REVIEW ON OPINION DATA SUMMARIZATION USING K-MEANS CLUSTERING AND LATENT SEMANTIC ANALYSIS

***Renu, **Miss Neha, ***Mr. Kunal**

[*] *Department of Computer Science, Shri Ram college of Engineering & Management, Palwal*

ABSTRACT

Text Summarization is condensing the source text into a shorter version preserving its information content and overall meaning. It is very difficult for human beings to manually summarize large documents of text. Text Summarization methods can be classified into extractive and abstractive summarization. An extractive summarization method consists of selecting important sentences, paragraphs etc. from the original document and concatenating them into shorter form. The importance of sentences is decided based on statistical and linguistic features of sentences. An abstractive summarization method consists of understanding the original text and re-telling it in fewer words. It uses linguistic methods to examine and interpret the text and then to find the new concepts and expressions to best describe it by generating a new shorter text that conveys the most important information from the original text document. Usually, the flow of information in a given document is not uniform, which means that some parts are more important than others. The major challenge in summarization lies in distinguishing the more informative parts of a document from the less ones. Though there have been instances of research describing the automatic creation of abstracts, most work presented in the literature relies on verbatim extraction of sentences to address the problem of single-document summarization. In this scheme, we describe some eminent extractive techniques. First, we look at early work from the aspect of research on summarization. Second, we concentrate on approaches involving machine learning techniques. In this dissertation, ontology based document summarization is proposed that provide efficient and accurate summary than other approaches. The main motivation for summarization is to identifying summary from a large document, that it is a data is beneficial for us or not. It is identify weather a product is purchasable or not. This make difficult for a potential customer to read them to make an informed decision on whether to purchase the product. It also makes it difficult for the manufacturer of the product to keep track and to manage customer opinion. In the scheme we proposed enhanced algorithm vide latent semantic kernel for better results.

Index Terms- ***Data or Text Summarization, Inverse Document Frequency, Document Clustering***.

INTRODUCTION

Text summarization: Text summarization has become an important and timely tool for assisting and interpreting text information in today's fast-growing information age. It is very difficult for human

beings to manually summarize large documents of text. There is an abundance of text material available on the internet. However, usually the Internet provides more information than is needed. Therefore, at twofold problem is encountered: searching for relevant documents through an overwhelming number of documents available, and absorbing a large quantity of relevant information. The goal of automatic text summarization is condensing the source text into a shorter version preserving its information content and overall meaning. A summary can be employed in an indicative way as a pointer to some parts of the original document, or in an informative way to cover all relevant information of the text. In both cases the most important advantage of using a summary is its reduced reading time. A good summary system should reflect the diverse topics of the document while keeping redundancy to a minimum. Summarization tools may also search for headings and other markers of subtopics in order to identify the key points of a document. Microsoft Word's AutoSummarize function is a simple example of text summarization.

Single-Document text Summarization: Usually, the flow of information in a given document is not uniform, which means that some parts are more important than others. The major challenge in summarization lies in distinguishing the more informative parts of a document from the less ones. Though there have been instances of research describing the automatic creation of abstracts, most work presented in the literature relies on verbatim extraction of sentences to address the problem of single-document summarization. In this section, we describe some eminent extractive techniques. First, we look at early work from the 1950s and 60s that kicked of research on summarization. Second, we concentrate on approaches involving machine learning techniques published in the 1990s to today. Finally, we briefly describe some techniques that use a more complex natural language analysis to tackle the problem

Multi-Document text Summarization: is an automatic procedure aimed at extraction of information from multiple texts written about the same topic. The resulting summary report allows individual users, such as professional information consumers, to quickly familiarize themselves with information contained in a large cluster of documents. In such a way, multi-document summarization systems are complementing the news aggregators performing the next step down the road of coping with information overload.

Text Summarization Early History: Interest in automatic text summarization, arose as early as the fifties. An important paper of these days is the one in 1958, suggested to weight the sentences of a document as a function of high frequency word disregarding the very high frequency common words. Automatic text summarization system in 1969, which, in addition to the standard keyword method (i.e., frequency depending weights), also used the following three methods for determining the sentence weights:

1. Cue Method: This is based on the hypothesis that the relevance of a sentence is computed by the presence or absence of certain cue words in the cue dictionary.
2. Title Method: Here, the sentence weight is computed as a sum of all the content words appearing in the title and (sub-) headings of a text.
3. Location Method: This method is based on the assumption that sentences occurring in initial position of both text and individual paragraphs have a higher probability of being relevant. The results showed, that the best correlation between the automatic and human-made extracts was achieved using a combination of these three latter methods.

