European Scientific Journal July 2017 edition Vol.13, No.21 ISSN: 1857 – 7881 (Print) e - ISSN 1857- 7431

Textual Data Mining For Knowledge Discovery and Data Classification: A Comparative Study

Nadeem Ur-Rahman
Wolfson School of Mechanical and Manufacturing Engineering,
Loughborough University, Loughborough, Leicestershire, LE11 3TU, UK.

doi: 10.19044/esj.2017.v13n21p429 URL:http://dx.doi.org/10.19044/esj.2017.v13n21p429

Abstract

Business Intelligence solutions are key to enable industrial organisations (either manufacturing or construction) to remain competitive in the market. These solutions are achieved through analysis of data which is collected, retrieved and re-used for prediction and classification purposes. However many sources of industrial data are not being fully utilised to improve the business processes of the associated industry. It is generally left to the decision makers or managers within a company to take effective decisions based on the information available throughout product design and manufacture or from the operation of business or production processes. Substantial efforts and energy are required in terms of time and money to identify and exploit the appropriate information that is available from the data. Data Mining techniques have long been applied mainly to numerical forms of data available from various data sources but their applications to analyse semi-structured or unstructured databases are still limited to a few specific domains. The applications of these techniques in combination with Text Mining methods based on statistical, natural language processing and visualisation techniques could give beneficial results. Text Mining methods mainly deal with document clustering, text summarisation and classification and mainly rely on methods and techniques available in the area of Information Retrieval (IR). These help to uncover the hidden information in text documents at an initial level. This paper investigates applications of Text Mining in terms of Textual Data Mining (TDM) methods which share techniques from IR and data mining. These techniques may be implemented to analyse textual databases in general but they are demonstrated here using examples of Post Project Reviews (PPR) from the construction industry as a case study. The research is focused on finding key single or multiple term phrases for classifying the documents into two classes i.e. good information and bad information documents to help decision makers or project managers

to identify key issues discussed in PPRs which can be used as a guide for future project management process.

Keywords: Textual Data Mining, Knowledge Discovery, Latent Semantic Analysis(LSA), Post Project Reviews (PPR)

1. Introduction

The efficiency and effectiveness of decision making are important factors in business process improvement. Business intelligence (BI) can be considered as a tool to improve many types of processes either related to manufacturing a product or improving the quality of services in the construction industry. BI can thus help an enterprise to remain competitive in a business environment by Hsieh (2007). Corporate information are available in many different data formats however it has been found that 80% of company information exists in some textual form detailed by Yu, Wang, and Lai (2005).

Data mining techniques in general and clustering in particular can be used in many contexts, for example to analyse structured, semi-structured or unstructured data formats. Frequent itemset mining methods are a means of presenting information and knowledge with desired support and confidence levels to decision makers. Text Mining methods can be used to give suitable representation to semi-structured or unstructured forms of data in order to find useful information which can then be transformed further into valuable information or knowledge. Text Mining methods help to find patterns in the data and share techniques from other areas such as statistical data analysis techniques, natural language processing (NLP) and graphical methods. Text mining can be defined as the "process of discovering useful information from textual databases through the application of computer techniques" by Fan, Wallace, Rich and Zhang (2006). Several different definitions can be found in the literature for the task of knowledge discovery but the definition used in this research work is, "the process of discovering valuable information or knowledge from textual data through the application of data mining techniques" by Han and Kamber (2000). The overall process of knowledge discovery from text (KDT) can be divided into three main steps that is data collection, pre-processing of documents and text mining described by Karanikas, and Theodoulidis (2002).

In Manufacturing or Construction Industry environments there are very few reported applications of TDM. However a couple of applications of TDM techniques to resolve the quality and reliability issues in manufacturing of new products have been reported by Menon, Tong, and Sathiyakeerthi (2005). Data Mining techniques have also been used to improve customer

service activities stated by Hui and Jha (2000). An associative classification based system was proposed to support personalisation in Business to Commerce (B2C) applications to help customers in identifying the product to best meet their need discussed by Zhang, and Jiao (2006). Text Mining techniques were also applied to identify the morphology of existing products and technology roadmap as found by Phaal and Probert (2008). In construction environments textual data mining techniques have been used to manage the information and knowledge resources to improve the service conditions discussed in (Gibbons, et al. (2000)) and (Caldas, and Soibelman (2003)). So a relatively new and promising area of research exists for the application of Text Mining techniques to explore and seek hidden or implied knowledge in these environments. This has therefore provided the motivation to explore new dimensions in terms of applying knowledge discovery techniques to these databases. A new method is proposed in this paper to discover knowledge in terms of multiple key term phrasal knowledge sequences from a textual document collection. The proposed method should help decision makers or knowledge workers to automatically identify the key issues from any text document. The effectiveness of this method is demonstrated by using Post Project Reviews (PPR) as an implementation of case study.

The rest of the paper is organised as follows: Section 2 outlines the proposed methodology or framework based on Text Mining and Knowledge generation modules. Section 3 discusses the applications of the framework on the case study data available in the form of PPR. The conclusion and future work is given in Section 4.

2. **Proposed Architecture and Framework**

In this section a framework is outlined to analyse textual databases. This framework consists of two main parts;

- A data and information handling section called the “Text Mining Module” and
- A knowledge discovery section for the discovery of multiple key term phrasal knowledge sequences. These knowledge sequences are used for the classification of data into two different classes of good and bad information documents.

The Text Mining Module starts its working on text documents available in the form of free formatted text documents. Information pre-processing and structuring techniques are then applied on this free formatted text data with the help of the Information Pre-processing Unit and Information Structuring Unit.

