Chapter 1

Introduction

1.1 Text Mining

The recent abundance of digital information available electronically has made the organization of textual information into an important task. Text mining is a burgeoning new technology for discovering knowledge from text data. With the fast growth of the number of pages on the World Wide Web, text mining plays a key role in managing information and knowledge, and is therefore has become an active research area. Text mining corresponds to the extension of the data mining approach to textual data and is concerned with finding useful or interesting patterns, model, directions, trends, or rules from unstructured text.

Text mining, also known as intelligent text analysis is the process of extracting interesting and non-trivial information and knowledge from unstructured text. Text mining is a young interdisciplinary field, which draws on information retrieval, data mining, machine learning, statistics and computational linguistics. Text mining is about looking for patterns in natural language text, and may be defined as the process of analyzing text to extract information from it for particular purposes. In simple words, Text mining is about uncovering patterns in data when the data is text. Text mining includes several text processing and classification techniques, such as text categorization, clustering and retrieval and information extraction. Traditional
information retrieval techniques become inadequate for the increasingly vast amount of text data. Typically only a small fraction of many available documents will be relevant to a given individual or user. Without knowledge of what could be in the documents, it is difficult to formulate effective queries for analyzing and extracting useful information from the data. Users need tools to compare different documents, rank the importance and relevance of the documents, or find patterns. Thus text mining has become an increasingly popular and essential theme in Data mining.

[41] describes a Text Data Mining (TDM) architecture that unifies information retrieval from text collections, information extraction from individual texts, knowledge discovery in databases, knowledge management in organizations, and visualization of data and information.

Figure 1.1 presents a similarly broad model for a six-step Text Data Mining process in which three main functions are performed:

- **Data Collection:** including Source Selection and Text Selection.

  *Source Selection* is the process of selecting sources to exploit. Source selection requires awareness of the available sources, domain knowledge, and an understanding of the goals and objectives of the data mining effort.

  *Text Retrieval* is the process of discovering, selecting, and obtaining individual texts from the selected sources. This process might be fully automated or it may be performed interactively by a domain expert.

- **Data Warehousing:** including Metadata Assignment and Data Storage.

  *Information Extraction* is the process of analyzing unrestricted text in order to extract information about pre-specified types of events, entities or relationships.
Data storage is the process of providing storage of and access to data. Data models that specify known relationships in the data can be stored with the data to facilitate subsequent processing.

- Data Exploitation: including Data Mining and Data Presentation.

Data Mining is the process of fitting models to data.

Presentation is the process of explaining and visualizing data mining results to support evaluation of data quality, assessment of whether the selected model is appropriate, and interpretation of the model.

Figure 1.1: Text Data Mining Architecture.

Two main learning paradigms in Text Data mining are supervised learning and unsupervised learning. Supervised learning refers to learning with a teacher, typically in situations where one has a set of training data whose class labels are known. This is called classification. On the other hand, learning without a teacher refers to unsupervised learning, the situation where training data are not labeled and the typical objective is to find natural grouping (or clusters) among the given patterns. This is called clustering.

- Classification: Classification is the process of finding a set of models (or functions) that describe and distinguish data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose
class label is unknown. The derived model is based on the analysis of a set of training data (i.e., data objects whose class label is known).

- Clustering: Clustering is the process of grouping a set of physical or abstract objects into classes of similar object. A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in the other clusters. Cluster analysis is an important human activity. Early in childhood, one learns how to distinguish between animals and plant by continuously improving subconscious clustering schemes.

There are many basic reasons for interest in unsupervised learning. Some are like collecting and labeling a large data set of sample patterns can be surprisingly costly, in early stages of an investigation it may be valuable to perform exploratory data analysis and thereby gain some insight into the nature or structure of the data. By clustering, one can identify dense and sparse regions and, therefore, discover overall distribution patterns and interesting correlations among data attributes.

