0	0	**A**
D6	2	0	0	1	0	**A**

The information handling in case of Decision Trees (C 4.5) algorithm, K-NN, Naïve Bayes Rule and Kernel Based matrix representation in SVMs in higher dimensional spaces are detailed as follows;

8.4.5.1 Decision Tree (C4.5)

The selection of information is based on entropy is given by the following criteria given in equation 8.1;

Step1:

$$H(T1) = - \sum_j p_j \log_2(p_j)$$

$$= -(\frac{4}{6})\log_2(\frac{4}{6}) - (\frac{2}{6})\log_2(\frac{2}{6}) = 0.91804$$

The information gain for choosing the term T1 as splitting node is calculated in Weka(3.4) as frequency of occurrence of terms shown in the Table 8.2, where the frequency ≤ 0 or frequency > 0. The information gain is calculated at each term T1, T2, T3 and so on and finally the algorithm will split tree on the term with maximum information gain.

In this example the information gain for term T1 defined by equation 8.3 is calculated as given below;

Step2:

For frequency ≤ 0

Entropy value is given as;

$$-\frac{1}{2}\log_2(\frac{1}{2})-\frac{1}{2}\log_2(\frac{1}{2})=1$$

For frequency > 0

The entropy value is given by;

$$-\frac{3}{4}\log_2(\frac{3}{4})-\frac{1}{4}\log_2(\frac{1}{4})=0.81145$$

The entropies are combined for both frequency ≤ 0 and frequency > 0 as;

$$\text{Hs (T1)} = \frac{2}{6}(1)+\frac{4}{6}(0.81145)=0.87430$$

Step3:

InfGain(T1) = H(T1) – Hs(T1) = 0.044

The formation of decision tree resulting in classification of documents into two classes (i.e. A and B) is shown in figure (8.6) and figure (8.7) and is discussed in detail in the same section.

8.4.5.2 K-Nearest Neighbouring Algorithm

Considering the same exemplary data given in above Table 8.2 the selection of nearest neighbour is done by following two step procedure i.e. selecting K and measuring the similarities among documents by calculating Euclidean distance measure given in equation 8.4. If the value of K is taken as K = 2 then classification of new document D7 is done based on training examples.

For example if K = 2 then documents D3 and D5 are considered as training examples based on which the decision is made for the new document D7(1,0,0,1,0) with unknown class variable.

The distances are given as under;

$$\text{Dist (D3,D7)} = \sqrt{(1-1)^2+(0-0)^2+(0-0)^2+(1-1)^2+(0-1)^2} = 1.0$$

Dist (D5,D7) = $\sqrt{(1-1)^2 + (0-1)^2 + (0-0)^2 + (1-0)^2 + (0-0)^2}$ = 1.41

So the new document D7 will be classified as Class B as the distance measure from document D3 is smaller than from the document D5 as shown in the figure 8.4 below;

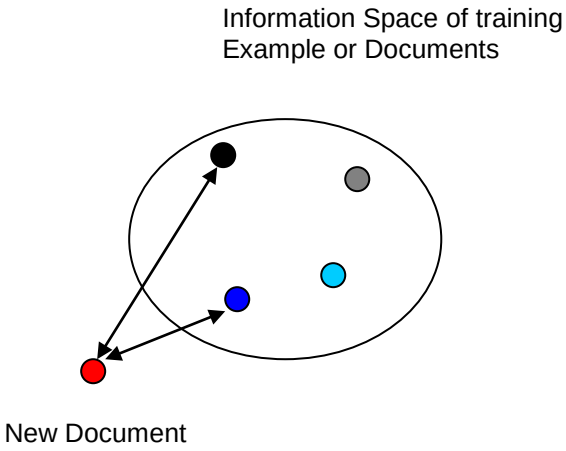

Figure 8.4: Classification of new document based on K-NN criteria

8.4.5.3 Naïve Bayes Algorithm

The information selection and classification of new document D7 by following the steps defined in the Naïve Bayes algorithms using the Table 8.2 are detailed as follows;

Step1:

If the term T1 is taken as the term for selecting information to classify the new document, then frequency ranges are described as [0,1],]1,2],]2,3] where both open interval (i.e.]]) and closed interval (i.e. [ ]) are used to define these ranges. The prior probabilities are calculated with class variables as A and B.

P(A) = (4/6)= 0.667

P(B) = (2/6)= 0.333

Step2:

Table 8.3: Probability calculated for two class variables A and B

Attribute	Value	Count A	Count B	Probabilities for variable A	Probabilities for variable B
Documents	Frequency	4	2	4/6 = 0.667	2/6 = 0.333
	[0,1]	3	2	3/4= 0.75	2/2= 1
	]1,2]	1	0	1/4= 0.25	0/2= 0
	]2,3]	0	0	0/4= 0	0/2= 0

The table above is used to find the probabilities of new document D7 taking T1 as a term of selecting information. The corresponding probabilities are given as under;

P(D7/ A) = 0.667x 0.75 = 0.50025

P(D7/B) = 1x 0.333 = 0.333

Step3:

Likelihood of Document D7 of being in Class A = 0.667x 0.50025 = 0.334

Likelihood of Document D7 of being in Class B = 0.333 x 0.333= 0.111

Sum of Likelihood for Document D7 = P(D7) = 0.334 + 0.111= 0.445

Step4:

Then posterior or actual probabilities for new Document D7 are calculated using equation 8.5 ;

P(A/D7)= 0.334/0.445 = 0.751

P(B/D7) = 0.111/0.445 = 0.249

So the new document will be termed as belonging to class A i.e. good information documents as the posterior value is higher than the class B value.

8.4.5.4 Support Vector Machines (SVMs)

The SVM methods use different types of kernel tricks for information selection. In the current research linear kernel methods have been tested for which information is structured by finding the inner product of each data point. Then a matrix representation is formed, called a Kernel Matrix which is then used to find the support vectors by performing SVM optimization task. The data given in above Table 8.2 is used here to form the corresponding Kernel Matrix which is the core of SVM classifiers optimisation procedure to find the support vectors and classifying the textual data into two different categories or classes. The linear Kernel matrix is used to find the inner product based distance representation. For example for document D1 in Table 8.2 the inner product value is calculated as;

<D1.D1> = <1,0,2,3,1> . <1,0,2,3,1>

= (1.1) + (0.0) + (2.2) + (3.3) + (1.1)

=1 + 4 +9 +1 = 15

The detailed representation of the matrix formed in given in the Table 8.4;

Table 8.4 : Gram Matrix based on Linear Kernel i.e. K(D, D')= <D, D'>

K(D1,D1)=15	K(D1,D2)=4	K(D1,D3)=5	K(D1,D4)=9	K(D1,D5)=1	K(D1,D6)=5
K(D2,D1)=4	K(D2,D2)=6	K(D2,D3)=2	K(D2,D4)=3	K(D2,D5)=1	K(D2,D6)=0
K(D3,D1)=5	K(D3,D2)=2	K(D3,D3)=3	K(D3,D4)=3	K(D3,D5)=1	K(D3,D6)=3
K(D4,D1)=9	K(D4,D2)=3	K(D4,D3)=3	K(D4,D4)=6	K(D4,D5)=0	K(D4,D6)=2
K(D5,D1)=1	K(D5,D2)=1	K(D5,D3)=1	K(D5,D4)=0	K(D5,D5)=2	K(D5,D6)=2
K(D6,D1)=5	K(D6,D2)=0	K(D6,D3)=3	K(D6,D4)=2	K(D6,D5)=2	K(D6,D6)=5

8.5 PPR Data for Classification

8.5.1 PPRs as Defining Good and Bad Information Documents

A sample data set consisting of post project reviews from the TrackStore Project (://www.lboro.ac.uk/departments/cv/projects/trackstore/index.htm) was used in these experiments. The data is in the form of free formatted text that has been composed of about 10-15 pages further divided into some main and sub headings of time, cost, planning, etc. The parts of the data set which were considered in the current analysis

were taken from the sub headings of '*Time*' and '*Cost*'. All the information available in these free formatted textual data or documents were then divided into two different classes of good or bad information documents. This task was performed by reading through every document with the help of domain experts. The meanings of all the text under the '*Time*' and '*Cost*' subheadings were well understood and then any identified knowledge was assigned to one of two different categories. The purpose of this exercise was to create a training set of data to test the proposed methodology for automatically classifying text into two different categories. The class attribute were assigned to each category of information as 'A' and 'B' for good and bad information documents after careful reading through the text.

An example of good information documents identified with the help of domain experts or knowledge worker is given as under;

"*The project took 51 weeks and was handed over on programme-giving good KPI for time. It was noted that the customer issued very few instrtuctions…***"**

Similarly an example of bad information document is given as under;

"The project was programmed to take 49 weeks, but finished four weeks late. Most of extra work was caused by the work starting late because…**"**

8.5.2 MKTPKS Based Matrix Model For PPRs Classification

This section discusses the method of preparing PPRs data for performing the automatic text classification task to classify it into two different categories using Decision trees (J48 or C4.5), K-NN, Naïve Bayes and SVMs. The data must first be transformed into a suitable format to test the algorithms. A candidate feature space is obtained through application of hybridised efforts of 1st Level Knowledge Processing and storing Unit and 2nd Level Knowledge Refinement Unit applications (as discussed in Chapters 5/6) . Single key term phrases are identified using a clustering technique and then the clustered instances are further processed to generate multiple key terms phrasal knowledge sequences . Consequently a feature space is prepared and used to represent the existence or non existence of key phrases in the documents. To adapt

classification procedures for this research work, each vector representation of documents is done through MKTPKS. The relationship between the list of key phrases and their representative classes is shown in the figure 8.5 using binary representation method is shown below;

List of key term phrases

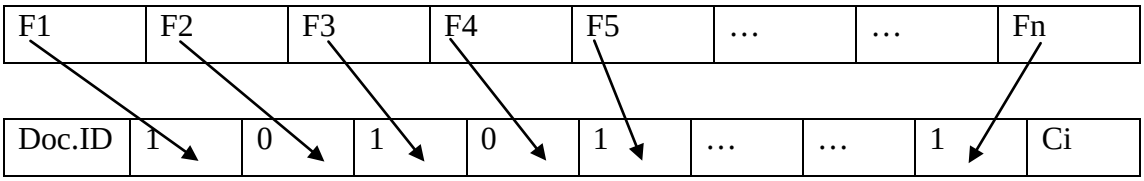

Figure. 8.5 Candidate termset representation for MKTPKS

Where Ci represents the class of labels given to the training data and taken as C1, C2, C3,…, Cn and F1,F2,…, Fn represent the corresponding frequent termset sequence (FTS). A matrix of representative key term phrases and their class labels is therefore formed and the (Decision tree C4.5, K-NN etc.) as data mining algorithm are used to classify data into their predefined categories. Thus matrix representation formed for the textual data is given in the Table 8.5.

Table 8.5: Matrix Representation of the textual data using MKTPKS

Docs IDs	F1	F2	F3	…	Fn	Class
D1	1	1	0	0	1	A
D2	0	1	0	1	0	B
D3	0	0	1	1	1	A
D4	1	1	0	0	1	A
…	…	…	…	…	…	…
Dm	0	1	0	0	1	B

Thus the whole information space is then transformed into a matrix representation with MKTPKS in column and documents in the rows where binary representation method is used to generate this information space. This new matrix model is then used to perform the classification task with class variables A and B representing good and bad information documents in the current research context. The representative terms

of F1, F2, …, Fn are used for frequent 3-termset sequences and Weka (3.4) was used to test the classification accuracies of classifiers defined under '*classify*' (drop down option available in Weka (3.4) software).

