Expert Systems with Applications 39 (2012) 4729–4739

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Textual data mining for industrial knowledge management and text classification: A business oriented approach

N. Ur-Rahman, J.A. Harding *

Wolfson School of Mechanical and Manufacturing Engineering, Loughborough University, Loughborough, Leicestershire LE11 3TU, UK

ARTICLE INFO

Keywords:
Textual data mining
Text mining
Post Project Reviews

ABSTRACT

Textual databases are useful sources of information and knowledge and if these are well utilised then issues related to future project management and product or service quality improvement may be resolved. A large part of corporate information, approximately 80%, is available in textual data formats. Text Classification techniques are well known for managing on-line sources of digital documents. The identification of key issues discussed within textual data and their classification into two different classes could help decision makers or knowledge workers to manage their future activities better. This research is relevant for most text based documents and is demonstrated on Post Project Reviews (PPRs) which are valuable source of information and knowledge. The application of textual data mining techniques for discovering useful knowledge and classifying textual data into different classes is a relatively new area of research. The research work presented in this paper is focused on the use of hybrid applications of text mining or textual data mining techniques to classify textual data into two different classes. The research applies clustering techniques at the first stage and Apriori Association Rule Mining at the second stage. The Apriori Association Rule of Mining is applied to generate *Multiple Key Term Phrasal Knowledge Sequences (MKTPKS)* which are later used for classification. Additionally, studies were made to improve the classification accuracies of the classifiers i.e. C4.5, *K*-NN, Naïve Bayes and Support Vector Machines (SVMs). The classification accuracies were measured and the results compared with those of a single term based classification model. The methodology proposed could be used to analyse any free formatted textual data and in the current research it has been demonstrated on an industrial dataset consisting of Post Project Reviews (PPRs) collected from the construction industry. The data or information available in these reviews is codified in multiple different formats but in the current research scenario only free formatted text documents are examined. Experiments showed that the performance of classifiers improved through adopting the proposed methodology.

1. Introduction

In the current digital based economy a large amount of information is available in the form of textual data which can often be used more easily if it is categorised or classified into some predefined classes (Miao, Duan, Zhang, & Jiao, 2009). In any business or industrial environment corporate information may be available in multiple different formats, about 80% of which is in text documents (Yu, Wang, & Lai, 2005). This information exists in the form of descriptive data formats which include service reports about repair information, manufacturing quality documentation and customer help desk notes (Kornfein & Goldfrab, 2007). It is also often in the form of concise text formats, containing many industry specific terms and abbreviations. Both technical and manual efforts are needed to handle these information sources, to unearth the patterns and discover useful knowledge hidden within these resources (Kornfein & Goldfrab, 2007). Transformation of these useful sources of information into usable formats will help to improve future product or service quality and provide solutions to project management issues. Decision makers or knowledge workers may therefore be assisted and business decisions improved through the discovery of useful knowledge patterns. Identified knowledge can also be transferred from one project to another. This will ultimately help to cut the overhead costs of product or service quality improvement and project management. Therefore the purpose of these studies is to try to improve any business context where useful knowledge of previous experience can be discovered in reports or other documents. For example, if customers' needs can be identified and classified then better future decisions can be made resulting in improved levels of customer satisfaction.

The overall process of knowledge discovery in databases (KDD) is the identification of valid, novel, potentially useful and ultimately understandable patterns in data (Fayyad, Piatetsky-Shapiro,

* Corresponding author.
E-mail address: J.A.Harding@lboro.ac.uk (J.A. Harding).

0957-4174/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2011.09.124

& Smyth, 1996). The term knowledge discovery from textual databases (KDT) is a little different to the general form of KDD and can be defined as discovering useful information and knowledge from textual databases through the application of data mining techniques (Han & Kamber, 2000; Karanikas & Theodoulidis, 2002). However it shares the common methods of collecting information as raw data and processing it through the application of data mining techniques. Indeed, a three step process of data collection, pre-processing and applications of text mining techniques (Karanikas & Theodoulidis, 2002) is required.

Text classification is an important approach to handling textual data or information in the overall process of knowledge discovery from textual databases. It has been a most promising area of research since the inception of the digital text based economy (Ikonomakis, Kotsiantis, & Tampakas, 2005). It is mainly used to classify text documents into predefined categories or classes based upon content and labelled training samples (Jinshu, Bofeng, & Xin, 2006). Text mining techniques have been widely used in various application fields like e-mail filtering, document management, customer needs identification, etc. It can therefore be concluded that the use of this technology can help to access information and manage it for better use in future applications.

Applications of data mining techniques have long been seen to improve predictive and classification methods and have widely been used in different subject areas ranging from finance to health sciences. Quite a few applications of these techniques have been reported in manufacturing or construction industry environments. There may however be problems of non-availability of data due to some confidentiality, proprietary and sensitivity issues (Wang, 2007). This leads to the exploitation of data mining techniques to handle textual databases being less frequently reported in the literature.

The research work reported in this paper proposes a new hybridised method of handling textual data formats and classifying the text documents into two different classes. The effectiveness of the proposed methodology is demonstrated with the help of a case study taken from a real industrial context. The new approach adopted within this research will help to uncover useful information in terms of *Multiple Key Term Phrasal Knowledge Sequences (MKTPKS)* which can then be used for the classification of Post Project Reviews (PPRs) into good or bad information-containing documents. Focus has been put on the application of different classifiers such as Decision Trees, Naïve Bayesian learner, *K*-NN classifiers and Support Vector Machines (SVMs) to test the usefulness of the proposed model. The results obtained are also compared with simple bag of words (BoW) representation models and the *F*-measure is used as the quantitative metric for measuring the effectiveness of the model.

The remainder of this paper is organised as follows: Section 2 provides the background for Text Classification methods and related work reported in the literature for industrial knowledge management solutions. Sections 3 discusses the proposed methodology and architecture, and different methods incorporated within this methodology. Section 4 discusses an implementation of the proposed methodology, based on real industrial data in the form of PPRs, and its classification results. Conclusions and future work are discussed in Section 5.

2. Text classification background and related work

Text classification methods were first proposed in the 1950s where the word frequency was used to classify documents automatically. In the 1960s the first paper was published on automatic text classification and from then until the early 1990s it became a major sub-field of information systems (Weiss, Indurkhya, Zhang, Damerau, 2005). Applications of machine learning techniques helped to reduce the manual effort required in analysis and the accuracy of the systems also improved through use of these techniques. Many text mining software packages are available on the market and these can be used to perform different tasks of handling textual databases and classifying them to discover useful information (Tan, 1999). Substantial 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, Youwen, & Shiziong, 2009).