The Trainable Document Summarizer [9] in 1995 performs sentence extracting task, based on a number of weighting heuristics. Following features were used and evaluated:

1. Sentence Length Cut-O Feature: sentences containing less than a pre-specified number of words are not included in the abstract
2. Fixed-Phrase Feature: sentences containing certain cue words and phrases are included
3. Paragraph Feature: this is basically equivalent to Location Method feature in
4. Thematic Word Feature: the most frequent words are defined as thematic words. Sentence scores are functions of the thematic words' frequencies
5. Uppercase Word Feature: upper-case words (with certain obvious exceptions) are treated as thematic words, as well.

A Corpus was used in this method, which contained 188 document/summary pairs from 21 publications in a scientific/technical domain. The summaries were produced by professional experts and the sentences occurring in the summaries were aligned to the original document texts, indicating also the degree of similarity as mentioned earlier, the vast majority (about 80%) of the summary sentences could be classified as direct sentence matches. The ANES text extraction system in 1995 is a system that performs automatic, domain-independent condensation of news data. The process of summary generation has four major constituents:

Algorithm:

1: Rank all the sentences according to their score.
2: Add the main title of the document to the summary.
3: Add the first level-1 heading to the summary.
4: While (summary size limit not exceeded)
5: Add the next highest scored sentence.
6: Add the structural context of the sentence: (if any and not already included in the summary)
7: Add the highest level heading above the extracted text (call this heading h).
8: Add the heading before h in the same level.
9: Add the heading after h in the same level.
10: Repeat steps 7, 8 and 9 for the next highest level headings.

An another query-specific summarization method views a document as a set of interconnected text fragments (passages) and focuses on keyword queries, since keyword search is the most popular information discovery method on documents, because of its power and ease of use. Firstly, at the preprocessing stage, it adds structure to every document, which can then be viewed as a labeled, weighted graph, called the document graph. Then, at query time, given a set of keywords, it performs keyword proximity search on the document graphs to discover how the keywords are associated in the document graphs. For each document its summary is the minimum spanning tree on the corresponding document graph that contains all the keywords.

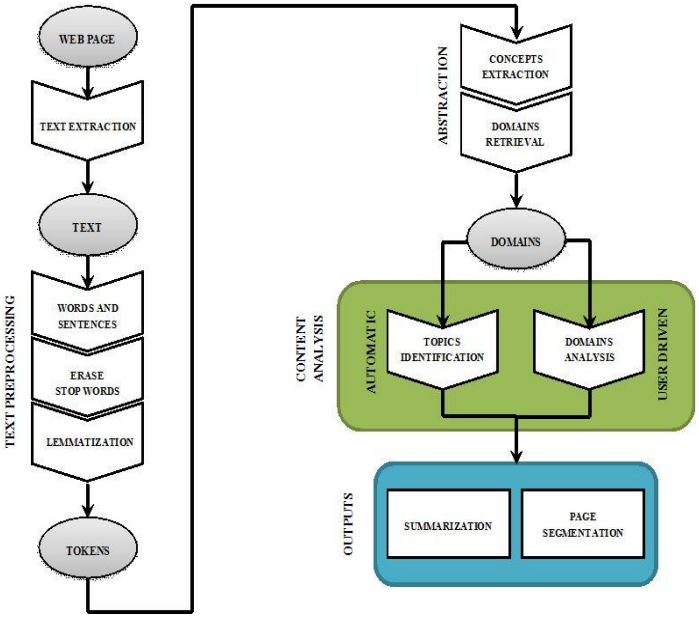

Figure 1: Workflow of Opinion Data Summarization

Text Summarization : With the proliferation of online textual resources, an increasing need has arisen to improve online access to data. This requirement has been partly addressed through the development of tools aimed at the automatic selection of portions of a document, which are best suited to provide a summary of the document, with reference to the user's interests. Text summarization has become one of the leading topics in informational retrieval research, and it was identified as one of the core tasks of computational linguistics and AI in the early 1970's. Thirty Five years later, though good progress has been made in developing robust, domain independent approaches for extracting the key sentences from a text and assembling them into a compact, coherent account of the source, summarization remains an extremely difficult and seemingly intractable problem. Despite the primitive state of our understanding of discourse, there is a common belief that a great deal can be gained for summarization from understanding the linguistic structure of the texts.