8.6 Applications of Methodology on PPRs for Text Classification and Results

8.6.1 Text Mining Module Application

This section discusses the application of the Text Mining Module for data pre-processing and structuring. In this part of the analysis the data is structured through the application of the different units detailed in the section 6.3. This step makes the data ready for applications of different data mining algorithms e.g. k-means clustering algorithm at first stage in the current research scenario. So the data is prepared by first consolidating it in a text file and then converting it into a suitable representation. Java code was written to count the terms and their corresponding frequencies for representation in the term frequency matrix (TF). Stop words are removed from the text and simple stemming was performed as detailed in the section 6.3.1. The resulting data was then saved into the comma separated (csv) file.

8.6.2 Knowledge Generation and Classification Module

The 1st Level Knowledge processing and Storing unit and then the 2nd Level Knowledge Refinement Unit are applied on the structured data obtained from the previous section. This is done by processing the csv file in Weka (3.4). There are different clustering techniques that can be used for the discovery of first level knowledge in the form of discovered single key term phrases. Weka (3.4) software is based on a java environment which is open source and allows users to customise or add new features to it. It offers different functionalities ranging from pre-processing to data classification and analysis. For further details and handling data or information see reference (Witten and Frank 2000).

The clustering algorithm splits the input space into a number of subspaces. A large number of experiments with other clustering techniques (i.e. Expected Maximization)

were made to find the number of clusters to reduce the effect of information loss. Two main factors highly affected the process of choosing the right number of clusters i.e. the data size under consideration and the need to retain useful information within each cluster. To handle very sparse data and reduce the effect of loosing key information captured within each cluster a number of experiments were made for selecting the most appropriate number of clusters in Weka (3.4) and between 2-10 clusters were considered. The simple matrix model used in these experiments at first stage of analysis was of dimensions (20x315) formed after applications of different units defined in the Text Mining Module. The number of experiments showed that the number of clusters should neither be too small nor too large as this could cause a great loss of information. Ultimately six clusters were selected for the current research work to split the whole information space into a number of subspaces. Three examples of clusters, their ID's and corresponding single term phrases identified through applications of the K-means clustering algorithm are shown in the Table 8.6 given below;

Table 8.6: Single key term phrase identification by K-means Clustering

Cluster ID's	Single Key term Phrases Identified
CL1	"business", "carried", "fitting", "interiors", "***X***", "less", "lift", "number", "out", "overroofing", "pit", "price", "project", "same", "shop", "slightly", "small", "two", "under", "unit"
CL2	"cause", "complete", "delay", "due", "extension", "fortyfive", "granted", "mortuary", "planned", "problems", "programme", "re-roofing", "significant", "still", "13-week", "time", "twentysix", "weeks"
CL3	"any", "approximately", "extra", "few", "figure", "financial", "give", "good", "KPI", "noted", "pounds", "request", "rigidly", "scheme", "six", "stuck", "within"

The letter X is used for company name.

The key information identified by this stage simply exists as different clusters which refer to multiple different sets of information contained in different documents

represented by different document IDs. Up to this point every piece of knowledge identified is treated in an equivalent way and no attempts have been made to interpret it as being good or bad information documents. Also, the business context is not clear from the single term phrases like "*business*" in the cluster and therefore the context cannot be used in the decision making process. For example, two different key term phrases like "business" and "unit" captured in the CL1 can refer to quite different concepts in the documents like "*business unit to supply the contribution*", "*business unit needed some work*" and also the concept of "*business unit target*". So it becomes difficult to map these single key term phrases to find key issues in the form of good or bad information documents. In order to overcome this difficulty of interpreting the key issue there is need to further refine the process of extracting useful information codified within these documents. This is done by applying 2nd Level of Knowledg Refinement unit.

The knowledge obtained through applications of the clustering algorithms needs to be stored in the form of relational database tables containing all fields from the cluster label, key term identified and their corresponding documents identification codes (IDs). ***This task is then performed by generating multiple key term phrasal knowledge sequences which at one end reduce the number of dimensions in the feature space and at other end are used to test the improvement made in the classification accuracy of the classifiers.*** Association Rule mining is used at this stage for preliminary analysis for generating multiple key term phrasal knowledge sequences . The input is given in the form of relational tables where documents are taken in the form of transactions and terms as items. The representative MKTPKS are shown in the Table 8.7 given below;

Table 8.7: Representative Frequent 3-termset sequences formed using Association Rule of Mining

Cluster's ID	MKTPKS 3-termsets
CL1	[T1 T13 T17, T1 T13 T20, T1 T17 T20, , …, T13 T18 T19]
CL2	[T1 T5 T6, T1 T5 T8, T1 T11 T16, T4 T8 T14,…,T12 T14 T16]
CL3	[T2 T8 T10, T4 T7 T8, T4 T7 T9, T5 T10 T15,…, T8 T9 T10]

The MKTPKS formed through application of Apriori Association Rule of Mining are shown in Table 8.7 in the form of T1, T2, T3 etc. which refer to the single key term phrases given in the Table 8.6. The co-occurrences among these terms are given in the form of T1 T2 T3 as a single entity referring to a set of multiple key terms occurring in the documents. These occurrence of terms together form a sequence of knowledge which is later used for the classification task following the criteria defined in the Section 6.4.3.

8.6.3 Classification Results and Accuracies of Models

This section illustrates the methods used for classifying the textual data into two different categories or classes. The data is first transformed into a model based on MKTPKS where the representative terms like (T1 T5 T6) , (T2 T8 T10) and (T4 T7 T8) etc. are taken as single units of information and where the binary representation detailed in Section 8.5.2 is used for performing the classification task. The new matrix model based on FTS was formed with dimensions (20 x 223). This matrix model was based on MKTPKS which were then loaded into Weka (3.4) in the form of a csv file and four different classifiers were tested to classify the data into their respective categories and the accuracies achieved by each are discussed in this section. The target variable set for this purpose was the class variable to determine the number of good or bad information documents. The objective was to train the system and determine the correct and incorrect classification rates. The results obtained through application of different classifiers on the FTS based matrix model are used to compare the classification accuracies against the simple term based representation model. A glimpse of the classifying textual data of PPRs using Decision Trees (J48 or C4.5) defined above using simple term based representation is shown in the figure 8.6. The tree diagram shows that each node (circular / elliptic) is divided into interpretable sub-nodes or leaves with the classified Good (A) and Bad (B) information documents. The terms t130, t262, and t66 are used as representative terms in the simple term based matrix model used for classification of the documents.

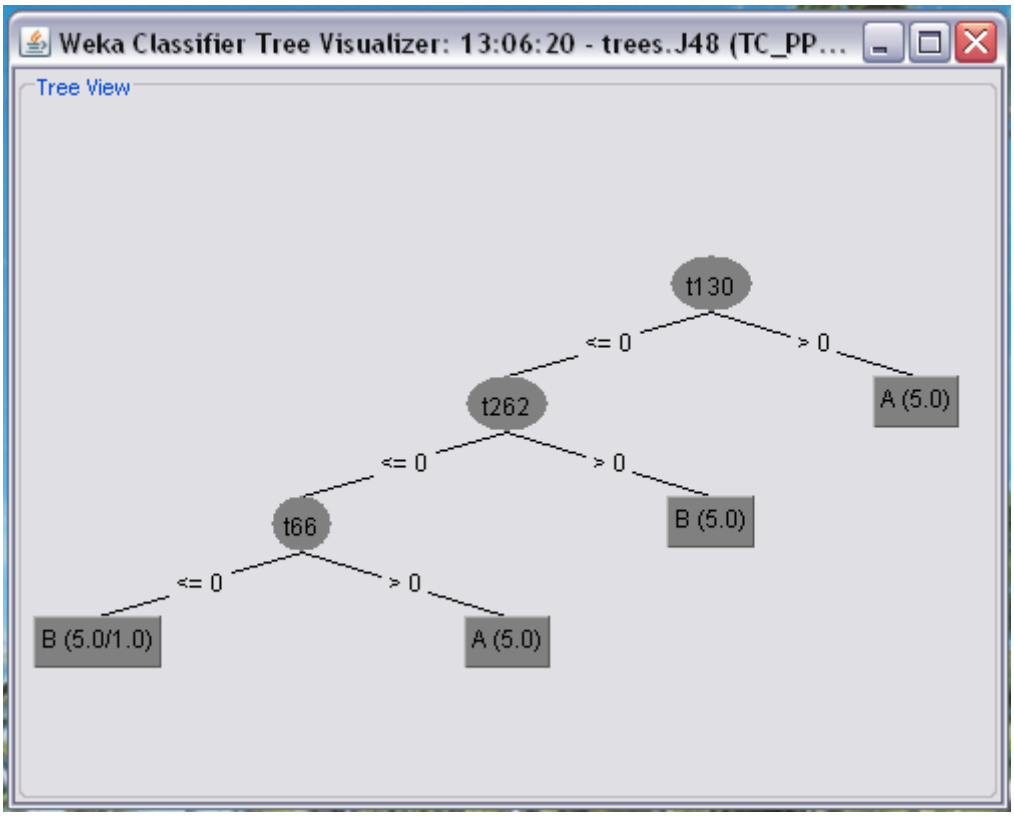

Figure 8.6: Snapshot of decision tree based classification

Each node (elliptic / circular) has been divided into its sub-nodes on the basis of maximum information gain. Each leaf node (rectangular) represents the finally classified information into Good or Bad information documents about project within PPRs. The corresponding A(5.0) shows that five documents are classified as good information documents at the deciding node (circular) of representative term t130. Similarly B(5.0) shows the number of documents classified as bad information is five whereas B(5.0 / 1.0) shows that four out of five documents were classified as bad information documents with an error of one as good information documents at the branch (circular) node of t66.

In terms of classification of data based on proposed MKTPKS system the Decision Tree results using Weka (3.4) based classifier is shown in the figure 8.7 below;

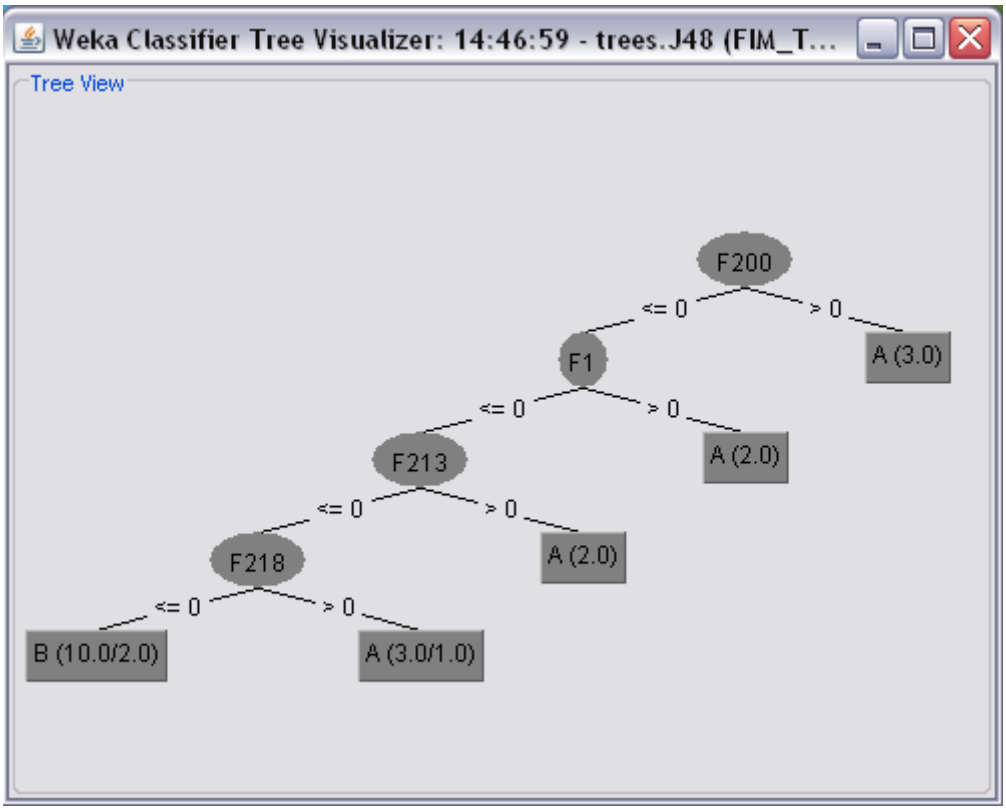

Figure 8.7: MKTPKS based Classification using C4.5

The information space is classified into two classes of good and bad information documents by selecting the information nodes and sub-nodes. The branch leaf represent the number of documents classified as good and bad information documents . The node (elliptic) with representative MKTPKS F200 splits the information with the frequency of occurrence either less than or equal to zero and greater than zero. The classification rule is illustrated as IF the frequency of occurrence is greater than Zero then the information given at this point is classified into Good Information documents which is represented as A(3.0) while for other case i.e. less than or equal to Zero the branch node (circular) is used to perform further classification procedures. Thus the process of forming the decision tree continues until the document space of information is fully classified into two different categories. The information node (elliptic) at representative MKTPKS F218 shows the binary classification leaves (rectangular) in the form of A(3.0/1.0) which means that two documents are classified as good information documents with the error of classification being one document as a bad information space document. Similarly the leaf node B(10.0/2.0) shows that the eight documents are classified as Bad

information documents with an error of two falling into the category of good information documents.

