similarity values, obtaining the raw value of this feature for s. The process is repeated for all sentences.[7]

L. Sentence-to-Centroid Cohesion:

For each sentence s as compute the vector representing the centroid of the document, which is the arithmetic average over the corresponding coordinate values of all the sentences of the document; then compute the similarity between the centroid and each sentence, obtaining the raw value of this feature for each sentence.[7]

M. Occurrence of non-essential information:

Some words are indicators of non-essential information. These words are speech markers such as "because", "furthermore", and "additionally", and typically occur in the beginning of a sentence. This is also a binary feature, taking on the value "true" if the sentence contains at least one of these discourse markers, and "false" otherwise.[7]

N. Discourse Analysis

Discourse level information, in a text is one of good feature for text summarization. In order to produce a overall discourse structure of the text and then removing sentences peripheral to the main message of the text. These features are important as, a number of methods of text summarization are using them. These features are covering statistical and linguistic characteristics of a language.[7]

4. Extractive Based Summarization Methods

This process can be divided into two steps: Pre-Processing step and Processing step. Pre-Processing is structured representation of the original text. It usually includes:

1) Sentences boundary identification. In English, sentence boundary is identified with presence of dot at the end of sentence.
2) Stop Word Elimination-Common words with no semantics.
3) Stemming-The purpose of stemming is to obtain the stem or radix of each word, which emphasize its semantics.

In Processing step, features influencing the relevance of sentences are decided and calculated and then weights are assigned to these features using weight learning method. Final score of each sentence is determined using Feature-weight equation. Top ranked sentences are selected for final summary. Most of the current automated text summarization systems use extraction method to produce a summary .Sentence extraction techniques are commonly used to produce extraction summaries. One of the methods to obtain suitable sentences is to assign some numerical measure of a sentence for the summary called sentence scoring and then select the best sentences to form document summary based on the compression rate. In the extraction method, compression rate is an important factor used to define the ratio between the length of the summary and the source text. As the compression rate increases, the summary will be larger, and more insignificant content is contained. While the compression rate decreases the summary to be short, more information is lost. In fact, when the compression rate is 5-30 %, the quality of summary is acceptable.

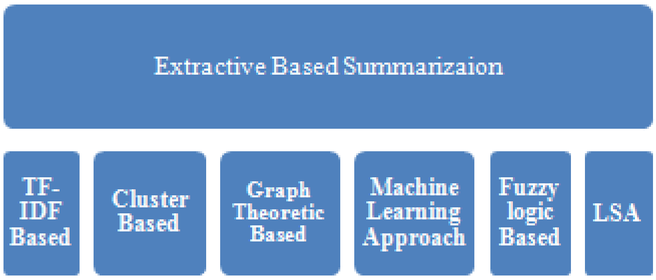

Figure 3: Extractive Based Summarization method

4.1 Term Frequency-Inverse Document Frequency

It is a numerical statistic which reflects how important a word is in a given document. The TF-IDF value increases proportionally to the number of times a word appears in the document. This method mainly works in the weighted term-frequency and inverse sentence frequency paradigm .where sentence-frequency is the number of sentences in the document that contain that term. These sentence vectors are then scored by similarity to the query and the highest scoring sentences are picked to be part of the summary. Summarization is query-specific. The hypothesis assumed by this approach is that if there are "more specific words" in a given sentence, then the sentence is relatively more important. The target words are usually nouns .This method performs a comparison between the term frequencies (tf) in a document -in this case each sentence is treated as a document and the document frequency (df), which means the number of times that the word occurs along all documents. The TF/IDF score is calculated as,

$$TF/IDF(w)=DN(\log(1 + tf)/\log(df))$$

where DN is the number of documents.

4.2 Cluster Based Method

In this method, the semantic nature of a given document is captured and expressed in natural language by a set of triplets (subjects, verbs, objects related to each sentence).Cluster these triplets using similar information. The triplets' statements are considered as the basic unit in the process of summarization. More similar the triplets are, the more the information is useless repeated; thus, a summary may be constructed using a sequence of sentences related the computed clusters [7].

