# TEXTUAL DATA MINING FOR NEXT GENERATION INTELLIGENT DECISION MAKING IN INDUSTRIAL ENVIRONMENT: A SURVEY

***N. Ur-Rahman***
Wolfson School of Mechanical and Manufacturing Engineering,
Loughborough University, Loughborough, Leicestershire, UK

**Abstract**

This paper proposes textual data mining as a next generation intelligent decision making technology for sustainable knowledge management solutions in any industrial environment. A detailed survey of applications of Data Mining techniques for exploiting information from different data formats and transforming this information into knowledge is presented in the literature survey. The focus of the survey is to show the power of different data mining techniques for exploiting information from data. The literature surveyed in this paper 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. Thus the research work presented in the paper uncover the importance and power of different data mining techniques supported by text mining methods used to exploit information from semi-structured or un-structured data formats. A great source of information is available in these formats and when exploited by combined efforts of data and text mining tools help the decision maker to take effective decision for the enhancement of business of industry and discovery of useful knowledge is made for next generation of intelligent decision making. Thus the survey shows the power of textual data mining as the next generation technology for intelligent decision making in the industrial environment.

**Keywords:** Data Mining**,** Text Mining, Business Intelligence, Decision Making, Knowledge Discovery,

## Literature Review

### 1. Business Intelligence in Manufacturing: An Introduction

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, D., F. Casati, U. Dayal, M-C., Shan (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 (Brezocnik,, Balic, and Brezocnik 2003), (Powell and Bradford,2000), (Reimer, Margelisch and Staudt 2000). 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 (Hsieh and Tong 2001), (Hsieh et. al., 2005), (Tong and Hsieh 2000).

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.

## 2. Knowledge and Information Management Need 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 $18^{th}$, $19^{th}$, and the first half of the $20^{th}$ 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 $20^{th}$ century assured the formal conditions necessary for the expansion of various organizational forms. The second half of the $20^{th}$ century can thus be regarded as an era when organizational aspects were prevailing Brezocnik, Balic and Brezocnik(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) (Brezocnik, Balic, and Kuzman 2002), (Mitchell,1997), (Gen and Cheng 1997).

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, Griffin and Frederick (2003). This information extends across both internal and external sources and includes the following points: Strategic, Financial, Project management, Customer, Customer need, 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, Lu and Loh (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 sorted 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 a multitude of 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, Ala-Risku, and Framling(2003) and transmission of product information may cause overflow of information at the downstream level of a supply chain Beulens, Jansen, and Wortmann(1999). It is therefore an established fact and has been widely accepted that the future of manufacturing organisations is information theoretic and knowledge driven where information about daily manufacturing operations to built a global manufacturing environments Rahman, Sarker, and Bignall(1999).

### *2.1. 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.

2.1.1 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, Balic, and Brezocnik(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."

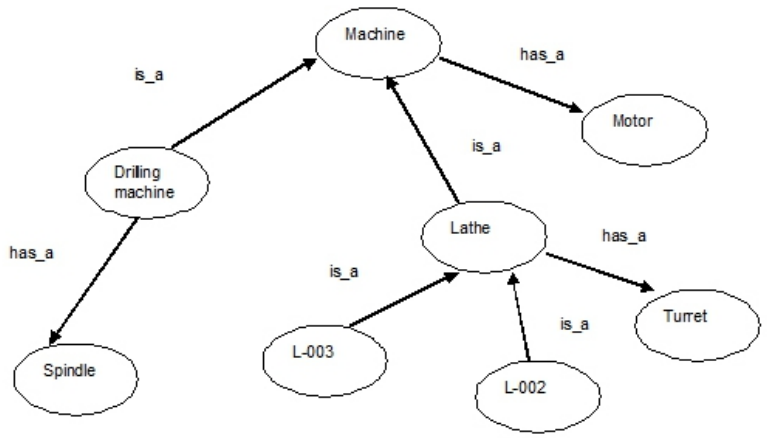


Figure1: Simple Semantic network partially representing the concept of a machine[adapted from (Kusiak, 1990)]

2.1.2 Learning through 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 up gradation, 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, Piramuthu, and Tsai (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 genetic algorithm was used at second level to select the most appropriate dispatching rules for each of the system's machines.