Identification of useful information from textual databases through the application of different data mining techniques has long been widely used in various application domains. However there are less reported applications in industrial contexts which implies that industrial databases have not been fully utilised to explore information and transform it into useful knowledge sources. A few instances of text mining and classification techniques have been reported in the engineering domain. For example, the application of classification techniques has been explored to classify manufacturing quality defects and service shop data sets (Kornfein & Goldfrab, 2007). A new probabilistic term weighting scheme was introduced in Liu, Loh, and Sun (2009) to handle an imbalanced textual data set of Manufacturing Corpus Version1 (MCV1) related to manufacturing engineering papers. The weighting scheme helped to classify data into predefined categories or classes with measurable accuracies regardless of the classifiers used. This ultimately helped to provide an effective solution to improve the performance of imbalanced text categorisation problems. An incremental algorithm was introduced in Sanchez, Triantaphyllou, Chen, and Liao (2002) to learn a Boolean function (i.e. positive or negative) in an environment where training data examples which have already been divided into two mutually exclusive classes are assumed to be available. The proposed function was combined with an existing algorithm OCAT (one clause at a time) and tested on the TIPSTER (a project lead by Defence Advanced Research Projects Agency (DARPA)) textual data set. The empirical results were found to be effective and efficient in such learning environments.

A TAKMI (Text Analysis and Knowledge Mining) system was proposed in Nasukawa and Nagano (2001) to handle PC help centres databases in order to detect the issues of product failures and identify customer behaviours related to particular products. Empirical studies were carried out to detect signals of interest from World Wide Web data to help an organisation to take effective decisions (Aasheim & Koehler, 2006). A combination of a vector space model, linear discriminant analysis, environmental scanning (a method for obtaining and using information from an organisations external environment) and text classification methods were studied to determine their effect in helping the decision making process of an organisation. A study was made on a textual database available in a pump station maintenance system with the aim of classifying it into scheduled and unscheduled repair jobs (Edwards, Zatorsky, & Nayak, 2008). Textual data mining techniques have also been used to resolve the quality and reliability issues in the manufacture of new products (Menon, Tong, & Sathiyakeerthi, 2005). Applications of text mining techniques, for developing a knowledge based product by considering the potential international, inter-cultural end user views, are discussed in Haravu and Neelameghan (2003). The study suggested that the concept terms from processed text can be linked to a related thesaurus, glossary, schedules of classification schemes, and facet structured subject representations. Text mining techniques were used to diagnose engineering problems in the automotive industry and to map them into their correct categories using text document classification and

term weighting schemes (Huang & Murphey, 2006). Diverse applications of text mining techniques have also been reported in Kasravi (2004), including predictive warranty analysis, quality improvements, patent analysis, competitive assessments, FMEA and product searches.

In the construction industry, a methodology based on hierarchical document classification was developed to improve information organisation and its accessibility to classify project documents according to their components (Caldas & Soibelman, 2003). A machine learning approach was used to classify construction project documents according to their related project components to improve the information organisation and inter organisational system (Caldas, Asce, Soibelman, Asce, & Han, 2002).

2.1. Textual data mining for industrial knowledge management

Data mining technology provides flexibility to exploit information from multiple different data formats or databases such as relational databases, data warehouse and transactional databases etc. Text based databases can contain information in the form of papers, reports, web pages, messages, notes etc. which can be unstructured, semi-structured or structured (Han & Kamber, 2000). Text mining can be defined as textual data mining or knowledge discovery from textual databases. Although the text mining process relies heavily on applications of data mining techniques for discovering useful knowledge, it is also focused on handling more unstructured data formats which pose more challenges for pattern discovery than numerical data formats do (Tan, 1999).

A standard data mining process modelling approach known as CRISP-DM was developed in 1996 by a group of analysts to discover valuable knowledge from data (Chapman et al., 2000). A standard text mining process mainly consists of three different stages i.e. text preparation, text processing and text analysis (Natarajan, 2005). The iterative procedures that have been adapted to discover valuable knowledge from textual data formats are shown in Fig. 1. The information available in the form of textual data 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 find 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.

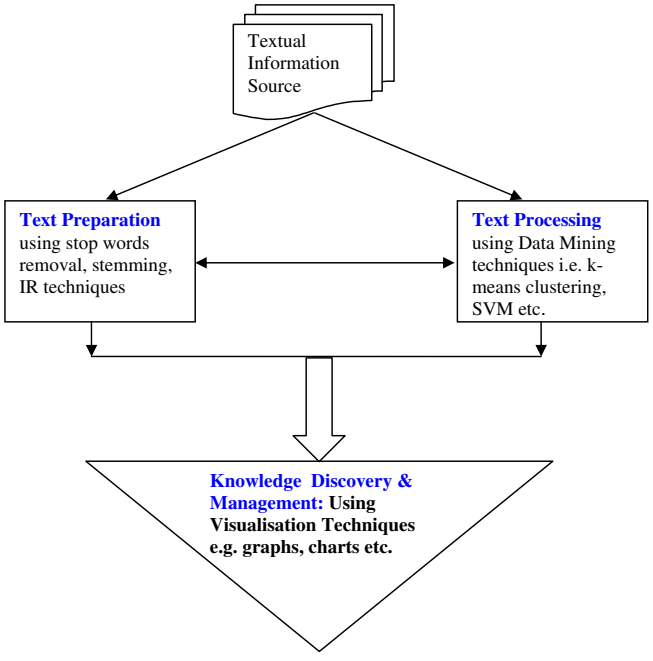

Fig. 1. Text mining process as interactive and iterative procedures.

Data mining techniques are less efficient at handling databases where information is available in unstructured data formats. However, solutions can be provided by other techniques such as data warehouse, multidimensional models and ad hoc reports, although these techniques are unable to cover the full scope of business intelligence (Berry & Linoff, 2004). Text mining methods can give the additional advantages of better management of knowledge resources and knowledge management activities (Hafeez, Zhang, & Malak, 2002).

Text mining is a term for discovering useful knowledge to help in processing information and improving the productivity of knowledge workers. It consequently can add value to a business by facilitating the process of decision making at less cost than other text processing techniques (Spinakis & Chatzimakri, 2005). To gain more competitive advantages and exploit multiple information sources, knowledge discovery techniques need to be considered. Therefore more attention should be paid to text mining techniques in business intelligence solutions (Gao et al., 2005; Nasukawa & Nagano, 2001). The downstream knowledge discovery and management process to gain competitive business advantages is shown in Fig. 2.

