Chapter 1: Introduction to Text Summarization

This chapter briefly introduces Text Mining, followed by its superset Data Mining and then Text Summarization. Next, the discussion is about the need for the Text Summarization followed by two basic types of summarization: abstractive and extractive. Subsequent sections discuss about the various approaches available for summarization along with advantages and disadvantages of the same. Finally, the chapter ends with the problem statement of the work that has been carried out and the objective to be met for the problem to be solved.

1.1 Introduction

Text Mining is a flourishing and thriving field that attempts to find meaningful information from textual or rather unstructured data. It may beloosely characterized as the processofanalyzing text to extract information that is useful for particular purposes. Compared with the kind of data stored in databases, text is unstructured, amorphous, and difficult to deal with algorithmically. Nevertheless, in modern culture, text is the most common vehicle for the formal exchange of information. The field of Text Mining usually deals with texts whose function is the communication of factual information or opinions, and the motivation for trying to extract information from such text automatically is compelling - even if success is only partial.

In 1999, Hearst wrote that the nascent fieldof 'TextData Mining' had a name and a fair amount of hype, but as yet almost no practitioners. Hearst defines Data Mining, information access, and corpus-based computational linguistics and discusses the relationship of these to Text Data Mining.

To understand Text Mining it was necessary to understand the concepts, theory and model of Data Mining first. Since the literature on Data Mining is far more extensive, and also more focused: there are numerous textbooks and critical reviews that trace its development from roots in machine learning and statistics.

The book ‘Data Mining Concepts’ by Han and Kamber served as a platform to comprehend the various aspects of Data Mining, its applications and methodologies. This book however contains only a few pages on the concept of Text Mining. There are many other good books which have a very extensive coverage of different Data Mining techniques. They have been mentioned in the bibliography.

There are a number good academic journals on Data Mining – some which are free and some are payable. The ‘Data Mining and Knowledge Discovery’ journal of SpringerLink allows abstracts to be accessed by guest. This journal has many latest research papers on Data Mining. Apart from this other journals like ‘Knowledge and Information Systems’, ‘Machine Learning’, ‘IEEE Transactions on Knowledge and Data Engineering’ etc. are other sources of Data Mining related material.

Text Mining emerged at an unfortunate time in history. Data Mining was able to ride the back of the high technology extravaganza throughout the 1990s, and became firmly established as a widely-used practical technology—though the dot com crash may have hit it harder than other areas. Text Mining, in contrast, emerged just before the market crash—the first workshops were held at the International Machine Learning Conference in July 1999 and the International Joint Conference on Artificial Intelligence in August 1999—and missed the opportunity to gain a solid foothold during the boom years.

1.2 Data Mining

Since Data Mining is the superset of Text Mining, it is important to understand Data Mining first. Data is increasing at an unimaginable rate every year. The area of Data Mining has arisen over the last decade to address this problem. Progress in digital data acquisition and storage technology has resulted in the growth of huge databases. This has occurred in all areas starting from simple applications like supermarket transactions, railway reservations to the more complex and complicated ones like space research, molecular databases, images and astronomical bodies etc. Using this data to discover hidden knowledge, unexpected patterns and unknown information is Data Mining.

Data Mining research and practice is in a state similar to that of databases in the 1960s. At that time the concept of databases was new and in the development stage where programmers were still trying to come out of the third generation of languages. Slowly the concept of relational databases was being developed, implemented and improvised upon. Presently we can say that databases are fully implemented and working efficiently all over the world.

The evolution of data warehouses from databases is slowly taking shape. The evolution of Data Mining techniques may take a similar path over the next few decades, making Data Mining techniques easier to use and develop.

Data Mining can be defined as follows:

"Data Mining is the non-trivial extraction of implicit, previously unknown and potentially useful information from data."