In the second step the information that has been structured with the help of Text Mining Module is passed to the Knowledge Generation Module for further processing and to discover key term phrasal knowledge sequences. The word 'term' is used to refer to the key information that has been coded in the textual data available in the form of Post Project Reviews (PPR). Every single word in the text has some special meaning in the context of project handling. Knowledge Generation Module mainly works with the hybrid applications of Clustering and Association Rule of Mining techniques to discover first level of knowledge and then process it to generate Multiple Key Term Phrasal Knowledge Sequences (MKTPKS). These MKTPKS are used to summarise key issues discussed in each particular document and thereby identify good or bad information documents[1] which are uniquely represented by each cluster. Figure1 illustrates the proposed framework showing the flow of information.

[1] Good or Bad Information Document: The term is used for set of information available in the form of Case Study data. For example "*The Project took 51 weeks and was handed over on programme giving good KPI…* " is taken as good information whereas "*The Project was programmed to take 49 weeks, but finished four weeks late because…*" is taken as bad information set.

Figure 1. illustrates the proposed framework showing the flow of information, its processing modules and transformation of processed information into a set of MKTPKS.

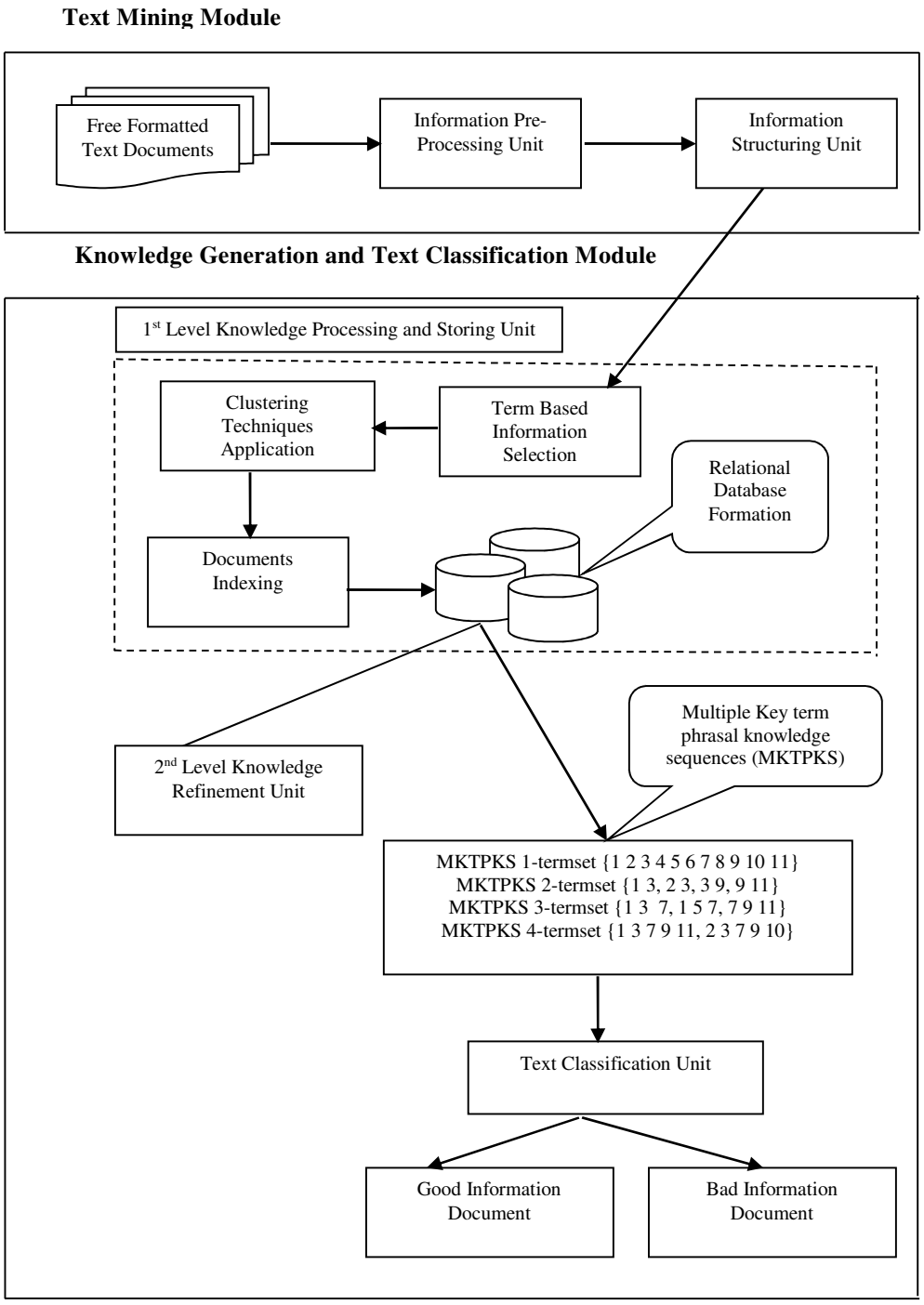

Figure 1: Multiple Key term Phrasal Knowledge Sequences Identification Framework

The proposed framework is based on applications of Textual Data Mining Techniques to analyse free formatted textual data and discover useful knowledge in terms of Multiple Key Term Phrasal Knowledge Sequences (MKTPKS) and thereby classifying it into two different classes of good and bad information documents.

