Novel Algorithm for summarizing the group of documents using Lexical Chains Analysis and Sentiwordnet

Kavita Jain[1], Sayar Singh[2]
[1]*M.Tech Scholar, Department of Computer Science and Engineering, Arya Institute of Technology & Engineering, Jaipur Rajasthan, INDIA*
[2] *Head of Department, Assistant Professor, Department of Computer Science and Engineering , Arya Institute of Technology & Engineering, Jaipur Rajasthan, INDIA*

***Abstract*- Summarizing the documents is the big issue and the reading of all the documents which are available over the internet is not possible. But is the simple summary of the document is available then we can easily get the meaningful information out of it., In this paper we presents the novel algorithm which by making use of the lexical chains and wordnet simplifies the summary generation process and generates the summary in the faster and the effective way as compared to the previous work in this field.**

***Index Terms*- Lexical chains, wordnet, document summary , extractive summary.**

I. INTRODUCTION

Today summarization technologies are used in large number of sectors, for example in search engines (Google), document summarization, image collection summarization and video summarization. By finding the most informative sentences document summarization automatically create a representative summary or abstract of the entire document. Similarly, in image summarization the system finds the most representative and important images. Likewise, in consumer videos one would want to remove the boring or repetitive scenes, and extract out a much shorter and abstract version of the video. In surveillance videos, extraction of important events in the recorded video is only considered, since most part of the video may be uninteresting with nothing going on. The problem of information overload grows, and the amount of data increases, the interest in automatic summarization is also increasing.

1.2 Classification of Summarization

Text summarization strategies are often classified into extractive and abstractive summarization [2]. An extractive summarization technique consists of choosing necessary sentences, paragraphs etc. From the original document and concatenating them into shorter kind. The importance of sentences is determined based on statistical and linguistic characteristics of sentences [2].

It uses linguistic strategies to look at and interpret the text and so to search out the new concepts and expressions to best describe it by generating a brand new shorter text that conveys the most necessary info from the initial text document.

Extractive summaries [2] are developed by extracting key text segments (sentences or passages) from the text, based mostly on statistical analysis of individual or mixed surface level options like word/phrase frequency, location or cue words to find the sentences to be extracted. The "most important" content is treated as the "most frequent" or the "most favorably positioned" content. Such an approach therefore avoids any efforts on deep text understanding. They're conceptually easy, simple to implement.

Extractive text summarization [2] methods are often divided into 2 steps:

1) Preprocessing step and

2) Processing step.

Preprocessing is structured illustration of the initial text.

It usually includes:

a) Sentences boundary identification [2]:- In English, sentence boundary is known with presence of dot at the end of sentence.

b) Stop-Word Elimination [2]:-Common words with no semantics and that don't combine relevant info to the task is eliminated.

c) Stemming [2]:-the purpose of stemming is to get the stem or base form of every word that emphasize its semantics.

In processing step, characteristics influencing the relevancy of sentences are determined and calculated and so weights are allotted to those features using weight learning technique. Final score of every sentence is decided using Feature-weight equation. Prime hierarchal sentences are elected for final summary.

Automatic Text Summarization can be characterized into single document text summarization and multi document summarization.

Single-Document Summarization: The biggest challenge in summarization is to identify or generalize the most important and informative sentences from a document because the information in the document is non-uniform usually [1].

There are certain ways for single document summarization:

