International Journal of Engineering Research and Management (IJERM) – Volume-4, Issue-1, March 2018

Data Summarization Using Android

Jyoti A. Avhad[#1], Ankita B. Hase[#2], Shubham R. Rane[#3], Kajal T. Tupe[#4]

#UG Students & Department of Information Technology, & Savitribai Phule Pune University
Udoji Maratha Boarding Campus, Gangapur Road, Nashik, Maharashtra, India

[1] jyo.avhad@gmail.com
[2] ankitah24@gmail.com
[3] shubhrane4896@gmail.com
[4] kajaltupe0@gmail.com

Abstract: *Document summarization which will automatically merge similar information across the text documents, and generate abstractive summary. Apply pre-processing on documents, implementing pre-processing techniques on documents. It interpret and examine the source text and create a concise summary. Summarization is the process of decreasing large source document to shorten version of summary which will be easy to read. Document summarization is an emerging technique which is used for understanding the main purpose of any kind of documents. Summarization can be either single or multi document summarization. If summary is to be created for multiple relevant documents then it is called as multi document summarization. The purpose of this project is to interpret and examine the source text and creates a concise summary. A Graph based approach for Multi Document Summarization is a graph based multi document summarization technique in which, set of documents is pre-processed, undirected graph will be constructed to calculate similarity between sentences, the word class is attached to each sentence, sentences are ranked according to word class and similarity of sentences and top ranked sentences are included in the summary.*

Keywords: *Summarization, Abstractive, Extractive, Concise Summary.*

I. INTRODUCTION

Currently lots of information present on internet in different format. Hence it is difficult to the user to find relevant information according to his need. Also to get that appropriate information user needs to read the whole document. Hence ample of time will also be waste. So to handle such problem, data summarization is helpful to collect useful and appropriate information without wasting ample of time. Summary is defined as a text which is generated from more texts which include specific information. Data summarization is the process of containing useful information from source data to produce a shorter information for a particular user. Summarization is the process of decreasing a large volume of information into abstractive format by preserving only the most essential information. Due to rapid growth of the Internet within low-cost, large storing capacity of devices, we exposed to a lot of online information in our daily life. It makes difficult to find and collect exact information which we need.

Automatic text summarization is a key technology to solve such a type of difficulty, with the properly summarized information. We can easily and quickly understand the major points of the original document. Summary helps to understand and to get right information without checking the documents entirely. Therefore we need a summary of document so that we can get the main information of the whole documents. There are two techniques of summarization that are:

1. Extractive Summarization
2. Abstractive Summarization.

Extractive Summarization extracts important sentences from the input document and group them together to generate summary without changing the input text. Whereas Abstractive Summarization interpret and examine the text. Its aim to produce a generalized summary. Text summarization process works in three steps:

1. Analysis Step
2. Transformation Step
3. Synthesis Step

Analysis step analyse source text and select attributes. Transformation step transforms the result of analysis step and last step that is synthesis step which will represent the result of summary.

II. LITERATURE SURVEY

Many researchers have try to develop and summarize the text in different areas. Techniques used by these researchers are summarized below:

Narendra Andhale, L.A. Bewoor describes as the amount of information on the web is increasing rapidly day by day in different formats such as text, video, and images. It has become difficult for individual to find relevant information of his interest. Suppose user queries for information on the internet he may get thousands of result documents which may not necessarily relevant to his concern. To find appropriate information, a user needs to search through the entire documents this causes information overload problem which leads to wastage of time and efforts. To deal with this dilemma, automatic text summarization plays a vital role. Automatic summarization condenses a source document into meaningful content which reflects main thought in the document without altering information. Thus it helps user to grab the main notion within short time span. If the user gets effective summary it helps to understand document at a glance without checking it entirely, so time and efforts could be saved. Text summarization process works in three steps analysis, transformation and synthesis. Analysis step analyses source text and select attributes. Transformation step transforms the result of analysis and finally representation of summary is done in synthesis step. Text summarization approaches generally categorized into extractive summarization and abstractive summarization. Extractive summarization extracts important

sentences or phrases from the source documents and group them to generate summary without changing the source text. However, abstractive summarization consists of understanding the source text by using the linguistic method to interpret and examine the text. The abstractive summarization aims to produce a generalized summary, conveying information in a concise way [1].