2.1 Text Mining Module

This module performs two different tasks on the information, using the Information Pre-processing unit and the Information Structuring Unit. These units are used to remove un-necessary information from the free formatted text documents and then structure it for the application of different data mining algorithms which will subsequently be used to analyse the text. Therefore the first step in the analysis is to process the information using the information pre-processing unit.

2.1.1 Information Pre-processing Unit

In any industrial business environment the process of handling textual data starts by considering the opinion of domain experts in order to thoroughly understand the key terms or words. These terms are sometime taken as decision variables or decision attributes and may be used to measure the effectiveness of the proposed knowledge discovery system. During this research work the key issues in terms of good or bad information documents identified by the domain experts with the help of single or multiple key term phrases and taken as to measure the performance measure of the proposed framework. Data mining experts and domain experts usually have to work interactively to better understand the relevant issues related either manufacturing or construction domain. Decisions made in the early stage of analysis highly affect the success of the Knowledge Discovery (KD) process. The task of selecting decision variables therefore needs to be performed carefully with the help of domain experts. The input information must then be codified in a format suitable for the TDM tasks.

The first step is to remove the un-necessary information available in the form of words which are less effective in textual data analysis. These words include some verbs, pronouns, conjunctions and disjunctions such as 'a', 'an', 'the', 'of ' , and, ' I ', etc. and are called stop words. These words must be removed from the list to help to assess and interpret the meaning of text that needs to be interpreted or conveyed more easily. Removal of these less informative words increases the efficiency and accuracy of results in processing the text and has been a common technique in text analysis by Rijsbergen (1979). Word stemming is also used and this is the process of reducing the inflected or derived words to their actual stem, base or root

form. For example, in English, the words like design, designing, and designed are stemmed to their root word, "design". In this study, a simple suffix stripping technique was used to reduce the loss of meaning or context. This method captures more explicit relationships among terms defined in the text. The information pre-processing unit uses three different methods to improve the overall process of text analysis i.e.

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

2.1.2 Information Structuring Unit

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

2.2 Knowledge Generation and Text Classification Module

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

- 1st Level Knowledge Processing and Storing Unit
- 2nd Level Knowledge Refinement Unit
- Text Classification Unit

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

2.2.1 1st Level Knowledge processing and Storing Unit

2.2.1.1 Term based Information Selection

At this stage of analysis the input information is available in different structural representations obtained through the application of the Information Structuring Unit. These structures are in the form of a term frequency (TF) matrix which consists of vectors describing the information in a document. Each word or term can additionally be weighted in the document collection using Inverse document frequency (IDF) and term frequency inverse document frequency (TF*IDF) matrices. A suitable representation must be

selected before performing further analysis of text data. To avoid losing some key information at this stage, a term based representation model is considered where the terms and their corresponding frequencies are counted. The information matrix so formed is taken as an input for application of clustering techniques.

2.2.1.2 Clustering Techniques Applications

Clustering is defined as a process of grouping data or information into groups of similar types of information using some physical or quantitative measures by Larose (2005). These quantitative measures are based upon a distance function, which is measured from some centre point i.e. termed as the centroid of the cluster. The Euclidean distance measure is used as a natural distance function which may be replaced with other similarity measures to find the similarities between documents. Within this research context the information is available as an input matrix based on term frequencies so the similarities are found between terms. Thus clustering techniques are then implemented on these information matrixes and output is generated in the form of correlated terms based on the natural relationships existing within each document.

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

2.2.1.3 Document Indexing

After performing the clustering task, information is obtained in terms of single key term phrases and these are used to index the documents. The documents corresponding to these single key term phrases are marked with their identification codes. Indexing the documents at this stage helps to store information in a useful format, so that new documents can be classified based on the information possessed within a cluster.

2.2.1.4 Relational Database Formation

The first level knowledge captured in the previous stages must be stored in the form of relational tables, so that it can be used further for discovery of useful relationships between terms by generating multiple key term phrasal knowledge sequences. This task is performed by storing information in a relational database. The tables are in the form of cluster labels, indexed documents identification (IDs) and their respective key single

term phrases. The documents are considered as transactions and terms captured as a result of clustering are considered as items (a commonly used term in market basket analysis). This helps to form an input space of information which is used in the next stage, i.e. using the 2nd Level Knowledge Refinement Unit.

2.3 2nd Level Knowledge Refinement Unit

In this part of the analysis, the input matrix, from the relational database tables are processed using Apriori Association Rule of Mining discussed by Agrawal, Lmielinski, and Swami(1993). These methods are used to form MKTPKS from the single term phrases identified with the help of clustering techniques. The key aspect in implementation of these techniques is to find co-occurrence of terms by searching through the whole provided document space of information. The most useful knowledge is found by using varying levels of support values and this is defined as finding the co-occurrence of terms with some defined percentage. This unit is therefore focused on finding the MKTPKS rather than on determining association rules which might occur in the textual databases. The MKTPKS might help to overcome the difficulty of populating the knowledge base with too many association rules which may occur whilst trying to discover useful knowledge from these knowledge bases.

2.4 Text Classification Unit

In this part of the analysis the text is classified into two different classes termed as the good and bad information documents classes. The input for analysing the textual data is taken in the form of frequent terms sets generated through applications of frequent term set mining algorithm. The new input matrix generated is used for analysing the data and results are compared with the Latent Semantic Analysis (LSA) based textual data classification method of text analysis. The information structured in the form of a term based matrix representation model is used to find the semantic relationships among terms defined in the textual data. This is done through applications of LSA method for structuring information in the form of a matrix where each term has been represented by the corresponding numerical values. The new structured information based matrix model is used as an input for clustering information into different clusters which at later stage is used for classifying data through applications of Decision Trees (C4.5), K-nearest neighbouring Algorithm, Naive Bayes and Support Vector Machines.