2.1.3 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, Balic, and Brezocnik ( 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 and Balic, 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.

### *2.2. 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 by Beckman(1999). Nowadays, the focus of researchers is to consider the emergent nature of knowledge processes Markus, Majchrzak, and Gasser (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).

#### 2.2.1 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).

2.2.2 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 (Ameri and Dutta, 2005).

2.2.3 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 & 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 & Dutta, 2005).

2.2.4 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 (Dutta & Wolowicz, 2005), (Amann, 2002), (Ameri & Dutta 2004). 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 & 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 & Siege, 2002).

2.2.5 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, Pan and Harrision(2001).

## 3. Data Mining Need for 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 & 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, Srinivas, Harding and Turner (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, Shahbaz, Srinivas and Kusiak (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.

### *3.1. 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:

In Annand, Bell and Hughes (1995) the importance of domain knowledge within Data Mining was discussed three classes of domain knowledge i.e. Hierarchical Generalization Trees (HG-Trees), Attribute Relationship Rules (AR-rules) and Environment-Based Constraints (EBC) were introduced. The research in the paper realised the importance of incorporating domain knowledge into the discovery process within the EDM (Evidential Data Mining). Djoko, Cook and Holder (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, Piramuthu and Shaw (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, Lawrence, Henschen, Park and Makki (1999) introduced a method to utilize three types of domain knowledge i.e. inter-field, 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 & 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, Hsu, Chen and Ma (2000) developed a new approach to find interesting rules (in particular unexpected rules) from a set of discovered association rules. Chia, Tan, and Sung(2006) developed a novel technique for neural logic networks (or neulonets) learning by composing net rules using genetic programming. (Lin & Tseng, 2006) proposed an automatic

support specification for efficiently mining high-confidence and positive-lift associations without consulting the users.

(Last & 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, Bell and Guan(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 contribute towards the discovery of knowledge which links rules (knowledge) and which is normally considered domain knowledge. Zhou, Nelson, Xiao and Tripak (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 & 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.

### *3.2. Decision Trees Applications for Knowledge Discovery*

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 &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, Cheng, Fayad and Qian (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, Gorodetski and Pop (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 & Liu, 2005) and rules are presented in terms of IF-THEN rules. Zhou, Nelson, Xiao and Tripak (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 PCBs and also could be used to other parts. (Kusiak & Shah, 2006) proposed a decision tree based algorithm for learning and prediction of incoming of faults of watery chemistry.

### *3.3. Clustering for Knowledge Exploration*

Clustering is a data mining technique which is used to analyse data and classify it into different classes. Chien, Wang and Chang (2007) developed a framework on the basis of k-means clustering to increase the yield of semi-conductor manufacturing. Liao, Li, and Li(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, Ting and Chang(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.

### *3.4. Association Rules for Knowledge Interpretation*

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 & Kusiak, 2004). 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, Agard and Kusiak (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, Tseng, and Wang(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 and Wu (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 & 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, Wang, Li and Feng (2006) proposed architecture to discover associations between clusters of product specifications and configurations alternatives.

### *3.5. Support Vector Machines for Knowledge Refinement*

Samanta, Al-Balushi, and Al-Arimi (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, Wong and Li (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, Asfour, Onar and Kaundinya (2005) proposed an intelligent tool breakage detection system which used Support Vector Regression (SVR) analysis to detect the process abnormalities and suggesting 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, Jeong, and Omitaomu(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, Mannan and Poo (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.

### *3.6. Hybrid Approaches for Knowledge Mining*

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.

