CHAPTER 1

INTRODUCTION

1.1 GENERAL

This chapter provides an introduction to the thesis, beginning with general explanation, followed by the motivation for the research, problem statement, the proposed solution, and an account of the main contributions made. Finally, the structure of the rest of the thesis is presented.

Rapid advances in digital data acquisition and storage technology have led to the collection of vast amount of data and huge databases. Terabytes or petabytes of data are available in this data age. Extracting useful information from these databases is a big challenge. The process of finding precious nuggets from a great deal of raw material is defined by the term, mining. Based on the thing which is being mined, the term differs. Mining data is termed as data mining, and mining a text is termed as text mining.

Data mining is defined as the process of automatically discovering useful information in large data repositories. Data is referred to as facts or numbers which are highly structured. Text mining, also known as intelligent text analysis or knowledge discovery in a text, refers to the process of extracting interesting and non trivial information and knowledge from an unstructured text. Examples of unstructured text include, Microsoft word, HTML, PDF documents etc. It is an interdisciplinary field that draws on information retrieval, data mining, machine learning, statistics and
computational linguistics. Text mining is defined as “the non trivial extraction of implicit, previously unknown, and potentially useful information from (large amount of) textual data”. It is an exploration and analysis of textual (natural-language) data by automatic and semi-automatic means, to discover new knowledge.

A substantial portion of information is stored as text, such as news articles, technical papers, books, digital libraries, email messages, blogs and web pages. This has made research in text mining very active. An important goal of text mining is to derive high quality information from a text, which is done through the discovery of patterns and trends by means of statistical pattern learning, topic modelling etc.

In text mining, the text is composed of two fundamental units, namely, the document and the term. A document is generally used as a name for any text, such as chapters, sections, paragraphs, or even e-mail messages. A term is defined as a word, word-pair, or phrase within a document.

The text mining process includes various processes, like text pre-processing, text transformation (feature generation), feature selection, pattern discovery, and finally, evaluation. The text pre-processing methods include tokenization, i.e., a text document is split into a stream of words by removing all punctuation marks, and by replacing tabs and other non-text characters by white spaces (Miloš Radovanovic & Mirjana Ivanovic 2008), filtering methods like stop word removal and stemming, i.e., to build the basic form of words.

The next process is text transformation. In this process, the text documents are represented in several ways. One of the representations is a bag of words. In this approach, words are considered independently, irrespective of the order in which they appear. In this technique, each word is
represented as a dimension in the document space, and the document is represented as a vector. The weight of each word is calculated, based on the frequency of its appearance. The document D is transformed into a matrix format in BOW representation (Miloš Radovanović & Mirjana Ivanović 2008), which enables various matrix decomposition techniques to be applied for various tasks, like clustering, classification, summarization etc. The degree of similarity between two documents is calculated, using various distance/similarity measures like cosine similarity, Euclidean distance etc. The words will be ranked based on their weights, and based on that, the top score words will be retrieved for further processing.

Another method of document representation is the vector space model (Andrew Skabar & Khaled Abdalgader 2010). This method represents the semantic similarity of the documents. The documents are represented as vectors and the unique elements in the corpus are represented as elements. The similarity between documents is calculated, using vector arithmetic operations. The weights of the terms are calculated using Term-Frequency and the Inverse Document method (TF-IDF). The similarity values between the documents are constrained between zero and one. The most similar documents have the cosine value closer to one, and the dissimilar documents have values closer to zero (Fox 2005).

The third phase of the text mining process is feature selection. Feature selection is the process of selecting a subset from the original set based on some criteria. Feature selection methods are divided into supervised and unsupervised, based on whether the class label information is required or not. Information Gain (IG) and CHI- Square statistic are the two supervised methods, and Document frequency (DF), Term- Strength (TS), Entropy-based ranking (En), and Term Contribution (TC), are the four unsupervised methods, which are used for feature selection (Tao Liu et al 2003).
After feature selection, various text mining tasks are applied to discover new patterns. The text mining tasks include text categorization, text clustering, concept/entity extraction, sentiment analysis, document summarization and entity-relation modelling. The discovered patterns will be displayed to the end user. Then they are evaluated, using various metrics like precision, recall, F-score etc.