Humans generate a summary of a text by understanding its deep semantic structure using vast domain/common knowledge. It is very difficult for computers to simulate these approaches. Hence, most of the automatic summarization programs analyze a text statistically and linguistically, to determine important sentences, and then generate a summary text from these important sentences. The main ideas of most documents can be described with as little as 20 percent of the original text. Automatic summarization aims at producing a concise, condensed representation of the key information content in an information source for a particular user and task. In addition to developing better theoretical foundations and improved characterization of
summarization problems, further work on proper evaluation methods and summarization resources, especially corpora, is of great interest. Research papers and results of investigation reported in literature over the past decade have been analyzed with a view to crystallize the work of various authors and to discuss the current trends especially for a legal domain.

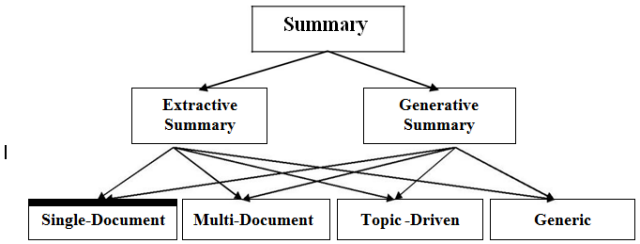

Figure 2: Classification of Summary Types

Taxonomically one can distinguish among the following types of summaries: Extractive/non-extractive, generic/query-based, single-document/multi-document, and monolingual/multilingual/cross lingual. Most existing summarizers work in an extractive fashion, selecting portions of the input documents (e.g. sentences) that are believed to be more salient. Non-extractive summarization includes dynamic reformulation of the extracted content, involving a deeper understanding of the input text, and is therefore limited to small domains. Query-based summaries are produced in reference to a user query (e.g. summarize a document about an international summit focusing only on the issues related to the environment) while generic summaries attempt to identify salient information in text without the context of a query. The difference between single- and multi-document summarization (SDS and MDS) is quite obvious; however some of the types of problems that occur in MDS are qualitatively different from the ones observed in SDS e.g. addressing redundancy across information sources and dealing with contradictory and complementary information. No true multilingual summarization systems exist yet; however, cross-lingual approaches have been applied successfully.

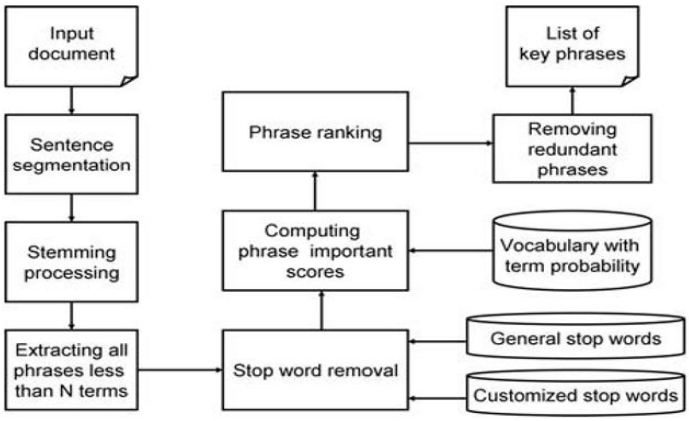

Figure 3: Analogy of Text Summarization

Term occurrence information: In addition to the evidence provided by the structural organization of the documents, the summarization system utilizes the number of term occurrences within each document to further assign weights to sentences. Instead of merely assigning a weight to each term according to its frequency within the document, the system locates clusters of significant words [20] within each sentence, and assigns a score to them accordingly. The scheme that is used for computing the significance factor for a sentence was originally proposed by Luhn [20]. It consists of defining the extent of a cluster of related words, and dividing the square of this number by the total number of words within this cluster.