1.2 Clustering

The ability to form meaningful groups of objects is one of the most fundamental modes of intelligence. Humans perform this task with remarkable ease. In early childhood one learns to distinguish, for example, between cats and dogs or apples and oranges. However, enabling the computer to do this task of grouping automatically is a difficult and often ill-posed problem. Cluster analysis is a tool for exploring the structure of data. The core of cluster analysis is clustering; the process of grouping objects into clusters such that objects from the same cluster are similar and objects from different clusters are dissimilar. Objects can be described in terms of
measurements (e.g., attributes, features) or by relationships with other objects (e.g., pairwise distance, similarity). Unlike classification, clustering does not require assumptions about category labels that tag objects with prior identifiers. Therefore, clustering is an unsupervised learning technique versus classification, which belongs to supervised learning.

The need to structure and learn from the vigorously growing amounts of data has been a driving force for making clustering a highly active research area. Humans are not able to easily discover knowledge from the glut of information in databases without the use of summarization techniques. Basic statistics (such as mean, variance or histograms) can provide an initial feel for the data. However, more intricate relationships among the objects, among the features, and between both can be discovered through cluster analysis.

In general, major clustering techniques can be classified into the following categories.

1. **Partitioning Methods** decompose a set of data objects into a given number of disjoint clusters which are optimal in terms of some predefined criteria functions. Some examples of this approach are k-means [33], PAM, CLARA [40], and CLARANS [55].

2. **Hierarchical Methods** creates a hierarchical decomposition of the given set of data objects. A hierarchical method can be classified as agglomerative or divisive based on how the hierarchical decomposition is formed. The agglomerative approach starts with each object forming a separate group. It successively merges the objects or groups close to each other until all of the groups are merged into one or until a termination condition holds. The divisive
approach starts with all the objects in the same cluster. In each successive
iteration, a cluster is split up into smaller clusters until eventually each object
is in one cluster or until a termination condition holds. Some examples of this
approach are BIRCH [66], CURE [64], and Chameleon [19].

3. **Density-Based Methods** use the notion of density (number of objects or data
points), where a given cluster continues to grow as long as the density in the
neighborhood exceeds some threshold; that is, for each data point within a
given cluster, the neighborhood of a given radius has to contain at least a
minimum number of points. Some examples of this approach are DBSCAN
[15], OPTICS [44], and DENCLUE [25].

4. **Grid-Based Methods** quantize the object space into a finite number of cells
that form a grid structure. All of the clustering operations are performed on the
grid structure. Some examples of this approach are STING [68], CLIQUE
[53], WaveCluster [20].

5. **Model-Based Clustering Methods** hypothesize a model for each of the clusters
and find the best fit of the data to the given model. Some examples of this
approach are COBWEB [16], SOM [65].

Some clustering algorithms integrate the ideas of several clustering methods.
Furthermore, some applications may have clustering criteria that require the
integration of several clustering techniques.
1.3 Applications

Clustering is used in many areas, including artificial intelligence, biology, customer relationship management, data compression, data mining, information retrieval, image processing, machine learning, marketing, medicine, pattern recognition, psychology, recommender systems and statistics. Below few examples are given:

- Biology

In biology, clustering is used, for example, to automatically build a taxonomy of species based on their features. More recently, biologists have applied clustering to analyze the large amounts of genetic information that are now available. For example, clustering has been used to find groups of genes that have similar functions. There is considerable interest in estimation of phylogenetic trees from gene sequence data. Another application of clustering is to better understand gene functions in the biological processes in a cell. A key step in the analysis of gene expression data is the detection of groups of genes that manifest similar expression patterns.

- Psychology and Medicine:

An illness or condition frequently has a number of variations, and cluster analysis can be used to identify these different subcategories. For example, clustering has been used to identify different types of depression. Cluster analysis can also be used to detect patterns in the spatial or temporal distribution of a disease. Cluster analysis of biomedical images is used for analysis and recognition in medical application.
- Business:
Cluster analysis is widely used in market research [21] to partition the general population of consumers into market segments and to better understand the relationships between different groups of consumers/potential customers. Another growing application area is customer relationship management, where data collected from multiple touch-points (e.g., web surfing, cash register transactions, call center activities) has become readily available. This data contains valuable knowledge of customer behavior that can help optimize marketing, bundling, and pricing strategies. Clustering can summarize data to a manageable level by forming, for example, groups of customers with similar profiles.

- Information Retrieval:
The World Wide Web consists of billions of web pages, and clustering is widely used for the organization of this huge collection of documents. Below few examples are given.

