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