Query-based summarization: Research on Question Answering (QA) is focused mainly on classifying the question type and finding the answer. Presenting the answer in a way that suits the user's needs has received little attention. A question answering system pinpoints an answer to a given question in a set of documents. A response is then generated for this answer, and presented to the user. Studies have shown however that the users appreciate receiving more information than only the exact answer. Consulting a question answering system is only part of a user's attempt to fulfill the information need: it's not the end point, but some steps along what has been called a 'berry picking' process, where each answer/result returned by the system may motivate a follow-up step. The user may not only be interested in the answer to a question, but also in the related information. The 'exact answer approach' fails to show leads to related information that might also be of interest to the user. This is especially true in the legal domain. Lin et al. show that when searching for information, increasing the amount of text returned to the users can significantly decrease the number of queries that they pose to the system, suggesting that users utilize related information from the supporting texts. In both the commercial and academic QA systems, the response to a question tends to be more than the exact answer, but the sophistication of their responses varies from system to system. Exact answer, answer plus context and extensive answer are the three degrees of sophistication in response generation. So the best method is to produce extensive answers by extracting the sentences which are

most salient with respect to the question, from the document which contains the answer. This is very similar to creating an extractive summarization: in both cases, the goal is to extract the most salient sentences from a document. In question answering, what is relevant depends on the user's question rather than on the intention of the writer of the document that happens to contain the answer. In other words, the output of the summarization process is adapted to suit the user's declared information need (i.e. the question). This branch of summarization has been called query-based summarization.

Fields of Application:

1. **Purchasing Product or Service:** While purchasing a product or service, taking right decision is no longer a difficult task. By this technique, people can easily evaluate other's opinion and experience about any product or service and also he can easily compare the competing brands.
2. **Quality Improvement in Product or service:** By Opinion mining and sentiment analysis the manufactures can collect the critic's opinion as well as the favorable opinion about their product or service and thereby they can improve the quality of their product or service.
3. **Marketing research:** By sentiment analysis techniques, the recent trend of consumers about some product or services can be analyzed. Similarly the recent attitude of general public towards some new government policy can also be easily analyzed. These all result can be contributed to collective intelligent research.
4. **Recommendation Systems**: By classifying the people's opinion into positive and negative, the system can say which one should get recommended and which one should not get recommended.

Cluster based method : The idea of clustering is to group similar objects into their classes. As far as multi documents are concerned, these objects refer to sentences and the classes represent the cluster that a sentence belongs to. By looking at the nature of documents that address different subjects or topics in the documents, some researchers try to incorporate the idea of clustering into their study. Using the concept of similarity, sentences which are highly similar to each other are grouped into one cluster, thus generating a number of clusters. The most common technique to measure similarity between a pair of sentences is the cosine similarity measure where sentences are represented as a weighted vector of tf-idf. Once sentences are clustered, sentence selection is performed by selecting sentence from each cluster. Sentence selection is then based on the closeness of the sentences to the top ranking tf-idf in that cluster. Those selected sentences are then put together to form the final summary.

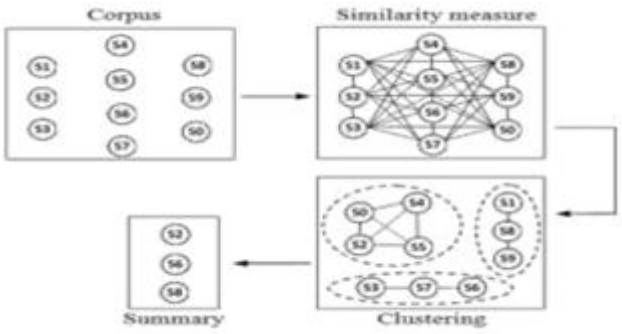

Figure 4: Various Clustering techniques.

Graph based method: The fundamental theory of graph representation is the connection or linking between objects. These connections exist based on their underlying relation. In the case of text documents, the underlying relation is usually the similarity between objects-in this case, sentences. Generally, a graph can be denoted in the form of G = (V, E), where V represents the graph's vertex or node and E is the edge between each vertex. In the context of text documents, vertex represents sentence and edge is the weight between two sentences. Using this approach, documents can therefore be represented as a graph where each sentence becomes the vertex and the weight between each vertex corresponds to the similarity between the two sentences. As in most literature concerning graph based approach, the most widely used similarity measure is the cosine similarity measure. An edge then exists if the similarity weight is above some predefined threshold. Figure 2.3 shows an example graph based document Summarization.