In a business context to provide better services to the customers the knowledge workers or decision makers have to consider their opinions to retain the customer to the industry. In the current research scenario the data under consideration defines the key phrase like "*customer issued very few instructions/ variations*" help the industrial workers to run the project smoothly and get it finished within time. This gave a good Key Performance Indicator (KPI) as time as the project was finished within stipulated time and thus company could easily retain its customer. In such a context if the decision maker or knowledge worker within an industry could identify and classify textual data on the basis of good or bad information documents then better decision could be made for future projects. This would ultimately help to enhance business by identifying ways of retaining their customers from experiences captured in earlier reports. To help the knowledge workers and to gain competitive advantages for improving the quality of the services the objective of this research was to accurately classify the textual data with a lower misclassification rate. To achieve this goal and better manage the knowledge resources different matrix models were considered to structure the textual data (i.e. term frequency and MKTPKS based methods) within this research context. The incorrect classification results obtained through application of different classifiers are calculated using the confusion matrix shown in the Table 8.8.

Table: 8.8 Confusion matrix for performance measure of classifier

Class variables	Predicted: a	b
Actual: a	TP	FN
b	FP	TN

The terms are defined as;

TP(True Positive): the number of documents correctly classified to that class

TN(True Negative): the number of documents correctly rejected from that class

FP(False Positive): the number of documents incorrectly rejected from that class

FN(False Negative): the number of documents incorrectly classified to that class

The classification accuracies are calculated by classifying information as good or bad information documents. For example the following information (i.e. document in this research context) ***"The customer suggested that the job should be done within twenty one (21) weeks and we agreed to that period. The work was completed on time."*** was originally marked as good information document by human experts but the system being tested here identified this as bad information and classified it into the category of B.

The calculated classification rates are given in the following Table 8.9.

Table 8.9: Incorrect classification rates using simple term based Matrix model

Method	Input	output	Incorrect classification Rate
C4.5(J48)	Simple term based Matrix	Class Variable	0.50
KNN(k=10)	Simple term based Matrix	Class Variable	0.60
Naïve Bayes	Simple term based Matrix	Class Variable	0.55
SVM(Linear Kernel)	Simple term based Matrix	Class Variable	0.55

The misclassification rates are calculated against each classifier to predict the correct category of the data. The lower the rate of misclassification the more accuracy is guaranteed in decision making process. The Table 8.10 below shows the rate of

incorrectly classifying the data when the MKTPKS based representation method is used for classifying textual data of PPRs into their respective categories.

Table 8.10: Incorrect classification rates using proposed MKTPKS based Matrix Model

Method	Input	output	Incorrect classification Rate
C4.5(J48)	MKTPKS	Class Variable	0.55
KNN(k=10)	MKTPKS	Class Variable	0.40
Naïve Bayes	MKTPKS	Class Variable	0.45
SVMs (Linear Kernel)	MKTPKS	Class Variable	0.45

The classification results obtained through applications of Decision Trees (C4.5), Naïve Bayes and K-NN algorithms are shown in Table 8.9 and Table 8.10. In terms of applications of decision trees (C4.5) the better classification results are obtained by using simple term frequency matrix where misclassification or incorrect classification rate is 5% lower than that of MKTPKS based method. While the accuracy of other classifiers i.e. K-NN (k=10) and Naïve Bayes are all better than the simple term based representation method (term frequency matrix). The results shown above in Table 8.10 have been obtained through wide range of experiments performed for getting the better accuracies of the classifiers. Different parametric setting were tested using the Weka (3.4) software where different values of K in terms of using K-NN classifier are tested. The optimal settings were found for K=10 for which better classification accuracies were obtained. Similarly in case of applications of Decision Trees (C4.5) algorithm different seed values were used for which the accuracies of the classifier did not change. Therefore optimal values of the classifier were obtained by keeping the parametric settings unchanged in the Weka (3.4) software. In case of implementation of SVMs technique different kernel functions i.e. Linear Kernel and Radial Basis Kernel functions (i.e. $k(x,x^{'}) = \exp(-\frac{\left\|x-x^{'}\right\|^2}{2\sigma^2})$ (Gunn 1997))are used to gain the better classification accuracies. In case of Radial Basis kernel function poor

precision and recall measures values were found which gave less classification accuracies in terms of F-measure. So ultimately Linear Kernel Function based classification methods were selected and found to be useful in classifying the textual data into two different categories. The classification accuracies were higher than that of Radial Basis function method. In case of Naïve Bayes classifier the optimal setting were found for the defined parametric values given in the Weka (3.4) software and simple Naïve Bayes techniques gave better classification model.

The results showed that the accuracy measure of the classifiers improved in case of Naïve Bayes, K-NN and SVMs except Decision Trees (C4.5) methods using MKTPKS based matrix model.

8.6.4 Evaluation Measure and Comparison

The final evaluation of the proposed system is made by comparing the results on the basis of F-measure which is defined as the harmonic mean of Precision and Recall. The precision is defined as the rate of correctly classified documents to the result of classifier and recall is defined as a measure of the rate of correctly classified documents to the documents to be classified correctly. The reason behind selection of F-measure is that both precision and recall ratios are considered in it (Miao et al. 2009). Mathematical representation of the formula for F-measure is based on the notation given in Table 8.8 given as follows;

$$\text{Recall (R)} = \frac{TP}{TP + FP} \tag{8.16}$$

$$\text{Precision (P)} = \frac{TP}{TP + FN} \tag{8.17}$$

$$\text{F- measure} = \frac{2 \times R \times P}{R + P} \tag{8.18}$$

The performance of the system is evaluated using 10-fold cross validation method which is more commonly used and gives more stable results (Mclachlan et al. 2004). The setting of the parameters available in the Weka (3.4) for each algorithm is changed and the optimal settings are selected. Some experiments were made to choose the optimal parameters to gain the better classification accuracies with the application of each classifier. Different parametric values are chosen for each classifier to find the best classification accuracy measure. In the case of Naïve Bayes

classifier the basic settings available in the Weka (3.4) software were unchanged as these gave better classification accuracies. In terms of application of K-NN classifier different settings are tested by varying the values of K for which the best possible accuracies are obtained while setting the value of K=10. Also in case of application of Support Vector Machines (SVMs) the better classification accuracies were obtained using Linear Kernel method. The results obtained are shown in the Table 8.11.

Table 8.11: Comparison of Performance of different classifiers

Classification Model	Term Based Classification Method (F-measure)	Proposed MKTPKS based Classification Method (F-measure)
Decision Trees (J48 or C4.5)	0.495	0.449
K-NN (k=10)	0.341	0.52
Naïve Bayes	0.374	0.561
SVMs (Linear Kernel)	0.374	0.475

The above Table 8.11 shows the performance of different classifiers based on simple term frequency matrix model and MKTPKS based classification model. The performance of the classifiers (i.e. K-NN, Naïve Bayes and SVMs (Linear Kernel) in terms of proposed method are better than the simple term based classification model.

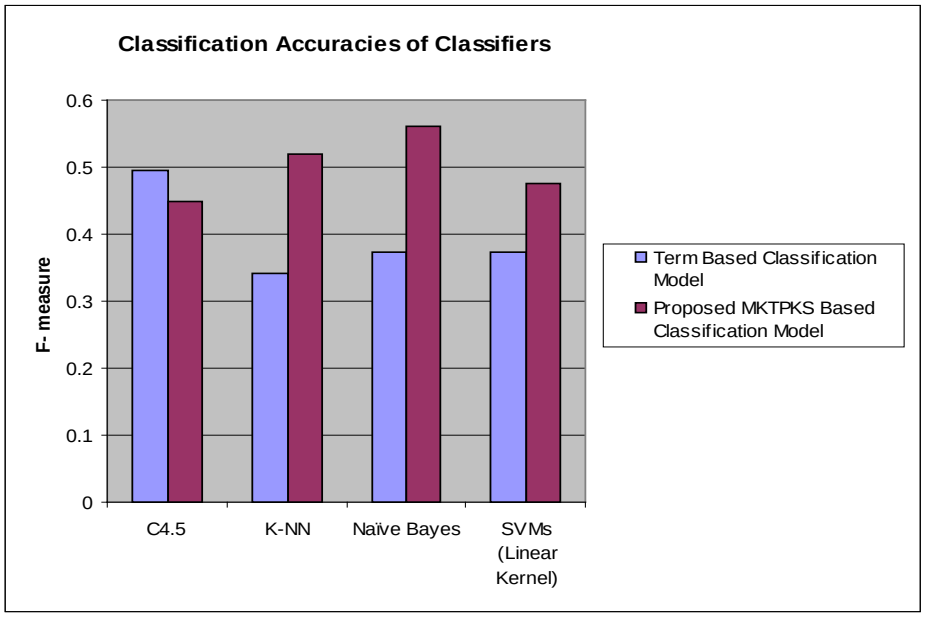

Figure 8.8: Comparison of Classification accuracies using F-measure

The figure 8.8 shows that the performance of the Decision Tree (C4.5) classifier using proposed method is slightly lower than the simple term based matrix model. However the performance values measured using F-measure for all other classifiers (i.e. K-NN, Naïve Bayes and SVMs) improved with the use of proposed methodology when compared with simple term based classification method. Thus overall accuracy of the classifiers are improved if the proposed method is used to classify the textual data into two different categories of **G*ood*** and ***Bad*** information documents.

8.7 Novelty of the Work and Discussion

To the author's best knowledge the work presented in this chapter is the first of its kind, where classification methods have been applied on textual data to divide it into two different categories or classes based on proposed MKTPKS based method of classification.

The research work presented in this chapter is focused on classification of textual data into two different categories to define good and bad information documents. A novel integration of textual data mining techniques is made to improve the classification accuracy of the classifier. In terms of classifying documents into their respective categories using decision tree (J48 or C4.5 algorithm) the accuracy of the classifier is reduced using the proposed methodology while in other classifiers' application there

is a significant improvement in the classification accuracies measured using F-measure. The reason behind loosing the accuracy of C 4.5 classifier may lie in the fact that information selection criteria in C 4.5 highly dependant on the terms and their corresponding frequencies.

The following points are therefore concluded from the research work presented in this part of application of proposed methodology;

- Single term based representation methods are useful sources of carrying information but these methods affect the classification accuracies of the textual data.
- Hybrid application of textual data mining techniques gives better results whereas in the current research scenario the information pruning and knowledge refinement is possible through use of Apriori Association Rule of Mining technique.
- Generating multiple key term phrasal sequences of knowledge and using these for performing the classification improved the accuracies of the classifiers.
-

8.8 Summary of the Chapter and Conclusion

This chapter discusses the implementation of different classifiers for textual data classification. A novel hybrid textual data mining approach is used for discovering useful knowledge from free formatted textual data and then using this knowledge in the form of MKTPKS to classify it into two different classes. The classification results are discussed which further serves the purpose of exploring other statistical or machine learning techniques to reduce the misclassification rate of the classifiers.