Most organizations have large databases that contain a wealth of potentially accessible information. However, it is usually very difficult to access this information. This uncontrolled growth of data will inevitably lead to a situation in which it becomes extremely difficult to access the desired information. In fact it would be like looking for a needle in a haystack.

The sudden rise of interest in Data Mining could be because of the following reasons:

- Most of the organizations have stored gigabytes of data about their products, customers, suppliers, competitors, etc. This database forms a potential gold or rather a diamond mine that can be explored to find hidden and extremely useful information. This information can be traced using simple queries. Data Mining algorithms typically zoom in on interesting sub-parts of the database and dig out the information.
- Since networks have developed extensively, it becomes easy to connect databases situated at remote places. Thus connecting a client's file to a file with demographic data may lead to unexpected views on the spending patterns of certain population groups.
- In the past few years, machine-learning techniques have expanded enormously. Neural networks, genetic algorithms and other techniques often make it easier to find connections in databases.
- The client-server revolution gives the individual access to central information systems. Marketing specialists and policy makers also want to

avail themselves of these newly acquired technical possibilities that would help them in making their strategies.

The terms ***Knowledge Discovery in Database (KDD)*** and Data Mining are often used interchangeably. In fact there are other names like knowledge extraction, information discovery, exploratory data analysis, etc. also given to Data Mining. However KDD is the most popular name.

KDD is a process that involves many different steps. The input to this process is the data and the output is the useful information desired by the users. To ensure the usefulness and accuracy of the results of the process, interaction throughout the process by domain experts and technical experts might be needed. Of the many steps in KDD, one of the steps is Data Mining. However, if Data Mining is considered separately, to perform Data Mining all these steps are required. So in a way both mean the same thing.

The different steps of KDD are as follows:

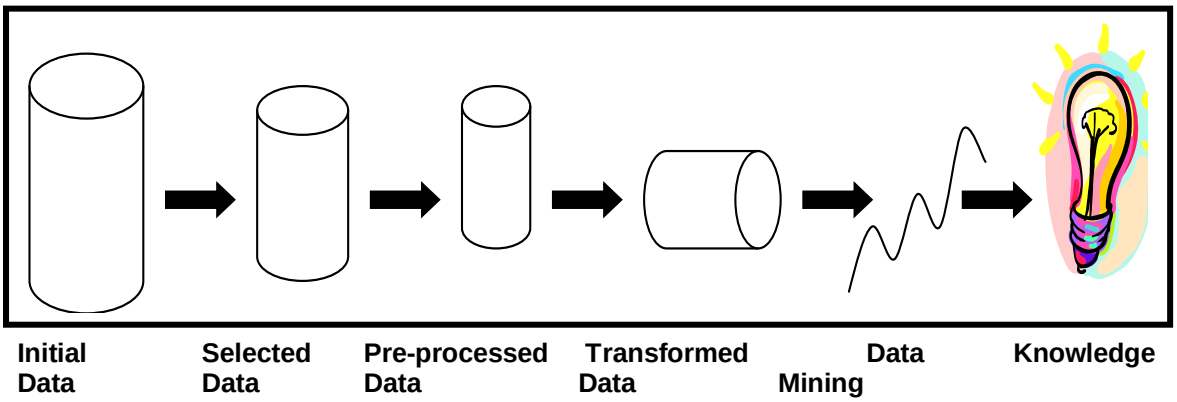

Figure 1-1Steps of Knowledge Discovery in Databases

Brief description of the steps:

Selection:

- The data obtained from heterogeneous data sources
- The data selected depends on objective of Data Mining
- The data are of different types like active, supplementary, shelf life etc.
- This step is also called the identification and extraction stage

Preprocessing:

- Erroneous data is removed i.e. data that is skewed and invalid
- Missing data is supplied i.e. usually by predicting the values

Transformation:

- The data from different sources is converted to a common format
- If required data is encoded
- Some data conversion is also done i.e. from simple format to more complex one
- If statistics is to be used, several variables may be grouped into one
- If neural networks is used values are changed to 1s and 0s

Data Mining:

- Applying algorithms to the transformed data
- Selection of the correct set of algorithms
- Each set could result in a different type of output

Knowledge:

- This step consists of interpretation and evaluation of results obtained
- This is a heuristic i.e. a self-learning approach
- The result, which is in the form of graphs and charts, is analyzed by experts giving knowledge

The steps shown above are those of Data Mining. When applied to textual data there is a slight change in the steps and the kind of work to be done on the textual data.