- Naïve-Bayes [2]: Here a classification function namely naïve-bayes is used to distinguish whether sentences are likely to be extracted or not.
- Rich Features and Decision Trees [3]: For the most part the text is depicted in an anticipated talk structure and the important sentences happen at particular areas. This strategy is known as "position technique" which demonstrates the position of sentences.
- Hidden Markov Model [4]: Conroy et al utilized concealed markov demonstrate (HMM) and recognized the problem of sentence extraction from a document.
- Log Linear Model [5]: Osborne utilized log-straight models and demonstrated that current methodologies utilized component autonomy and these models create preferable concentrates over credulous bayes show.
- Neural Networks [6]: Because of its outflanking measurable criticalness, neural system defeats the problem of extractive summarization.
- Deep Natural Language Analysis Method [7]: Here a set of heuristics are used to make document extracts. Also they model the discourse structure of texts.
- Multi-Document Text Summarization: Since 1990's, single document extraction has moved to numerous document extractions in the area of news articles. In spite of the fact that solitary document puts opposing outcomes by covering the information due to various documents availability [1]. So the significant concentrate on summary is that summary ought to take after the culmination, accuracy, incorrect property.
- Abstraction and Information Fusion [11, 12]: Here a summary is built by fusing multiple documents by giving input to process the text and then extracting the important information to produce a well-structured summary.
- Topic-driven Summarization and MMR [13]: Here the main focus is on the query and the information retrieved from text retrieval to topic-driven summarization. In maximal marginal relevance (MMR), the redundant sentences are less rewarded by some similarity measures.
- Graph Spreading Activation [14]: In this a document is dealt with as a diagram and every hub speaks to the word with its position. Also a node can have various links like adjacency links (ADJ) which shows the adjacent words, same links which shows the number of occurrences of a word, Alpha links encodes the meanings. Also Phrase links binds the arrangement of adjoining hubs in an expression while Name and Core links checks the event of co-referential name.
- Centroid-based Summarization [15]: Here articles are grouped together which describes the same event. Every cluster constitutes of 2-10 articles from different sources and are arranged in chronological order. This step is called as topic detection. An agglomerative clustering algorithm adds documents to clusters by using TF-IDF vector and recomputed the centroids.
- Multilingual Multi-document Summarization [16]: Here multiple documents are there in multiple languages. First, a translation system is applied for translation of document in a single preferable language. Then similar sentences are searched in the documents. If found relevant then they are included in summary directly rather than translating. This is useful for news applications

that take information from other agencies of different language.

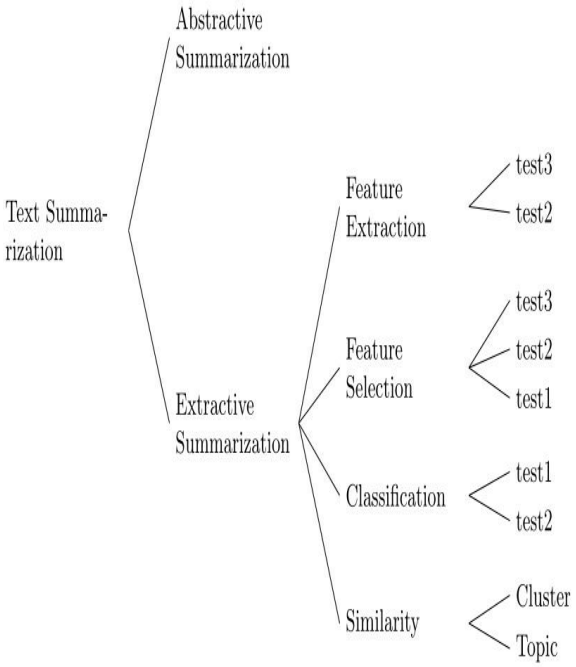

Fig 1. Classification of Text Summarization

II. LITERATURE REVIEW

Xu Han etal [19] Text summarization is to generate a condensed version of the original document. The major issues for text summarization are eliminating redundant information, identifying important difference among documents, and recovering the informative content. This paper proposes a Semantic Graph Model which exploits the semantic information of sentence using FSGM. FSGM treats sentences as vertexes while the semantic relationship as the edges. It uses FrameNet and word embedding to calculate the similarity of sentences. This method assigns weight to both sentence nodes and edges. After all, it proposes an improved method to rank these sentences, considering both internal and external information. The experimental results show that the applicability of the model to summarize text is feasible and effective.

Mohsen Pourvali etal [20] The technology of automatic document summarization is developing and may give an answer for the information over-burden problem. These days, document summarization assumes an important part in information recovery. With a vast volume of documents, giving the client a summary of each document enormously encourages the assignment of finding the coveted documents. Document summarization is a process of automatically making a compressed variant of a given document that gives helpful information to clients, and multi-document summarization is to create a summary conveying the larger part of information substance from an arrangement of documents around an express or understood principle point. The lexical union structure of the text can be abused to decide the significance of a sentence/expression. Lexical chains are valuable apparatuses to dissect the lexical attachment structure in a text. In this paper we consider the impact of the utilization of lexical union components in Summarization, and displaying an algorithm base on the information base. Our own algorithm at first locate the right feeling of any word, then develops the lexical chains, expel Lexical chains that less score than other, identifies themes generally from lexical chains, sections the text concerning the points and chooses the most important sentences. The exploratory outcomes on an open benchmark dataset from DUC01 and DUC02 demonstrate that our proposed approach can enhance the execution contrasted with satisfy of-the-craftsmanship summarization approaches.