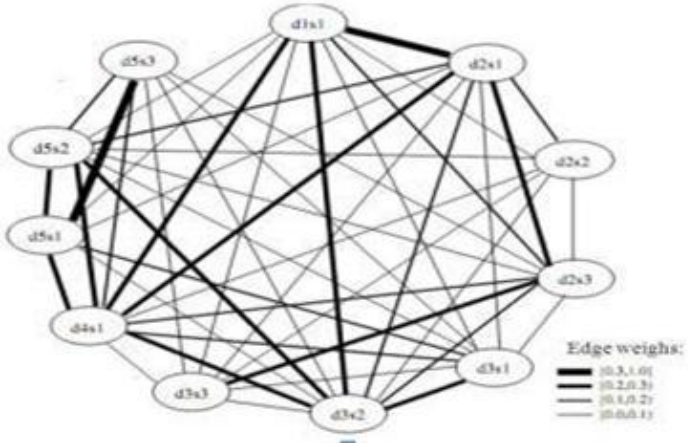

Figure 5: Graph based Summarization

Latent Semantic Analysis (LSA) is the combination of algebraic and statistical methods and this technique brings out the hidden structure of words, between words, sentences or document. The main ideas of LSA is that it extracts the input document and convert to sentence – term matrix and process it through an algorithm called singular value decomposition(SVD). The purpose of the SVD is to find

relationship between word and sentences, reduce noise and also model the relationship among sentences and words. Finally, output is obtained from SVD algorithm. LSA main algorithm to text summarization is divided into three steps: creation of sentence - term matrix, applying SVD to matrix and selection the sentence for the summary.

Latent Semantic Analysis approach The system generates latent semantic analysis for four existing systems and two proposed systems were used as an input document given by DUC as peer summaries. The input documents used abstract and extract as part of the library for the process of the input and an amalgamations of similarity index file is produced as output summary for all the documents in the document set were retrieved and stored for use by all summarization system.

LITERATURE REVIEW

R. Baeza-Yates, C. Hurtado, and M. Mendoza [6] suggests that, the search engine gives the list of related results. These results are based on the previously searched queries or such technique can be used to tune or redirect the user. In this method the clustering algorithm is used. The clustering is done on the basis of previously fired queries. It clusters the semantically similar queries. It does not only give the clustered data but it also ranks the suggested list of result. The ranking is done on the basis of two conditions, 1. Similarity of queries to the input query 2. Observation that measures the attention of the user attracted towards the result of the query. The combination of both these conditions measures the user interests. In the given algorithm, query session is considered for giving the result. The query session also considers the rank of clicked URL.The relevance ranking is measured by using two components similarity of query and support of query

Joel larocca Neto, Alex A. Freitas and Celso A.A.Kaestner, "Automatic Text Summarization using a Machine Learning Approach" **[3]** discuses that user search goals for a query by clustering feedback sessions. For that, we use a concept of pseudo document, which is the revised version of feedback session. At the end, we cluster these pseudo-documents to infer user search goals and represent them with some keywords. Since the evaluation of clustering is also an important problem, we used evaluation criterion classified average precision (CAP) to evaluate the performance of the restructured web search results. The clustering is done by bisecting k means where in the existing system it is done by k means clustering. The new algorithm increases the efficiency of result. After the segmented result formation, the result in the every segment is reorganized as per number of clicks of URLs. The link which is clicked more number of times will appear at first location in the segment. This reduces the time requirement for searching.

Dasari Amarendra, Kaveti Kiran Kumar[10] suggest that user's information needs due to the use of short queries with uncertain terms. thus to get the best results it is necessary to capture different user

search goals. These user goals are nothing but information on different aspects of a query that different users want to obtain. The judgment and analysis of user search goals can be improved by the relevant result obtained from search engine and user's feedback. Here, feedback sessions are used to discover different user search goals based on series of both clicked and unclicked URL's. The pseudo-documents are generated to better represent feedback sessions which can reflect the information need of user. With this the original search results are restructured and to evaluate the performance of restructured search results, classified average precision (CAP) is used. This evaluation is used as feedback to select the optimal user search goals.

PROPOSED WORK

The main idea of this approach is to classify sentences to a hierarchical Clustering along with K-Means which captures the theme of the sentence and then calculate a similarity measure between the sentence and the document that it belongs to. Our approach uses IF and IDF using K-Means clustering. The proposed approach involves Latent Semantic Analysis Approach, Document Segmentation, very less amount of work has been done in this area and few algorithms have been proposed. However, in this scheme we will ensure to increase accuracy and efficiency of the summary obtained for the document. In this paper, a novel approach for text summarization using clustering is presented below diagram depicts the same.