2.5 Expected Benefits Associated with Framework

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

- Implementation of Clustering techniques divide whole document space of information into multiple subspaces. It reduces the number of objects to be used for analysis of information codified in text data which is otherwise difficult to decipher.
- A new document or set of information can be handled easily by assigning it to the corresponding cluster.
- Text is classified using frequent termset mining is useful to decode the information and generated knowledge sequences which are helpful to decision makers to discover valuable knowledge for improving the business in an industrial setup.

3. Case Study Data and Applications of Proposed Methodology

The framework proposed has been implemented on case study data available in the from Post Project Reviews (PPR) taken from the construction industry. In construction industry the PPRs provide a mechanism to keep people working together and get their experiences and knowledge to be stored in structured format for future use shown in URL (2010). PPR are also useful source of capturing knowledge and transmitting it to participating organisations in Kamra, Anumba, Carrillo and Bouchlaghem (2003). Initially the decision variables were selected to start the process of analysing the textual data available in the form of PPR and good or bad information documents were located with the help of domain experts by identifying single or multiple term phrases. These key single or multiple term phrases were considered to represent the important areas of knowledge which might be covered during PPR e.g. project lead time (i.e. 45 weeks), Financial Issues (low cost, within the budget) etc. The consideration of these key words or phrases is used to measure the performance of the systems.

The business processes of the construction industry are based on the efficient use of resources in terms of time, cost, planning and customer's satisfaction levels. Identifying information related to these issues and tracing the causes and effects of problems in previous projects should help to reduce the repetition of these issues and improve the chances of success in current and future projects.

3.1 PPRs For Information Handling and Business Intelligence

The previous knowledge and experience of a project manager is critical to the overall success of a project and satisfaction of the customer. Experiences and knowledge from previous projects can be used by decision makers or knowledge workers to improve effective decisions in future projects. PPR are useful form of information available in a construction industry environment. These reports have huge potential as sources of

knowledge made available at the desks of workers on subsequent or similar projects. PPRs are also a necessary tool for knowledge management and a valuable source of shared knowledge across the boundaries of an enterprise discussed in Tan et el.,(2006). These reviews help in learning collective lessons showed by Carrillo (2005) and the lessons learnt might then be used to prevent similar mistakes being made in the future by Pitman (1991). Thus discovery of useful information or valuable knowledge from these reviews will provide solutions to improve future business processes of an industry. Since the form of PPR are collection of multiple information coded with key phrases either single or multiple terms therefore, handling with these type of information need special handling to structure it. Text Mining methods have various techniques that will be used to decode these information. Thus the combined efforts of data and text mining methods will give a chance to learn useful lessons from these reports beneficial for future use.

The examples of PPR used in this study consist of a large number of documents in different data formats containing about 10-15 pages in each review. These PPR were first saved in a text file by removing the headings and their subheadings to make the data suitable for application of different modules of proposed framework.

The example reports were already divided into sixteen different headings as given below;

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

These headings are further divided into sub-headings e.g. cost, time, prelims, subcontractors etc. The documents used for the implementation and evaluation of the proposed framework, were all taken from sections of the PPR with the sub-headings of "Time". The topics discussed in these reviews ranged from general outcomes in terms of cost and time, to general levels of satisfaction acquired during the whole project. The knowledge contained in these reviews therefore covers different stages of interaction with design teams, clients, errors and mistakes and health& safety status etc. observed during the project. The topics or issues discussed in these reviews were identified by domain experts. The identified key words, phrases or sentences all refer to some particular topic discussed in the PPR. Some examples of the topics and useful knowledge phrases identified in the sample PPR are shown in the Table 1;

Table 1: Key topics or knowledge areas identified by domain experts in PPR

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

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

This study will focus on issues associated with time in the PPR e.g. identification of key information that shows that a project has been completed late, early or on time and facts relating this to the associated causes which could help to trace the reasons of its delay or that are likely to provide useful knowledge for future projects.

A Phrase like "causing an additional delay" is a time related example of a knowledge phrase that might occur in the text of PPR. It may also give information about some issue causing the delay in the project being handed over to the client. Identification of such key phrases will help to uncover the hidden information in the text and their corresponding causes.

3.2 Framework Implementation on Case Study Data

The application of Text Mining Module as shown in Figure 1 requires the data to be made available in the form of free formatted text documents. This task is done by removing the headings or sub-headings from the reviews and storing the data in a text file with document IDs. This file is then passed to the information pre-processing unit where different functions are performed e.g. setting the aim or objectives of data analysis by thoroughly understanding the terms and selecting decision attributes or variables, removing redundant information in the form of stop words e.g. 'a', 'an' , 'the' and also performing a simple stemming process to conflate the words to their original stems by removing the suffix 'ing', 'ed' and 'ly' etc.

The next step towards analysis of data in the Text Mining Module is to structure the information into different representation (i.e. Term Frequency (TF), Term Frequency Inverse Document Frequency (TF*IDF) and Binary Representation) with the help of the information structuring unit. Since there was a great loss of information when using TF*IDF to analyse the data with the help of Clustering techniques, the documents are represented in terms of a TF matrix., and this task is carried out using java code to count words and their corresponding frequencies in the documents. The output is generated in the form of comma separated values (CSV) which are then loaded into Weka (3.4) for further analysis through applications of Clustering Techniques.