4.3 Graph Theoretic Approach

In this technique, there is a node for every sentence. Two sentences are connected with an edge if the two sentences share some common words, in other words, their similarity is above some threshold. This representation gives two results: The partitions contained in the graph (that is those sub-graphs that are unconnected to the other sub graphs), form distinct topics covered in the documents. The second result by the graph- theoretic method is the identification of the important sentences in the document. The nodes with high cardinality (number of edges connected to that node), are the important sentences in the partition, and hence carry higher preference to be included in the summary.

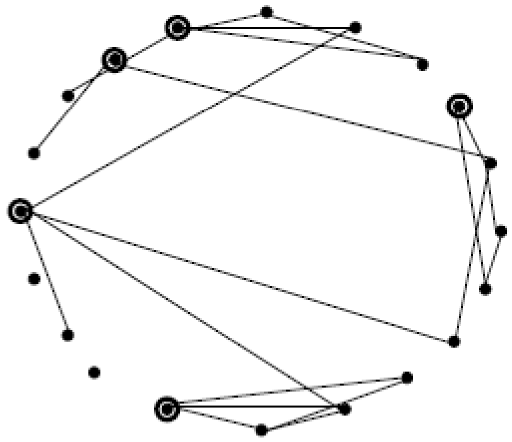

Figure 3.3: Graph Theoretic Approach

Figure (3.3) shows an example graph for a document. It can be seen that there are about 3-4 topics in the document; the nodes that are encircled can be seen to be informative sentences in the document, since they share information with many other sentences in the document. The graph theoretic method may also be adapted easily for visualization of inter and intra document similarity[7].

4.4 Machine Learning approach

In this method, the training data set is used for reference and the summarization process is modeled as a classification problem: sentences are classified as summary sentences and non-summary sentences based on the features that they possess. The classification probabilities are learnt statistically from the training data, using Bays' rule.[7]

4.5 Text summarization with neural networks

In this method, each document is converted into a list of sentences. Each sentence is represented as a vector [f1,f2,...,f7], composed of 7 features. Seven Features of a Document 1) f1 Paragraph follows title2) f2 Paragraph location in document 3) f3 Sentence location in paragraph 4) f4 First sentence in paragraph 5) f5 Sentence length 6) f6 Number of thematic words in the sentence 7) f7 Number of title words in the sentence. The first phase of the process involves training the neural networks to learn the types of sentences that should be included in the summary. Once the network has learned the features that must exist in summary sentences, we need to discover the trends and relationships among the features that are inherent in the majority of sentences. This is accomplished by the feature fusion phase, which consists of two steps: 1) eliminating uncommon features; and 2) collapsing the effects of common features[7].

4.6 Automatic text summarization based on fuzzy

This method considers each characteristic of a text such as sentence length, similarity to little, similarity to key word and etc. as the input of fuzzy system .Then, it enters all the rules needed for summarization, in the knowledge base of system. Afterward, a value from zero to one is obtained for each sentence in the output based on sentence characteristics and the available rules in the knowledge base.

