TABLE OF CONTENTS

LIST OF FIGURES

ABSTRACT

Text Analysis is important, emerging, research area, because plenty of text resources growing rapidly through the internet and digital world. In the text data analysis text categorization is one of the vital techniques. Traditional text categorization methods are not able to handle well with learning across different domains. Cross-domain classification is more challenging problem than single domain classification .In this project cross domain text categorization is implemented using EM(expectation maximization) algorithm.

Key words: pre-processing, word to vector, K-Means, EM algorithm.

CHAPTER 1

1.1 INTRODUCTION

In practical use, the Web has become the largest source of training data; however, it is cost-ineffective to hand label each extracted documents from the Web. Thus many algorithms using few labeled data and large unlabeled data to train a text classifier were presented proposed a method without requiring labeled training data. The purpose of the paper is extended from the previous work and focuses much on using EM techniques to extract more reliable training data.

Data mining is the extraction of useful knowledge from large amount of data. Data mining tools can provide solution to the business problems that were to too time consuming when done manually. Classification is one of the important aspect which comes under data mining and is a predictive modeling technique.

It is used to group the data or documents in some pre-defined classes i.e. to classify them according to attribute matching approach. Most companies have huge amount of data available which has to be refined, to apply classification on them. Data mining techniques can be implemented rapidly on existing software and hardware platforms to enhance the value of existing information resources, and can be integrated with new products and systems as they are brought on-line. Classification techniques are used in various real world problems with respect to application domain as well as for various research purposes relevance to today's business environment as well as a basic description of how data warehouse architectures can evolve to deliver the value of data mining to end users.

Text mining is the task of routinely sorting a set of documents into categories. When two or greater domains are concerned in a selected text document then it is called CROSS DOMAIN. Internet is a tremendous repository of disparate information growing at an exponential rate. The dynamic growth of net generates not only huge wide variety of text documents but additionally wide sorts of text documents in a end result of documents being generated in various domain. Efficient and effective document retrieval and classification structure are required to turn the huge amount of facts into useful information and eventually to knowledge.

Applications of textual content mining are publishing and media, telecommunications, energy and other offerings industries, data era region and internet, banks, coverage and financial markets, political institutions, political analysts, public administration and prison documents,

pharmaceutical and research organizations and healthcare Clustering is a procedure which partitions a given facts set into homogenous businesses primarily based on given capabilities such that similar gadgets are kept in a set whereas dissimilar gadgets are in different agencies. It is the most essential unsupervised studying problem. It offers with finding structure in a group of unlabeled data.

The number of text documents are growing with the arrival of the web and development of world wide web. The huge growth of text documents are incredible to manually classify. In general statistical approaches are applied in single domain for text classification. These approaches are based on the word occurrence i.e. frequency of one or more words during a given document. But these approaches don't work well with multiple domains. So to achieving this goal one of the most important challenges is the problem of learning topics in text documents that belong to different domains. In this paper cross domain text categorization is implemented using expectation maximization algorithm.

CHAPTER 2

LITERATURE SURVEY

This chapter gives the overview of literature survey. This chapter represents some of the relevant work done by the researchers.

Many existing techniques have been studied by the researchers on text categorization problem, few of them are discussed below.

The enormous amount of data stored in unstructured texts can-not simply be used for further processing by computers, which usually handle text as simple sequences of character strings. Therefore, specific (pre-)processing methods and algorithms are required in order to extract useful patterns. Text mining refers generally to the process of extracting interesting information and knowledge from unstructured text. This article discuss text mining as a young and interdisciplinary field within the intersection of the related area information retrieval, machine learning, statistics, linguistics and particularly data mining. It describes the main analysis tasks preprocessing, classification, clustering, information extraction and visualization. In addition, it[3] briefly discuss a number of successful applications of text mining.