In the Knowledge Generation Module, the analysis of the PPR data, which is now in the form of a term frequency matrix which contains terms and their respective frequency counts is continued. Since the intention is to find the similarities between terms the matrix representation of textual data is done showing terms along rows and documents along columns. The similarities are determined by calculating the Euclidean Distances between terms using the formula given in equation 1;

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

After selecting the suitable matrix representation for the term based information the data is ready for application of the clustering techniques. Weka (3.4) was used to support the activities defined within 1st Level Knowledge Processing and Storing Unit before the application of 2nd Level Knowledge Refinement Unit to form the MKTPKS (see Figure 1). Weka (3.4) software is based upon an open source java environment which allows the user to customise or add new features to it. It offers different functionalities ranging from pre-processing to data classification and analysis detail of which can be found in Witten and Frank (2000). There are variants of clustering algorithms available in Weka (3.4) software but in the present work the k-means clustering algorithm has been used due to its linearity (i.e. it has linear memory requirements for storing documents, cluster membership of the documents plus the cluster centroids). Clustering is also an essential part of the proposed methodology because it helps to divide the information space into multiple sub-spaces and to capture useful information based on single key term phrases.

The K-means Clustering algorithm was applied on the comma separated values (CSV) data file which was obtained from the Information Structuring Unit in the Text Mining Module. A large number of experiments were made to find the suitable number of clusters to capture information in terms of single key term phrases. These experiments showed that if a large number of clusters are used there is a danger of more information loss due to partitioning information into multiple different clusters. So careful handling and extensive experimentation was done to select an appropriate number of clusters, and this was eventually determined to be six for ten sets of PPR data related to Time, Cost, and Planning domains where the number of terms within each PPR range from hundreds to thousands of terms. The resultant matrix, after the applications of the units defined in the Text Mining Module, contains hundred of terms i.e. (100-300). The application of the K-means Clustering algorithm helped to identify single term phrases within each cluster and example results obtained through the application of these clustering techniques on documents from the 'Time' domain are shown in the Table 2 due to less availability of space;

Table 2: Single Key Term Phrases Identified by Clustering Technique

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

The letters (i.e. *A*-*G*) have been used to represent some company names.

The application of clustering is an essential part of the proposed framework and results shown in Table 2 give useful information as the information contained within each cluster is based on single key term phrases that refer to key issues discussed in the PPRs. However this is not immediately clear from the results and there are difficulties in how to present the results to the user for them to be correctly interpreted. e.g. the information captured in CL1 comes from multiple documents and the human observer may interpret a cluster description as being that the single key term phrase "time" may refer to the issue of "delivered on time" or "completed on time" or "extension of time" where these key multiple key terms phrases refer to three different issues discussed in the PPR. So it is difficult to map these key phrase information or knowledge to some particular document defining some good or bad practiced information. Similarly the key term phrase "job" may be used to define the concept of "job finished late", "job took just under one year" or "job should be done within twenty one weeks". So defining these structures are not an easy task to be performed using these tables and it is quite difficult to help a user with the knowledge discovery

within each cluster. Similarly terms defining the concept are identified in cluster CL2 and CL3 which are not specific to some exact information. Thus there becomes a difficulty of defining the concept on the basis of these single key term phrases. So if this model is used on its own the terms captured within each cluster are of some importance but the importance of these terms depends highly on the correct definition of the concept with the help of these terms. So there is a need to use some other technique to help to restrict the domain of key information or knowledge defined in these clusters.

The pruning of the first level discovered knowledge is done through the application of the 2nd Level of Knowledge Refinement Unit. Before passing the information to the next stage of analysis the set of identified key information are used to index the documents and store this in the form of a relational table using Relational Database Formation function. Then this key information is passed to the next stage of processing to generate the MKTPKS. These MKTPKS will ultimately serve the purpose of representing key information captured within each cluster and map it to identify the good or bad information documents. The process of refining key information with the help of the 2nd Level Knowledge Refinement Unit is discussed in detail in the next section.

3.2.1 2nd Level Knowledge Refinement Unit

This unit works through the application of the Apriori Association Rule of Mining for generating MKTPKS. The term frequent itemset associated with Apriori Association Rule of Mining comes from supermarket transactional databases where each frequent itemsets is regarded as products which are most likely to be purchased together in one transaction. For example if a person buys “milk” then he might be interested to buy “egg”. Thus the purpose of finding MKTPKS in the current research context is to serve the purpose of finding those terms which co-occur together in documents.

The simple Apriori algorithm is used to scan through the whole document space of information which are identified within each cluster to generate MKTPKS. At first it will generate all frequent 1-termset with the minimum support. Then the data is scanned to generate frequent 2-termset from the discovered frequent 1-termsets. This process continues until the required MKTPKS termsets are obtained. If all termsets are represented as an ordered set then the Apriori Algorithm exploits the property that all subsets of frequent termsets are also frequent. The MKTPKS formed through the application of the Apriori Association Rule of Mining are shown in the Table 3.

Table 3 : Identification of Key Term Phrasal Knowledge Sequences

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

Thus knowledge discovered with the help of MKTPKS termsets shown in Table 3 is used to identify good or bad information documents. Since these MKTPKS are considered as a single unit of knowledgeable termsets therefore it helps to easily map them to a unique document with time related information and this can then be marked as good or bad information document (within this research context). For example, the MKTPKS in the cluster CL1 refer to a unique information document representing some good practice information which explains that the customer suggested that the job should be done in some week's time and that the work was completed on time.