NimishaDheer etal [21] The current technology of automatic text summarization imparts an important role in the information retrieval (IR) and text classification, and it provides the best solution to the information overload problem. Text summarization is a process of reducing the size of a text while protecting its information content. When taking into consideration the size and number of documents which are available on the Internet and from the other sources, the requirement for a highly efficient tool on which produces usable summaries is clear. They present a better algorithm using lexical chain computation &WordNet. The algorithm one which makes lexical chains that is computationally feasible for the user. Using these lexical chains, the user will generate a summary, which is much more effective compared to the solutions available and also closer to the human generated summary.

H. Gregory Silber etal [22], The increased in the growth of the net has resulted in huge amounts of information that has become tougher to access with

efficiency. Web users need tools to manage this immense amount of information. The main goal of this analysis is to form an economical and effective tool that's able to summarize quite large documents quickly. This analysis presents a linear time algorithmic rule for finding out lexical chains that could be a technique of capturing the "abruptness" of a document. They additionally give different strategies for extracting and evaluation lexical chains. They show that their technique provides similar results to previous analysis, however is considerably quite more efficient. This efficiency is important in web search applications where several quite large documents might have to be summarized promptly, and where the reaction time to the end user is very vital.

Form this paper, we have learned and inspired by the concept of the lexical chains,and how they are created and applied in the field of the text summarization.

From this paper, we have also learned the concept of how to score the chain and find the usability of the chains.

ReginaBarzilay , They investigate one technique to supply a summary of an original text while not requiring its full semantic interpretation [23], however instead hoping on a model of the topic progression within the text derived from lexical chains. They present a new algorithmic program to find out lexical chains in a text, merging many robust knowledge sources: The WordNet thesaurus, a part-of-speech tagger, shallow parser for the identification of nominal teams, and a segmentation algorithmic program. Summarization is carried out in four steps: the initial step is, text is segmented, lexical chains are made, strong chains are marked or identified and vital sentences are extracted.

They present in this paper empirical results on the identification of strong chains [11] and of important sentences. Preliminary results indicate that quality indicative summaries are made. Unfinished issues are then identified. Plans to deal with these short-comings are concisely presented.

NikitaMunot , Text summarization is among one application of natural language processing and is now becoming much common for info condensation. Text summarization could be a method of reducing the size of original document and results a summary by holding necessary info of original document. This paper provides comparative study of varied text summarization strategies based on differing kinds of application. The paper discusses well 2 main classes of text summarization strategies these are extractive and abstractive summarization strategies [24]. The paper conjointly presents taxonomy of summarization systems and statistical and linguistic approaches [12] for summarization.

Natural language processing (NLP) could be a field of computer science, artificial intelligence and linguistics involved with the interactions between computers and human language. Natural language processing could be a method of developing a system process and results language pretty much as good as human can turn out. Document summarization is one attainable solution to the present problem.

Text summarization could be a method to precise the content of a document in a very condensed form that meets the requirements of the user. More and more electronic data is out there on the net and it's impracticable to read everything and therefore some sort of info condensation is required. Summarization is a tool that helps the user to expeditiously find required info from vast amount of information.

III. PROPOSED CONCEPT

Multi-document 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.

We have extended our research in summarizing the multiple documents at a time, so that it will reduce the work load and will save the time it getting the complete meaning or summary of the large documents.

And for improving the recall we have improve the coverage criteria, including the Noun, Proper Noun, Verbs ,Adjectives etc..