1. Clustering can be used to group the search result, so that similar documents appear together. For example the clustered results panel returned by the Vivisimo search engine (http://vivisimo.com) is a more effective user interface for understanding what is in the search result than a simple list of documents.

2. When a user enters a query into a search engine, the system often brings back many different pages. The Scatter/Gather interface uses text clustering as a way to group document according to the overall similarities in their content. Scatter/Gather is so named because it allows the user to
scatter documents into clusters, or groups, then gather a subset of these groups and re-scatter them to form new groups. This process is repeated until a cluster of interest is found.

III. Cluster-based navigation is an interesting alternative to keyword searching where users prefer browsing over searching because they are unsure about which search terms to use. As an alternative to the user-mediated iterative clustering in Scatter-Gather, we can also compute a static hierarchical clustering of a collection that is not influenced by user interactions. Google News is an example of this approach. In the case of news, we need to frequently recompute the clustering to make sure that users can access the latest breaking stories. Clustering is well suited for access to a collection of news stories since news reading is not really search, but rather a process of selecting a subset of stories about recent events.

IV. The fourth application of clustering exploits the cluster hypothesis directly for improving search results, based on a clustering of the entire collection. A standard inverted index is used to identify an initial set of documents that match the query, but then other documents are added from the same clusters even if they have low similarity to the query. For example, if the query is car and several car documents are taken from a cluster of automobile documents, then documents can be added from this cluster that use terms other than car (automobile, vehicle etc). This can increase recall since a group of documents with high mutual similarity is often relevant as a whole.
1.4 Requirements of a Clustering Algorithm

Many clustering algorithms have been proposed in the literature, and they should satisfy the following requirements.

- Should be *scalable* to deal with real world data sets which can have a very huge number of samples. Many clustering algorithms work fine on small data sets, but fail to handle large data sets efficiently.

- Should classify *accurately* so that there is high intra-cluster similarity and low inter-cluster similarity, i.e., objects within the same cluster should be similar but are dissimilar to objects in other clusters. An external evaluation method is commonly used for examining the accuracy of a clustering algorithm.

- Should be able to deal with *different types of attributes* such as binary, categorical (nominal), and ordinal data, or mixtures of these data types.

- Many clustering algorithms find spherical clusters with similar size and density. However, a cluster could be of any shape and the clustering algorithm should be able to detect clusters with *arbitrarily shape*.

- Should require *minimum domain knowledge* to determine input parameters. Many clustering algorithms require the user to specify some input parameters, e.g., the number of clusters. However, the user often does not have such prior domain knowledge. Clustering accuracy may degrade drastically if an algorithm is too sensitive to these input parameters.

- Should be able to deal with *noise and outlier* in the dataset. Outliers may have significant importance. Finding these outliers is highly non-trivial, and removing them is not necessarily desirable.
• Should be able to deal with high dimensionality of feature set. Data such as image collection or document collection have a very high number of features. Here one has to face the curse of dimensionality [17].

• The resulting clustering solutions should be interpretable and usable.

1.5 Contributions

The goal of this thesis is to develop efficient clustering algorithm to deal with complex, high-dimensional, and sparse data which are characteristics of document datasets. The specific contributions of this thesis are as follows:

• Proposal of a new similarity measure for document dataset based on the concepts of fuzzy sets. The impact of different similarity functions on the clustering results is studied.

• Development of a new document clustering algorithm using a fuzzy approach. The algorithm is agglomerative and requires just one pass through the dataset and only the compact representations of the clusters are kept in the memory at any given time.

• Development of a two phase clustering approach for large data. In the first phase, a single pass over the database is used to produce an in-memory summary of the data set. In the second phase, the in-memory summary of the data set obtained in the previous phase is merged based on the concepts of neighbors and links.
1.6 Organization

In the following chapter, background and related work are discussed. In Chapter 3 [36], we proposed a new similarity measure and compared with other commonly used similarity function for document datasets. In Chapter 4 [2, 37] we present our proposed document clustering algorithm where the concepts of fuzzy sets have been used. In Chapter 5 [39], we present our two-phase clustering approach to deal with large datasets. Chapter 6 concludes the thesis with a summary of the contributions made and presents some future directions of research.