Text classification using ai essay

Business, Decision Making

Text classification is the primary requirement of text retrieval systems, which retrieve texts in response to a user query, and text understanding systems, which transform text in some ay such as producing summaries, answering questions or extracting data. Existing supervised learning algorithms for classifying text need sufficient documents to learn accurately. This paper presents a new algorithm for text classification using artificial intelligence technique that requires fewer documents for training. Instead of using words, word relation I. . Association rules from these words is used to derive feature set from pre-classified text documents. The concept of naive Bases classifier is then used on derived features and finally only a single concept of genetic algorithm has been added for final classification.

A system based on the proposed algorithm has been implemented and tested. The experimental results show that the proposed system works as a successful text classifier. 1. INTRODUCTION There are numerous text documents available in electronic form. More and more are becoming available every day. Such documents represent a massive amount of information that is easily accessible Seeking value in this huge collection, organization requires much work to organize documents, but this can be automated through data mining-an artificial intelligence technique. The accuracy and understanding of such systems greatly influence their usefulness. The task of data mining is to automatically classify documents into predefined classes based on their content.

Many algorithms have been developed to deal with automatic text classification. The most common techniques used for this purpose including naive Bases classifier, association rule mining, genetic algorithm, decision

tree etc. Association rule mining [1] finds interesting association or correlation among a large set of data items. The discovery of these relationships among huge amounts of transaction records can help in many decision making process. On the other hand, the naiveBases classifier uses the maximum a posteriori estimation for learning a classifier. It assumes that the occurrence of each word in a document is conditionally independent of all other words in that document given its class.

Although the naive Bases works well in many studies [3], [6] it requires a large number of training documents for training accurately. Genetic algorithm starts with an initial population which is created consisting of randomly generated rules. Each rule can be represented by a string of bits.

Typically, the fitness of a rule is assessed by its classification accuracy on a set of training examples. This paper presents a new algorithm for text classification. Instead of using words, word relation I. E. Association rules is used to derive feature set from pre-classified text documents. The concept of naive Bases classifier is then used on derived features and finally a concept of genetic algorithm has been added for final classification.

A system based on the proposed algorithm has been implemented and tested. The experimental results show that the proposed 2. RELATED WORK Because of the strong research interest, as shown in the literature, a number of algorithms for text classification have been developed During this work it was tried to consider the established efficient approaches that were found.

This section presents a study on some research works to give a view on some techniques used to classify text. Bayesian classifiers have been gaining popularity lately, and have been found to perform surprisingly well. Classification on new examples is performed with Bases rule by selecting the class that is most likely to have generated the example. Accuracies found on several research works using Ana; e Bases classification were 41% to 74%.

Research on text classification using the concept Of association rule of data mining where naive Bases classifier was used to classify text and finally showed the dependability of the naive Bases classifier with associated rules But since this method ignores the negative calculation for any specific class determination in some cases, accuracy may fall. Accuracies found up to 60%. In text classification using decision tree, authors showed an acceptable accuracy using 76% training data [9], while it is possible to achieve good accuracy using only 40 to 50% of total data as training data. Text classification based on genetic algorithm showed satisfactory performance using 69% training ATA, but this process requires time-consuming steps to classify the texts [4]. Text classification in conference management research has emphasized on the effectiveness of supervised machine learning techniques on experimental data sets. In contrast, the amount Of research in the area of real-world application is sparse especially in the domain of text classification. In the area of spam detection most of the widely used spam detection products incorporate some kind of personalized text classifier with the majority using incrementally trained naive Bases classifiers.

In traditional conference management systems problems an occur in the phase where the author has to choose the research topic under which the paper should be filed. The author can be irresolute and uncertain in selecting the topic Of the paper. This feeling can be reinforced when the categories are not precisely described. In this system the classifier is trained with examples from previous conferences and so the author should be guided to the correct category. This work shows the average amount of correct classification over all classes reached a value of 74. 46% [10]. 3. BACKGROUND STUDY 3.

1 Association Rule Association rule mining finds interesting association r correlation among a large set of data items. In short, association rule is based on associated relationships. Association rules are generated on the basis of two important terms namely minimum support threshold and minimum confidence threshold. 3. 2 Naive Bases Classifier Bayesian classification is based on Bases theorem. A simple Bayesian classification namely the naive classifier is comparable in performance with decision tree and neural network classifiers.

Naive Bases classifier assumes that the effect of an attribute value on a given class is independent of the values of the other attributes. This assumption is called class conditional independence. It is made to simplify the computations involved and, in this sense, is considered “ naive”. While applying naive Bases classifier to classify text, each word position in a document is defined as an attribute and the value of that attribute to be the

word found in that position. Here naive Bases classification can be given by: VAN = ragman P (Vs.

) n P (as Vs.). Here VAN is the classification that maximizes the probability of observing the words that were actually found in the example documents, subject to the usual naive Bases independence assumption. The first term an be estimated based on the fraction of each class in the training data. The following equation is used for estimating the second term: ink+l n + I vocabulary I where n is the total number of word positions in all training examples whose target value is Vs.

, ink is the number of items that word is found among these n word positions, and vocabulary I is the total number of distinct words found within the training data. 3. Genetic Algorithm Genetic algorithms are a part of evolutionary computing which is a rapidly growing area of artificial intelligence. As we can guess, genetic algorithms are inspired by Darning’s theory of evolution. Simply said, problems are solved by an evolutionary process resulting in a best (fittest) solution (survivor) in Other words, the solution is evolved. In general, genetic algorithm starts with an initial population which is created consisting of randomly generated rules. Each rule can be represented by a string of bits.

Based on the notion of survival of the fittest, a new population is formed to consist of the fittest rules in the current population, as well as offspring of these rules. Typically, the fitness of a rule is assessed by its classification accuracy on a set of training examples. Offspring are created by applying genetic operators such as crossover and mutation. 4. PROPOSED ALGORITHM

The proposed method for classifying text is an implementation of a hybrid method consisting of association rule, naive Bases classifier, and genetic algorithm.