Chapter 9 Semantic Analysis Methods for Text Classification

9.1 Introduction

In this chapter a semantic text analysis method called the Latent Semantic Analysis (LSA) is tested. This is a well known method in the area of information retrieval (IR). The method is used to cluster information on the basis of different ranking measures and the results will be compared and tested against the classification techniques used in Chapter 7 to classify the data into two different classes of good or bad information documents. The study made in this chapter is different from the ***normal trend*** of applications of LSA methods which are for feature selection and better information retrieval tasks. The purpose of this chapter is therefore to use the LSA based semantic model to classify the documents and then to compare the results with the MKTPKS based classification model proposed and implemented in previous chapters.

9.2 LSA Models for Text Classification and Effectiveness

Latent Semantic Analysis (LSA) methods are used to consider the semantic relationships among terms defined with textual documents and retrieve useful information from the textual data. These methods were first proposed in (Deerwester et al. 1990) and are used to automate the process of retrieving useful information from documents. The information contained in the documents is represented in the form of a matrix called a term by document matrix (i.e. t x d) where the 't' is used to represent terms and 'd' stands for the documents. The whole information space is then divided into a semantic space based on Singular Value Decomposition (SVD) where SVD is used to decompose the terms by documents (t x d) matrix into linearly independent spaces or sub dimensional vector spaces. LSA methods were originally used for performing the task of information retrieval based on semantic relationships existing among different terms or words used in the textual databases.

The general representation form of decomposing information into subspaces is given in the form (Berry et al. 1995; Grossman 1998):

$$A = U \sum V \qquad (9.1)$$

Where U and V are orthogonal matrices with the property of

$$UU^T = VV^T = I \quad (9.2)$$

with values as left and right singular vectors of A and $\sum$ is the diagonal matrix with entries as d1,d2,d3,…,dn where di > 0 for $1 \leq i \leq r$ and dj = 0 for $j \geq r+1$ which are called the singular values of A.

In terms of approximating the matrix A different values of k are taken where the largest values are kept along with the corresponding columns in the matrices U and V. Different approximations taken for different values of k are given as Uk, Vk and represented using the formula;

$$A \approx A_k = U_k \sum_k V_k^T \quad (9.3)$$

The matrix A_k so formed through multiplication of k approximated values of U , V and $\sum$ is uniquely closest to A. The geometrical representation of forming two dimensional SVD matrix is shown in the figure 9.1.

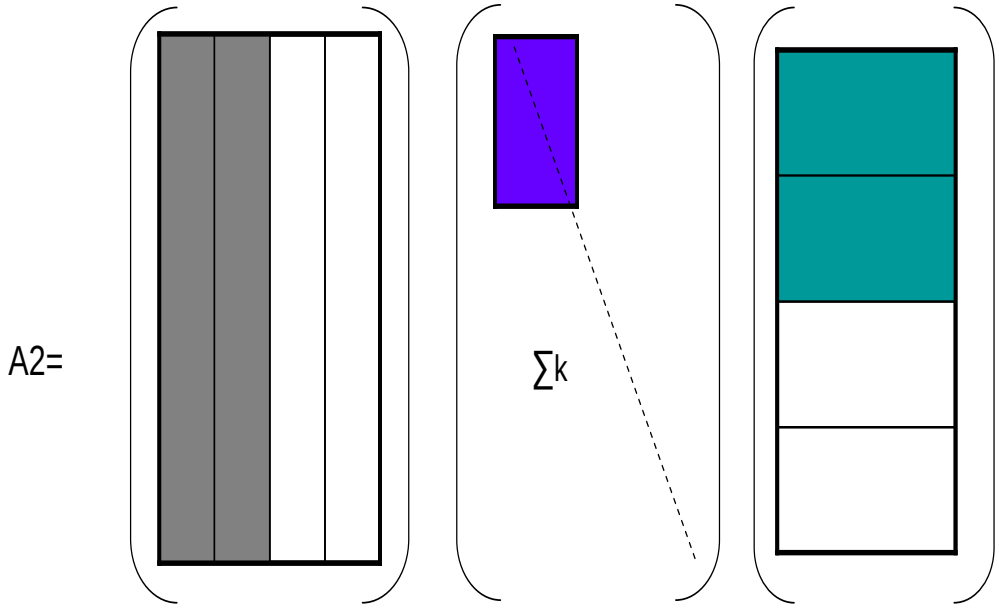

Figure 9.1: Geometrical Representation of SVD Matrix Model for k=2

The figure 9.1 shows that to form an approximated matrix using SVD method for k=2 first two columns are selected in case of U, while the first two diagonal entries are

used from the diagonal matrix ($\sum$) and two rows are from the matrix formed by taking the transpose of V_2.

Different approximations have been used to classify the textual data where different values of k have been tested i.e. k = 2,3,4. The target attribute for classification is set as good or bad information documents i.e. A as good and B for bad information documents.

9.2.1 Defining Knowledgeable Relationships through Matrix Models

The structured representation of information in the form of a matrix carrying the simple term based information is shown in Table 9.1. The matrix is formed using a simple term based structured representation of a textual data set and was used to define the semantic relationships among terms. Representing the key information in matrix form helps to capture the relationships among terms and identify their corresponding documents in the textual data available from PPRs.

Table 9.1: Simple Term Based Matrix Model

Terms	D1	D2	D3	D4	D5	D6	D7
About	0	0	0	0	0	1	1
Above	0	0	0	0	0	2	0
Account	0	3	1	1	1	0	2
Accurate	0	0	0	0	0	1	0
Achieve	0	0	0	0	0	1	0
Actual	0	0	0	0	1	1	0
Additional	0	0	0	0	1	0	1
Adjusted	0	0	2	0	0	1	0
Administer	0	0	0	0	0	1	0
Against	0	0	0	0	0	1	0

The Table 9.1 above shows the relationships among terms and their frequency of occurrence in the corresponding documents. This shows that the terms '*accurate*' and '*achieve*' occurred in the same document giving some meaning to the text like

'*accurate targets for future work*' and '*achieve the maximum gain*'. The relationship among these terms could be explained that if targets were measured accurately then the company could gain maximum profit. These types of relationships include some human bias and this needs to be reduced where possible. Different techniques are used in the literature to capture semantic relationship, but in the current research context the focus is made to find these relationships by reconstructing the simple term based matrix model using the latent semantic analysis based matrix model. Later this model was used to classify the textual data into two different classes of good or bad information documents. The corresponding reconstructed example matrix from Table 9.1 using the Latent Semantic Analysis model is shown in Table 9.2.

Table 9.2: LSA based Matrix model based on simple term based data structuring

Terms	D1	D2	D3	D4	D5	D6	D7
About	0.0618	0.224	0.1696	0.13399	0.18384	0.89579	0.35267
Above	0.0542	-0.12	0.105	0.131	0.1299	1.9243	-0.133
Account	0.2285	1.763	0.7561	0.4538	0.7715	-0.057	2.6085
Accurate	0.0271	-0.06	0.0524	0.066	0.065	0.9622	-0.066
Achieve	0.0271	-0.06	0.0524	0.066	0.065	0.9622	-0.066
Actual	0.043	0.041	0.1024	0.0986	0.1169	1.0418	0.085
Additional	0.053	0.382	0.172	0.1066	0.1764	0.0850	0.5674
Adjusted	0.0271	-0.06	0.0524	0.0657	0.065	0.9622	-0.066
Administer	0.0271	-0.06	0.0524	0.0657	0.065	0.9622	-0.066
Against	0.0271	-0.06	0.0524	0.0657	0.065	0.9622	-0.066

The Table 9.2 shows the new representation of matrix model where the frequencies of the matrix model given in Table 9.1changed gives an indication of terms probabilities of occurring together in the documents D1-D7. The representative matrix given in Table 9.2 showed the term frequency of '*account*' given as '0' in D1 as shown in Table 9.1 is replaced with '0.2285'. Similarly the frequencies of other terms such as accurate, adjusted and administer are also changed which are used to measure he closeness of one document to other. The proximity of closeness of one document 'd_i'

to other document 'd_j' can be measured by using Cosine similarity measure (Grossman 1998) given in equation 9.4.

$$\cos(d_i, d_j) = \frac{d_i^t d_j}{\|d_i\| \|d_j\|} \quad (9.4)$$

Thus a new approximated matrix of transformed values is used to cluster the information first by selecting different values of *k* (i.e.2,3,4) and then used for performing the classification task. The process of using the approximated matrix for text classification is detailed in the next section.

9.2.2 Classification Methods, Accuracies and Comparative Analysis

The input data has been transformed into information as shown in the Table 9.3 and each document now carries a class variable i.e. A and B to indicate whether it is a good or bad information documents. Three different approximated matrices for different values of k were formed by multiplying the k approximated values of $U, \sum$ and transpose of V using Equation 9.3 as shown in Figure 9.1. The formal representation of matrix model formed for performing the classification task into good (A) and bad (B) information documents is shown in the Table 9.3. The new approximated matrices are then loaded into Weka (3.4) in the form of a comma separated value (csv) file. The classification functions available under the heading classify were used to classify data into two different categories of good or bad information documents. The details of these functions and their information selection criteria have already been given in Section 8.4.

Table 9.3: LSA Based Matrix Model for Text Classification

Docs Id	T1	T2	T3	T4	…	Tn	Class
D1	0.0618	0.0542	0.2285	0.0271	…	0.0271	A
D2	0.224	-0.12	0.105	0.131	…	-0.06	B
D3	0.1696	0.105	0.7561	0.0524		0.0524	A
…	…	…	…	…	…	…	…
Dm	0.35267	-0.133	2.6085	-0.066	…	-0.066	A

The effectiveness of the method is measured using the percentage of correctly classified textual data while the final comparison is made using the average values of F-measure. The lower the misclassification rate of the classifier the better the accuracy would be of classifying data and identifying knowledge that may improve the business intelligence solutions. For different values of k the classification accuracies were found and given in the form of tables obtained for k = 2,3,4 as shown in the Tables 9.4, 9.5 and 9.6 respectively.

Table 9.4: Classification Accuracies of LSA Model for k =2

Classifier	Correct classification rate	Incorrect classification rate	F-measure
Decision Trees (C4.5 or J48)	0.30	0.70	0.271
KNN (k=10)	0.50	0.50	0.479
Naïve Bayes	0.50	0.50	0.405
SVM (Linear Kernel)	0.55	0.45	0.355

Table 9.5: Classification Accuracies of LSA Model for *k*=3

Classifier	Correct classification rate	Incorrect classification rate	F-measure
Decision Trees (C4.5 or J48)	0.50	0.50	0.479
KNN (k=10)	0.50	0.50	0.479
Naïve Bayes	0.45	0.55	0.3105
SVM (Linear Kernel)	0.55	0.45	0.436

Table 9.6: Classification Accuracies of LSA Model for *k*=4

Classifiers	Correct classification rate	Incorrect classification rate	F-measure
Decision Trees (C4.5 or J48)	0.35	0.65	0.2595
Naïve Bayes	0.50	0.50	0.405
KNN (K=10)	0.45	0.55	0.3105
SVM (Linear Kernel)	0.35	0.65	0.2595

The accuracy of the proposed method of classifying textual data using the simple term based and MKTPKS based models are shown in the figure 9.2-9.4 shown below;

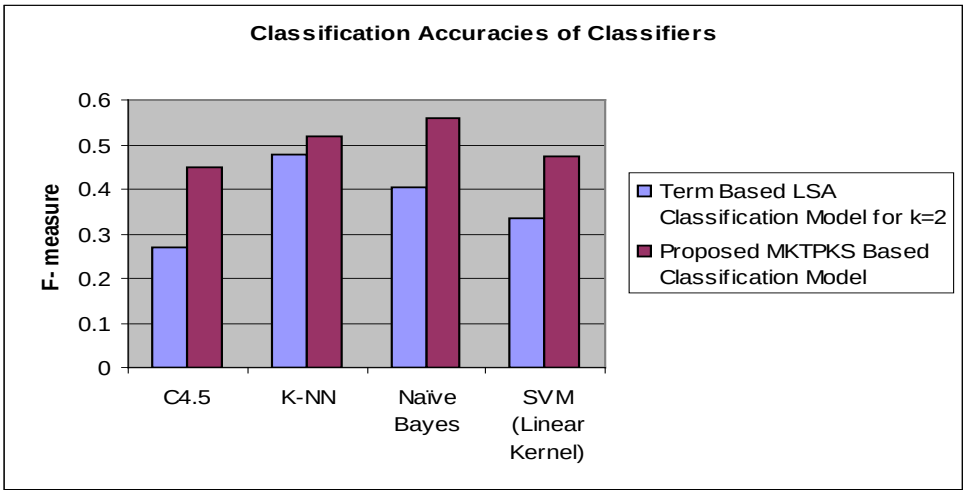

Figure 9.2: Comparing LSA (k =2) with MKTPKS Classification Models

The figure 9.2 shows that the proposed MKTPKS based classification model gave comparable values over the simple term based LSA model in the case of all the classifiers.