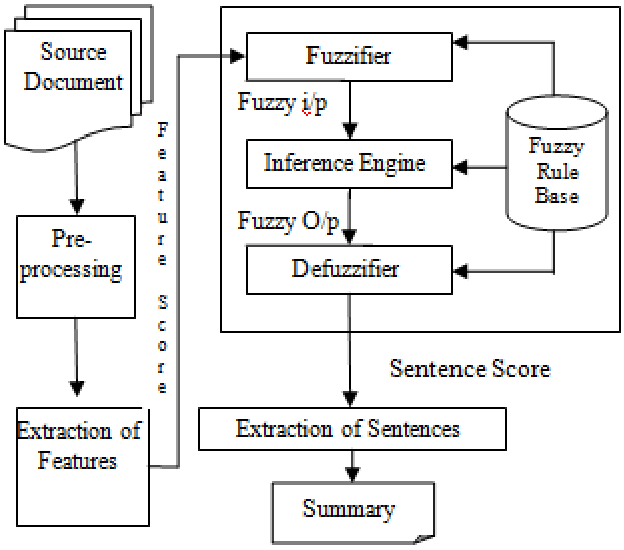

Figure 3.6: Fuzzy logic based method

The obtained value in the output determines the degree of the importance of the sentence in the final summary. The input membership function for each feature is divided into three membership functions which are composed of insignificant values (low L), very low (VL), medium (M), significant values (High) and very high (VH). The important sentences are extracted using IF-THEN rules according to the feature criteria. The fuzzy logic system consists of four components: Fuzzifier, Inference engine, Defuzzifier, and the Fuzzy knowledge base. In the fuzzifier, crisp inputs are translated into linguistic values using a membership function to be used to the input linguistic variables. After fuzzification, the inference engine refers to the rule base containing fuzzy IF-THEN rules to derive the linguistic values. In the last step, the output linguistic variables from the inference are converted to the final crisp values by the defuzzifier using membership function for representing the final sentence score. Fig 3.6 shows the fuzzy logic based method [7].

4.7 LSA

Singular Value Decomposition (SVD) is a very powerful mathematical tool that can find principal Orthogonal dimensions of multidimensional data. It has Applications in many areas and is known by different names: Karhunen-Loeve Transform in image processing, Principal Component Analysis (PCA) in signal processes and Latent Semantic Analysis (LSA) in text processing. It gets this name LSA because SVD applied to document word matrices, groups' documents that are semantically related to each other, even when they do not share common words. Words that usually occur in related contexts are also related in the same singular space. This method can be applied to extract the topic-words and content-sentences from documents.

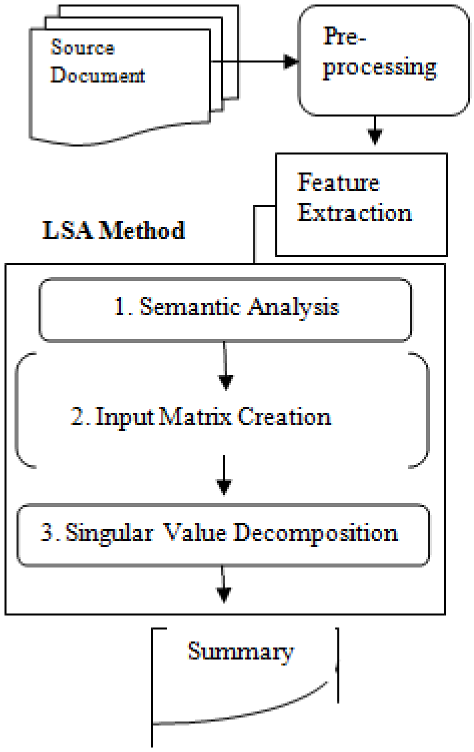

Figure 4(f): LSA Based

The advantage of using LSA vectors for summarization rather than the word vectors is that conceptual (or semantic) relations as represented in human brain are automatically captured in the LSA, while using word vectors without the LSA transformation requires design of explicit methods to derive conceptual relations. Since SVD finds principal and mutually orthogonal dimensions of the sentence vectors, picking out a representative sentence from each of the dimensions ensures relevance to the document, and orthogonality ensures non-redundancy. It is to be noted that this property applies only to data that has principal dimensions inherently—however, LSA would probably work since most of the text data has such principal dimensions owing to the variety of topics it addresses.[7]

5. Evaluating the Summarization Systems

Evaluation methods are useful in evaluating the usefulness and trustfulness of the summary. In summary, evaluating the qualities like comprehensibility, coherence, and readability is really difficult. System evaluation might be performed manually by experts who compare different summaries and choose the best one. A problem with this approach is that the individuals who perform the evaluation task normally have very different ideas on what a good summary should contain. In a test, Hassel (2003) found that at best there was a 70% agreement between summaries created by two individuals. A further problem with manually performed evaluation is that it is an extremely time consuming task. Automatic system evaluation is another way for evaluating summarization systems which is still an open research topic. Since there is not a base standard for evaluating systems, different criteria are being used for evaluation. In the following paragraph two major and most practical methods are discussed. Two main criterion for evaluating the proficiency of a system is *precision* and *recall* which are used for specifying the similarity between the summary which is generated by a system versus the one generated by human. These terms are defined by following equations:

