CREATION OF PRESENTATIONS WITH FEATURE SPECIFIC MULTI-DOCUMENT TEXT SUMMARIZATION

Vrishali Bhor[1], Keven Sebastian[1], Varun Saxena[1], Pravin Hodge[1]
Mrs. Poonam Bari[2]
[1.] B.E. Department Of Information Technology, FCRIT, Vashi
[2.] Asst. Professor FCRIT, Vashi

***ABSTRACT*:**
The goal of the proposed system is condensing the source text in a shorter version to be accommodated topic wise in a presentation, preserving its informational content and overall meaning. The proposed system extracts 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.

***Index Terms*:**
Clustering, feature specific, presentations, summarizing and summary.

I. INTRODUCTION

A. 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.

B. 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.

C. 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.

D. Feature Specific Summarization

Feature based text summarization is used for orderingsentences in a cluster by analyzing the features of

eachsentence. 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.

E. 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.

II. LITERATURE SURVEY

A. Existing Systems Related to Text Summarization

1) *H.P.LUHN*

This system could only have a single-document input and has a domain specific to technical articles. Term filtering and word frequency is carried out in which the words with low frequency were removed. Luhn proposed that the importance of a particular word is based on the frequency of the word in that article. The output of Luhn's work was called abstract or 'auto abstract' but the correct term is extract because sentences from the source were included in the summary.

2) *SUMMARIST*

This system creates a robust automated text summarization system based on the equation summarization = topic identification + interpretation + generation. This system is used for both a research tool and as an engine to produce summaries. It produces both extracts and abstracts for arbitrary English and other-language text.

3) *MEAD*

This system was developed at the University of Michigan in 2001. It can produce both single and multi-document extractive summaries. The system is based on the idea of centroid-based feature and other features such as position and overlap with the first sentence. It uses CIDR Topic Detection and Tracking system to identify all the articles related to an emerging event. This system was only domain specific to news articles.

4) *CATS*

This system was developed in the year 2005. It is uses single as well as multi-document input. This system is domain specific only to news articles. This system analyzes which information in the document is important in order to include it in the summary. In this system statistical techniques are used to compute a score for each sentence as well as temporal expressions and redundancy is solved.

B. Existing Systems Related to Presentation Makers

1) *Microsoft Office PowerPoint*

PowerPoint is a part of the Microsoft Office Suite ofproductivity programs. It allows people to create a series of single page slides that contain information to be presented. It allows the presenter to include graphs, images and movies in the presentation making the information both understandable and memorable.

2) *PREZI*

Prezi is a cloud-based presentation software and storytelling tool for presenting ideas on a virtual canvas. The product employs a zooming user interface (ZUI), which allows users to zoom in and out of their presentation media, and allows users to display and navigate through information within a 2.5D or parallax 3D space on the Z-axis.

III. METHODOLOGY

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.

Objectives

1) To automate the process of creating the presentation
2) Increased accuracy in regarding summarization of text
3) Multipurpose features of text summarization and presentation creation.
4) To ease the work of people who deals with presentations and summaries daily like in seminars, classrooms, etc

Features and Advantages

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

Initially, the user has to provide the system with input documents, which 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. There are seven steps to be covered in the design phase. They are as follows:

1. Data preprocessing
2. Document representation
3. Clustering of data
4. Score generation on the basis of feature profile
5. Sentence ordering based on clustering
6. Summary generation
7. Presentation creation

1) *Data preprocessing*

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

- 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.

- 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. [4]

- 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. [4]

2) *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.

3) *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.

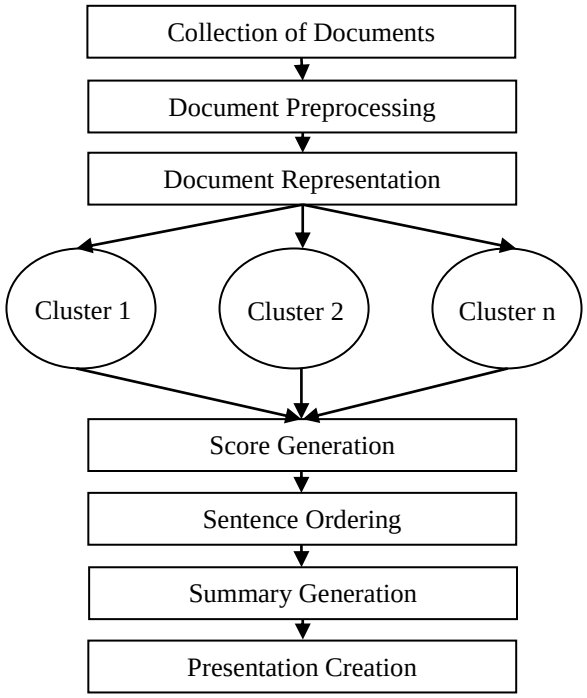

Fig. 6.1 Design of proposed system

4) *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 position, presence of proper noun and numerical data. Higher the score of a sentence more is the importance of the sentence.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) *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 weight 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.

8) *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.

9) *Presentation creation*
In this stage, for each slide to be generated, the system takes the title of the slide as input from the user. The user is also prompted for inclusion of image in the slide. If the use provides a positive response, images are searched from Google and the user is asked to choose one. The user may also provide images from the directory. Depending upon the requirement of the user, Sentences from the summary are pasted on the slides. For creation and manipulation of presentations, Apache POI is used.

VIII CONCLUSION

Information plays an important role in various sectors like business, education and other private, public and government sectors. Due to the evolution of Internet, huge amount of information can be accessed today. However, dealing with such a huge amount of information is not an easy task and this leads to the problem of information overloading. The proposed system of the project attempts to solve this problem by automating the process of text summarization as well as presentation creation. However, the major focus of the proposed system is to ease the work of professionals who deal with presentations daily such as business professionals, teachers and professors, students and the speakers participating in various seminars. Thus, the project deals with extracting the useful, non-redundant information with the process of automatic text summarization. The summary is then put into the format of presentations automatically, which are more interactive and easy to understand.

REFERENCES

[1] Hovy, E. H. Automated Text Summarization. In R. Mitkov (ed), The Oxford Handbook of Computational Linguistics, chapter 32, pages 583-598.Oxford University Press, 2005.

[2] A. Kogilavani and Dr. P. Balasubramani InClustering and Feature specific sentence extraction based summarization of multiple documents, International journal of computer science & information Technology (IJCSIT) Vol.2, No.4, August 2010 DOI : 10.5121/ijcsit.2010.2409 99.

[3] LaddaSuanmaliet. al. Automatic Text Summarization Using Feature-Based Fuzzy Extraction

[4] LendenevaYuliaet. al. Effect of Preprocessing on Extractive Summarization with Maximal Frequent Sequences National Polytechnic Institute Center for Computing Research