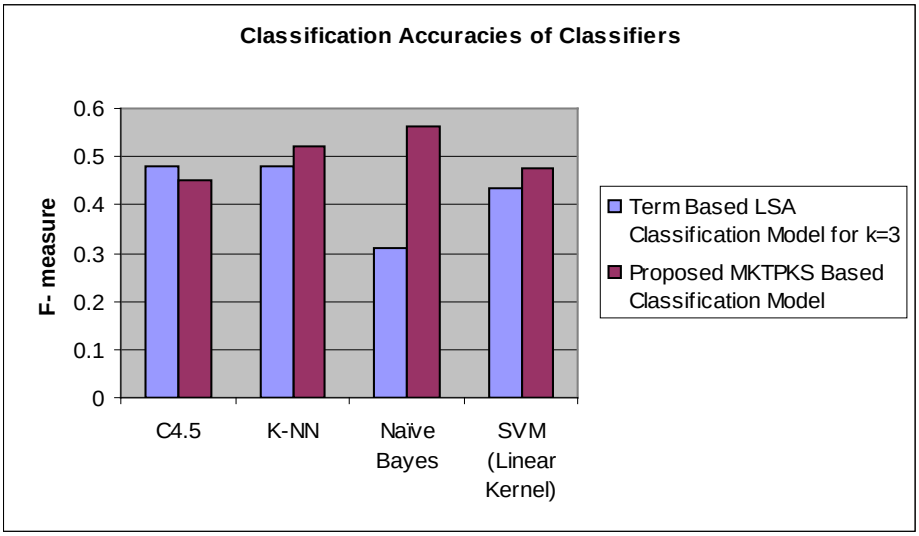

Figure 9.3: Comparing LSA (k =3) with MKTPKS Classification models

The figure 9.3 shows that the classification accuracies using MKTPKS based model are higher than the term based LSA model except the case of C4.5 where the proposed method is less efficient where the difference in the accuracies is (0.479-0.449= 0.03).

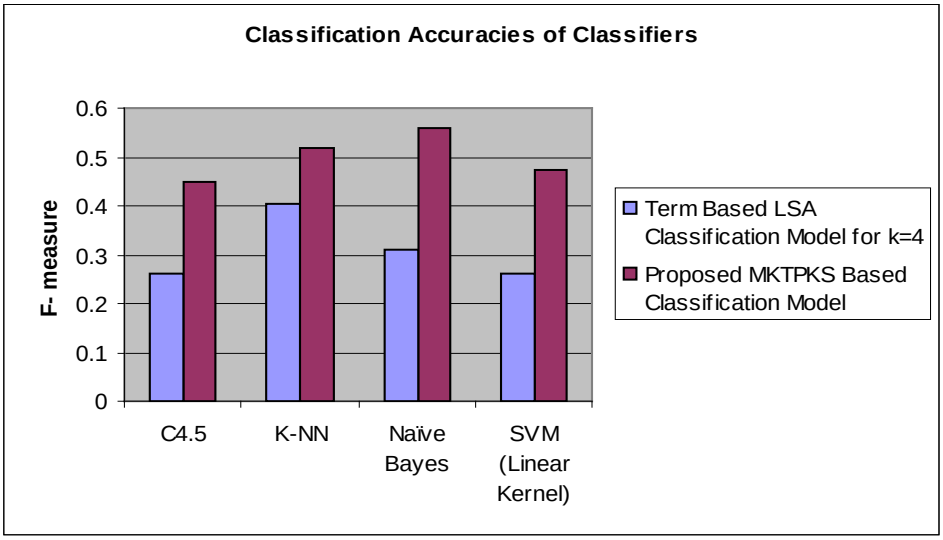

Figure 9.4: Comparing LSA (k=4) with MKTPKS Classification Models

The figure 9.4 shows that the proposed FTS based classification model performs well when compared with term based LSA model. The accuracy the MKTPKS method is greater for all classifiers when compared to the LSA results.

Thus overall this study shows that the accuracies of the classifiers improves using proposed MKTPKS model for classification of textual data into two different classes.

9.3 Summary of the Chapter and Conclusion

In this chapter LSA model was used to classify the data into two different classes and the results are presented. The model was used to compare the accuracies against the proposed MKTPKS based model to further verify the hypothesis that better classification is achieved through the proposed method of classifying textual data into two different categories. Thus the work reported in this chapter should benefit knowledge workers or decision makers to better classify textual data into their corresponding categories and identify the issues discussed in the data and help to improve the overall business of the industry.

Chapter 10 Conclusions and Future Work

10.1 Conclusion

The last three Chapters (i.e. Chapter 7,8,9) have highlighted different methodologies that have been developed and adopted to meet the objectives of the research introduced in Chapter 1. Based on the past literature reported in the areas of textual data mining applications in manufacturing or construction industry, the conceptual development of theory to achieve the objective of the research was mainly focused on application of clustering and then applications of Apriori Association Rule of Mining which has been widely used in various areas of applications. Hence substantial time and efforts have been made in this research in developing and applying these techniques for better selection of number of clusters and then using MKTPKS based matrix model to better classify textual data.

In the first stage Clustering was used to discover the first level of knowledge in terms of finding natural term based relationships defined in the textual data. The discovered knowledge in terms of single key term phrases was difficult to interpret and to use in identifying good or bad information documents available in the form of PPRs. Therefore the 2nd level knowledge refinement process is done with the application of Apriori Association Rule of Mining techniques to find more useful multiple key term knowledge sequences (MKTPKS) using varying level of support values. The one reason for generating the MKTPKS was to find those sequences of terms which refer to the terms which co-occur in the documents to identify good or bad information documents. The results obtained in the form of sequence of terms are useful to map information to the particular document space of information and then these sequences (i.e. MKTPKS) are compared with those identified by the domain experts. F-measure was used to measure the accuracy where the results were (37%) accurate where the value of recall measure was better than the precision measure.

The second part of the implementation of the methodology was to implement the different classification techniques on the discovered MKTPKS based matrix model. The purpose of this implementation was to study the affect on classification accuracies of different classifiers to help the knowledge workers or decision makers to better classify the data into their predefined classes. Since natural relationships in

terms of finding MKTPKS were captured by applying clustering and Apriori Association Rule of Mining techniques, the discovered knowledge must be used properly. This discovered knowledge was used for the classification task of textual data on the basis of its representation in terms of MKTPKS. There is a significant improvement in the classification accuracies obtained and these results are shown in the chapter 8.

A novel aspect of this research is the discovery of knowledge in terms of single or multiple key term phrases which were used to discover the relationships among terms defined in the textual data of PPRs. The discovery of these natural relationships could be used to improve business intelligence solutions as this research provides a means of reducing misclassification errors. The lower the error the better the classification accuracies would be and the previous knowledge thus stored in the form of textual databases would effectively be used for finding the solutions to new unclassified problems. The kind of work presented in terms of application of methodologies in chapter 7 and chapter 8 is a novel integration of several techniques and gave good results so the knowledge discovered might be used on other textual databases which are available in the free formatted text.

Another big advantage of the proposed methodology and its implementation is that it provides help in finding term based relationships among terms and this would be a big advantage to industry in terms of storing information in terms of different clusters where natural relationships among terms is stored. New information could then easily be compared with previously stored information in terms of clusters and its analysis would be easier as it could be put it into different information subspaces. The analysis of a large corpus of information available in textual data formats was made easy by first putting the whole information into multiple subspaces and efforts in analysing would be reduced.

The research proposed in this thesis also concludes with the important fact that the selection of the most appropriate data mining techniques in terms of classifying textual data also depends upon the information selection criteria which vary from simple distance measure to probabilistic methods. So the choice of classifier also

depends on the form of data available and its quality which helps the classifier to govern the rules for classification.

The main contribution of this research are enumerated as under;

- Developing of a generic method of discovering useful knowledge in terms of single key term phrases specific to some key issues discussed in the textual databases.
- Using the single term phrases sequence, multiple key term phrases are generated within each cluster to produce more valuable knowledge sequences and then mapping information to some specific set of documents as good or bad information documents.
- A novel integration of methods for generating multiple key term phrasal knowledge sequences are used to reduce the classification error when compared to simple single term representation methods.
- An introduction of novel integration to perform the text classification task where the path is followed from unsupervised learning to supervised learning for identifying key knowledge areas from textual databases and classifying documents. This technique is a hybridization of different methods and techniques which ultimately supports the generation of useful classification results for textual data.
- The proposal of novel integration of textual data mining techniques to capture key information or knowledge and disseminate it in terms of classification of textual data within any industrial setups.

10.2 Future Work

10.2.1 Next Generation Knowledge Based Product or Service Quality Improvement System

Structured methods of data analysis which identify and store knowledge in terms of useful patterns in the databases are required so that the knowledge can be reused for future business activities which are important in many product or service improvement contexts. These techniques enable the lessons learnt through different stages of product or project management to be made accessible to future team

members so that better collaboration can be achieved among different team members in distributed manufacturing industrial environments.

The work conducted during this research has been focused on finding the term based relationships and exploiting these to form the multiple key term phrasal knowledge sequences. These multiple key term phrasal knowledge sequences are used to classify the textual data into two different classes of good or bad information documents. Once the knowledge is available in the form of MKTPKS it can be used either to retrieve project related documents defined with some key issues or classifying these documents into two different categories. However there is much wider potential for the application and development of these techniques in any business intelligence context where information has been recorded in semi-structured text based documents which can range from Emails, operator notes, customer feedback and reports. Using the methodology proposed here, knowledge can be extracted from the original sources and stored in the form of a knowledge base that can further provide solutions to perform different activities in product or service quality improvement scenarios as shown in the figure 10.1. The figure shows that how the activities of storing information about some product or service histories or project reports in terms of post project reviews or other documents and sequence of activities defined within data mining technology work together to solve the issues and further improve the process of customer service support and retain the customer to the industry. However following points can be considered to improve the proposed methodology which are part of the future work:-

- A comparative study of simple term based and multiple key term phrasal knowledge or MKTPKS based representations has been made but further analysis is still required in terms of application to different databases (i.e. product/ process information, customer reviews and sentiment analysis databases) as future work to test the methodology.
- The current research has focused on simple term based data structuring methods whereas the sentence based and n-gram based representation methods could be used to improve the proposed methodology as part of future research work.

- Syntactic and Semantic structures of textual data have not been considered during the current research work, but these could also help to improve the methodological development of the proposed framework, again, as a part of future research work.