2.2. Problem description and objective of research

Text analysis and classification can help to identify key issues which can ultimately play an effective role in the future decision making process in many industrial contexts. The information specific to a particular product or service quality issue may be available in the form of numerical or textual data formats or through colour coding, abbreviations and special text features. Better project management, reduced product lead time to the market and improved customer satisfaction levels or service quality are major reasons for revisiting data stored in current manufacturing systems. Revisiting manufacturing systems using text analysis and classification techniques can enable more cost effective and efficient decisions to be made to better achieve the current and future requirements.

Text classification methods help to classify textual data into different classes and classification of data into two different classes is called binary classification. Training sets are identified manually to support the application of different data mining algorithms or classifiers to automate the system of classification. Thus manual efforts help to make the data ready for the task of learning, which requires the help of domain experts.

The hypothesis made within this research work is that, "The generation and use of multiple key term phrasal knowledge sequences (MKTPKS) for classifying documents into two different classes will improve the classification accuracy compared to the single term based classification models."

Therefore the objectives of the current research work can be enumerated as follows:

- To apply textual data mining techniques for capturing a first level of knowledge in terms of single key term phrases and then generate MKTPKS to represent useful knowledge discovered through the application of clustering techniques.
- To study the effects of different textual representation models on the classification of data and suggest improvement methods to acquire better classification accuracies from the classifiers. Further, to validate the hypothesis that the MKTPKS based classification model gives better classification accuracies.

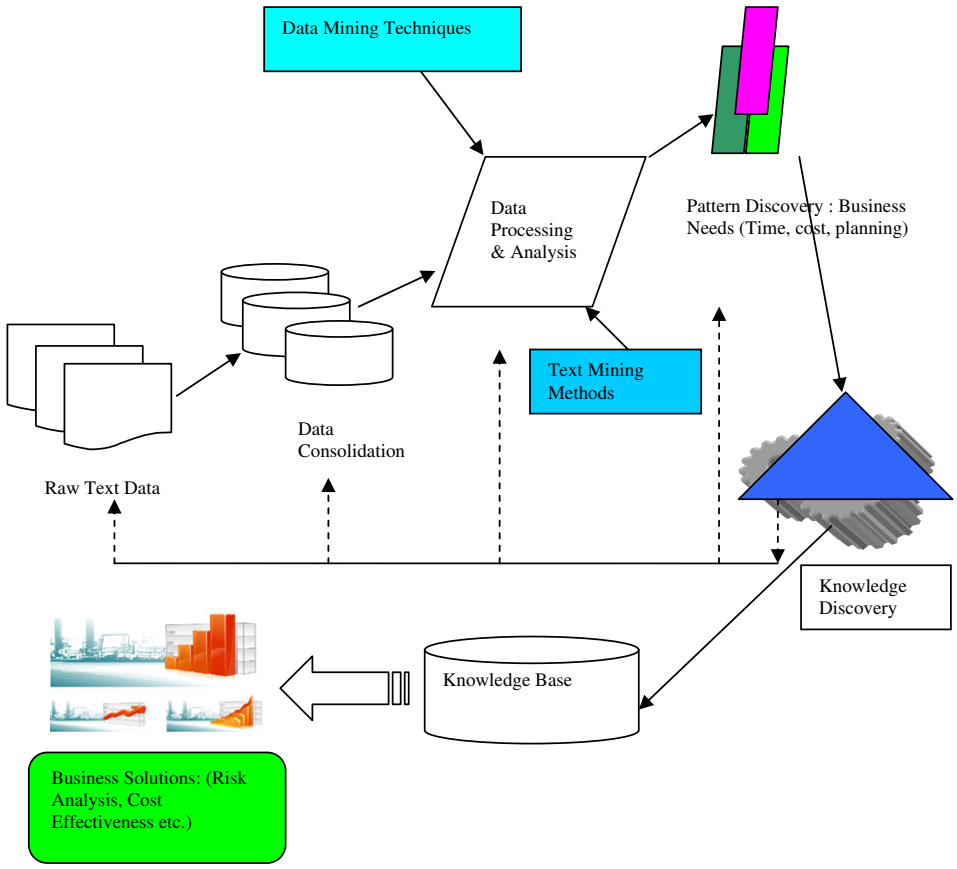

Fig. 2. Textual data mining for downstream knowledge discovery and management solutions.

3. Proposed methodology and architecture

In this section a system is proposed to analyse textual databases and classify their content into two different classes. The proposed three level system incorporates the different functionalities of a Text Mining (Data or Information Processing) Unit, a 1st Level Knowledge Processing and Storing Unit, a 2nd Level Knowledge Refinement Unit and finally a 3rd Level Knowledge Utilisation and Text Classification Unit. These are illustrated in Fig. 3, which also shows the flow of information and knowledge from different parts of the systems to generate summaries of the text in terms of finding multiple key term phrasal knowledge sequences (MKTPKS) and then classify the documents based on these MKTPKS.

The detailed description of the sequence of activities is given in the following subsections.

3.1. Text Mining (Data or Information Processing) Unit

The first step towards handling and analysing textual data formats in general is to consider the text based information available in free formatted text documents. Commonly this information would be processed manually by reading thoroughly and then human domain experts would decide whether the information was good or bad (positive or negative). This is expensive in relation to the time and effort required from the domain experts. To begin the automated text classification process the input data needs to be represented in a suitable format for the application of different textual data mining techniques, this includes stop words removal and simple stemming functions as explained below.

To achieve the objective of making data suitable for further analysis through the applications of different data mining techniques, the first step is to remove the un-necessary information available in the form of stop words. These include some verbs, conjunctions, disjunctions and pronouns, etc. (e.g. is, am, the, of, an, we, our). These words need to be removed as they are less useful in interpreting the meaning of text. Stemming is defined as the process of conflating the words to their original stem, base or root. In this study simple stemming is applied where words e.g. '*deliver*', '*delivering*' and '*delivered*' are stemmed to '*deliver*'. This method helps to capture a whole information carrying term space and also reduces the dimensions of the data which ultimately affects the classification task.

The next step is to represent the textual data in a matrix form where each row vector contains the terms and each column vector contains the corresponding document identification codes (IDs). To reduce the effect of loosing key information at this stage of text data representation, well known bag of words (BoW) approaches have been used as these methods consider the whole information space for the analysis. These methods are independent of text structures and each word is considered as an independent entity containing some information. They are also commonly reported in literature and have been adopted in many studies due to their simplicity and effectiveness (Salton, 1989).