Atif Khan, Naomie Salim, Yogan Jaya Kumar describes although fully abstractive summarization is a big challenge, our proposed semantic graph based approach shows the feasibility of this new direction for summarization research. Existing graph based approaches treat sentence as bag of words and cannot capture redundant sentences that are semantically equivalent as they mostly rely on content similarity measure. The proposed approach assumes semantic structure of sentence - predicate argument structure as graph node, and establish semantic relationships between PASs using Jiang semantic similarity measures. The semantic similarity measures assists in detecting redundancy by capturing semantically equivalent predicate argument structures. The proposed graph based approach incorporates PAS-to-PAS semantic similarity and PAS-to-document set relationship into the graph-based ranking algorithm, and experimental results demonstrate that modified ranking algorithm improves summarization results. The approach is promising enough to be applicable to any domain and does not require any intervention of human experts [2].

Daan Van Britsom, Antoon Bronselaer, Guy De Tre explains that they have managed to design and implement an automatic multi-document summarization algorithm with fairly decent results, but the work does not end here. Despite the fact that there are multi-document summarization systems out there using sentence extraction, such as the NEWSUM algorithm and the Multi-Gen algorithm used in the Columbia News-Blaster system, there are other possibilities. The next step is to use sentence compression as illustrated so we can provide the same information even briefer. The final step of summarizing is sentence generation, but this requires an entirely different approach [3].

Giuseppe Di Fabbrizio, Ahmet Aker, Robert Gaizauskas describes the addresses extractive summarization for reviews containing opinions on multiple aspects of the product or services being reviewed. We propose a method called STARLET. It uses aspects as features to score sentences in the input documents. The features are weighted linearly and summaries are generated using A* search. We trained the weights using MERT and use "best" reviews as gold standard summaries. We performed both automatic and manual evaluations in the restaurant reviews domain. In both evaluations the results show that STARLET summaries contain more review information than alternative baselines [4].

III. PROPOSED SYSTEM

Data Summarization is the process of fetching the important data from the huge amount of data. So firstly, document is given to the system as an input. For the given data, the input document is pre-processed and features like features extraction, features vectors, also sentence scoring are extracted from the source. In the Pre-processing some key functions are apply on document for summarization. Such functions are Punctuation Marks, Removal of Words and Top Sentence Ranking. In the first function of Pre-processing the punctuation marks are summarized using same punctuations. Punctuation Marks in documents also indicates importance of words, sentence as well as paragraph like hyphens Uses in adjective or sentence connectivity, brackets, Quotations (), Question mark (?), exclamation mark (!) etc. For Quotations (), Question mark (?) and exclamation mark (!) we have assigned more score for considering in final summary. The next function of Pre-processing is Removal of Words. This function is used for removed the same words in document. After this two functions Top Sentence Ranking is apply. The system assigns the rank as per the score assigned to the sentences in document and then sorting of sentence is done in top sentence ranking.

The significance of learner-dependent features in features extraction is used for sentence extraction. In this Sentence Position (SP), Title Similarity (TS), Centrality (Cen), Term Frequency (TF), Positive or Negative Keywords are used for extracting the important sentences and text dependent features like Sentence Length, Trigger word, Noun Occurrences is used for extracting important data which is easy to understand.

The feature vector is used to calculate the sentence scoring. A sentence scoring method is totally based on the concepts of stop-word removal, Semantic and Statistical relationship.

In existing system for the summarization process some tools, websites are available. So if we want to summarize some large amount of data then we need to find tools or the websites.

But this is not enough to summarize the data efficiently. But also if we try to summarize the data we didn't get the expected output. The main drawback of the existing system is that it is not that much user friendly which we need.

So to overcome this problem, we introduce new summarization technique that is "Data Summarization using Android". In this we are developing an application which will help us to summarize the large amount of data in just a few minutes or seconds. As it is an android application everyone can access it. Also we get the proper and expected output.

A. Advantages of Proposed System:

- Time and efforts get reduce.
- More user friendly.
- Effective GUI.
- Appropriate output.

B. Outcomes of proposed System:

- As we tried for the summarization of IEEE paper it is massively useful for the people who are doing there PHD course.
- Also for Under Graduates, Post Graduates students can also going to use this application for better and expected output.

- So mainly students of UG, PG and PHD level are the important outcomes of this newly introduced proposed system.