The area of textual content mining seeks to extract useful statistics from unstructured textual statistics via the identification and exploration of interesting patterns. The strategies employed commonly do no longer contain deep linguistic analysis or parsing, but depend upon simple "bag-of-words" textual content representations based on vector space. Several approaches[4] to the identification of styles are discussed, such as dimensionality reduction, automated class and clustering. Pattern exploration is illustrated thru two programs from our latest work: a category-based Web meta-seek engine and visualization of co-authorship relationships robotically extracted from a semi-structured series of files.

Document mining is the process of deriving splendid records from huge collections of documents like news feeds, databases, or the Web. Document mining tasks consist of cluster evaluation, category, era of taxonomies, data extraction, trend identity, sentiment analysis. The Challenges in Document Mining that customers can expect are as many clusters as they become aware of topics within the end result set, the files within each cluster are semantically similar to every other, every cluster is classified intuitively. In order to obtain a satisfying solution, the state-of-artwork of concepts and algorithms from facts retrieval, unsupervised learning, information extraction, and herbal language processing need to be combined[5] in a user - focused manner.

Text category is one of the core applications in facts mining because of the big amount of uncategorized textual records available. Training a text classifier effects in a class version that

reflects the characteristics of the area it was discovered on. However, if no training facts is available, labelled information from a related but distinctive area might be exploited to perform cross-domain class. Authors aim to appropriately classify unlabelled weblogs into usually agreed upon newspaper categories the usage of labelled records from the information area. The labelled information and the unlabelled blog corpus are incredibly dynamic and hourly developing with a topic drift, so the class wishes to be efficient. Experiments showed that this algorithm achieves a comparable accuracy than k-Nearest Neighbour (k-NN) and Support Vector Machines (SVM), yet at linear time cost for education and category. We[2] inspect the classifier performance and generalization ability the usage of a unique visualization of classifiers. This technique is to apply a quick novel text class algorithm to carry out efficient cross-domain category.

CHAPTER 3

ANALYSIS

3.1 SYSTEM ANALYSIS

Systems analysis is a problem solving technique that decomposes a system into its component pieces for the purpose of the studying how well those component parts work and interact to accomplish their purpose.

3.1.1 PROBLEM STATEMENT

The number of text documents are growing with the advent of the internet and development of world wide web. The huge growth of text documents are incredible to manually classify. In general statistical approaches have been applied in single domain for text classification. These approaches are based on the word occurrence i.e. frequency of one or more words in a given document. But these approaches don't work well with multiple domains. So to achieving this goal one of the most important challenges is the problem of learning topics in text documents that belong to different domains.

3.1.2 EXISTING SYSTEM

Rocchio's learning algorithm is in the classical IR tradition. It was originally designed to use relevance feedback in querying full-text databases, Rocchio's Algorithm is a vector space method for document routing or filtering in informational retrieval, build prototype vector for each class using a training set of documents, i.e. the average vector over all training document vectors that belong to class ci, and calculate similarity between test document and each of prototype vectors, which assign test document to the class with maximum similarity.

K-NN classifier is a case-based learning algorithm that is based on a distance or similarity function for pairs of observations, such as the Euclidean distance or Cosine similarity measure's This method is try for many application. Because of its effectiveness, non-parametric and easy to implementation properties, however the classification time is long and difficult to find optimal value of k .The best choice of k depends upon the data; generally, larger values of k reduce the effect of noise on the classification, but make boundaries between classes less distinct.

A major drawback of the similarity measure used in k-NN is that it uses all features in computing distances. In many document data sets, only smaller number of the total vocabulary may be useful in categorizing documents.

Some machine learning algorithms like Naive bayes (NB), Support Vector machine (SVM) also used.

3.1.3 PROPOSED SYSTEM

With the initial data, we apply a greedy EM algorithm to determine the proper number of relevant concepts of each class. Then, for each mixture, refine its corresponding distribution depending on more ”augmented” training data extracted via sending several queries into the search engine. The queries are those with the highest mutual information in distributions, i.e. keywords. In this step, an augmented EM algorithm is applied to iteratively update parameters in each distribution with the increasing of likelihood. Experimental results have shown the great potential of the proposed approach in creating text classifiers without the pre-request of labeled training data.