3.2. 1st Level Knowledge Processing and Storing Unit

This part of analysing textual data helps the application of different data mining algorithms, as the data is given a suitable representation based on words or terms defined in the text. Different data representations methods i.e. term frequency (TF), in-

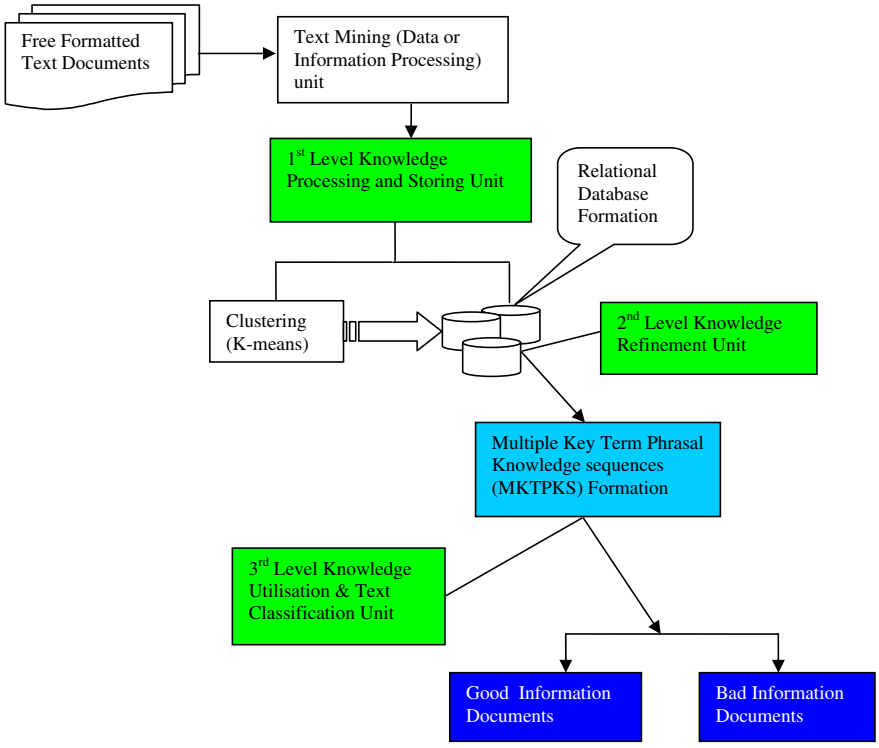

Fig. 3. MKTPKS based knowledge management and text classification system.

verse document frequency (IDF) and term frequency and inverse document frequency methods may be used at this level of information processing. The selection of a suitable representation should be made through extensive experimentation to retain the whole information space and overcome the possible loss of useful information. Therefore a term frequency based matrix representation was used which helped to retain useful information and reduced the effect of key information loss. Then the whole document space of information was represented in the matrix form. Different data mining algorithms may be applied on the selected representation of the textual data to discover 1st level knowledge. However, the focus in the current research is on the application of Clustering Techniques to partition the data into useful subsets of information within each cluster.

3.2.1. Clustering

Clustering is defined as a process of grouping data or information into groups of similar types using some physical or quantitative measures (Larose, 2005). These quantitative measures are based on different distance functions measured from a point called the centroid of the cluster. Different clustering techniques were tested to find the similarities between terms and the *K*-means clustering technique was found to be good for dividing the useful information into multiple subspaces. *K*-means clustering was ultimately used to discover the natural relationships between terms to further capture an initial level of knowledge. The similarities between terms are measured on the basis of Euclidean distance given by Eq. (1):

$$D(x,y) = \sqrt{\sum_{i=1}^{m}(x_i - y_i)^2} \tag{1}$$

3.2.2. Relational database formation

The output from the application of *K*-means clustering is stored in a usable format in the form of different relational tables. The tables consist of columns with clusters IDs, the terms identified within each cluster and cluster labels which are further used in the process of pruning the key information or knowledge discovered. This function helps to store and manage the information for further analysis.

3.3. 2nd Level Knowledge Refinement Unit

The input to this unit is in the form of relational tables where documents are taken as transactions and terms are taken as items. The process of refining the key information or knowledge and generating MKTPKS is performed through applications of Apriori Association Rule of Mining (Agrawal, Lmielinski, & Swami, 1993). Generating these MKTPKS forms an essential part of the text analysis used in the classification of text documents. The generation of MKTPKS is preferred over finding the association rules because identification of too many rules may overpopulate the knowledge bases. MKTPKS can also help to discover more valuable relationships among terms defined in the text. The co-occurrence of terms helps to find the relationships among the different concepts defined in the text documents. Mapping the discovered MKTPKS to a particular set of documents helps to identify document sets containing good or bad information. The output of this unit is therefore obtained in terms of MKTPKS which are the consequences of key single term based knowledge refined through the application of the Apriori Association Rule of Mining technique within the 2nd Level of Knowledge Refinement Unit.

3.4. 3rd Level Knowledge Utilisation & Text Classification Unit

Data is mainly stored in semi-structured databases i.e. neither fully structured nor unstructured in nature (Han & Kamber, 2000). To classify textual data into predefined classes it is necessary to partition it 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 attribute or target variable. The given data is therefore first divided into two different classes with the help of a domain expert who has sufficient knowledge to interpret the terms defined in the textual databases. In this research this task was performed with the help of domain experts who had clear understanding of the context of the textual data and the meaning of the terms defined within the text documents.

In this 3rd Level unit different classifiers were used to study their effect in terms of the classification of textual data into two different classes so that improvements could be made to more accurately classify the documents. The particular classifiers considered were Decision Trees (C4.5), *K*-nearest neighbour (*K*-NN), Naïve Bayes and Support Vector Machines (SVMs) algorithms. The reason for testing these different classifiers is their variability in selecting information variables based on different distance measures ranging from a simple Euclidean measure to kernel based methods. The purpose of classification is to validate the hypothesis that the proposed method based on MKTPKS improves the classification accuracies of the classifiers over the simple term based data classification method. A brief introduction to these classifiers is given in the following subsections.

3.4.1. Decision tree analysis

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. The decision tree algorithm C 4.5 is an extension of Quinlan's algorithm ID3 which generates decision trees (Quinlan, 1992) based on splitting each decision node by selection of an optimal split and continuing its search until no further split is possible. It uses the concept of information gain or entropy reduction to select the optimal split. Suppose there is a variable X for which k possible values have probabilities $p_1, p_2, \ldots, p_k$. Then the entropy of X is defined as:

$$H(X) = -\sum_j p_j \log_2(p_j) \tag{2}$$

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

$$H_s(T) = \sum_{i=1}^{k} P_i H_s(T_i) \tag{3}$$

where P_i represents the proportion of records in subset i. The Information gain is defined as:

$$\text{Information gain } \mathrm{IG}(S) = H(T) - H_s(T) \tag{4}$$

That is, the increase in information produced by partitioning the training data T according to the candidate split S. The selection of an optimal split at each decision node is based on the greatest information gain, IG(S).