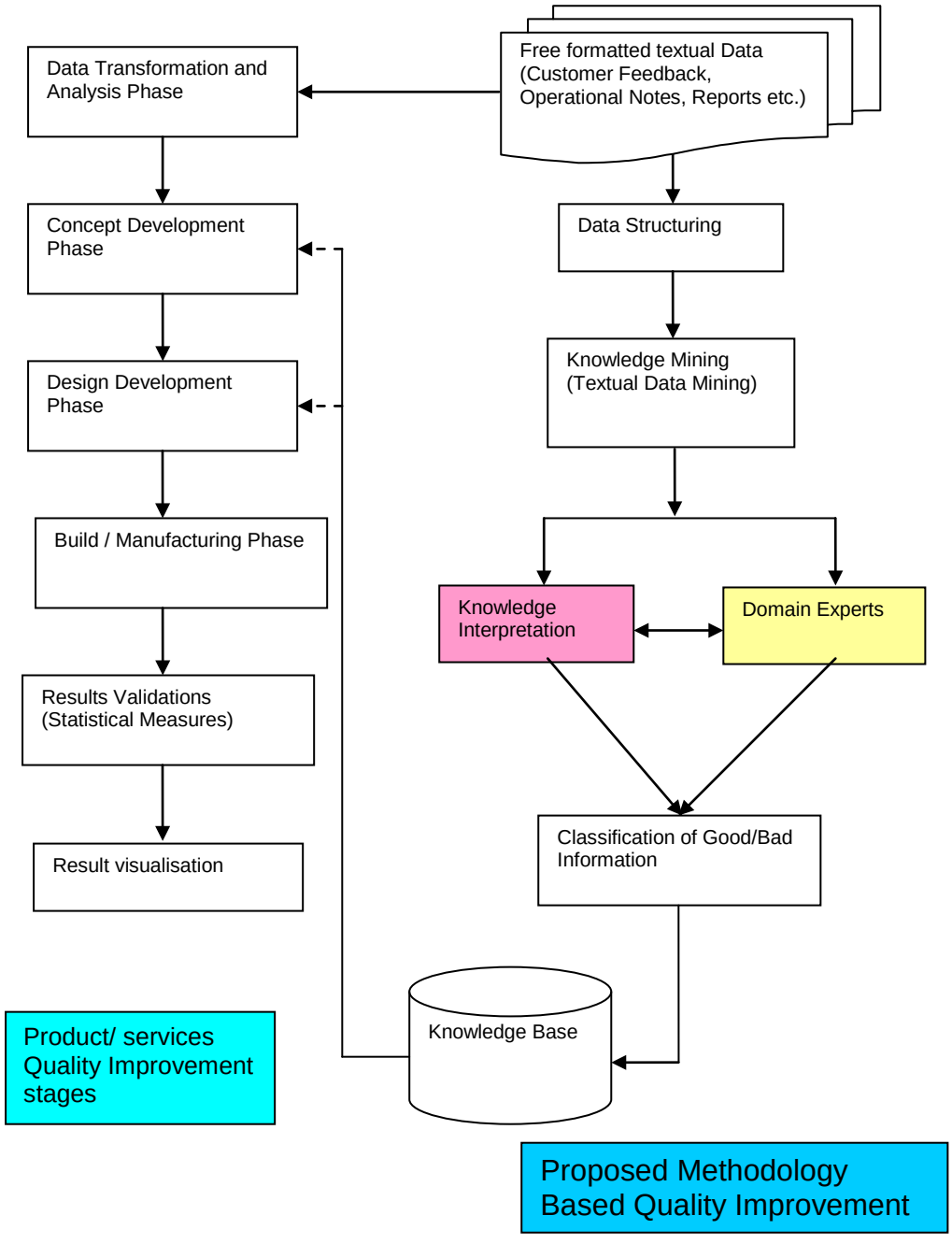

Figure 10.1: Knowledge Based Product or Service Quality Improvement System

10.3 Summary of the Chapter and Conclusion

In this chapter the conclusions of the research work and future improvements of the proposed system are suggested which could be used to accommodate other areas of information and knowledge seeking communities in an industrial environments.

References:

Abramovici, M. and Siege, O. C. (2002). Status and Development Trends of Product Lifecycle Management Systems. In Proceedings of IPPD, Wroclaw, Poland.

Agard, B. and Kusiak, A. (2004a). "Data Mining-based Methodology for Product Families." International Journal of Production Research **42**(15): 2955-2969.

Agard, B. and Kusiak, A. (2004a). "Data Mining for Subassembly Selection." Journal of Manufacturing Science and Engineering **126**: 627-631.

Agrawal, R., Lmielinski, T. and Swami, A. (1993). Mining Association Rule between Sets of Items in Large Databases. In Proceedings of 1993 International Conference on Management of Data (SIGMOD 93).

Aizermann, M. A., Braver, E. M. and Rozonoer, L. I. (1964). "Theoretical Foundations of the Potential Function in Pattern Recognition Learning." Automation and Remote Control **25**(821-837).

Amann, K. (2002). Product Lifecycle Management: Empowering the Future of Business, CIM Data, Inc.

Ameri, F. and Dutta, D. (2004). Product Lifecycle Managment needs, concepts and components, University of Michigan, Ann Arbor, MI**:** 2.

Ameri, F. and Dutta, D. (2005). "Product Lifecycle Management: Closing the Knowledge Loops." Computer Aided Design and Applications **2**(5): 577-590.

Annand, S. S., Bell, D. A. and Hughes, J. G. (1995). The Role of Domain Knowledge in Data Mining. Knowledge and Information Management, Baltimore MD USA.

Apte, C., Damerau, F. and Weiss, S. (1994). "Automatted Learning of Decision Rules of Text Categorization." ACM Transactions on Information Systems **12**(3): 233-251.

Baeza-Yates, R. and Ribeiro-Neto, B. (1999). Modern Information Retrieval, Addison Wesley Longman Publishing Company.

Beckman, T. J. (1999). The Current State of Knowledge Management in Liebowitz CRC Press.

Berry, M. J. A. and Linoff, G. (2004). Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management, Wiley Computer Publishing.

Berry, M. W., Dumais, S. T. and O'Brien, G. W. (1995). "Using Linear Algebra for Intelligent Information Retrieval." SIAM Review **37**(4): 573-595.

Beulens, A. J. M., Jansen, M. H. and Wortmann, J. C. (1999). The Information de-coupling point in Global Production Management, IFIP WG 5.7. International Conference on Advances in Production Management Systems, Boston, Kluwer Academic Publishers.

Bloor, M. S. and Owen, J. (1991). "CAD/CAM Product-data Exchange: the next step." Computer Aided Design **23**: 237-243.

Bodon, F. (2003). A Fast Apriori Implementatio. Proceedings of the IEEE ICDM Workshop on Frequent Itemset Mining Implementations (FIMI'03), CEUR Workshop Proceedings, Melbourne, Florida, USA.

Bolasco, S., Canzonetti, A., Capo, F.-M., Ratta-Rinaldi, F. D. and Singh, B.-K. (2005). "Understanding Text Mining: A Pragmatic Approach." Studies in Fuzziness and Soft Computing **185**: 31-50.

Bolasco, S., Canzonetti, A., Ratta-Rinaldi, F. D. and Singh, B.-K. (2002). Understanding Text Mining, Roma, Italy.

Bordogna, G. and Pasi, G. (1995). "Controlling Retrieval Through a User-adaptive Representation of Documents." International Journal of Approximate Reason **12**: 317-339.

Braha, D. (2002). Data Mining for Design and Manufacturing, Sringer.

Brezocnik, M., and Balic, J. (2001). "Genetic-based Approach to Simulation of Self-Organizing Assembly." Robotics Computer Integrated Manufacturing **17(1-2)**: 113-120.

Brezocnik, M., Balic, J. and Brezocnik, Z. (2003). "Emergence of Intelligence in Next Generation Manufacturing Systems." Robotics Computer Integrated Manufacturing **19**: 55-63.

Brezocnik, M., Balic, J. and Kuzman, K. (2002). "Genetic Programming Approach to Determining of Metal Material Properties." Journal of Intelligent Manufacturing **13**: 5-17.

Bykowski , A. and Rigotti, C. (2001). A Condensed Representation to find frequent Patterns. In Proceedings of ACMPODS Conference, C A,USA.

Carrillo, P. M. (2005). "Lessons Learned Practices in the Engineering, Procurement and Construction Sector." Journal of Engineering, Construction and Architectural Management **12**(3): 236-250.

Chang, C.-W., Lin, C.-T. and Wang, L.-Q. (2009). "Mining the Text Information to Optimize the Customer Relationship Management." Expert System with Application **36**: 1433-1443.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. and Wirth, R. (2000). CRISP-DM Step-by-Step Data Mining Guide**:** 1-78.

Chen, M. C., and Wu, H.P. (2005). "An Association Based Clustering Approach to Order Batching Considering Customer Demand Pattern." The International Journal of Management Science **33**: 333-343.

Chen, W. C., Tseng, S. S. and Wang, C. Y. (2005). "A Novel Manufacturing Detection Method Using Association Rule Mining Techniques." Expert System with Applications **29**: 807-815.

Cheng, K., Pan, P. Y. and Harrision, D. K. (2001). "Web-based Design and Manufacturing Support Systems: implementation perspectives." International Journal of Computer Integrated Manufacturing **14**(1): 14-27.

Chia, H. W. K., Tan, C. L. and Sung, S. Y. (2006). "Enhancing Knowledge Discovery via Association-based Evolution of Neural Logic Networks." IEEE Transactions on Knowledge and Data Engineering **18**(7): 889- 901.

Chien, C. F., Wang, W.C., and Chang, J.C. (2007). "Data Mining for Yield Enhancement in Semi-conductor Manufacturing and Empirical Study." Expert System with Applications **33**(1): 192-198.

Cho, S., Afsour, S., Onar, A. and Kaundinya, N. (2005). "Tool Breakage Detection Using Support Vector Machine Learning in a Milling Process." International Journal of Machines Tools & Manufacture **45**: 241-249.

Choudhary, A. K., Harding, J. A., Carrillo, P. M., Oluikpe, P. and Rahman, N. (2008). Text Mining Post Project Reviews to Improve Construction Project Supply Chain Design. International Conference Data Mining, DIM'08. Las Vegas Nevada, USA.

Christiniani, N. and Shawe-Taylor, J. (2000). An Introduction to Support Vector Machines and Other Kernel-based Learning Methods, Cambridge University Press.

Cladas, C. H. and Soibelman, L. (2003). "Automatic Hierarchical Document Classification for Construction Management Information Systems." Automation in Construction **12**: 395-406.

Cloza, C. M. and Ziemba, E. (2006). "Business Intelligence Systems in the Holistic Infrastructure Development Supporting Decision-Making in Organisations." Interdisciplinary Journal of Information, Knowledge and Management **1**: 47-58.

Cooper, L. G. and Giuffrida, G. (2000). "Turning Data Mining into a Management Science Tool: New Algorithms and Empirical Results." Management Science **46**(2): 249-264.

Cunha, D., Agard, B., and Kusiak, A. (2006). "Data Mining for Improvement of Product Quality." International Journal of Production Research **44**(18-19): 4027-4041.

Davenport, T. H. and Prusak, L. (1997). Information Ecology: Mastering the Information and Knowledge Environment, Oxford Universtiy Press.

Davenport, T. H. and Prusak, L. (1998). Working Knowledge: How Organisations Manage What they Know, Harvard Business School Press.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K. and Harshman, R. A. (1990). "Indexing by Latent Semantic Analysis." Journal of American Society of Information Science **41**(6): 391-407.

Dhaliwal, J. S. and Benbasat, I. (1996). "The Use and Effects of Knowledge-based System Explanations: Theoretical Foundations and a Framework for Emperical Evaluation." Information System Research **7**(3): 342-362.