1.3 Text Mining

Marti Hearst was one of the first researchers who talked about Text Mining and presented a paper on it.(Hearst, M. Untangling Text Data Mining .In the Proceedings of ACL 1999). According to him, Text Mining is the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources. A key element is the linking together of the extracted information together to form new facts or new hypotheses to be explored further by more conventional means of experimentation.

Text Mining is different from what we're familiar with in web search. In search, the user is typically looking for something that is already known and has been written by someone else. The problem is pushing aside all the material that currently isn't relevant to your needs in order to find the relevant information. In Text Mining, the goal is to discover unknown but useful information from documents or rather unstructured data.

To the uninitiated, it may seem that Google and other Web search engines do something similar, since they also pore through reams of documents in split-second intervals. But, as experts note, search engines are merely retrieving information, displaying lists of documents that contain certain keywords.

Text-mining programs go further, categorizing information, making links between otherwise unconnected documents and providing visual maps (some look like tree branches or spokes on a wheel) to lead users down new pathways that they might not have been aware of.

Thus, Text Mining can be defined as:

'The discovery by computer of new, previously unknown information, by automatically extracting information from a usually large amount of unstructured textual resources.'

Text Mining can be compared in a simple form to different concepts like Data Mining, Web Mining, Natural Language Processing (NLP) etc. as follows:

Data Mining

- In Data Mining the data is structured and generally located in databases and in Text Mining, patterns are extracted from unstructured data in documents and text files
- In Data Mining the information is implicit in the input – data i.e. unknown and not possible to extract without automatic techniques. In Text Mining the information is clearly stated in the input text but it not implied in a manner that is open to automatic processing. Text Mining strives to bring out the text in a form that is suitable for computer processing directly without human intervention

Web Mining

- The source of data is the Web – the largest source of data in the world where in Text Mining, the input is not necessarily the web – it could be textual data from any source (local or otherwise)
- Data on the web is dynamic and rich in features and patterns and the data is text, audio, video, graphics, hyperlinks, tags etc.

Information Retrieval (IR)

- No genuinely new information is found.
- The desired information merely coexists with other valid pieces of information.

Computation Linguistics (CPL)& Natural Language Processing

- An extrapolation from Data Mining on numerical data to Data Mining from textual collections
- CPL computes statistics over large text collections in order to discover useful patterns which are used to inform algorithms for various sub-problems within NLP, e.g. Parts Of Speech tagging, and Word Sense Disambiguation