This thesis focuses on feature selection methods, and the text clustering process. Feature selection based on words and sentences has been developed. Also, clustering of the text documents, using the association rule, Must Link and Cannot Link algorithm, and finally based on WordNet has been successfully implemented.

1.2 MOTIVATION FOR THIS THESIS

The objective of this thesis is centred on the feature selection technique and text clustering. Feature selection methods improve the efficiency and performance of the text clustering algorithm. To partition a given document collection, into clusters of similar documents, a choice of good features along with good clustering algorithms is of great importance. The feature selection method sorts out terms on the basis of a numerical measure computed from the document collection to be clustered, and selects a subset of the terms. For example, term frequency (tf) is a simple term extraction technique based on the occurrences of terms in a document, by assigning a weight to each term, to indicate its importance. In addition, more complex techniques of keyword extraction involve using statistical techniques (Matsuo & Ishizuka 2004), or linguistic techniques “Natural Language Processing (NLP)” (Sado et al 2004) and others, based on the specific domain, like domain specific ontologies (Hulth et al 2001). Liu et al (2009) concluded that even though the unsupervised approaches were simple and effective, their system performance compared to the human performance has
been low, and suggested that more work was needed for this domain. This motivated to drive this thesis towards feature selection methods, which led to the development of a genetic based keyword extraction technique and sentence based extraction technique.

Next, text clustering is defined as a technique that organizes a large quantity of unordered texts into a small number of meaningful and coherent clusters (Anna Huang 2008). In general, clustering organizes a collection with a substantial number of data objects into a much smaller number of coherent groups (Anna Huang 2008) termed as a cluster, such that the objects in a cluster are similar to one another, but dissimilar to objects in other clusters. Text clustering is an unsupervised process, because the process has unlabeled training examples, and the class label information is not present. Clustering has been widely used in many applications such as business intelligence, web search, pattern recognition etc. The best known methods of text clustering are partitioning methods and hierarchical methods.

These methods are found to have high efficiency only in the case of low dimensionality. But the text documents will not have a definite dimension to describe the content of the document. Therefore, it is preferred to reduce the content and to increase the efficiency of the text clustering process.

These also form the basic motivation for this thesis. To overcome the disadvantage, an association rule based text clustering method has been developed. Also, a Must Link and Cannot Link algorithm has been developed to cluster documents based, on the relationship between words. Word meaning is an important factor to be considered while clustering text documents. Based on this, a WordNet based text clustering algorithm has also been developed and implemented successfully.
Figure 1.1 Overall System Architecture
Figure 1.1 shows the block diagram of the entire system. In this, first, the features are selected using various methods, like term frequency method, genetic based method, and sentence based method. Then, the extracted features are clustered, using various methods like association rule mining, Must Link and Cannot Link algorithm, and WordNet based method.

1.3 PROBLEM STATEMENT AND PROPOSED SOLUTION

The main objective of this thesis is to provide an efficient feature selection technique, and improved clustering techniques. Most of the feature selection methods concentrate on the number of occurrence of the word and not on the position of the word. So taking this into consideration, an improved genetic keyword extraction technique has been developed. Also, most of the clustering algorithms are based on the keywords that are extracted. As an improvement of this, instead of keywords extraction, sentences are extracted based on some nine variables.

In the clustering phase, it is been analyzed that the performance of the clustering algorithm degrades as the size of the document increases. So, an association based hierarchical and K-means algorithm has been developed. Also, based on the similarity between the words, a Must Link and Cannot Link algorithm has been developed to cluster the documents using the keywords that are extracted using genetic based keyword extraction algorithm. Also, it is been found that the clustering performance will improve when the documents are clustered based on semantics. So, a Wordnet based clustering algorithm has been developed based on the sentences that are previously extracted.

The various modules of this research are given below:

1. Feature Selection
i. Term Frequency method

ii. Genetic based keyword Extraction

iii. Sentence based Extraction

2. Clustering

i. Association Rule based Hierarchical algorithm

ii. Modified K-Means algorithm

iii. Must Link and Cannot Link algorithm (MLCL)

iv. WordNet based clustering algorithm.

Based on this, the thesis is divided into two parts. The first part deals with keyword extraction and the second part is on the various text clustering procedures. Keyword extraction is done using three methods, namely, the Term Frequency method, genetic based method, and sentence based method.