Djoko, S., Cook, D. J. and Holder, L. B. (1997). "An Empirical Study of Domain Knowledge and Its Benefits to Substructure Discovery." IEEE Transactions on Knowledge and Data Engineering **9**(4).

Dong, A. and Agogino, A. (1997). "Text Analysis for Constructing Design Representations " Artificial Intelligence in Engineering **11**(2): 65-75.

Drewes, B. (2005). Some Industrial Applications of Text Mining. Berlin Heidelberg, Springer-Verlag. **185:** 223-232.

Dutta, D. and Wolowicz, J. P. (2005). An Introduction of Product Lifecycle Management(PLM). 12th ISPE International Conference on Concurrent Engineering: Research and Applications. Dallas.

Duverge, B., Ehlers, M., Lawrence, K. and Deptowicz, D. (2005). Warranty. Management: Transforming Aftermarket Processes to Improve Product Quality. SIAM International Conference on Data Mining,, Newport Beach, CA,. USA

Earl, M. (2001). "Knowledge Management Strategies Towards a Taxonomy." Journal of Management Information Systems **18**(1): 215-233.

Edwards, B., Zatorsky, M. and Nayak, R. (2008). Clustering and Classification of Maintenance Fault Logs Using Text Data Mining Australian Conference on Data Mining Aus DM.

El Wakil, M. M. (2002). Introducing Text Mining, Information System Department, Faculty of Computers and Information,Cairo University.

Fan, W., Wallace, L., Rich, S. and Zhang, Z. (2006). "Tapping into the Power of Text Mining." Communication of the ACM(Accepted for publication) **49**(9): 77-82.

Fong, A. C. M. and Hui, S. C. (2001). "An Intelligent Online Machine Fault Diagnosis Systems." Computing and Control Engineering Journal **12**(5): 217-223.

Freitas, A. A., (2006), ”Are We Really Discovering Interesting Knowledge from Data?”, Proceedings of Second UK Knowledge Discovery and Data Mining Symposium,(UKKDD’06). (2006). Are We Really Discovering Interesting Knowledge from Data? (UKKDD’06) Proceedings of Second UK Knowledge Discovery and Data Mining Symposium. Norwich.

Gao, L., Chang, E. and Han, S. (2005). Powerful Tool to Expand Business Intelligence: Text Mining. Proceedings of World Academy of Science, Engineering and Technology.

Gardner, M. and Bieker, J. (2000). Data mining solves tough semiconductor manufacturing problems. Proceedings of the sixth ACM SIGKDD international conference on Knowledge Discovery and Data mining, Boston, Massachusetts, United States

Gascoigne, B. (1995). "PDM: The Essential Technology for Concurrent Engineering." World Class Design to Manufacture **2**(1): 38-42.

Gen, M. and Cheng, R. (1997). Genetic Algorithms and Engineering Design. Canada, Jhon Wiley and Sons.

Gordon, M. and Pathak, P. (1999). "Finding Information on the World Wide Web: The Retrieval Effectiveness of Search Engines." Information Processing and Management **35**: 141-180.

Grievel, L. (2005). Customer Feedbacks and Opinion Surveys Analysis in the Automotive Industry. Text Mining and its Applications to Intelligence, CRM and Knowledge Management. A. Zanasi, WIT Press.

Grigori, D., Caati, F., Castellanos, M., Dayal, U., Sayal, M. and Shan, M. C. (2004). "Business Process Intelligence." Computers in Industry **53**: 321-343.

Grigori, D., Casati, F., Dayal, U. and Shan, M.-C. (2001). Improving Business Process Quality Through Exception Understanding, Prediction and Prevention. Proceedings of the 27th VLDB Conference, Roma, Italy.

Grossman, D. A. (1998). Information Retrieval and Heuristics. Boston, London, kluwer Publishers.

Guh, R. S. (2005). "Real Time Pattern Recognition in Statistical Process Control: A Hybrid Neural Networks/ Decision Tree-based Approach." Proceedings IMechE, PartB: Journal of Engineering Manufacturing **219**: 283-298.

Gunn, S. R. (1997). Support Vector Machines for Classification and Regression. Southampton, University of Southampton.

Gupta, V. and Lehal, G. S. (2009). "A Survey of Text Mining Techniques and Applications." Journal of Emerging Technologies in Web Intelligence **1**(1).

Hafeez, K., Zhang, Y. and Malak, N. (2002). "Identifying Core Competence." IEEE Potencials **49**(1): 2-8.

Han, J. and Kamber, M. (2001). Data Mining: concepts and Techniques, Morgan Kaufmann Publishers.

Harding, J. A., Shahbaz, M., Srinivas and Kusiak, A. (2006). "Data Mining in Manufacturing: A Review." Journal of Manufacturing Science and Engineering, Transactions of the ASME **128**: 969-976.

Hauser, J. (2002). Challenges and Visions for Marketing's Role in Product Development Processes, Marketing Science Institute.

Hearst, M.-A. (1999). Untangling Text Data Mining. 37th Annual Meeting of the Association for Computational Linguistics.

Hill, A., Song, S., Dong, A. and Agogino, A. (2001). Indentifying Shared Understanding in Design using Document Analysis. ASME Design Engineering Technical Conference.

Hsieh, K. L. and Tong, L. I. (2001). "Optimization of Multiple Quality Response Involving Qualitative and Quantitative Characteristics in IC Manufacturing Using Neural Networks." Computers in Industry **46**: 1-12.

Hsieh, K. L., Tong, L. I., Chiu, H. P. and Yeh, H. Y. (2005). "Optimization of a Multiresponse Problem in Taguchi's Dynamic System." Computers & Industrial Engineering **49**: 556-571.

Huang, C.-C., Tseng, T.-L., Chuang, H.-F. and Liang, H.-F. (2006). "Rough-set-based Approach to Manufacturing Process Document Retrieval." International Journal of Production Research **44**: 2889-2911.

Huang, L. and Murphey, Y. L. (2006). Text Mining with Application to Engineering Diagnostics. IEA/ AIE.

Irani, K. B., Cheng, J., Fayyad, U.M., and Qian, Z. (1993). "Applying Machine Learning to Semi Conductor Manufacturing." IEEE Expert: Intelligent Systems and Their Applications archive
8(1): 41-47.

Jiao, J. and Zhang, Y. (2005). "Product Portfolio Identification Based on Association Rule Mining " Computer Aided Design **37**: 149-172.

Jiao, J., Zhang, Y., Pokharel, S. and He, Z. (2007). "Identified Generic Routings for Product Families Based on Text Mining and Tree Matching." Decision Support Systems **43**: 866-883.

Jing, H., Barzilay, R., McKeown, K. and Elhadad, M. (1998). Summarization Evaluation Methods: Experiments and Analysis. In Proceedings of AAAI'98 Spring Symposium on Intelligent Text Summarization, Stanford University, CA, AAAI Press.

Joachims, T. (1998). Text Categorization with Support Vector Machines: Learning with Many Relevant Features. Proceedings of the European Conference on Machine Learning, Dortmund, Germany.

Kalakota, R. and Robinson, M. (1999). E-business: Roadmap for Success, Addison-Wesley

Kang, H. K. and Choi, K. S. (1997). "Two-level Document Ranking Using Mutual Information in Natural Language IR." Information Processing and Management **33**: 289-306.

Karanikas, H. and Theodoulidis, B. (2002). Knowledge Discovery in Text and Text Mining Software. Manchester, UK.

Karkkainen, M., Ala-Risku, T. and Framling, K. (2003). "The Product Centric Approach: A Solution to Supply Network Information Management Problem." Computers in Industry **52**(2): 147-159.

Kasravi, K. (2004). Improving the Engineering Processes with Text Mining Proceedings of the ASME Design Engineering Technical Conferences and Computers and Inforamtion in Engineering Conference. Salt Lake City, Utah, USA.

Kusiak, A. (1990). Intelligent Manufacturing Systems, Englewood Cliffs, New Jersey: Prentice Hall,.

Kusiak, A., and Shah, S. (2006). "Data Mining Based Systems for Prediction of Water Chemistry Faults." IEEE Transactions on Industrial Electronics **15** (2): 593-603.

Kwak, C. and Yih, Y. (2004). "Data Mining Approach to Production Control in the Computer Integrated Testing Cell." IEEE Transactions on Robotics and Automation **20**(1): 107-116.

Kwon, Y., Jeong, M. K. and Omitaomu, O. A. (2006). "Adaptive Support Vector Regression Analysis of Closed-loop Inspection Accuracy." International Journal of Machine Tools& Manufacture **46**: 603-610.

Lagus, K. (2000). Text Mining with the WEBSOM. Finland, Helsinki University of Technology

Larose, D. T. (2005). Discovering Knowledge in Data: An Introduction to Data Mining. Hoboken, New Jersey, Jhon Wiley and Sons, Inc.

Last, M. and Kandel, A. (2004). "Discovering Useful and Understandable Patterns in Manufacturing Data." Robotics and Autonomous Systems **49**: 137-152.

Laudon, K. C. and Laudon, J. P. (2002). Essential of Management Information Systems. New Jersey, Prentice Hall.

Lee, C.-Y., Piramuthu, S. and Tsai, Y.-K. (1997). "Job Shop Scheduling with a Genetic Algorithm and Machine Learning." International Journal of Production Research **35**(4): 1171-1191

Leopold, E. and Kindermann, J. (2002). "Text Categorisation with Support Vector Machines.How to Represent Texts in Input Space?" Macine Learning **46**: 423-444.

Ler, W. L. (1999). "Business @ the Speed of Thought: Using a Digital Nervous Systems." Business Journal ROC.

Liao, T. W., Li, D. M. and Li, Y. M. (1999). "Detection of Welding Flaws from Radiographic Images with Fuzzy Clustering Methods." Fuzzy Sets and Systems **108**: 145-158.

Liao, T. W., Ting, C.F., and Chang, P.C. (2006). "An Adaptive Genetic Clustering Method for Explorator Mining of Feature Vector and Time Series Data " International Journal of Production Research **44**(15): 2731-2748.

Lin, F. and Hsueh, C. (2002). Knowledge Map Creation and Maintenance for Virtual Communities of Practice. Proceedings of the 36th Hawaii International Conference on System Sciences(HICSS'03), Hawaii, IEEE Computer Society, ©2002 IEEE.

Lin, F., Huang, K. and Chen, N. (2005). "Integrating Information Retrieval and Data Mining to Discover Project Team Co-ordination Patterns." Decision Support Systems.

Lin, W.-Y. and Tseng, M.-C. (2006). "Automated Support Specification for Efficient Mining of Interesting Association Rules." Journal of Information Science **32**(3): 238-250.

Liu, B., Hsu, W., Chen, S. and Ma, Y. (2000). "Analyzing Subjective Interestingness of Association Rules." IEEE Intelligent Systems **15**(5): 47-55.

Liu, Y., Lu, F. W. and Loh, H. T. (2006). A framework of Information and Knowledge Management for product design and development-A text mining Approach, The International Federation of Automatic Control (IFAC).

Loh, H.-T., Koh, W.-L., Menon, R. and Leong, C.-K. (2002). A Study of Service Centre Records Using Data Mining. Innovation in Manufacturing Systems and Technology (IMST).

MacQueen, J. (1967). Some Methods for Classification and Analysis of Multivariate Observations. Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, CA, University of California Press.

Markus, M. L., Majchrzak, A. and Gasser, L. (2002). "A Design Theory for Systems and Support Emergent Knowledge Processes." MIS Quarterly **26**(3): 179-213.

Marwick, A. D. (2001). "Knowledge Management Technology." IBM Systems Journal **40**(4): 814-830.

McErlean, F. J., Bell, D. A. and Guan, J. W. (1999). "Modification of Belief in Evidential Causal Networks." Information and Software Technology **41**: 597-603.

Mclachlan, G. J., Do, K. A. and Ambroise, C. (2004). Analysing Microarray Gene Expression Data, Jhon Wiley & Sons.