The MKTPKS uniquely represent this set of information in cluster CL1 "{agree, complete, customer, time, week, within, work}" and identifies the relevant document. Similarly the MKTPKS representing cluster CL2 uniquely represents the information available in the document as good practice carrying information terms sequences showing that the both client and industry men were agreed on the time to finish the job. Finally in the cluster CL3 the key term frequent termsets refer to the document where all these key terms occur uniquely. The occurrence of these key terms phrasal knowledge sequence in this document show that information coded within this document is some good practice information as well. So identification and storing of these key phrases within the knowledgebase and the retrieval of the relevant documents can be used to improve the future activities of a project. Further benefits may also be obtained such as reducing lead time of a project by taking effective decisions based on issues and their respective causes previously reported in the PPR. Thus this research may ultimately help to improve the competitive edge of business. The preliminary results obtained from the experiments carried out at this stage of the framework are highly promising to identify the documents carrying key information using MKTPKS.

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

- Firstly by forming the sequences of terms with useful knowledge to compare with those identified by the domain experts in the PPR.

- Secondly to map these knowledge sequences to identify the good or bad information documents and thereby classifying the documents into two classes detail of which can be found in Ur-Rahman and Harding (2012). However the study made in this research work is to discover useful knowledge for decision maker and conduct the comparative analysis of MKTPKS based text classification model with that of Latent Semantic Analysis (LSA) based matrix model.

The discovery of MKTPKS is based on using varying levels of support values and the formation of these MKTPKS ultimately generate a lattice of concepts as shown in figure 2 below;

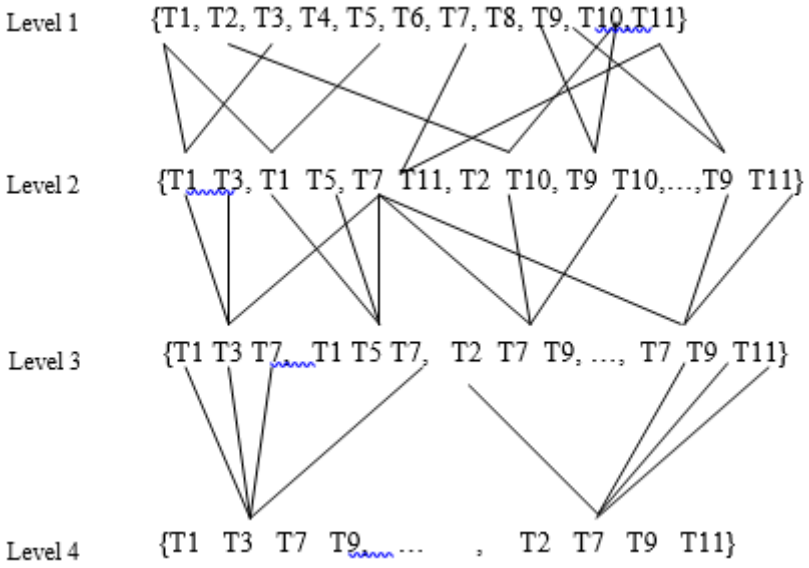

Figure 2: Levels of forming multiple key term phrasal knowledge sequences

The figure 2 shows that at the first level (or Level 1) MKTPKS 1-termsets are identified by searching through whole document space of information and then at second level (or level 2), the data is scanned again to form MKTPKS 2-termsets from the discovered MKTPKS 1-termsets. This process can be continued to generate the knowledge sequence as shown in figure.

3.2.2 Text Classification Unit Application

The matrix model generated in the form of simple term based representation is used to transform it into two different models matrices one based on MKTPKS and other based on Latent Semantic Analysis (LSA) method. Latent Semantic Analysis (LSA) methods are used to consider the semantic relationships among terms defined with textual documents and retrieve useful information from the textual data. These methods were first

proposed in Ur-Rahman and Harding (2012)and used to automate the process of retrieving useful information from documents. These methods used to represent the information contained in the documents in the form of matrix termed as term by document matrix (i.e. t x d) where the 't' is used to represent terms and 'd' stands for the documents. The whole information space is then divided into semantic space based on Singular Value Decomposition (SVD) where SVD is used to decompose the whole information space available in the form of t x d into linearly independent spaces or sub dimensional vector spaces. These methods were originally used for performing the task of information retrieval based on some semantic relationships existing among different terms or words used in the textual databases.

The LSA model is used to capture the latent relationships among different terms defined in the textual data based on the representation in the form of numerical values. The simple term based model shown in Table 1 shows the frequency of occurrences of terms in their respective documents which gives the importance of information coded with the help of these terms defined in the information space of documents.

Table 1: Simple Term Based Matrix Model

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

This shows that the terms 'accurate' and 'achieve' occurred in the same document giving some meaning to the text like 'accurate targets for future work' and 'achieve the maximum gain'. The relationship among these terms could be explained that if targets were measured accurately then the company could gain maximum profit. The matrix model shown in Table 1 is then transformed into the new representation of model through application of LSA method of text analysis as shown in Table 2.