## 4. Information Retrieval 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 users 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, Ruiz, Kraft and Chen (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, Buckley and Mitra (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 by Srinivasan et el. (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 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, Michael, and Praveen(2004), thus reducing their satisfaction of search result. Therefore, increasing the effectiveness of retrieval algorithms remains an important goal discussed by 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.

### *4.1. Data Mining to Support IR Based Solutions for Textual Data Analysis*

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, Huang and Chen (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, Tseng, Chuang and Liang (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, Raj and Raman (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.

### *4.2. Textual Data Analysis for Manufacturing Knowledge*

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 , Damerau and Weiss (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.

### *4.3. Manufacturing Product or 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 (2004). 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, Zhang, Pokharel and He (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, Song, Dong and Agogino (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 (www.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.

In Yang, Wood and Cutkosky (1998) it is discussed that 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, Yang and Cutkosky (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, Zatorsky and Nayak (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, Tong, Sathiyakeerthi

and Brombacher (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, Phaal, and Robert (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.

The diverse nature of problems in industrial manufacturing environments e.g. predictive warranty analysis, quality improvements, patent analysis, competitive assessments, FMEA, and product searches text mining methods can help to uncap the vast amount of information buried in textual data. Kasravi (2004) discussed various application domains in the engineering sector and specifically improvement of processes through tracking of information from top to downstream levels and complex data analysis.

### *4.4. Business Improvement through Customer Knowledge Management*

The business process quality improvement issue has been handled in Grigori, Casati, Castellanos, Dayal, Sayal and Shan (2004) 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 el 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, Lin, and Wang (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.

## 5. Summary and Conclusion

This paper describe the detailed applications of textual data mining techniques for discovering useful knowledge from different data formats. The paper detailed about handling of information available in different data formats and this information has been exploited through applications of different data mining and text mining techniques. Thus paper discusses about the prevalent technological efforts in the industrial setups and their limitations in handling with the requirements of next generation intelligent manufacturing. The need of textual data mining in handling multiple different data formats to extract useful information and meeting with the requirements of IT based manufacturing or industrial environment has been discussed in detail. The core concept of sustainable knowledge management lies in exploiting information from data available in any industrial setup either in semi-structured or unstructured data formats and therefore these information need to be uncovered by textual data mining techniques. This concept is supported through defining different functionalities of data mining and their power to handle the data or information sources by surveying the literature.

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 designs, operational and manufacturing contexts. However, to date, most of these applications have used numerical and / or structured knowledge sources for which the detailed survey was done by Harding,

Shahbaz, Srinivas and Kusiak (2006). 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 so if these technological efforts are used to exploit the information then better decision would be made possible and better knowledge management solutions are expected to be achieved.

**Acknowledgements**

The author acknowledges the Loughborough University for awarding the PhD scholarship to conduct research activities in the Wolfson School of Mechanical and Manufacturing Engineering.

**References:**

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

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

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

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

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

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

Hsieh, K.L., et al.(2005), *Optimization of a Multiresponse Problem in Taguchi's Dynamic System.* Computers & Industrial Engineering, **49**: p. 556-571.