Precision = Correct / (Correct + Wrong)......... (1)

Recall = Correct / (Correct + Missed) .... (2)

Where, *Correct* is the number of sentences that are the same in both summary which are produced by human and system.; *Wrong* is the number of sentences presented in summary and produced by system but is not included in human generated summary; *Missed* is the number of sentences which are not appeared in system generated summary but presented in the summary produced by human. Therefore, *Precision* specifies the number of suitable sentences which are extracted by system and *Recall* specify the number of suitable sentences that the summarization system missed. There are also two other criteria for evaluating system which are *compression ratio* and *retention ratio* And defined as follows:

$$Compression\ Ratio: CR = \frac{Length\ S}{Length\ T} \qquad (3)$$

$$Retention\ Ratio: RR = \frac{Information\ in\ S}{Information\ in\ T} \qquad (4)$$

Where *S* is the summarized text and *T* is the main text. So we can conclude that a good summary is the one with low *CR&* high *PR.[8]*

6. Applications of Automatic Text Summarization

The very first application area for automatic text summarization was to create abstracts/extracts from articles without abstracts to be stored in library systems together with the title and author name, (Luhn 1959). At that time one could not store the whole article digitally in the library system due to storage constraints. Today there is a wide range of application areas for automatic text summarization, the most common and obvious one is in information retrieval. We can already observe it in the result list *of search engine*s where a summarized part of each retrieved document is presented interweaved together with the search terms of the user, the so-called snippets. We can consider these snippets to be a crude form of user adapted text summaries. Another possible application is in the *mass media area*. Today a news article is written by a journalist, but when typesetting the newspaper the article is shorten manually to the appropriate size so that it can fit in the layout, in between the advertisement. In parallel the same article is also typeset for the web, WAP or SMS text messages. An experiment is described in Dalianis et al. (2004) where both manual editors and the SweSum text summarizer (Dalianis 2000) where given the task to summarize 334 news texts written in Swedish to the appropriate format for the newspaper Sydsvenska Dagbladet. The manually cut down texts were compared with the automatically summarized texts and it was found that the texts where almost identical. Both the editors and the SweSum text summarizer cut down and summarized the texts mainly from the end. The same experiment was carried out for SMS format (maximum 160 characters) and the results from SweSum were considered suitable to be used directly in news paper production. Business Intelligence systems or news monitoring systems are today very common where one surveys a large flow of news media, this news flow can be summarized so the user can obtain an overview of the stream before deciding if she should click on the news summary and read the complete news article. One nice live application is the Columbia News Blaster, which takes several news articles describing the same topic and summarizes it to one single news flash (McKeon & Radev 1999, McKeown et al. 2003). Further one might need a multilingual multi-document automatic text summarizer, in which case one could use MEAD (Radev et al., 2004).If we go to the area of medicine and biomedicine we find several attempts to use automatic text summarization and also the closely related area natural language generation to adapt both text and data to different user groups such as patients, physicians, nurses and scientists. In Hirst et al. (1997) a system is presented that from medical digital libraries produces user adapted information towards individual patients' specific needs, summarized from information on surgery of breast cancer to living with diabetes but also general health education. If we look at generation of text from source data Portet et al. (2009 describe a system that takes survey data from a baby at a neonatal clinic and generates a textual description for several different user groups such as the clinicians, the parents or even the relatives and friends of the patient. The textual description contains information that is adapted to the interest and needs of each user group. Another system is PERSIVAL, which is described in McKeown et al. (2001). PERSIVAL generates user-adapted information both for patients and physicians, and uses as input the patient record of the patient to find what topics the generated text should contain. PERSIVAL then searches for the relevant information in external resources and summarizes it to the relevant level of the user. The text that is constructed for patients' origins from several consumer health texts, while the text constructed for physicians is collected from medical journal articles.[9].