Menon, R., Tong, L. H., Sathiyakeerhi, S., Brombacher, A. and Leong, C. (2004). "The Needs and Benefits of Applying Textual Data Mining within the Product Development Process." Quality and Reliability Engineering International, **20**: 1-15.

Menon, R., Tong, L. H., Sathiyakeerithi, S. and Brombacher, A. (2003). Automated Text Classification for Fast Feedback - Investigating the Effects of Document Representation. Knowledge Based Intelligent Information & Engineering Systems. University of Oxford, United Kingdom.

Mercer, J. (1909). "Foundations of Positive and Negative Type and their Connection with the Theory of Integral Equations." Philosophical Transactions of the Royal Society: 415-446.

Miao, D., Duan, Q., Zhang, H. and Jiao, N. (2009). "Rough set based hybrid algorithm for text classification." Expert System with Applications **36**: 9168-9174.

Mitchell, T. M. (1997). Machine Learning Singapore, McGraw-Hill.

Nasukawa, T. and Nagano, T. (2001). "Text Analysis and Knowledge Mining System." IBM Systems Journal **40**(4).

Natarajan, M. (2005). "Role of Text Mining in Information Extraction and Information Management." DESIDOC Bulletin of Information Technology **25**(4): 31-38.

Ng, H. S., Toukourou, A. and Soibleman, L. (2006). "Knowledge Discovery in a Facility Condition Assessment Database Using Text Clustering." Journal of Infrastructure Systems: 50-59.

Ngu, D. S. W. and Wu, X. (1997). "SiteHelper: A Localized Agent that helps Incremental Exploration of the World Wide Web." Computer Networks and ISDN Systems **29**: 1249-1255.

Nguyen, T. T. and Skowron, A. (2004). Rough Set Approach to Domain knowledge Approximation. The 9th International Conference on Rough Sets, Fuzzy Sets, Data Mining and Granular Computing (RSFDGrC 2003).

Nonaka, I. and Takeuchi, H. (1995). The Knowledge Creation Company, Oxford University Press.

Padmanabhan, B. and Tuzhilin, A. (1998). A Belief-Driven Method for Discovering Unexpected Patterns. Fourth International Conference on Knowledge Discovery and Data Mining, ACM Press.

Park , S. C., Piramuthu, S. and Shaw, M. J. (2001). "Dynamic rule refinement in knowledge-based data mining systems." Decision Support Systems **31**: 205-222.

Peace, G. S. (1993). Taguchi Methods: A Hands-on Approach. Massachusetts, USA, Addison Wesley Reading

Pitman, B. (1991). "A System Analysis Approach to Reviewing Completed Projects." Journal of Systems Management **42**(6): 6-37.

Pohle, C. (2003). Integrating and Updating Domain Knowledge with Data Mining VLDB, Ph.D. Workshop.

Polanyi, M. (1967). The Tacit Dimension. London, U.K., Routledge & Kegan Paul.

Popescu, A.-M. and Etzioni, O. (2005). Extracting Product Features and Opinions from Reviews. Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processings (HLT/ EMNLP), Vancouver, Association for Computational Linguistics.

Powell, J. H., and Bradford, J.H. (2000). "Targetting Intelligence Gathering in Dynamic Competitive Environment." International Journal of Information Management **20**: 181-195.

Quinlan, J. R. (1992). C4.5: Programs for Machine Learning. San Francisco, CA, Morgan Kaufmann.

Rahman, S. M., Sarker, R. and Bignall, B. (1999). "Application of multimedia technology in manufacturing: a review." Computers in Industry **38**: 43-52. .

Ramesh, R., Mannan, M. A., Poo, A. N. and Keerthi, S. S. (2003). "Thermal Error Measurement and Modelling in Machine Tools. Part II. Hybrid Bayesian Network-support vector machine model." International Journal of Machine Tools and Manufacture **43**: 405-419.

Reimer, U., Margelisch, A. and Staudt, M. (2000). "EULE: A Knowledge-based System to Support Business Process." Knowledge Based Systems **13**: 261-269.

Reinschmidt, J. and Francoise, A. (2000). Business Intelligence Certification Guide. IBM, International Technical Support Organisation**:** 164.

Rezyat, M. (2000). "Knowledge-based Product Development Using XML and KCs." Computer Aided Design **32**: 299-309.

Romanowski, C. J. and Nagi, R. (2004). "A Data Mining Approach to Forming Generic Bill of Materials in Support of Variant Design Activities " ASME Journal of Computing and Information Science in Engineering **4**(4): 316-328.

Romanowski, C. J. and Nagi, R. (2005). "On Comparing Bill of Materials: A Similarity / Distance Measure of Unordered Trees." IEEE Transactions on Systems Manufacturing and Cybernatics-PartA **35**(2): 249-260.

Salton, G. (1989). Automatic Text Processing-the Transformation, Addison-Wesley: Reading, MA.

Salton, G. and Buckley, C. (1988). "Term Weighting Approaches in Automatic Text Retrieval." Information Processing and Management **24**(5): 513-523.

Samanta, B., Al-Balushi, K. R. and Al-Arimi, S. A. (2003). "Artficial Neural Network and Support Vector Machines with Genetic Algorithm for Bearing Fault Detection " Engineering Applications of Artificial Intelligence **16**: 657-665.

Saravanan, P. C., Raj, R. and Raman, S. (2003). "Summarization and Categorization of Text Data in High-level Data Cleaning for Information Retrieval." Applied Artificial Intelligence **17**: 461-474.

Scholkopf, B. and Smola, A. j. (2001). Learning with Kernels. Cambridge MA, MIT Press.

Sha, D. Y. and Liu, C. H. (2005). "Using Data Mining for Due Date Assignment in a Dynamic Job Shop Environment." International Journal of Advance Manufacturing Technology **25**: 1164-1174.

Shahbaz, M., Srinivas, Harding, J. A. and Turner, M. (2006). "Product Design and Manufacturing Process Improvement Using Association Rules." PartB: Journal of Engineering Manufacture **220**: 243-254.

Shao, X.-Y., Wang, Z.-H., Li, P.-G. and Feng, C.-X. J. (2006). "Integrating Data Mining and Rough Set for Customer Group-based Discovery of Product Configuration Rules." International Journal of Production Research **44**(14): 2789-2811.

Shen, W. and Norrie, D. H. (1999). "Agent Based Systems for Intelligent Manufacturing: A State-of-the-art survey." International Journal Knowledge and Information System **1**(2): 129-156.

Sholom, M. W., Indurkhya, N., Zhang, T. and Damerau, F. J. (2005). Text Mining Predictive Methods for Analyzing Unstructured Information New York, USA, Springer Science and Business Media, Inc.

Singh, N., Hu, C. and Roehl, W. S. (2007). "Text Mining a Decade of Progress in Hospitality Human Resource Management Research:Identifying Emerging Thematic Development." International Journal of Hospitality Management **26**(1): 131-147.

Singhal, A., Salton, G., Mitra, M. and Buckley, C. (1996). "Document Length Normalization." Information Process Management **32**: 619-633.

Skormin, V. A., Gorodetski, V.I. and Pop, Y.I.J. (2002). "Data Mining Technology for Failure of Prognostic of Avionics." IEEE Transactions on Aerospace and Electronic Systems **38**(2): 388-403.

Spertus, E. (1997). "Parasite: Mining Structural Information on the Web." Computer Networks and ISDN Systems **29**: 1205-1215.

Spinakis, A. (2001). Text Mining: A Powerful Tool for Knowledge Management**:** .

Spinakis, A. and Chatzimakri, A. (2005). "Comparative Study of Text Mining Tools." STUDIES IN FUZZINESS AND SOFT COMPUTING **185**: 223-232.

Spinakis, A. and Peristera, P. (2003). Text Mining Tools: Evaluation Methods and Criteria, Athens, Greece.

Srinivasan, P., Ruiz, M. E., Kraft, D. H. and Chen, J. (2001). "Vocabulary Mining for IR: Rough Sets and Fuzzy Sets." Information Process Management **37**: 15-38.

Tan, A.-H. (1999). Text Mining: The State of Art and the Challanges. In Proceedings, PAKDD'99 Workshop on Knowledge Discovery from Advanced Databases (KDAD'99). Beijing**:** 71-76.

Tan, S. (2005). "Neighbour-weighted K-nearest neighbour for Unbalanced Text Corpus." Expert Systems with Applications **28**: 667-671.

Tan, S. (2006). "An Effective Refinement Strategy for k-NN Text Classifier." Expert System with Applications **30**: 290-298.

Tong, L. I. and Hsieh, K. L. (2000). "A Novel Means of Applying Neural Networks to Optimize the Multiple Response Problem." Quality Engineering **13**: 11-18.

Toyryla, I. (1999). Realising the Potential of Traceability- A Case Study Research on Usage and Impacts of Product Traceability. Espoo(Finland), Helsinki University of Technology**:** 216.

van Rijsbergen, C. J. (1979). Information Retrieval (2nd edition). London, Butterworths,UK.

Vapnik, V. (1998). The Statistical Learning Theory. New York, Springer.

Vapnik, V. and Chernonenkis, A. (1964). "A Note on Class of Perceptron." Automation and Remote Control **25**.

Vapnik, V. and Lerner, A. (1963). "Pattern Recognition Using Generalised Portrait Method." Automation and Remote Control **24**.

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

Vong, C., Wong, P. and Li, Y. (2006). "Prediction of Automotive Engine Power and Torque using Least Squares Support Vector Machines and Bayesian Inference." Engineering Applications of Artificial Intelligence **19**: 277-287.

Wang, C.-H. (2008). "Outlier Identification and Market Segmentation Using Kernel-based Clustering Techniques." Expert System with Applications.

Wang, K. (2005). Applied Computational Intelligence in Intelligent Manufacturing Systems.

Wang, K. (2007). "Applying Data Mining to Manufacturing: the nature and implications." Journal of Intelligent Manufacturing **18**: 487-495.

Weiguo, F., Michael, D. G. and Praveen, P. (2004). "A Generic Ranking Function Discovery Framework by Genetic Programming for Information Retrieval." Information Processing and Management **40**: 587-602.

Weiss, S. M., Indurkhya, N., Zhang, T. and Dameran, F. J. (2005). Text Mining: Predictive Methods for Analyzing Unstructured Information, Springer Science and Business Media,Inc.

Witten, I. H. and Frank, E. (2000). Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. San Diego, California, Morgan Kaufmann Publishers.

Wood, W. H., Yang, M. C. and Cutkosky, M. R. (1998). Design Information Retrieval: Improving Access to the Informal Side of Design. In Proceeding ASME, DETC Design Theory and Methodology Conference.

Yang, M. C., Wood, W. H. and Cutkosky, M. R. (1998). Data Mining for Thesaurus Generation in Informal Desing Information Retrieval In Proceeding Conference on Computing in Civil Engineering Boston, MA, USA.

Yong, Z., Youwen, L. and Shiziong, X. (2009). "An Improved KNN Text Classification Algorithm Based On Clustering." Journal of Computers **4**(3).

Yoon, B., Phaal, R. and Robert, D. (2008). "Morphology Analysis for Technology Road Mapping: Application of Text Mining." R&D Management **38**(1): 51-68.

Yoon, S.-C., Henschen, L. J., Park, E. K. and Makki, S. (1999). Using Domain Knowledge in Knowledge Discovery. Eighth International Conference on Information and Knowledge Management.

Zahay, D. L., Griffin, A. and Frederick, E. (2003). Exploring Information use in Detail in the New Product Development Process. In Proceedings of PDMA Research Forum.

Zhou, C., Nelson, P. C., Xiao, W., Tirpak, T. M. and Lane, S. A. (2001). "An Intelligent Data Mining System for Drop Test Analysis of Electronic Products." IEEE Transactions on Electronics Packaging Manufacturing **24**(3).