Table 2: LSA based Matrix model based on simple term based data structuring

Terms	D1	D2	D3	D4	D5	D6	D7
About	0.0618	0.224	0.1696	0.13399	0.18384	0.89579	0.35267
Above	0.0542	-0.12	0.105	0.131	0.1299	1.9243	-0.133
Account	0.2285	1.763	0.7561	0.4538	0.7715	-0.057	2.6085

Accurate	0.0271	-0.06	0.0524	0.066	0.065	0.9622	-0.066
Achieve	0.0271	-0.06	0.0524	0.066	0.065	0.9622	-0.066
Actual	0.043	0.041	0.1024	0.0986	0.1169	1.0418	0.085
Additional	0.053	0.382	0.172	0.1066	0.1764	0.0850	0.5674
Adjusted	0.0271	-0.06	0.0524	0.0657	0.065	0.9622	-0.066
Administer	0.0271	-0.06	0.0524	0.0657	0.065	0.9622	-0.066
Against	0.0271	-0.06	0.0524	0.0657	0.065	0.9622	-0.066

Thus the terms defined in simple term based representation have got new semantic relationships adjusted through their numeric values. This matrix model is used to generate a new classification matrix model for application of data mining algorithm i.e. Decision Trees (C4.5), K-nearest neighbouring Algorithm, Naive Bayes and Support Vector Machines. The matrix model is shown in Table 3 below;

Table 3: LSA Based Matrix Model for Text Classification

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

4. Results and Discussion

The application of model proposed as Multiple Key term Phrasal Knowledge Sequences (MKTPKS) for knowledge discovery and text classification into two different classes of good and bad information documents showed promising results as compared with LSA based text classification model. The classification accuracies recorded for MKTPKS and LSA model for different values of *k* are given in table 4-6.

Table 4: Comparison of Performance of different classifiers

Classification Model	Term Based LSA Classification Model (F-measure) for k=2	Proposed FTS Classification Method (F-measure)
Decision Trees (J48 or C4.5)	0.271	0.449
K-NN (k=10)	0.479	0.52
Naïve Bayes	0.405	0.561
SVMs (Linear Kernel)	0.355	0.475

Table 5: Comparison of Performance of different classifiers

Classification Model	Term Based LSA Classification Model (F-measure) for k=3	Proposed FTS Classification Method (F-measure)
Decision Trees (J48 or C4.5)	0.479	0.449

K-NN (k=10)	0.479	0.52
Naïve Bayes	0.311	0.561
SVMs (Linear Kernel)	0.436	0.475

Table 6: Comparison of Performance of different classifiers

Classification Model	Term Based LSA Classification Model (F-measure) for k=4	Proposed FTS Classification Method (F-measure)
Decision Trees (J48 or C4.5)	0.260	0.449
K-NN (k=10)	0.405	0.52
Naïve Bayes	0.311	0.561
SVMs (Linear Kernel)	0.260	0.475

The accuracy of the proposed method of classifying textual data using the simple term based LSA and MKTPKS based models are shown in the figure 4-6 shown below;

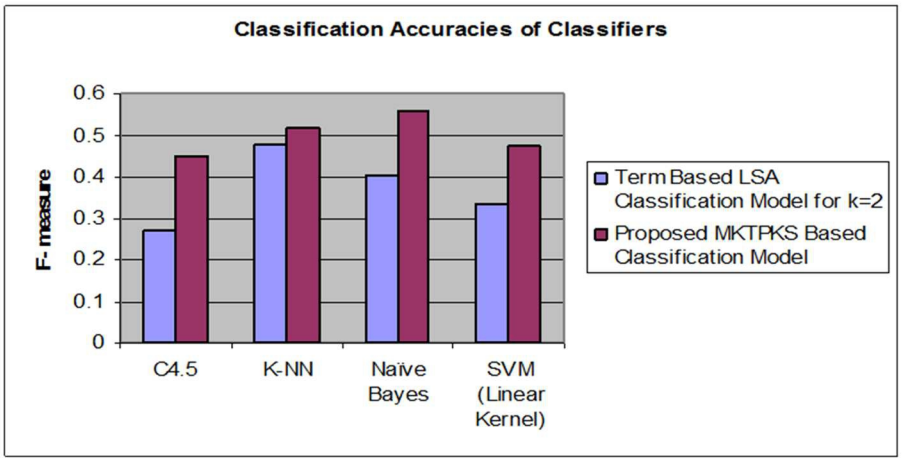

Figure 4: Comparing LSA (k =2) with MKTPKS Classification Models

The figure 4 shows that the proposed MKTPKS based classification model gave comparable values over the simple term based LSA model in the case of all the classifiers.