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

Brezocnik, M., J. Balic, and K. Kuzman(2002), *Genetic Programming Approach to Determining of Metal Material Properties.* Journal of Intelligent Manufacturing, **13**: p. 5-17.
Mitchell, T.M., *Machine Learning* (1997), Singapore: McGraw-Hill.
Gen, M. and R. Cheng(1997), *Genetic Algorithms and Engineering Design*., Canada: Jhon Wiley and Sons.
Zahay, D.L., A. Griffin, and E. Frederick(2003). *Exploring Information use in Detail in the New Product Development Process*. in *In Proceedings of PDMA Research Forum*.
Liu, Y., F.W. Lu, and H.T. Loh.(2006) *A framework of Information and Knowledge Management for product design and development-A text mining Approach*: The International Federation of Automatic Control (IFAC).
Toyryla, I.(1999), *Realising the Potential of Traceability- A Case Study Research on Usage and Impacts of Product Traceability*., Helsinki University of Technology: Espoo(Finland). p. 216.
Karkkainen, M., T. Ala-Risku, and K. Framling(2003), *The Product Centric Approach: A Solution to Supply Network Information Management Problem.* Computers in Industry, **52**(2): p. 147-159.
Beulens, A.J.M., M.H. Jansen, and J.C. Wortmann. (1999)*The Information de-coupling point in Global Production Management, IFIP WG 5.7*. in *International Conference on Advances in Production Management Systems*. Boston: Kluwer Academic Publishers.
Rahman, S.M., R. Sarker, and B. Bignall,(1999) *Application of multimedia technology in manufacturing: a review.* Computers in Industry, **38**: p. 43-52.
Kusiak, A.,(1990) *Intelligent Manufacturing Systems*.: Englewood Cliffs, New Jersey: Prentice Hall,.
Lee, C.-Y., S. Piramuthu, and Y.-K. Tsai, (1997) *Job Shop Scheduling with a Genetic Algorithm and Machine Learning.* International Journal of Production Research, **35**(4): p. 1171-1191
Brezocnik, M., and Balic, J., (2001) *Genetic-based Approach to Simulation of Self-Organizing Assembly.* Robotics Computer Integrated Manufacturing, **17(1-2)**: p. 113-120.
Davenport, T.H. and L. Prusak,(1998) *Working Knowledge: How Organisations Manage What they Know*: Harvard Business School Press.
Polanyi, M., *The Tacit Dimension*. (1967), London, U.K.: Routledge & Kegan Paul.
Nonaka, I. and H. Takeuchi, *The Knowledge Creation Company*. (1995): Oxford University Press.
Beckman, T.J.,(1999) *The Current State of Knowledge Management in Liebowitz* Knowledge Management Handbook.: CRC Press.

Markus, M.L., A. Majchrzak, and L. Gasser, (2002) *A Design Theory for Systems and Support Emergent Knowledge Processes.* MIS Quarterly, **26**(3): p. 179-213.
Earl, M., (2001) *Knowledge Management Strategies Towards a Taxonomy*. Journal of Management Information Systems, **18**(1): p. 215-233.
Dhaliwal, J.S. and I. Benbasat, (1996) *The Use and Effects of Knowledge-based System Explanations: Theoretical Foundations and a Framework for Emperical Evaluation.* Information System Research, **7**(3): p. 342-362.
Laudon, K.C. and J.P. Laudon, (2002) *Essential of Management Information Systems*. 5 ed., New Jersey: Prentice Hall.
Shen, W. and D.H. Norrie, (1999) *Agent Based Systems for Intelligent Manufacturing: A State-of-the-art survey.* International Journal Knowledge and Information System, **1**(2): p. 129-156.
Ameri, F. and D. Dutta, (2005) *Product Lifecycle Management: Closing the Knowledge Loops.* Computer Aided Design and Applications, **2**(5): p. 577-590.
Bloor, M.S. and J. Owen, (1991) *CAD/CAM Product-data Exchange: the next step.* Computer Aided Design **23**: p. 237-243.
Dutta, D. and J.P. Wolowicz, (2005) *An Introduction of Product Lifecycle Management(PLM),* in *12th ISPE International Conference on Concurrent Engineering: Research and Applications*.: Dallas.
Gascoigne, B., (1995) *PDM: The Essential Technology for Concurrent Engineering.* World Class Design to Manufacture, **2**(1): p. 38-42.
Amann, K.,(2002) *Product Lifecycle Management: Empowering the Future of Business*.: CIM Data, Inc.
Ameri, F. and D. Dutta, (2004) *Product Lifecycle Managment needs, concepts and components*, University of Michigan, Ann Arbor, MI. p. 2.
Abramovici, M. and O.C. Siege.(2002) *Status and Development Trends of Product Lifecycle Management Systems*. in *In Proceedings of IPPD*. Wroclaw, Poland.
Rezyat, M.,(2000) *Knowledge-based Product Development Using XML and KCs.* Computer Aided Design **32**: p. 299-309.
Cheng, K., P.Y. Pan, and D.K. Harrision, (2001)*Web-based Design and Manufacturing Support Systems: implementation perspectives.* International Journal of Computer Integrated Manufacturing, **14**(1): p. 14-27.
Han, J. and M. Kamber, (2001) *Data Mining: concepts and Techniques*, Morgan Kaufmann Publishers.
Shahbaz, M., Srinivas, J.A. Harding, and M. Turner (2006) *Product Design and Manufacturing Process Improvement Using Association Rules.* PartB: Journal of Engineering Manufacture, **220**: p. 243-254.
Wang, K., (2007) *Applying Data Mining to Manufacturing: the nature and implications.* Journal of Intelligent Manufacturing, **18**: p. 487-495.