As the result of which we have saved the time as well as get the better recall

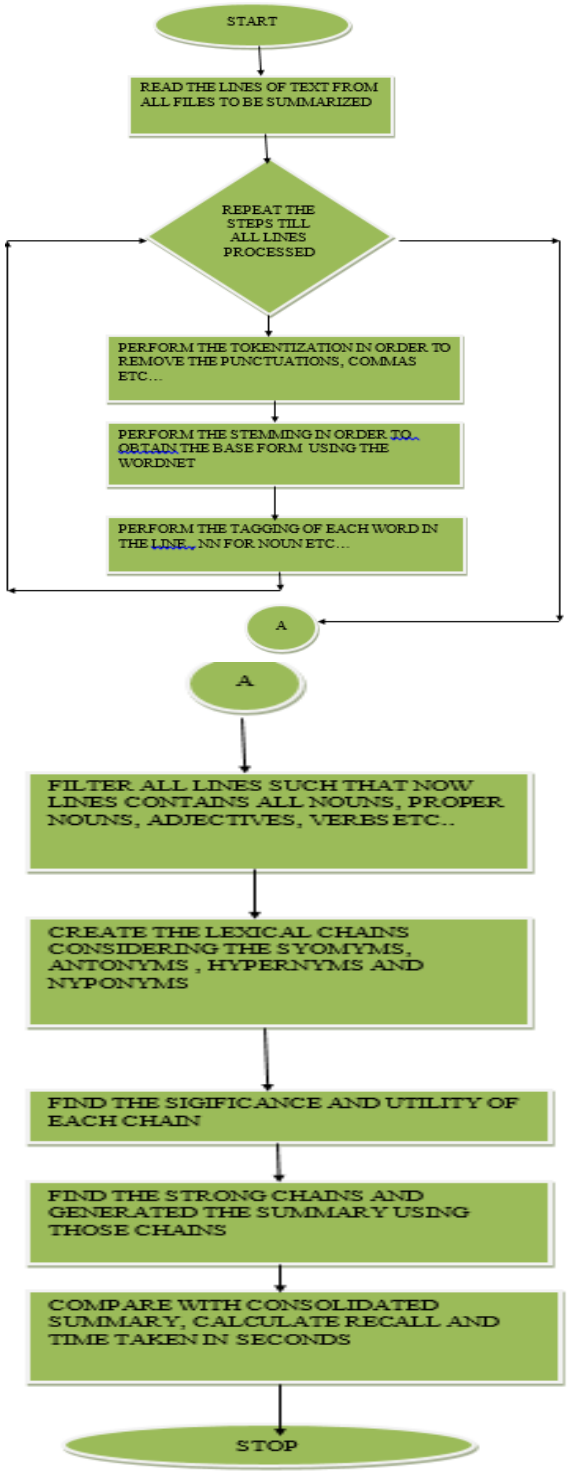

Fig 2. Flowchart of Proposed Concept

IV. IMPLEMENTATION

Java

Java is a general-purpose computer programming language that is concurrent, class-based, object-oriented, and specifically designed to have as few implementation dependencies as possible. It is intended to let application developers "write once, run anywhere" (WORA), meaning that compiled Java code can run on all platforms that support Java without the need for recompilation. Java applications are typically compiled to bytecode that can run on any Java virtual machine (JVM) regardless of computer architecture. As of 2016, Java is one of the most popular programming languages in use,particularly for client-server web applications, with a reported 9 million developers. Java was originally developed by James Gosling at Sun Microsystems (which has since been acquired by Oracle Corporation) and released in 1995 as a core component of Sun Microsystems' Java platform. The language derives much of its syntax from C and C++, but it has fewer low-level facilities than either of them.

Eclipse

Eclipse is an integrated development environment (IDE) used in computer programming, and is the most widely used Java IDE. It contains a base workspace and an extensible plug-in system for customizing the environment. Eclipse is written mostly in Java and its primary use is for developing Java applications, but it may also be used to develop applications in other programming languages via plug-ins, including Ada, ABAP, C, C++, COBOL, D, Fortran, Haskell, JavaScript, Julia,Lasso, Lua, NATURAL, Perl, PHP, Prolog, Python, R, Ruby (including Ruby on Rails framework), Rust, Scala, Clojure, Groovy, Scheme, and Erlang. It can also be used to develop documents with LaTeX (via a TeXlipse plug-in) and packages for the software Mathematica. Development environments include the Eclipse Java development tools (JDT) for Java and Scala, Eclipse CDT for C/C++, and Eclipse PDT for PHP, among others.