3.4.2. K-nearest neighbouring algorithm

The *K*-nearest neighbouring (*K*-NN) algorithm is a technique that can be used to classify data by using distance measures. It assumes the training set includes not only the data in the set but also the desired classification for each item. 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 & Kamber, 2000). 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 the most items of the K closest items. The distance between the new or incoming item to the existing one is calculated by using a distance measure, and the most common distance function is the Euclidean distance given in equation1.

3.4.3. Naïve Bayes algorithm

A Naïve Bayes algorithm is a simple and 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 within a given set of records (Witten & Frank, 2000). This approach is based on the Naïve Bayes Theorem. In the context of text classification the probability of a class c, for a given document d_j, is calculated by using the Bayes Theorem shown below:

$$P(c/d_j) = \frac{p(d_j/c)p(c)}{p(d_j)} \tag{5}$$

As $p(c)$ is constant for all classes, only $P(c/d_j)p(d_j)$, where $j = 1,2,3,\ldots,m$, need be maximised. It is assumed that classes are independent of each other and this is called the Naïve assumption of class conditional independence and it is made while evaluating the classifier. The classification task is done by considering prior information and likelihood of the incoming information to form a posterior probability model of classification.

3.4.4. Support Vector Machines

The Support Vector Machine was first developed in Russia in the 1960s (Vapnik & Chernonenkis, 1964; Vapnik & Lerner, 1963). 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 objective of the SVM algorithm is to choose the optimal separating hyperplane that maximises the margin between two classes (Vapnik, 1995). For a binary classification problem where w_1 and w_2 represent two classes for a training data set, given as $X = \{x_1, x_2, \ldots, x_n\}$with class labels. The hyperplanes which separate data into two classes are described by:

$$f(x) = \mathrm{sgn}(\langle w.x \rangle + b) \tag{6}$$

where w is the coefficient vector and b is the bias of the hyperplane and sgn[.] stands for the bipolar 'sign' function. The optimisation problem which yields the hyperplane (Christiniani & Shawe-Taylor, 2000) can be written as:

$$\text{minimise}_{w,b} \frac{1}{2}\|w\|^2 \tag{7}$$

Subject to

$$y_i(\langle w.x_i \rangle + b) \geqslant 1, \quad \text{for } i = 1,2,\ldots,N \tag{8}$$

The larger the margin the better the generalisation abilities expected.

4. Applications of proposed methodology

4.1. Post project reviews as defining good or bad information

To demonstrate the application of the proposed methodology, an example using a sample data set which has been collected from the construction industry will now be considered. This data is in the form of Post Project Reviews (PPRs) where each review consists of 10–15 pages which have been divided into main and sub headings of time, cost, planning, etc. This dataset and the relevance of PPRs has previously been explained in Choudhary, Oluikpe, Harding, and Carrillo (2009), Carrillo et al. (2008). The data set for current analysis is taken from the subheadings of 'Time' and 'Cost' in the PPRs. The data was then divided into two different classes of documents, each containing good or bad information. This task

was done by careful reading through every review with the help of domain experts to ensure that the meanings of each item of text under the 'Time' and 'Cost' subheadings was understood. Then these reviews were divided into two different classes to test the proposed methodology. The classes were marked as 'A' for good information documents or 'B' for bad information documents.

An example of an identified good information document is:

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

Similarly an example of a bad information document is:

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

4.2. Classification of documents as good or bad information-containing documents

Decision trees (C4.5), *K*-NN, Naïve Bayes and Support Vector Machines (SVMs) defined above were all used to classify the textual data into two different classes in this research work. The algorithms were applied on the transformed data set in the form of attributes or candidate feature sets obtained through the hybrid application of the 1st Level Knowledge Processing and Storing Unit and the 2nd Level Knowledge Refinement Unit. As a result of the applications of these hybridized approaches a list of terms was generated in the form of MKTPKS termsets. Single key term phrases were identified using the clustering approach and then the clustered instances were further processed to generate MKTPKS termsets. Consequently a feature space was made to represent the existence or non-existence of key phrases in the documents. Each vector representation of the documents was done using MKTPKS 3-termsets. The relationship of the list of key phrases formed and their existence in documents and corresponding representative classes is shown in Fig. 4.

In Fig. 4, Ci represents the class of labels given to the training data, composed of C1, C2, C3, ..., C*i*, and F1, F2, ..., F*n* represent the corresponding MKTPKS 3-termsets. This, therefore, forms a matrix with the list of representative key term phrases and their class labels to which the data mining algorithms of Decision Tree (C4.5 or J48), *K*-NN, Naïve Bayes and SVMS are applied in order to classify the data into predefined classes.

Weka (3.4) was used to test the classification accuracies on the transformed dataset where the representation of the data is given by considering MKTPKS 3-termsets. The information contained within each cluster and represented in the form of MKTPKS termsets is then used to form the matrix shown in the Table 1.

Thus the whole document space is now transformed into the form of MKTPKS 3-termsets carrying key information defined in the textual data. This new representation matrix is then used to perform the classification task by dividing documents into two different classes (i.e. good or bad information-containing documents).

4.3. Text Mining (Data or Information Processing) Unit applications

In this part of the analysis the data needs to be made ready for the application of different data mining algorithms, e.g. *K*-means clustering, to the present scenario. So the data needs to be prepared in a suitable format. The data is therefore consolidated in a single text file for further processing. To parse the text into tokens a Java code was written to count the terms so that their corresponding frequencies could be represented in a term frequency matrix (TF). The stop words were also removed from the text data and a simple stemming method was applied. These text mining methods helped to reduce the dimensions of the data whilst retaining a useful information space without losing key information. The data was then saved in a comma separated file (csv) which was ready to be used in the 1st level knowledge processing and relational database formation.

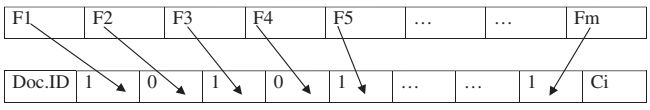

Fig. 4. Candidate termset representation if exists 1 otherwise 0.

Table 1
Matrix representation of the textual data using MKTPKS.

Doc IDs	*F*1	*F*2	*F*3...	Class attribute
D1	1	0	0	A
D2	1	0	1	B
D3	0	1	0	A
D4	0	1	1	B

4.4. 1st Level Knowledge Processing and Storing Unit applications

The 1st Level Unit was then applied on the saved csv file by loading it into Weka (3.4) where different clustering techniques could be used to discover an initial understanding of the contained knowledge and capture the key term phrases. The Weka (3.4) software is based on 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 about handling data or information see reference (Witten & Frank, 2000).