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

Wang, K., (2005) *Applied Computational Intelligence in Intelligent Manufacturing Systems*. International Series on Natural and Artificial Intelligence. 2005.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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): p. 388-403.

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

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

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): p. 192-198.

Liao, T.W., D.M. Li, and Y.M. Li, (1999) *Detection of Welding Flaws from Radiographic Images with Fuzzy Clustering Methods.* Fuzzy Sets and Systems,**108**: p. 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): p. 2731-2748.

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

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

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

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**: p. 333-343.

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

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

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

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

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

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

Ramesh R.,M.A. Mannan and A.N. Poo (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**: p. 405-419.

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

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**: p. 283-298.

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

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

Salton, G., (1989) *Automatic Text Processing-the Transformation*. Analysis and Retrieval of Information by Computer: Addison-Wesley: Reading, MA.
Gordon, M. and P. Pathak, (1999) *Finding Information on the World Wide Web: The Retrieval Effectiveness of Search Engines.* Information Processing and Management, 1999. **35**: p. 141-180.
Kang, H.K. and K.S. Choi, (1997) *Two-level Document Ranking Using Mutual Information in Natural Language IR.* Information Processing and Management,**33**: p. 289-306.
Weiguo, F., D.G. Michael, and P. Praveen, (2004) *A Generic Ranking Function Discovery Framework by Genetic Programming for Information Retrieval.* Information Processing and Management,**40**: p. 587-602.
Bordogna, G. and G. Pasi, (1995) *Controlling Retrieval Through a User-adaptive Representation of Documents.* International Journal of Approximate Reason,**12**: p. 317-339.
Lin, F., K. Huang, and N. Chen, (2005) *Integrating Information Retrieval and Data Mining to Discover Project Team Co-ordination Patterns.* Decision Support Systems.
Lin, F. and C. Hsueh. (2002) *Knowledge Map Creation and Maintenance for Virtual Communities of Practice*. in *Proceedings of the 36th Hawaii International Conference on System Sciences(HICSS'03).* Hawaii: IEEE Computer Society, ©2002 IEEE.
Tan, S., (2005) *Neighbour-weighted K-nearest neighbour for Unbalanced Text Corpus.* Expert Systems with Applications,**28**: p. 667-671.
Tan, S., (2006) *An Effective Refinement Strategy for k-NN Text Classifier.* Expert System with Applications, **30**: p. 290-298.
Huang, C.-C., T-L. Tseng, H-F. Chuang and H-F. Liang (2006) *Rough-set-based Approach to Manufacturing Process Document Retrieval.* International Journal of Production Research, **44**: p. 2889-2911.
Saravanan, P.C., R. Raj, and S. Raman, (2003) *Summarization and Categorization of Text Data in High-level Data Cleaning for Information Retrieval.* Applied Artificial Intelligence, **17**: p. 461-474.
Spertus, E., (1997) *Parasite: Mining Structural Information on the Web.* Computer Networks and ISDN Systems, **29**: p. 1205-1215.