Java Free Charts

JFreeChart is an open-source framework for the programming language Java, which allows the creation of a wide variety of both interactive and non-interactive charts.

JFreeChart supports a number of various charts, including combined charts:

- X-Y charts (line, spline and scatter). Time axis is possible.

- Pie charts
- Gantt charts
- Bar charts (horizontal and vertical, stacked and independent). It also has built-in histogram plotting.
- Single valued (thermometer, compass, speedometer) that can then be placed over map.
- Various specific charts (wind chart, polar chart, bubbles of varying size, etc.).

It is possible to place various markers and annotations on the plot.

JFreeChart also works with GNU Classpath, a free software implementation of the standard class library for the Java programming language.

JFreeChart automatically draws the axis scales and legends. Charts in GUI automatically get the capability to zoom in with mouse and change some settings through local menu. The existing charts can be easily updated through the listeners that the library has on its data collections.

Fig 3 shows the multi-document summarization implementation. It will scan the complete directory and compare the obtained summary with the consolidated summary of all the document.

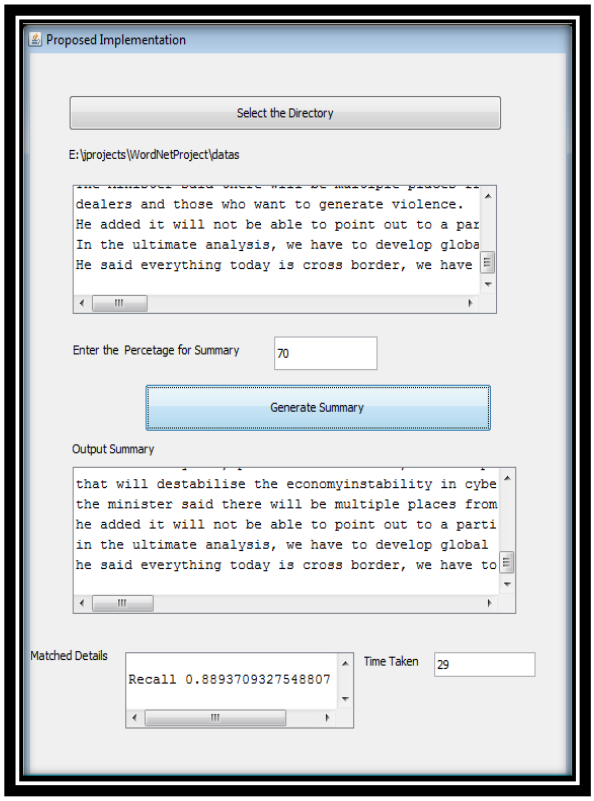

Fig 3. Multi-document summarization

V. TEST RESULTS

In this we have taken the set of various documents together with the manual generated summary and tested in both the implementations to get the result related to recall and time taken in the summary generation process.

Result comparison for DataSet1

1. Sample Document 1


Most San Francisco-area homeowners may have to pay for damage from Tuesday's earthquake out of their own pockets, while insurance companies may reap long-term benefits from higher rates, industry spokesmen and analysts said Wednesday. Only 15 percent to 20 percent of California homeowners have earthquake insurance, which typically requires a 10 percent deductible and costs between $200 to $400 a year for a $100,000 home, according to industry spokesmen. The Association of California Insurance Cos. in Sacramento said that in the San Francisco area roughly 25 to 30 percent of the homes have earthquake insurance. The organization estimated residential damages from Tuesday's earthquake at $500 million in the Bay area, with between $100 million to $150 million insured. Insured homeowners without earthquake protection will get reimbursed only if their homes were ravaged by fire, which is covered under basic homeowner insurance polices, said Hugh Strawn, director of catastrophe services at the Property Loss Research Bureau in Schaumburg, Ill. Insurance companies attempted Wednesday to assess the amount of quake-related damages they're likely to have to pay. In addition to home damage, the companies likely will get claims for automobile damage, broken glass, theft and burglary, business interruption due to electrical outages, water damage and, possibly, workers compensation. Some estimated that insurers might face bills totaling $1 billion or more from the quake. But industry observers said they don't expect any company to suffer serious financial damage from quake-related claims. ''We don't think any company is going to have problems paying claims,'' said Elisa Siegal, public affairs manager for the American Insurance Association, a washington-based trade group. The insurers actually could benefit. Industry analysts predicted insurers would be able to reverse three years of declining rates and win rate hikes from state regulators due to the quake damages and the estimated $4 billion in damages from Hurricane Hugo, which hammered South Carolina and other parts of the southeastern United States earlier this month. An increase in insurance rates could translate into greater profitability in the long term, the analysts said. ''There's a perception that this could turn the cycle ...'' that this could be enough to firm pricing,'' said Gloria L. Vogel, an analyst with Bear, Stearns & Co. Inc. in New York. Despite their predictions for the long run, the analysts warned that fourth-quarter earnings among insurance companies are likely to be disappointing. ''It's going to be a meaningful loss, perhaps as big as Hugo,'' said Robert Glasspiegel, an analyst with Hartford-based Langen McAlenney. On the New York Stock Exchange, some insurance company stock rose on wednesday. Aetna Life & Casualty Co. rose $2.37 to $59.50 a share; ITT Corp., parent of The Hartford, was up 37 cents to $59; and the Travelers Cos. rose $1 to $40.87. Reinsurance companies, which absorb risk from policy writers, did especially well. General Re Corp. rose $2.75 to $86.50 and Marsh & McLennan Cos. rose $3.12 to $75.87. In California, earthquake insurance is an emotional and politically-charged issue.