In the first part of the thesis (Chapter 3), clustering based on the association rule has been described. The Association rule is combined with two major fundamental clustering methods such as partitioning methods and hierarchical methods, and the algorithms are analysed for their performance.

The hierarchical method creates a hierarchical decomposition of the given set of objects. A hierarchical method can be classified as being either agglomerative or divisive, based on how the decomposition is formed. The hierarchical method suffers from the fact that once a step (i.e. merge or split) is done, it can never be undone (Pankaj Jajoo 2008).
So, considering this disadvantage, the association algorithm is combined with the hierarchical algorithm. In this, the association algorithm is first applied to find the associated documents. Then the Hierarchical Clustering algorithm is applied to those documents and they are clustered. This approach has been proved to provide improved cluster quality and reduced time complexity.

The partitioning method constructs ‘k’ partitions of the data, given a set of n objects. Each partition center in space is represented using centroids and the elements are assigned to them based on their closeness. Every time, the centroids are updated based on the mean distance of these elements.

The disadvantage of the K-means algorithm is the initial selection of the K value. To overcome this, the association rule is integrated with the K-means algorithm. The algorithm is evaluated for its efficiency using Reuters-21578 data set.

In the second part (Chapter 4), a genetic based method is described for extracting keywords from the documents. The algorithm initially assigns weights to the terms based on some equations. Then, genetic procedures like crossover, mutation and fitness functions are applied to those keywords and the final set of keywords is retrieved. The algorithm is analysed for the efficient and effective retrieval of terms, using both synthetic and real data sets.

In the latter half of the second part of the thesis, the clustering process (Chapter 5) based on the Must Link and Cannot Link algorithm is described. This algorithm consists of three phases. In the first phase, related words are identified using the MLCL algorithm. In the second phase, the words are grouped to form clusters, using three main equations. Finally, in the third phase, the clusters are optimized using Gaussian parameters. The output is visualised using scatter plots and the results are evaluated.
Finally, a sentence based clustering method using WordNet is described (Chapter 6). In this, sentences are extracted using various formulae. All the sentences are ranked based on all these parameters and the weightage is calculated for each sentence. Finally, a certain percentage of the topmost sentences are extracted for each document. Then the meaning of the documents are found using WordNet. The semantic similarity between all the documents is also found, using WordNet. If the semantic similarity between the documents is above 80%, then the documents are clustered. The process is continued for the entire document set, and the final results are displayed.

1.4 CONTRIBUTION OF THE THESIS

The main contribution of the thesis is the development of an efficient clustering algorithm. The original contributions are:

Developing a Genetic based keyword extraction to extract the keywords from the text documents more efficiently.

Implementing an association rule based hierarchical algorithm to overcome the disadvantage of the hierarchical algorithm.

Implementation of a modified K-means algorithm based on the association rule algorithm, to solve the initial centroid selection problem.

Implementation of a Must Link and Cannot Link algorithm to cluster text documents, based on the keywords extracted using the genetic based keyword extraction technique.

Implementation of a WordNet based clustering algorithm, to cluster documents based on semantics.
1.5 ORGANIZATION OF THE THESIS

The research work presented in the thesis is organized into six chapters. The contents and goal of each of them are as follows:

Chapter 1: This chapter provides an introduction to the thesis, beginning with the problem definition, the objectives and contributions of the thesis. Finally, the structure of the rest of the thesis is presented.

Chapter 2: This chapter reviews the existing work related to keyword extraction, sentence extraction, and clustering.

Chapter 3: This chapter discusses the implementation of the Genetic based keyword extraction technique, which extracts the important keywords from the text documents, which serves as the input for the algorithm discussed in Chapter 5.

Chapter 4: This chapter discusses the implementation of the hierarchical algorithm and the K-means algorithm, based on the association rule mining algorithm. The test results are also discussed.

Chapter 5: In this chapter, the keywords extracted in chapter 3 are clustered, using the Must Link and Cannot Link algorithm (MLCL). The evaluation results are also discussed.

Chapter 6: This chapter defines how effectively text documents can be clustered, based on semantics using WordNet. The final clusters are evaluated and the results are also tabulated.

Chapter 7: This chapter consists of the Conclusion and Future Work; it summarizes the results, draws conclusions, and offers suggestions for future work.