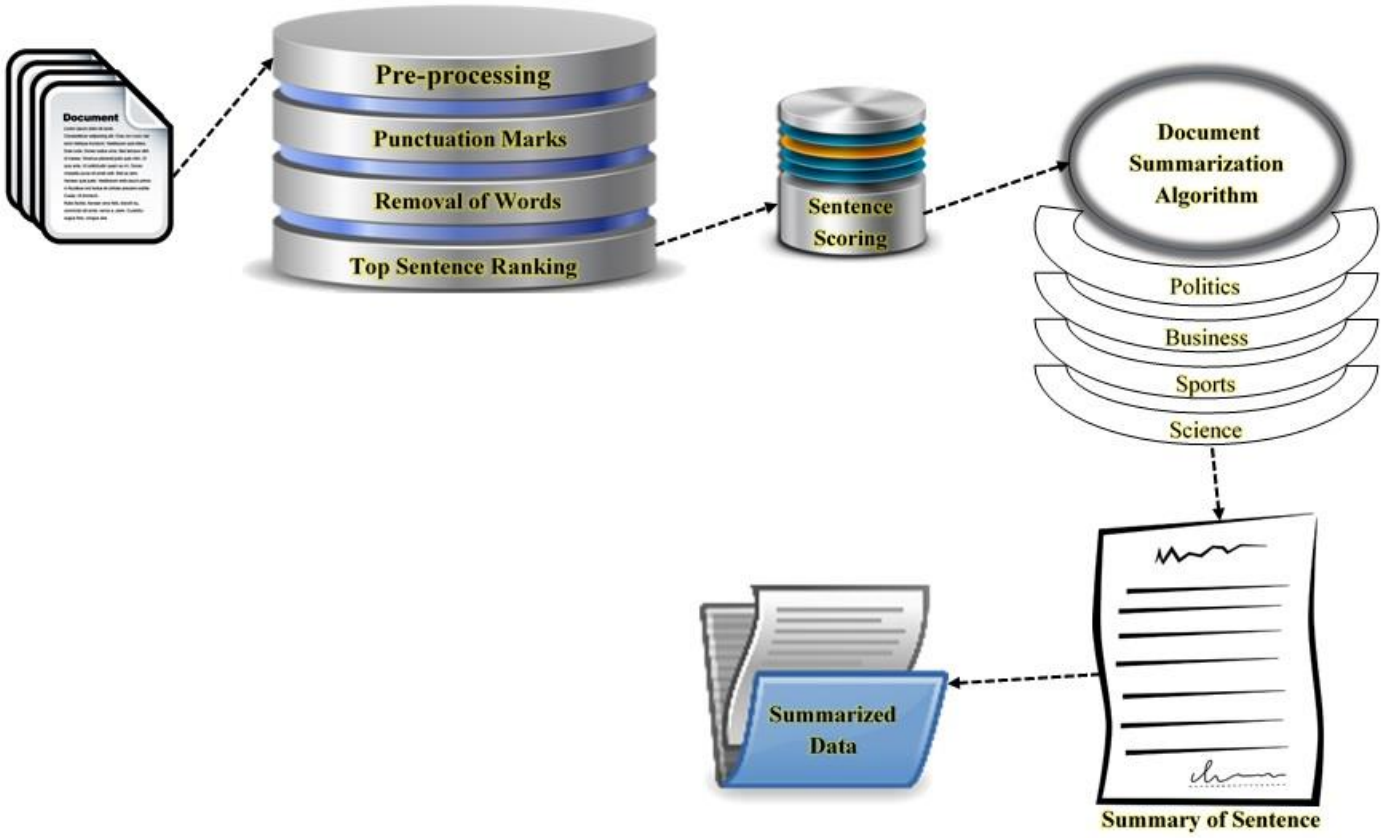

Fig. 1: Summarization Process

C. Document Summarization Algorithm:

This is the main process of summarization. On the basis of this summarization algorithm the important data is fetched. For the collection of important data various terms are used like politics, b usiness, sports, science, etc. According to this terms of the document which is given as an input contains information related to politics term then all the important information related to politics is taken so that we get the specific but important data. Same as this terms other terms are used and the data is get summarized.

D. Summary of Sentence:

After applying the summarization algorithm the summarized data is collected into the proper formation i.e. in summary of sentence. Summary of sentence stores the summarized data into a document. And finally we get the proper necessary data in document format.

IV. ALGORITHM

A. Functional Components of System:

Following is a list of the functional components of the tool:

1) ***Data pre-processing:*** This will work on the documents and convert them to plain text for processing by the rest of the system.
2) ***Sentence separator:*** In this components checks the documented data and separates the sentences based on some rules (like a sentence ending is determined by a dot and a space etc).
3) ***Word separator:*** This separates the words based on the process like space denotes the end of the word.
4) ***Stop-words eliminator:*** This eliminates the regular English words like 'a, an, the, of, from' etc for next processing.
5) ***Duplicate remover:*** This ensures the uniqueness in the list of keywords. It removes all duplicates and gives single copy of the words.
6) ***Word-frequency calculator*:** This component calculates the number of times a word used in the document. This calculator calculates the frequency of occurrence of various keywords in each sentence.
7) ***Scoring algorithm:*** This algorithm calculates the score of each sentence. The score can be made to be proportional to the sum of frequencies of the different words.

