**ABSTRACT**

In this technological era, almost all the educational institutes and corporate industries make use of presentations to convey information to the targeted audience. However, extracting useful information and making presentations is a time consuming process. There is abundance of data available on the internet. Choosing useful sources of data and then summarizing the provided contents requires sincere effort. The goal of our project is condensing the source text in a shorter version to be accommodated topic wise in the slides, preserving its informational content and overall meaning. It is an attempt to develop an automatic procedure designed to extract the information from multiple text documents written about the same topic using clustering and feature specific algorithm. Resulting summary allows individual users, such as professional information consumers, to quickly familiarize themselves with information contained in a large collection of documents. The software takes the documents containing information from the user as the input for summarization. The output is in the form of a presentation containing the gist of the topic under consideration. The user can also intervene in the process by providing specific headings of the topics to be covered in the presentation. Additionally, the user is also provided with the option of inserting images by choosing from a range of images retrieved from Google by the software.

1. INTRODUCTION

A presentation, as the name suggests, is a tool used for presenting and explaining an idea or a particular topic to the targeted audience. It is typically a demonstration, lecture, or speech meant to inform or persuade the listeners. The corporate and educational institutes use presentations extensively as a tool to convey their knowledge. However, creating a good presentation is a time consuming process. It requires a huge amount of information to be extracted and analyzed which can sometimes be tedious and tiring. To address this problem, we have proposed to create software that automates the process of creating presentation with the help of data mining concepts. It works in works two phases- multi-document summarization and creation of presentation from summarized information. This chapter highlights the basic concepts of text summarization using clustering and feature specific methodologies and creation of presentations. It also puts light on the objective and overview of our proposed system.

1. AUTOMATIC TEXT SUMMARIZATION

Due to the growth of internet, the amount of information available at one’s expense has increased, which in turn, has increased the problem of information overload. Thus, Summarization is the need of the era. According to [1], ‘A summary can be defined as a text that is produced from one or more texts, that contains a significant portion of the information in the original text(s), and that is no longer than half of the original text(s)’. Automatic text summarization is the technique, where a computer summarizes a text. A text is entered into the computer and a summarized text is returned, which is a non redundant extract from the original text. Summaries should preserve the essence of the subject of its source(s). A summary should typically be short, precise and to the point. There are two methods of summarization, extractive and abstractive.

1. Extractive methods work by selecting a subset of existing words, phrases, or sentences in the original text to form the summary. Extractive summarization is mainly concerned with what the summary content should be, usually relying solely on extraction of sentences.
2. Abstractive methods build an internal semantic representation and then use natural language generation techniques to create a summary that is closer to what a human might generate. Such a summary might contain words not explicitly present in the original.
3. **CLUSTERING**

Cluster analysis is a multivariate method which aims to classify a sample of subjects (or objects) on the basis of a set of measured variables into a number of different groups such that similar subjects are placed in the same group. It can be defined as “Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups [1].” Due to this feature, clustering plays an important role in data mining concept and to extract the important and summarized information from a huge amount of data. Clustering is used to identify themes or subtopics of common information, because multiple documents relating to a particular topic are likely to contain redundant information in addition to information unique to each document. Once themes have been known, a representative passage in each theme is selected and included in the summary.

A brief overview on clustering used in text summarization is:

1. There are various algorithms used for clustering. In the process of clustering, similar object are grouped together.
2. Thus it leads to increase in inter cluster distances and minimizes the intra cluster distances.
3. As a result various clusters are formed.
4. These clusters are arranged in an order.
5. Further these clusters are required for mining the data and creating a summarized text from a document.
6. **FEATURE SPECIFIC SUMMARIZATION**

Feature based text summarization is used for ordering sentences in a cluster by analyzing the features of each sentence. On capturing sentence-specific features we compute the feature profile of each sentence [2]. In the proposed method we consider 6 features for ranking sentences. They include title feature, sentence length, term weight, sentence position, proper noun, and numerical data.