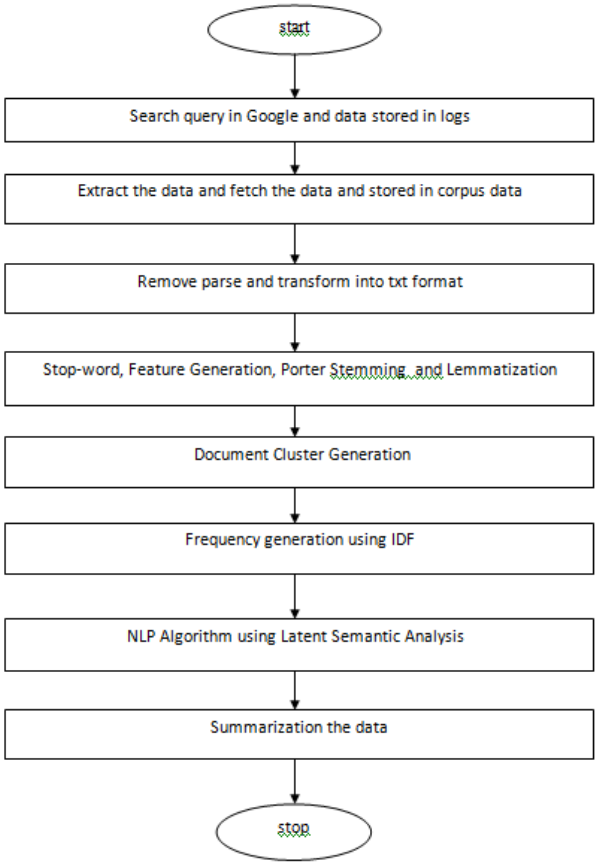

Figure 6 : Proposed Technique

REFERENCES

[1] R. Baeza-Yates, C. Hurtado, and M. Mendoza, "Query Recommendation Using Query Logs in Search Engines,"Proc. Int'l Conf. Current Trends in Database Technology (EDBT '04), pp. 588-596, 2004

[2] Harshada P. Bhambure, Mandar Mokashi, "Inferring User Search Goals Using Feedback Session", International Journal of Science and Research (IJSR), www.ijsr.net, Volume 4 Issue 6, June 2015, 2880 - 2884 - See more at: http://www.ijsr.net/archive/v4i6/v4i6_03.php#sthash.1TMBdOKC.dpuf

[3] Joel larocca Neto, Alex A. Freitas and Celso A.A.Kaestner, "Automatic Text Summarization using a Machine Learning Approach", Book: Advances in Artificial Intelligence: Lecture Notes in computer science, Springer Berlin / Heidelberg, Vol 2507/2002, 205-215, 2002.

[4] Weiguo Fan, Linda Wallace, Stephanie Rich, and Zhongju Zhang, “Tapping into the Power of Text Mining”, Journal of ACM, Blacksburg, 2005.

[5] Fang Chen, Kesong Han and Guilin Chen, “An Approach to sentence selection based text summarization", Proceedings of IEEE TENCON02, 489-493, 2002.

[6] 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.

[7] H. P. Luhn, “The Automatic Creation of Literature Abstracts”, Presented at IRE National Convention, New York, 159-165, 1958.

[8] H. P. Edmundson.,” New methods in automaticextracting”, Journal of the ACM, 16(2):264-285, April 1969.

[9] J. Kupiec, J. Pedersen, and F. Chen, “A trainable document summarizer”, In Proceedings of the 18th ACMSIGIR Conference, pages 68-73, 1995

[10] Ronald Brandow, Karl Mitze, and Lisa F. Rau. “Automatic condensation of electronic publications by sentence selection. Information Processing and Management”, 31(5):675-685,1995.

[11] E. Mittendorf and P. Schauble, “ Document and passage retrieval based on hidden markov models”, In Proceedings of the 17th ACM-SIGIR Conference, pages 318-327,1994.

[12] A. Bookstein, S. T. Klein, and T. Raita, “Detecting content-bearing words by serial clustering”, In Proceedings of the 18th ACM-SIGIR Conference, pages 319-327, 1995.

[13] Madhavi K. Ganapathiraju, “Overview of summarization methods”, 11-742: Self-paced lab in Information Retrieval, November 26, 2002.

[14] Klaus Zechner, “A Literature Survey on Information Extraction and Text Summarization”, Computational Linguistics Program, Carnegie Mellon University, April 14, 1997.