7. Conclusion

Now days, Automatic Text Summarization is one of the hot areas of research and attracts lots of attentions from different field. It consists of automatically creating a summary from one or more texts. There are three main steps for producing a summary from an input text (Topic Identification, Interpretation and summary generation). Most of summarization systems follow these steps in order to create summary. In this paper we discuss types of summarization methods which might be used in a system for generating a summary. First is abstractive based summarization and second is Extractive based summarization method. Abstractive summary method produces highly coherent, cohesive, information rich and less redundant summary. Abstractive text summarization is a challenging area because of the complexity of natural language processing. We only mention the Abstractive based text summarization methods. This paper is focussing on extractive summarization methods . An extractive summary is selection of important sentences from the original text. The importance of sentences is decided based on statistical and linguistic features sentences

Extractive based text summarization approaches are based on Neural Network, Graph Theoretic, LSA , Fuzzy and cluster have to an extent, succeeded in making an effective summary of a document .

References

[1] Karel Jezek and Josef Steinberger, "Automatic Text summarization", Vaclav Snasel (Ed.): Znalosti 2008, pp.1-12, ISBN 978-80-227-2827-0, FIIT STU Brarislava, Ustav Informatiky a softveroveho inzinierstva, 2008.
[2] Farshad Kyoomarsi, Hamid Khosravi, Esfandiar Eslami and Pooya Khosravyan Dehkordy, "Optimizing Text Summarization Based on Fuzzy Logic", In proceedings ofSeventh IEEE/ACIS International Conference on Computer and Information Science, IEEE, University ofShahid Bahonar Kerman, UK, 347-352, 2008.
[3] L. Suanmali , N. Salim and M.S. Binwahlan, "Fuzzy Logic Based Method for Improving Text Summarization" , International Journal of Computer Science and Information Security, 2009, Vol. 2, No. 1,pp. 4-10.
[4] Oi Mean Foong, Alan Oxley and Suziah Sulaiman , " Challenges and Trends of Automatic Text Summarization" IJITT, Vol. 1, Issue 1, 2010
[5] Archana AB, Sunitha. C " An Overview on Document Summarization Techniques "International Journal on Advanced Computer Theory and Engineering (IJACTE),Volume 1, Issue-2, 2013
[6] Atif Khan, Naomie Salim," A Review On Abstractive Summarization Methods" JATIT, 10th January 2014. Vol. 59 No.1
[7] Vishal Gupta, Gurpreet Singh Lehal" A Survey of Text Summarization Extractive Techniques" Journal Of Emerging Technologies In Web Intelligence, Vol. 2, No. 3, August 2010
[8] Saeedeh gholamrezazadeh, Mohsen Amini Salehi ,Bahareh Gholamzadeh "A Comprehensive Survey On Text Summarization Systems "2009 IEEE
[9] M. Hassel "Portable Text Summarization", (www.diva-portal.org/smash/get/diva2:423456/FULLTEXT01.pdf)
[10] Weiguo Fan, Linda Wallace, Stephanie Rich, and Zhongju Zhang, "Tapping into the Power of Text Mining", Journal of ACM, Blacksburg, 2005.
[11] Fang Chen, Kesong Han and Guilin Chen, "An Approach to sentence selection based text summarization", Proceedings of IEEE TENCON02, 489-493, 2002.
[12] Mohamed Abdel Fattah and Fuji Ren, "Automatic Text Summarization", Proceedings of World Academy of Science, Engineering and Technology, Vol 27,ISSN 1307- 6884, 192-195, Feb 2008.

Author Profile

Patil Pallavi D. Received the B.E degree in Information Technology From D.A.C.O.E. Karad from Shivaji University and Now,M.E. Scholar at Pune University , ZEAL education's "Dnyanganaga College of Engineering And Research ", Narhe, Pune.

Mr. Mane Prashant M. (Assistant Professor) at Pune University, ZEAL education's "Dnyanganaga College of Engineering And Research ", Narhe, Pune. Qualification: M.E. (IT), Teaching Experience: 6 Years