Fig 4. Sample Document 1[8]

2. Manual Summary DataSet 1

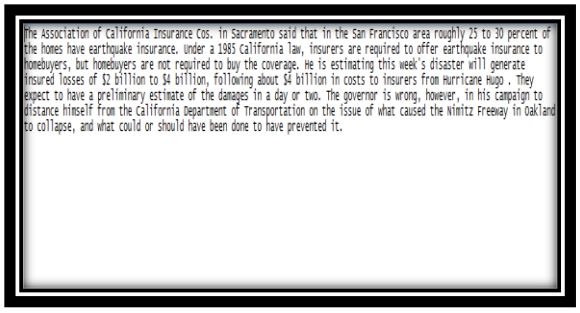
The Association of California Insurance Cos. in Sacramento said that in the San Francisco area roughly 25 to 30 percent of the homes have earthquake insurance. Under a 1985 California law, insurers are required to offer earthquake insurance to homebuyers, but homebuyers are not required to buy the coverage. He is estimating this week's disaster will generate insured losses of $2 billion to $4 billion, following about $4 billion in costs to insurers from Hurricane Hugo . They expect to have a preliminary estimate of the damages in a day or two. The governor is wrong, however, in his campaign to distance himself from the California Department of Transportation on the issue of what caused the Nimitz Freeway in Oakland to collapse, and what could or should have been done to have prevented it.

Fig 5. Standard Summary DataSet 1[9]

Result of Comparison

	Base Implementation	Proposed Implementation
Recall	.65	.7092
Time Taken	16	14

Table 1. Recall and Time Comparison For Dataset 1.

In the table 1., we have compare the efficiency of the both the base and the proposed algorithm on the basis of the recall and time taken. In the Dataset 1, the percentage match with the standard summary is .65 i.e. 65% similarity and that for the proposed is .70.92 i.e. 70.92%. And the time taken by base is 16 seconds to complete the process and proposed work complete that in the 14 seconds.

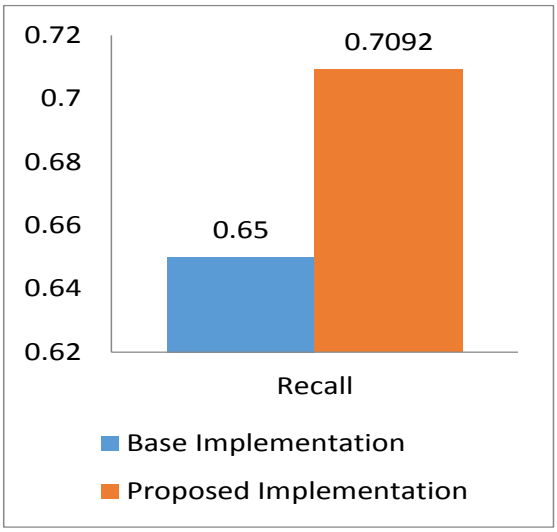

Fig 6. Graphical Comparison for Dataset 1 on Recall basis

VI. CONCLUSION

The document summarization problem is a very important problem due to its impact on the information retrieval methods as well as on the efficiency of the decision making processes, and particularly in the age of Big Data Analysis. Though good kind of text summarization techniques and algorithms are developed there's a requirement for developing new approaches to supply precise and reliable document summaries that may tolerate variations in document characteristics. This paper presents the innovative approach of summarizing the multiple documents which will let us to speed the process of the automatic text summarization concept.