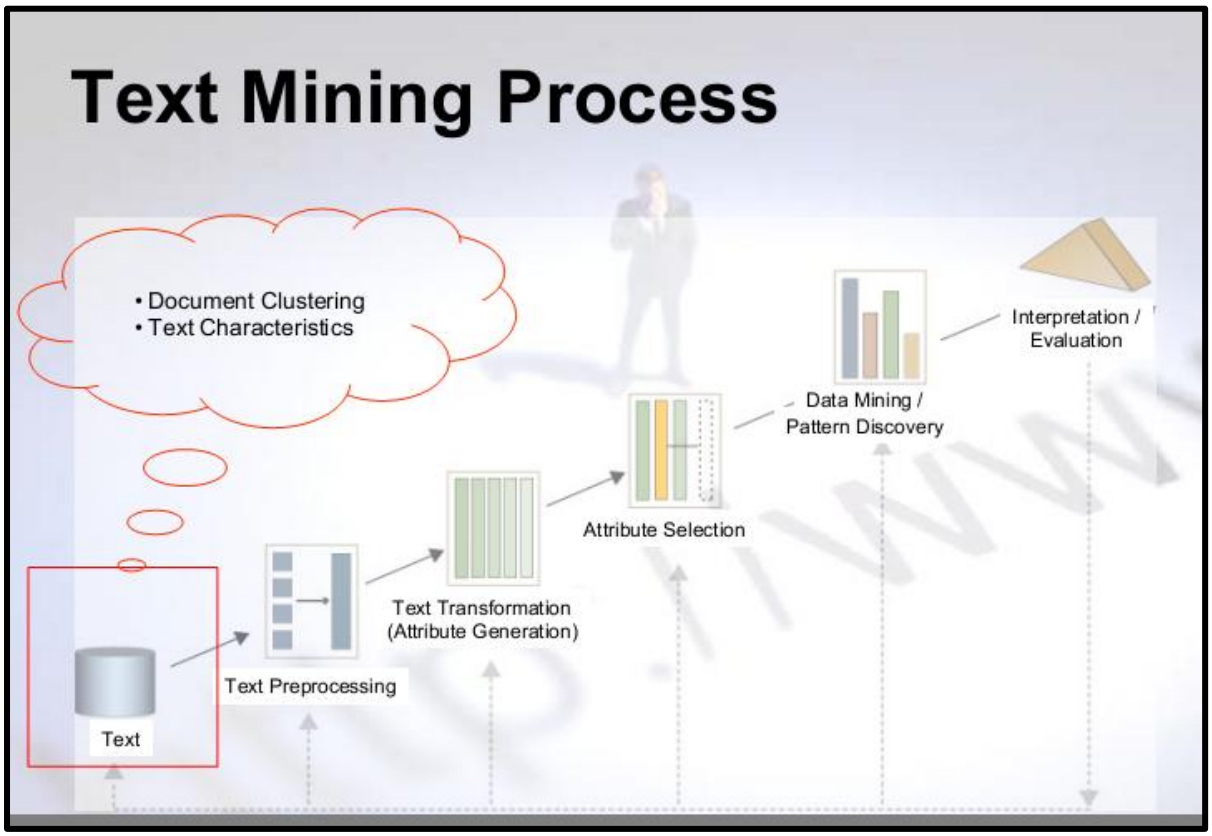

Figure 1-2 The Text Mining Process

Although a lot of work has been done on Text Mining, it is still a field of pure delight for researchers. As observed from the material – books, papers, online articles, journals, periodicals etc. there is a lot of scope for text miners to compare and contrast the different Text Mining methods and put forth the comparatives in a well-organized form. This type of work would be useful to researchers, students and people involved in the Decision Support System of their organizers to mine the large amount of textual information that is available with them.

Researchers and students would find this work very useful for the understanding and detailed study of Text Mining algorithms and methods. Though a number of books are available, the topics covered are so vast and in some cases too detailed to grasp the real hub and core of Text Mining.
Text Mining itself has sub-divisions like Text Clustering, Text Classification and Text Summarization. The work done over here is focused on Text Summarization.

1.4 Motivation for Text Summarization

In recent days, automatic Text Summarization has drawn a considerable interest in the research communities in the field of the Text Mining and Natural Language Processing. During the late 1960s, a large number of scientific papers in American research libraries were to be digitally stored to make them searchable for some purpose. This created the interest for automated Text Summarization. However, locating relevant text was an arduous task in the earlier days when the personal computers were not invented, and the World Wide Web (WWW) had not emerged as a global digital repository.

The structure and way of text searching have been altered after the creation of the WWW to facilitate the academicians, researchers and laymen to browse the contents online and use it for apt purpose. In spite of the reduced burden of information retrieval, it has remained a challenging issue to acquire relevant information in a concise and precise manner. This issue can be addressed by Text Summarization. It is a technique in which the system automatically produces summary or an abstract textual description from one or more textual documents.