The *K*-means clustering algorithm was applied to split the input space of information into a number of subspaces. A large number of experiments were made to find an appropriate number of clusters to reduce the effect of information loss. Ultimately six clusters were selected as best for the current research work. The application of the clustering technique helped to capture key information or the first level of knowledge in terms of single key term phrases. The identified clusters IDs and their corresponding entries were identified as shown in Table 2 below. Only three representative clusters have been selected here in consideration of the length of this paper.

The key information captured in the different clusters refers to different sets of information contained within each document. It is therefore difficult to interpret this key information and exactly define good or bad information-containing documents. Hence it is not clear in what business context single term phrases like "business" identified in the cluster CL1 are used. Two different key term phrases like "business" and "unit" captured in the CL1 may refer to different concepts in the documents like "*business unit to supply the contribution*", "*business unit needed some work*" or "*business unit target*". It therefore becomes difficult to map these single key term phrases to find key issues discussed in the PPRs and thereby classify documents into good or bad information-containing documents. In order to overcome this difficulty, the process of extracting useful information codified within these documents needs to be further refined. This refinement of key information or knowledge discovered at the 1st Level is done by applying the Apriori Association Rule of Mining. Before applications of this, the key information discovered in terms of single key term phrases, needs to be stored. This activity is performed by creating a relational database with tables containing fields from the cluster labels, key terms identified and their corresponding document identification codes (IDs). These relational tables are then used to

Table 2
Single key term phrase identification by *K*-means clustering

Cluster IDs	Single key term phrases identified
CL1	"business", "carried", "fitting", "interiors", "***A***", "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"

Letter A is used to represent a company's name.

form MKTPKS which firstly reduce the number of dimensions in the feature space and secondly validate the hypothesis of acquiring better classification accuracies of the classifiers.

4.5. 2nd Level Knowledge Refinement Unit applications

The Apriori Association Rule of Mining was used at this stage for MKTPKS. The input is given in the form of relational tables where documents are taken as transactions and terms as items. The output in the form of MKTPKS 3-Termsets are shown in the Table 3.

The representative terms, i.e. T1, T2, T3 etc., refer to the single key term phrases identified within each cluster through applications of clustering as shown in Table 2. The co-occurrences of these single key term phrases to generate MKTPKS 3-termsets are given in the form of [T1 T2 T3] as a single entity for representing key issues discussed in the PPRs documents. These MKTPKS 3-termsets are taken as sequences of knowledge which are later used in the classification task following the criteria defined and implemented in Section 4.6.

4.6. 3rd Level Knowledge Utilisation and Text Classification Unit applications and results

This section illustrates the methods used for classifying the textual data into two different classes. The results obtained by the application of the 2nd Level Knowledge Refinement Unit to form a new matrix model with dimensions (20×223) based on the MKTPKS 3-termsets are discussed in Section 4.2. This new matrix was then loaded into Weka (3.4) and four different classifiers used to classify the data into their respective classes. The target variable set for this purpose was the class variable to determine the number of good or bad information-containing 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 MKTPKS 3-termsets based matrix model are then used to compare the classification accuracies against the simple term based representation model. A glimpse of the classification of textual data of PPRs using Decision Trees (J48 or C4.5) based on single term based representation is shown in Fig. 5. 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 representative of elements in the simple term based matrix model for classification of the documents.

Each node has been divided into its sub-nodes on the basis of maximum information gain. Each leaf node (rectangular) represents the final classification of the information into documents containing good or bad information about a project within the PPRs. A (5.0) shows that five documents are classified as good information documents at the deciding node (elliptic) 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 good information document at the branch (circular) node of t66.

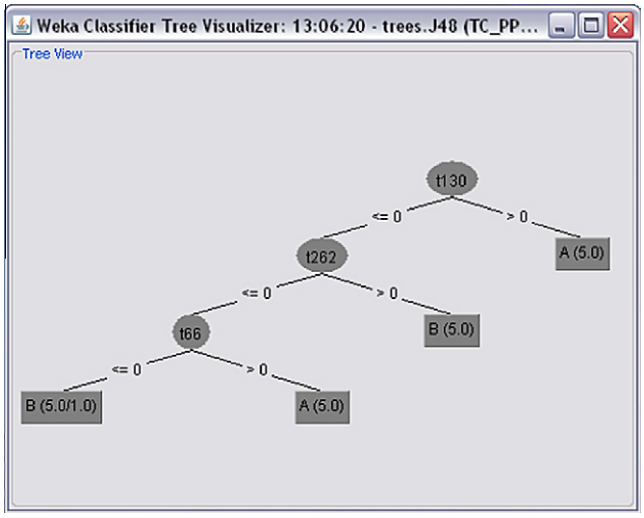

Fig. 5. Snapshot of decision tree based classification.

In terms of classification of data based on the proposed MKTPKS 3-termsets system, the Decision Tree results obtained using the Weka (3.4) based classifier are shown in Fig. 6.

The information space is classified into two classes of good and bad information-containing documents by selecting the information nodes and sub-nodes. The branch leaf represents the number of documents classified as good and bad. The node (elliptic) with representative MKTPKS 3-termset F200 splits the information into frequencies of occurrence of either less than or equal to zero and greater than zero. The classification rule is specified 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 classes. The information node (elliptic) at representative MKTPKS 3-termset 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

Table 3
MKTPKS 3-termsets 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]

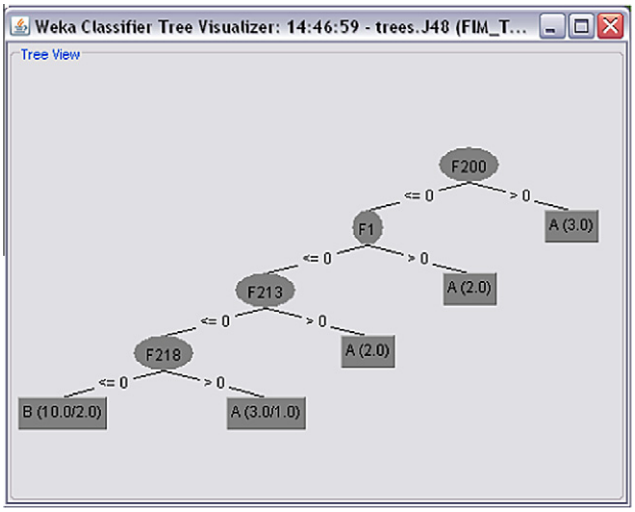

Fig. 6. MKTPKS 3-termeset based classification using C4.5.