REFERENCES

[1] SurajitKarmakar, Tanvi Lad, HitenChothani,"A Review Paper on Extractive Techniques of Text Summarization",International Research Journal of Computer Science (IRJCS),2015

[2] Kupiec, J., Pedersen, J., and Chen, F, "A trainable document summarizer. In Proceedings SIGIR", USA,1995.

[3] Lin, C.-Y. andHovy, E.,"Identifying topics by position", In Proceedings of the Fifth conference on Applied natural language processing, USA, 1997.

[4] Conroy, J. M. and O'leary, D. P. , "Text summarization via hidden markov models",In Proceedings of SIGIR ,USA, 2001.

[5] Osborne, M., "Using maximum entropy for sentence extraction" ,In Proceedings of the ACL Workshop on Automatic Summarization, May 2015

[6] Nenkova, A. , "Automatic text summarization of newswire: Lessons learned from the document understanding conference". In Proceedings of AAAI 2005,USA, 2005.

[7] Barzilay, R. and Elhadad, M. , "Using lexical chains for text summarization." ,In Proceedings ISTS,1997.

[8] https://github.com/albanie/text_summariser/find/master

[9] https://github.com/albanie/text_summariser/find/masterhttps://github.com/albanie/text_summariser/find/master

[10] http://NewsInEssence.com.

[11] McKeown, K. R. and Radev, D. R. , "Generating summaries of multiple news articles", In Proceedings of SIGIR ,1995.

[12] Radev, D. R. and McKeown, K., "Generating natural language summaries from multiple on-line sources." Computational Linguistics,2004

[13] Carbonell, J. and Goldstein, J., "The use of MMR, diversity-based reranking for reordering documents and producing summaries." ,In Proceedings of SIGIR ,1998

[14] Mani, I. and Bloedorn, E., "Multi-document summarization by graph search and matching." , In AAAI/IAAI, 1997

[15] Radev, D. R., Jing, H., Stys, M., and Tam, D. , "Centroid-based summarization of multiple documents." Information Processing and Management 40 ,2004.

[16] Evans, D. K., "Similarity-based multilingual multi-document summarization." Technical Report CUCS-014,2005.

[17] Luhn, H. P. , "The automatic creation of literature abstracts." , IBM Journal of Research Development,1958

[18] Shweta Saxena , AkashSaxena, PhD ,"An Efficient Method based on Lexical Chains for Automatic Text Summarization",International Journal of Computer Applications (0975 – 8887) Volume 144 – No.1, June 2016

[19] Xu Han, Tao Lv, Zhirui Hu, XinyanWang,and Cong Wang, "Text Summarization Using FrameNet-Based Semantic Graph Model",Scientific ProgrammingVolume 2016

[20] Mohsen Pourvali and Mohammad SanieeAbadeh "Automatic text summarization using lexical chains [microform] : algorithms and experiments.",IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 1, No 3, January 2012

[21] NimishaDheer Mr. Chetan Kumar ,"Extractive Automatic Text Summarization through Lexical Chain Method using WordNet Dictionary", IEEE 2016

[22] H. Gregory Silber Kathleen F. McCoy,"Efficient Text Summarization Using Lexical Chains",International Journal of Research in Engineering and Technology ,2013

[23] ReginaBarzilay and Michael Elhadad,Using Lexical Chains for Text Summarization „University of Israil , 2013

[24] Nikita Munot,Sharvari S. Govilkar ,Comparative Study of Text Summarization Methods,International Journal of Computer Applications (0975 – 8887) Volume 102– No.12, September 2014

[25] A.R.Kulkarni,S.S.Apte , "An Automatic Text Summarization Using Lexical Cohesion And Correlation Of Sentences ",Ijret: International Journal of Research in Engineering and Technology ,2014