There are many definitions available for the process of summarization. The definitions themselves reflect the motivation in working in this fascinating area.

According to Mani and Maybury (1999), “Summarization is the process of distilling the most important information from the source (or sources) to produce an abridged version of the same for a particular user (or others) and task (or tasks)”. According to Mani (2001) “The goal of an automated Text Summarization system is to read through the information source, extract and then present the

most important contents to the user in a condensed form and sensitive manner based on the user's or application's need".

1.5 Problem Statement and Objectives

1.5.1 Problem Statement

The aim of this study is to design and develop a robust model(s) to generate the qualitative and readable summary by extracting the most important sentences from the given document.

1.5.2 Objectives

To achieve the given aim, following objectives have been set:

1. To implement existing approaches for extraction based automatic summarization to prove the difficulties in automatic summarization and to obtain a high-quality summarizer.
2. To investigate the role of Latent Semantic Analysis in summarization to improve the quality of summarization.
3. To incorporate the classic Naïve Bayes classification method for summarization.
4. To analyze the role of machine learning and deep learning algorithms to generate effective extractive Text Summarization.
5. Comparing the model generated summary against human generated summary and summaries from existing summarizers to prove the effectiveness of the proposed model.

1.6 The Research Contribution

The models developed in this research would enhance the single document Text Summarization process by generating a better and concise summary. The models have been evaluated using the Rouge Toolkit. The Research Contribution from the two models designed and developed can be explicitly mentioned as follows:

1. The models improvise upon the summaries generated by implementing Latent Semantic Analysis as the initial step towards summarization after pre-processing the document.

2. The Naïve Bayes based model can be used to generate single document summary wherein the metrics calculated – Accuracy, Precision and F-score display an improved output.
3. The Hybrid model incorporates Deep Learning technique to create an extractive summary which again exhibits a succinct output.
4. Both the models can be implemented forbetter Information Extraction.
5. Since Text Summarization is a lucrative area for Data miners, these models would aid in designing tools for Information Retrieval also.

1.7 Layout of the Thesis

The overall thesis has been divided into seven chapters with a brief introduction of the chapter in the beginning and summary of the chapter at the end of each chapter.

Chapter 1, i.e. this chapter, was about the evolution of Data Mining, Text Mining and then Text Summarization. It was followed by the motivation behind this work and a brief introduction to the work, which has been carried out, in terms of the problem statement, objectives, research contributions and various possible applications of this work. The rest of the chapters are organized as under:

Chapter 2 discusses the Literature Study carried out for this research work. It contains an in depth study of the existing techniques, methods and models that have been developed in the context of document summarization. The first section takes a look about history of summarization. The next section goes deeper into summarization techniques with different approaches for the same.

Chapter 3 is related to the process of Text Pre-processing. The steps related to pre-processing are discussed along with the details of the models, measures and concepts used in pre-processing. Then it discusses feature selection techniques and the proposed method of this research. Finally, the chapter ends with evaluation of the summary with several metrics description.

Chapter 4is based on detailed discussion on the concept of Latent Semantic Analysis which is the baseline of this research work. The thorough conceptual working of Singular Value Decomposition is explained. It also contains different existing methods and limitations of LSA.

Chapter 5 discusses the first model which implements Naïve Bayes classification. The model description, datasets used, implementation and result discussion is covered in this chapter.

Chapter 6 describes the Hybrid method for Text Summarization. Here too the detailed model description, datasets used, implementation followed by result discussion is given.

Chapter 7 briefly describes the future enhancements possible in this field of Text Summarization.

Each chapter ends with the summary of the content of the chapter.

After all chapters are completed, there are the Appendices, Publication Details and Bibliography.

The Appendix – A contains details of results for each document of the dataset.

The Appendix – B contains – sample input and output.