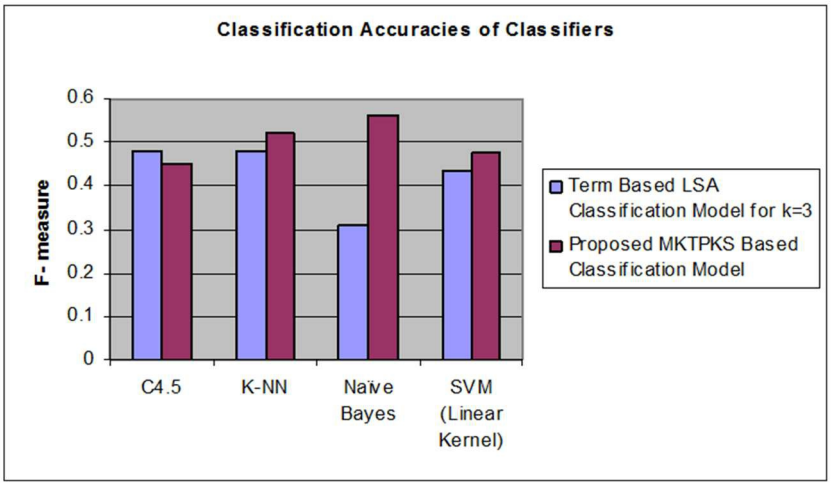

Figure 5: Comparing LSA (k =3) with MKTPKS Classification models

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

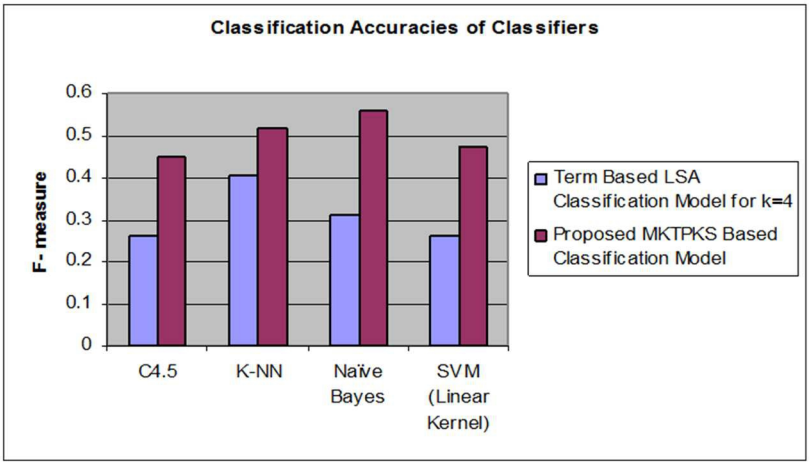

Figure 6: Comparing LSA (k=4) with MKTPKS Classification Models

The figure 6 shows that the proposed MKTPKS based classification model performs well when compared with term based LSA model. The accuracy the MKTPKS method is greater for all classifiers when compared to the LSA results.

Thus overall this study shows that the classification accuracies of the classifiers improved using proposed MKTPKS model as compared with simple term based LSA classification model.

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

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 and also the industrial collaborators for providing the data in the form of PPR to conduct the experimental studies.

References:

1. Hsieh, K-L., (2007). Employing Data Mining Technique to Achieve the Parameter Optimization Based on Manufacturing Intelligence, Journal of the Chinese Institute of Industrial Engineers 24, p.309-318.
2. Tan, H. C., P.M. Carrillo, C. Anumba, J. M. Kamara, D. Bouchlaghem, and C. Udeaja, (2006). Live capture and reuse of project knowledge in construction organizations, Knowledge Management Research and Practice 4, p.149-161.
3. Carrillo, P.M. , (2005). Lessons learned practices in the engineering, procurement and construction sector, Journal of Engineering, Construction and Architectural Management 12(3), p. 236-250.

4. Pitman, B., (1991). A system analysis approach to reviewing completed projects, Journal of Systems Management 42(6), p. 6-37.
5. Fan, W. , L. Wallace, S. Rich, and Z. Zhang, (2006). Tapping the power of text mining, Communication of the ACM 49 (9), p. 77-82.
6. Salton, G. (1989), Automatic text processing: the transformation, analysis, and retrieval of information by computer, Addison Wesley Reading.
7. Witten, I. H. and E. Frank, (2000). Weka machine learning algorithms in java, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations © 2000 Morgan Kaufmann Publishers.
8. Agrawal, R., T. Lmielinski, and A. Swami,(1993). Mining association rule between sets of items in large databases, In Proceedings of 1993 International Conference on Management of Data (SIGMOD 93) p. 207-216.
9. Han, J. and M. Kamber, (2000). Data mining: concepts and techniques, Morgan Kaufmann, San Francisco.
10. Karanikas, H. and B. Theodoulidis, (2002). Knowledge discovery in text and text mining Software , Technical Report, Centre for Research in Information Management (CRIM) UMIST, Manchester.
11. Yu, L. , S. Wang, and K. K. Lai, (2005). A Rough Set Refined Text Mining Approach for Crude Oil Market Tendency Forecasting, International Journal of System Sciences 2(1), p.33-46.
12. Menon, R. , L.H. Tong, and S. Sathiyakeerthi, (2005). Analyzing textual databases using data mining to enable fast product development processes, Reliability Engineering and System Safety 88, p.171-180.
13. Hui, S. C. and G. Jha, (2000). Data mining for customer service support, Information & Management 38, p.1-13.
14. Zhang, Y. and J. Jiao, (2006). An associative classification-based recommendation system for personalization in B2C e-commerce applications, Expert Systems with Applications (Article in Press).
15. Phaal, R. and D. Probert, (2008). Morphology analysis for technology road mapping: application of text mining, R& D Management 38 (1), p. 51-68.
16. Gibbons, W.M., M. Ranta, T. M. Scott, and M. Mantyla, (2000). Information management and process improvement using data mining techniques, International Problem Solving: Methodologies and Approaches, Proceedings Lecture Notes in Artificial Intelligence 18 (21), p. 93-98.

17. Caldas, C.H. and L. Soibelman, (2003). Automating hierarchical document classification for construction management information systems, Automation in Construction 12, p. 395-406.
18. Ur-Rahman N. and J.A. Harding, (2012). Textual data mining for industrial knowledge management and text classification: a business oriented approach, Expert System with Applications, p.2719-29.
19. Rijsbergen, C.J.V., (1979). Information Retrieval (2nd ed.), London UK: Butterworths.
20. Larose, D.T., (2005). Discovering knowledge in data : an introduction to data mining, Hoboken, New Jersey, Jhon Wiley and Sons, Inc.
21. URL: http://www.constructingexcellence.org.uk/pdf/document/KM_Guide.pdf dated: 20/07/2010
22. Kamra, J.M., C.J. Anumba, P.M. Carrillo, and N.M. Bouchlaghem, (2003). Conceptual framework for live capture project knowledge, In: R. Amer (Ed.), Construction IT: Bridging the Distance, Proc. CIBW 078 International Conference on Information Technology for Construction, Waiheke Islan, New Zealand 23-25, p.178-185.