A brief overview of the above features is as follows:

1. Title feature: It represents the number of words in a sentence that match with the words in the title. Higher the score of title feature, higher is the ranking of the sentence [3].
2. Sentence length: It is used to identify normalized length of sentences in a source. Shorter sentences are not expected to be a part of the summary.
3. Term weight: It depicts the number of occurrences of a term in a document. The term weight corresponds to the importance of the term in context with the subject under consideration.
4. Sentence Position: According to this feature, the opening and closing sentences usually contain the gist of the topic and hence are given more importance.
5. Proper noun: It is the number of proper nouns the sentence contains. Having a higher score in the feature denotes higher importance of the sentence.
6. Numerical data: Sentences having numerical data are usually necessary to be included in the summary. This feature is the ration of numerical data in a sentence and the length of the sentence.
7. OBJECTIVE

Today, the current system of text summarization technique is a complex issue that uses the concept of data mining. Thus creating a summary is not an easy task. Also, the current system creates only the summary of the document but does not provides anything in creating the presentation. The objective of the system proposed, is to develop software that accepts multiple documents as input to provide a concise and to-the-point presentation (as well as a summary) that conveys the content in the source documents in a presentable manner. The source documents are initially clustered to obtain groups of related sentences. Feature profiles are generated for ranking of these sentences which then form a part of the presentation. The software completely automates the process of creating presentations by choosing the topic for each slide on its own. However, it also allows the user to personalize the presentation in case the user wants to.

1. OVERVIEW OF THE PROPOSED SYSTEM

In our proposed system, the input to the system is a collection of documents. Depending on the user’s choice, the system provides output in the form of a concise summary or a presentation covering the important points of the subject. The proposed approach produces an extractive summary by selecting salient sentences from the documents cluster wise. All the relevant documents are grouped together into clusters by using threshold-based document clustering approach. Based on feature profile salient sentences from each cluster are identified and ranked according to their weights of importance. Based on the ranking of sentences, sentences are selected and ordered. The system then iteratively extracts one sentence at a time, and pastes it on the appropriate slide. The user is also provided with a facility to insert an image(s) in the slide by simply browsing in the directories or selecting from a range of pictures that the software retrieves from Google.

The proposed approach can be decomposed into seven sub processes:

1. **Data preprocessing:** From the collection of documents, the boundaries of sentences are identified and the documents are split into sentences. Sentences are in turn split into words. Frequently occurring insignificant words called functional words or stop words like “a”, “the”, “of” are removed because they do not contribute to the meaning of the sentence. [2]
2. **Documents representation:** In this step, the cleaned and noise free data is subjected for documentation representation which is further subjected for data clustering.
3. **Clustering of data:** In this, the represented documents are clustered on the basis of the objects that form the same group.
4. **Score generation based on feature profile:** Based on feature profile salient sentences from each cluster are identified and ranked according to their weights of importance.
5. **Sentence ordering on the basis of clustering:** Based on the ranking of sentences based on feature profile, sentences are selected and ordered.
6. **Summary generation:** In this stage, based on the ranking of the clusters, summary is generated.
7. **Presentation creation:** It is the final stage in which the summarized text is subjected to creation of presentation. During this process, the user is also provided with a facility to insert an image(s) in the slide by simply browsing in the directories or selecting from a range of pictures that the software retrieves from Google.

PROBLEM STATEMENT

A presentation plays an important role when it comes to represent an idea or to explain the topic to the audience. However, the current system of creating a presentation technique is a complex issue. Knowing the importance of creating a presentation, it has become important to modify it and make the system automated. Till the present situation, creating a presentation for a presenter is not an easy task. It is time consuming, tedious and effort making. The presenter has to go through multiple documents, create a small summary in which all the highlighted points are covered and then create the presentation. This problem can be solved by developing a software tool that automates the method of creating a presentation. Thus, a new system is proposed in which the software accepts multiple documents as input in order to provide a summary and from that summary a concise and to-the-point presentation is created (that conveys the content in the source documents in a presentable manner). From the multiple documents, various clusters of data or sentences are created and the feature profiles are generated for ranking of these sentences which then form a part of the presentation. The software completely automates the process of creating presentations by choosing the topic for each slide on its own. However, it also allows the user to personalize the presentation in case the user wants to.