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

In a business context, to provide better services to customers, decision makers have to use their experience and opinions. In the current research scenario the data under consideration defines key phrases like "*customer issued very few instructions/ variations*" which can help the workers to run the project smoothly and get it finished within time. Completion time gives a good Key Performance Indicator (KPI) as if a project was finished within a stipulated time the customer would be satisfied and the company is likely to retain this customer. Hence, if decision makers can easily identify and classify textual data on the basis of good or bad information-containing documents then better decisions may 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. The objective of this research was therefore to accurately classify the textual data (i.e. with a reduced 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 3-termset based methods). The incorrect classification results obtained through the application of different classifiers were calculated using the confusion matrix as shown in Table 4.

The classification accuracies are calculated by classifying information as good or bad information-containing 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 class of B.

Table 4
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.

4.7. Evaluation of the proposed system

The final evaluation of the proposed methodology has been made on the basis of average *F*-measure which is defined as the harmonic mean of Precision and Recall. The reason behind selection of the *F*-measure is that both precision and recall ratios are considered in it (Miao et al., 2009). The performance of the system was evaluated using the 10-fold cross validation method given in the Weka (3.4). The setting for each algorithm is varied to a certain level and the best possible accuracies are observed. In terms of Naïve Bayes classifiers, better classification accuracies are obtained by keeping the settings unchanged. In terms of the other classifiers the optimal parameters settings are chosen. In the case of Decision Tree (J48 or C4.5) algorithms, different seed ratios were used and the best results were obtained using a seed ratio of ten (10). Similarly for *K*-NN the optimal setting were taken as $K = 10$ and a Linear Kernel gave better results for SVMs based classification models. The results obtained are shown in the Table 5.

Table 5 shows the classification accuracies of the simple term based and MKTPKS 3-termset based classification models. The accuracies of the classifiers i.e. *K*-NN, Naïve Bayes and SVM (Linear Kernel) are better than the simple term based classification model (comparisons are depicted in the graph shown in Fig. 7).

Fig. 7 shows that the classification accuracy of the decision tree (C4.5) using the proposed MKTPKS 3-termset based classification model is lower than the simple term based classification model. However, the accuracies of the other classifiers (i.e. *K*-NN, Naïve Bayes and SVMs) are all improved using the proposed methodology when compared with the simple term based classification model. Hence, if the proposed methodology is used to classify the data, better classification accuracies are achieved for classification of the data into two different classes of good and bad information-containing documents.

4.8. Novelty in the work

To the best of the authors' knowledge the work presented in this paper is the first of its kind where classification tasks have been done through hybrid applications of textual data mining techniques by using MKTPKS 3-termsets. The research work presented in this paper is focused on classification of textual data into two different classes to define good and bad information-containing documents. A novel integration of textual data mining techniques is made to improve the classification accuracy of the classifier. In most cases the proposed approach gives a significant improvement in the classification accuracies measured using the *F*-measure. However with regard to classifying documents into their respective classes using Decision Tree (J48 or C4.5 algorithm) the accuracy of the classifier is reduced. The reason for loss of accuracy for the C 4.5 classifier is that the information selection criteria is highly dependant on the term frequencies. Also the nature of the data is highly sparse and this might affect the accuracy of the classifier.

The following points can thus be concluded the proposed methodology presented in this paper:

Table 5
Comparison of performance of different classifiers.

Classification model	Term based classification model (*F*-measure)	Proposed MKTPKS based classification model (*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

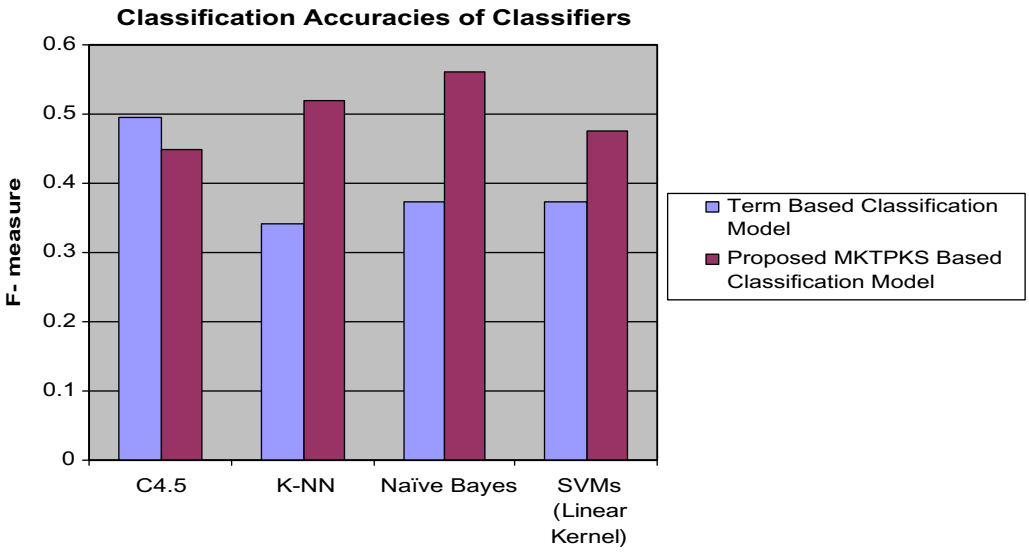

Fig. 7. Comparison of classification accuracies using *F*-measure

- Single term based representation methods are useful ways of discovering information but these methods affect the classification accuracies of the textual data.
- Hybrid application of textual data mining techniques give better results in the current research scenario and information pruning and knowledge refinement is made possible through use of Apriori Association Rule of Mining technique.
- Generating MKTPKS 3-termsets and using these to perform the classification improved the accuracies of the classifiers.
- In some business contexts if the misclassification rates are reduced then better decisions are made possible.

5. Conclusion and future work

The research work presented in this paper is focused on classification of textual data into two different classes of defining good and bad information-containing documents. A novel integration of textual data mining techniques is made to improve the classification accuracies of the classifier. In terms of classifying documents into their respective classes using a decision tree (J48 or C4.5 algorithm) the accuracy of the classifier is reduced using the proposed methodology while in the other classifiers there is significant improvement in the classification accuracies measured using the *F*-measure. Further research is needed into improvement strategies for the proposed methodology which could be in terms of introducing more refined methods for matrix representation models. The reason behind the loss of accuracy in the case of C4.5 classifier may lie in the use of entropy based information selection criteria whereas other classifiers uses simple distance, probabilistic and kernel based distance measures to find the similarities between documents.