Ngu, D.S.W. and X. Wu, (1997) *SiteHelper: A Localized Agent that helps Incremental Exploration of the World Wide Web.* Computer Networks and ISDN Systems,**29**: p. 1249-1255.
Apte, C., F. Damerau, and S. Weiss, (1994) *Automatted Learning of Decision Rules of Text Categorization.* ACM Transactions on Information Systems, **12**(3): p. 233-251.
Romanowski, C.J. and R. Nagi, (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): p. 316-328.
Romanowski, C.J. and R. Nagi, (2005) *On Comparing Bill of Materials: A Similarity / Distance Measure of Unordered Trees.* IEEE Transactions on Systems Manufacturing and Cybernatics-PartA, **35**(2): p. 249-260.
Agard, B. and A. Kusiak, (2004a) *Data Mining-based Methodology for Product Families.* International Journal of Production Research, **42**(15): p. 2955-2969.
Jiao, J., L. Zhang, S. Pokharel and Z. He (2007) *Identified Generic Routings for Product Families Based on Text Mining and Tree Matching.* Decision Support Systems,**43**: p. 866-883.
Hill, A., S.Song, A. Dong and A. Agogino (2001) *Indentifying Shared Understanding in Design using Document Analysis,* in *ASME Design Engineering Technical Conference*. 2001.
Yang, M.C., W.H. Wood, and M.R. Cutkosky. (1998) *Data Mining for Thesaurus Generation in Informal Desing Information Retrieval* in *In Proceeding Conference on Computing in Civil Engineering.* Boston, MA, USA.
Wood, W.H., M.C. Yang, and M.R. Cutkosky. (1998) *Design Information Retrieval: Improving Access to the Informal Side of Design*. in *In Proceeding ASME, DETC Design Theory and Methodology Conference.*
Dong, A. and A. Agogino, (1997) *Text Analysis for Constructing Design Representations* Artificial Intelligence in Engineering, 1997. **11**(2): p. 65-75.
Fong, A.C.M. and S.C. Hui, (2001) *An Intelligent Online Machine Fault Diagnosis Systems.* Computing and Control Engineering Journal,**12**(5): p. 217-223.
Popescu, A.-M. and O. Etzioni. (2005) *Extracting Product Features and Opinions from Reviews*. in *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processings (HLT/ EMNLP)*. Vancouver: Association for Computational Linguistics.
Edwards, B., M. Zatorsky, and R. Nayak, (2008) *Clustering and Classification of Maintenance Fault Logs Using Text Data Mining* in *Australian Conference on Data Mining*: Aus DM.
Menon, R., L.H. Tong, S. Sathiyakeerthi and A. Brombacher (2003) *Automated Text Classification for Fast Feedback - Investigating the Effects of Document Representation*, in *Knowledge Based Intelligent Information & Engineering Systems*: University of Oxford, United Kingdom.
Yoon, B., R. Phaal, and D. Robert, (2008) *Morphology Analysis for Technology Road Mapping: Application of Text Mining.* R&D Management,**38**(1): p. 51-68.

Huang, L. and Y.L. Murphey, (2006) *Text Mining with Application to Engineering Diagnostics*, in *IEA/ AIE*.
Kasravi, K., (2004) *Improving the Engineering Processes with Text Mining* in *Proceedings of the ASME Design Engineering Technical Conferences and Computers and Inforamtion in Engineering Conference*: Salt Lake City, Utah, USA.
Grigori, D.,F. Casati, M. Castellanos, U. Dayal, M. Sayal and M-C. Shan (2004) *Business Process Intelligence.* Computers in Industry, **53**: p. 321-343.
Chang, C.-W., C.-T. Lin, and L.-Q. Wang, (2009) *Mining the Text Information to Optimize the Customer Relationship Management.* Expert System with Application,**36**: p. 1433-1443.
Grievel, L., (2005) *Customer Feedbacks and Opinion Surveys Analysis in the Automotive Industry*, in *Text Mining and its Applications to Intelligence, CRM and Knowledge Management*, A. Zanasi, Editor , WIT Press.