1. SCOPE

The application is developed on the basis of two types of requirements of the people that is summarization of text and creation of presentations. Summarization of the text deals with knowing the basic idea of the entire topic. It is like giving brief information over the entire subject. On the other hand, creating a presentation involves presenting information in an effective manner. In a presentation, only the important points are highlighted which help in the understanding of the topic. Our system deals with both these concepts thus providing an easy way for gathering and delivering information.

1. OBJECTIVES
2. To automate the process of creating presentations
3. Increased accuracy in summarization of text
4. Multipurpose features of text summarization and presentation creation
5. To ease the work of people who deal with presentations and summaries on a day to day basis.

FEATURES AND ADVANTAGES

Easy to use. Makes use of internet as per user’s requirement. Maintains a separate database to store important details of the presentation or summarized text within the system. Cost effective and can be easily downloaded from various site.

ARCHITECTURAL DESIGN

Initially, the user has to provide the system with input documents. The documents can be of any file format, including docx, pdf, html etc. These files have to be first converted to text files. For this purpose, Java provides various facilities for reading, writing and creating files.

DATA PREPROCESSING

A preprocessing procedure consists of various steps. We include the following steps for our system:

1. Lexical analysis:

The aim of this step is to identify words in the documents. Digits, hyphens, punctuation marks and case of the letters are the primary concern of this step.

1. Elimination of stop words:

A stop word can be a word without meaning in a specific language or it can be token that does not have any linguistic meaning. Examples of stop words in English are ‘a’, ’the’, ‘is’ etc. [5]

1. Stemming:

A stem is the portion of the word which is left after removal of its affixes. Stems are thought to be useful for improving searching of terms because they reduce variants of the same root word to common concept. [5]

DOCUMENT REPRESENTATION

After preprocessing, we obtain a collection of words from the set of documents. This collection of words is known as vocabulary. It is represented as a two dimensional matrix wherein rows represent documents and columns represent words. The concept of TF-IDF (Term Frequency-Inverse Document Frequency) is used to fill the matrix to reflect the count of a particular word in a specific document. Each entry in this matrix is an integer count.

CLUSTERING OF DATA

After the process of document representation, the sentences are spilt into a matrix representation. In clustering, data is grouped on the basis of Euclidian distance method of clustering. In this method, the clusters are created on the basis of dissimilarities or the distances between the data. This is the most straightforward way of computing distances between objects in a multi-dimensional space. If we had a two or three-dimensional space this measure is the actual geometric distance between objects in the space.

SCORE GENERATION BASED ON FEATURE PROFILE

In this step, the documents in a particular cluster are split into sentences. A two-dimensional array with rows representing sentences and column representing words is created. TF-ISF (Term Frequency-Inverse Document Frequency) values are computed for each sentence-word pair. This score is further incremented based on the features of title, sentence length term weight, sentence tion, presence of proper noun and numerical data. Higher the score of a sentence more is the importance of the sentence.

SENTENCE ORDERING BASED ON CLUSTERING

In this process, sentences are ordered in a chronological manner. In feature specific algorithm, the score of the sentences are generated using the concept of inverse sentence frequency. In sentence ordering, sentences are selected on the basis of their chorological order irrespective of their scores. Consider a situation in which there are two sentences A and B. A has a weigh higher than B. However, if B comes before A, then B will be given a higher priority than A for sentence ordering. These sentences are then stored in the databases in order.

SUMMARY GENERATION

On the basis of clustering and feature profile system, score is generated which leads to ordering of the sentences. After ordering the sentences, the next step is summarization of text. In this, a summary is generated by extracting the highly ranked sentences and extraction is done until the summary length is met. To avoid redundancy, if the extracted sentence is already exists in the summary, then the sentence is eliminated and the next highest sentence is considered to form the summary. This process is repeated for each cluster and summary is generated depending on the rank of the sentence.