Acknowledgements

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

References

Aasheim, C., & Koehler, G. J. (2006). Scanning world wide web documents with the vector space model. *Decision Support Systems, 42*, 690–699.

Agrawal, R., Lmielinski, T., & Swami, A. (1993). Mining association rule between sets of items in large databases. In *Proceedings of international conference on management of data (SIGMOD 93)*. (pp. 207–216).

Berry, M. J. A., & Linoff, G. (2004). *Data mining techniques for marketing, sales and customer relationship management*. Hoboken, NJ: Wiley Computer Publishing.

Caldas, C. H., Asce, S. M., Soibelman, L., Asce, M., & Han, J. (2002). Automated classification of construction project documents. *Journal of Computing in Civil Engineering, 16*(4), 234–243.

Caldas, C. H., & Soibelman, L. (2003). Automating document classification for construction management information systems. *Automation in Construction, 12*, 395–406.

Carrillo, P. M., Oluikpe, P. I., Harding, J. A., & Choudhary, A. K. (2008). *Text mining of post project reviews*. Performance and Knowledge Management Joint CIB Conference, CIB 2008 Helsinki, Finland, June 3–4.

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

Choudhary, A. K., Oluikpe, P. I., Harding, J. A., & Carrillo, P. M. (2009). The needs and benefits of text mining applications on post project reviews. *Computers in Industry, 60*, 728–740.

Christiniani, N., & Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods*. Cambridge, UK: Cambridge University Press.

Edwards, B., Zatorsky, M., & Nayak, R. (2008). *Clustering and classification of maintenance logs using text data mining in Seventh Australasian Data Mining Conference (AusDM 2008)*. Glenelg, Australia.

Fayyad, U. M., Piatetsky-Shapiro, G., & Smyth, P. (1996). *From data mining to knowledge discovery: An overview*. *Advances in Knowledge Discovery and Data Mining*. California, USA: American Association for Artificial Intelligence (AAAI) (pp. 1–34).

Gao, L., Chang, E., & Han, S. (2005). Powerful tool to expand business intelligence: Text mining. *Proceedings of World Academy of Science, Engineering and Technology, 8*, 110–115.

Hafeez, K., Zhang, Y., & Malak, N. (2002). Identifying core competence. *IEEE Potentials, 49*(1), 2–8.

Han, J., & Kamber, M. (2000). *Data mining: Concepts and techniques*. San Francisco: Morgan Kaufmann.

Haravu, L. J., & Neelameghan, A. (2003). Text mining and data mining in knowledge organisation and discovery: The making of knowledge-based products. *Cataloging and Classification Quarterly, 37*(1/2), 97–113.

Huang, L., & Murphey, Y.L. (2006). Text mining with application to engineering diagnostics. In *19th international conference on industrial engineering and other applications of applied intelligence. IEA/AIE 2006* (pp. 1309–1317). LNAI.

Ikonomakis, M., Kotsiantis, S., & Tampakas, V. (2005). Text classification using machine learning techniques. *WSEAS Transactions on Computers, 4*(8), 966–974.

Jinshu, S., Bofeng, Z., & Xin, X. (2006). Advances in machine learning based text categorisation. *Journal of Software, 17*(9), 1848–1859.

Karanikas, H., & Theodoulidis, B. (2002). *Knowledge discovery in text and text mining software*. Manchester: Centre for Research in Information Management (CRIM), UMIST.

Kasravi, K. (2004). Improving the engineering processes with text mining. In *Proceedings of DTE'04, ASME 2004*. Salt Lake City, Utah, USA: Design Engineering Technical Conferences and Computers and Information in Engineering Conference.

Kornfein, M.M., & Goldfrab, H. (2007). A comparison of classification techniques for technical text passages. In *World congress on engineering*. London, UK: Proceedings of the World Congress on Engineering.

Larose, D. T. (2005). *Discovering knowledge in data: An introduction to data mining*. Hoboken, NJ: John Wiley & Sons, Inc.

Liu, Y., Loh, H. T., & Sun, A. (2009). Imbalanced text classification: A term weighting approach. *Expert Systems with Applications, 36*, 690–701.

Menon, R., Tong, L. H., & Sathiyakeerthi, S. (2005). Analysing textual databases using data mining to enable fast product development processes. *Reliability Engineering and System Safety, 88*, 171–180.

Miao, D., Duan, Q., Zhang, H., & Jiao, N. (2009). Rough set based hybrid algorithm for text classification. *Expert Systems with Applications, 36*, 9168–9174.

Nasukawa, T., & Nagano, T. (2001). Text analysis and knowledge mining systems. *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.

Quinlan, J. R. (1992). *C.45: Programs for machine learning*. San Francisco, CA: Morgan Kaufman.

Salton, G. (1989). *Automatic text processing: The transformation, analysis and retrieval of information by computer*. Reading, MA: Addison Wesley.

Sanchez, S. N., Triantaphyllou, E., Chen, J., & Liao, T. W. (2002). An incremental learning algorithm for constructing Boolean functions from positive and negative examples. *Computers and Operations Research, 29*, 1677–1700.

Spinakis, A., & Chatzimakri, A. (2005). Comparative study of text mining tools. *Studies in Fuzziness and Soft Computing, 185*, 223–232.

Tan, A.-H. (1999). Text mining: The state of art and the challenges. In *Workshop on knowledge discovery from advanced databases (KDAD'99)* (pp. 71–76). Beijing: Proceedings of PAKDD'99.

Vapnik, V. N. (1995). *The nature of statistical learning theory*. New York: Springer.

Vapnik, V., & Chernonenkis, A. (1964). A note of class of perceptron. *Automation and Remote Control, 25*.

Vapnik, V., & Lerner, A. (1963). Pattern recognition using generalised portrait method. *Automation and Remote Control, 24*.

Wang, K. (2007). Applying data mining to manufacturing: The nature and implications. *Journal of Intelligent Manufacturing, 18*, 487–495.

Weiss, S. M., Indurkhya, N., Zhang, T., & Damerau, F. J. (2005). *Text mining: Predictive methods for analyzing unstructured information*. Springer Science and Business Media, Inc.

Witten, I. H., & Frank, E. (2000). *Data mining: Practical machine learning tools and techniques with java implementations*. San Francisco: Morgan Kaufman.

Yong, Z., Youwen, L., & Shiziong, X. (2009). An improved KNN text classification algorithm based on clustering. *Journal of Computers, 4*(3), 230–237.

Yu, L., Wang, S., & Lai, K. K. (2005). A rough-set-refined text mining approach for crude oil market tendency forecasting. *International Journal of System Sciences, 2*(1), 33–46.