8) ***Ranking:*** The sentences will be ranked according to the score of their frequency.
9) ***Summarizer:*** Based on the user input of the size of the summary required, the sentences will be picked from the ranked list and concatenated. The resulting summary will be displayed on screen.

B. Data Summarization Algorithm:

Steps followed in the process of summarization are:

1. Data is given as input to the pre-processor.
2. Stop words and replicas are removed from the data.
3. The full document is then extracted.
4. Stop words are removed from the document data.
5. All the sentences of the input data are separated using sentence separator.
6. The frequency of each word present in the data is evaluated for each sentence using the frequency evaluator.
7. Each data is now checked for its relevancy.
8. Only relevant data are considered for further processing and rest are discarded.
9. Relevant data are stemmed duplicates and stop words are removed.
10. The total frequency (f2) of all keywords of data is evaluated for each sentence present in body.
11. Overall frequency (F) for each sentence is calculated and for this frequency measure is used to rank the sentences.
12. The number of sentences for summary is input by the user.
13. The top ranked sentences are collected in document.
14. Finally output will be displayed.

V. CONCLUSION

A data summarization using android gives the idea to focus only on specific approach (like some terms of algorithm) to improve and generate better summary in less effort and construct new procedure for next generation. In summarization algorithm, document summarization techniques sentences are pre-processed, class is attached to each sentence, sentence length is calculated, and each sentence is given rank based on class and then top ranked sentences has selected in summary, therefore its more efficient than other technique.

REFERENCES

[1] Narendra Andhale and I. Goldberg, "*An overview of Text summarization,*" Springer, 2015, pp. 158–172.

[2] Atif Khan, Naomie Salim and Yogan Jaya Kumar, "*Genetic Semantic Graph Approach for Multi-document Abstractive Summarization,*" Digital Information Processing and Communications (ICDIPC), 2015 Fifth International Conference on, 7-9 Oct. 2015.

[3] Daan Van Britsom, Antoon Bronselaer and Guy De Tre, "*Automatically Generating Multi-Document Summarizations,*" Intelligent Systems Design and Applications (ISDA), 2011 11th International Conference on, 22-24 Nov. 2011.

[4] Giuseppe Di Fabbrizio, Ahmet Aker, Robert Gaizauskas, "*STARLET: Multi-document Summarization of Service and Product Reviews with Balanced Rating Distributions,*" Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on, 11-11 Dec. 2011

[5] M. Srinivas and L. M. Patnaik, "*Genetic algorithms: A survey,*" Computer, vol. 27, pp. 17- 26, 2016.

[6] G. Carenini, R. Ng, and A. Pauls, "*Multi-document summarization of evaluative text,*" in 11th Meeting of the European Chapter of the Association for Computational Linguistics (EACL 2006), 2014.

[7] Rafeal Ferreira, D. W.-l. Cheung, B. Kao, and N. Mamoulis, "*Context based summarization,*" in SIGMOD. ACM, 2013.

[8] K. Sankar and L. Sobha, "*An approach to text summarization,*" in Proceedings of Third International Cross Lingual Information Access Workshop- CLIAWS3, pp. 53–60, Association for Computational Linguistics, Jun 2012.

[9] R. Mihalcea, "*Graph-based ranking algorithms for sentence extraction, applied to text summarization,*" in Proceedings of the ACL 2012on Interactive poster and demonstration sessions ACLdemo 04, no. 20, pp. 20–es, Association for Computational Linguistics, 2012.

[10] V. K. Gupta and T. J. Siddiqui, "*Multi-document summarization using sentence clustering,*" in 4th International Conference on Intelligent Human Computer Interaction, pp. 1–5, IEEE, Dec 2012.

[11] D. Das and A. F. T. Martins, "*A Survey on Automatic Text Summarization,*" Carnegie Mellon University, pp. 131, 2011.

[12] Dragomir R. Radev, Eduard Hovy and Kathleen McKeown, "*Introduction to the special issue on summarization, Association for Computational Linguistics,*" Volume 28, Number 4, 2002.

[13] Ani Nenkova and Kathleen McKeown, "*Automatic summarization,*" Foundations and Trends in Information Retrieval, Vol. 5, Nos. 2–3, pp.103–233, 2011.

[14] R. Ferreira et al., "*Assessing sentence scoring techniques for extractive text summarization,*" Expert Systems with Applications, vol. 40, No. 14, pp.5755 – 5764, 2013.

[15] M.A. Fattah and Fuji Ren, "*Automatic text summarization,*" International Journal of Computer Science, Volume 3, Number 1, 2009.

[16] Mahak Gambhir, Vishal Gupta, "*Recent automatic text summarization techniques: a survey,*" Artif Intell Rev DOI 10.1007/s10462-016-9475- 9, © Springer science + Business Media Dordrecht 2016.