# Review on Summarization using Mapreduce Framework

**Akshay Lahu Sawant[1] Prof. K.T. Belerao[2]**
[1, 2] Department of Computer Engineering
[1, 2] Trinity College of Engineering and Research, Pune

***Abstract***- *Summarization refers to the shorter version of the source text keeping information content and overall meaning of the document as it is. Manual summarization of any document requires huge human efforts and it is difficult. Collecting required information from a collection of document may not be a feasible for any system. In multi-document summarization the major challenge is that document set contains different information which is relevant or non-relevant from the topic and there-fore to generate effective abstract summary needs analysis of information. Automatic text summarization resolves this problem by providing summary of the document set. MapReduce is a programming model which handles massive amounts of data. MapReduce provides distributed approach which can help in generating faster summary result. In proposed approach DBSCAN is used for clustering which partitions data into various cluster and Hidden Markov model is used for summarization.*

***Keywords***- Big Data Analysis, Clustering, MapReduce, Text Summarization

## I. INTRODUCTION

Internet contains information which has tons of documents with useful and not useful information. Summarization techniques are used by search engines like Google, yahoo, etc. to provide brief snippet of the documents retrieved by the user. These query based summaries can be more focused, as the users query terms are known to the retrieval system and it can be used to target summaries for the query. Document contains information which may or may not be in uniform format, this means that some information is more important than others. The challenge in summarization is to pick up the important information from the document than less ones. Shorter and preservable summary provides valuable information. Summarizing multiple documents requires intensive text processing and computations. Machine learning approaches need to be applied in more distilled way for generating abstract summaries. Information overload problems can be resolved using automatic document summarization.

Single and Multi-Document Summarization: Single doc-ument and multi document summarization are two main categories of summarization. Single document summarization built summary of single document only while in multi document summarization multi documents summary is generated from all the documents. Multiple documents summarization generates summary from multiple documents where single document refers single ones. Mutlidocument summarization is an extension to the single document summarization for the same topic. It differs with single document summarization on degree of redundancy, temporal dimension and compression ratio. Different strategies have been used to produce multi-document summary and usually these methods are abstractive or extractive.

Abstract and Extract Summarization: Abstract summary contains most important words from the document set. Sen-tences are constructed using paraphrasing keeping meaning same. Sentences are reformulated and compressed in abstract summaries with information fusion is implemented. Ranking of sentences, paragraphs of the document are important in case of extractive summary. Term frequency that is tf*idf calculates the frequency of word occurrences. High frequency terms are selected and extracted from the document set. In Extraction subset of sentences are picked up and used for summary.

MapReduce Model for summarization: As summarization of big data is a complex task but with help of distributed programming approach it can be resolved. MapReduce is new framework which handles large data using distributed com-putting on large-scale clusters. HDFS a distributed file system stores the input data, MapReduce applies a divide and conquer technique to divide the input data sets into small data sets, and then processed on different machines, which has achieved parallelism. MapReduce follows three step workflow which is Map, Shuffle and Reduce. Data in MapReduce framework is seen as a series of key-value pairs. User uses map and reduce functions. In map phase, each node has assigned a map task in a cluster and multiple map tasks are running in parallel at the same time in cluster. A key-value pair is given to map call and it produces list of key-value pairs. The output is transferred to the reduce node which is shuffle phase. All the intermediate records with the same key are sent to the same reducer node. At reduce nodes all the received records are sorted and grouped. In a single call each group is processed.

## II. LITERATURE REVIEW

Automatic text summarization based on sentences clustering and extraction [1], proposed a sentence similarity computing method based on the three features of the sentences, on the base of analyzing of the word form feature, the word order feature and the semantic feature, using the weight to describe the contribution of each feature of the sentence, describes the sentence similarity more preciously. It uses K-means clustering for sentence clustering which clusters sentences based on semantic similarity between them.

In Improving Performance of Text Summarization [2] it is proposed that, the importance of text summarization for saving time is more. Extractive summarization technique extracts terms from the document and these terms are weighted. Sentences are weighted as per importance in the document. Fuzzy logic technique is used for extractive summarization. Feature extraction manipulates the extracted data from document like title, sentence length, similar words, and sentence position. Fuzzy logic scoring is used as summarization subtask. The generated summary using this has higher quality and better results than online summarizer tools.

In WordNet-based Document Summarization [3] it is proposed that, extractive summarization technique in which most relevant sentences from original document are extracted. The technique is proposed on the basis of key sentences from document using statistical methods and using WordNet library. The two approaches are combined to get the candidate sentences from document. It follows the step where semantic similarity of sentences is conducted and redundancy is reduced. Refining sentences for similarity measures improves the accuracy of summarization and correct summary can be generated.

In A Comparative study of Clustering Algorithms using MapReduce in Hadoop [4] it is proposed that, three clustering algorithms are compared, K-means, canopy clustering and Fuzzy k-mean. These are implemented using Mapreduce and Sequential approach. These algorithms works with data in a portioned form and also need to consider distributed nature of data. Result and Conclusion shows that the Clustering algorithm should scale according to increasing amount of data. The choice of clustering algorithm is based on the type of data and purpose of application.

In Multidocument extraction based Summarization [5], three different techniques are proposed for extraction based summary generation and which also includes a graph based formulation to improve former methods. In the first method it uses sentence importance scoring based on semantic similarity score to select the sentences. The selected sentences would be most informative and representative. Stack-decoder algorithm is used for building summaries closer to the optimal. The second approach creates a cluster of similar sentences based on above step for summary generation. The third approach explained in this paper is novel graph problem based where summaries are generated based on cliques found in the constructed graph.DUC 2004 dataset is used for this summarization. ROUGE score is comparable for other techniques.

In Research of parallel DB-SCAN clustering algorithm based on MapReduce [6], it is proposed that, As DBSCAN algorithm lacks in dealing with large datasets MapReduce programming model is used for clustering of DBSCAN. Data analysis is completed in map functions and clustering rules are in different data objects; Reduce functions merge the clustering rules to get the final result. On cloud computing platform MapReduce with DBSCAN provide better performance. Results from this paper shows that DBSCAN algorithm based on MapReduce with large data sets have good timeliness.

In Summarizing Speech without Text Using Hidden Markov Models [7], the method proposed for summarizing speech documents is using Hidden Markov Model framework. Hidden variables in the model represent whether sentence needs to be included in summary or not. Sequences of segments are predicted by model and that best summarize the document. It does not use any lexical information for summary creation. The result shows that speech can be summarized well using acoustic/prosodic features without lexical features.

In Abstractive Multi-Document Summarization via Phrase Selection and Merging [8], an abstraction-based multi document summarization framework is proposed which can construct new sentences with noun/verb phrases. It first constructs a pool of concepts and facts represented by phrases from the input document. New sentences are generated by merging phrases and keeps sentence construction as per the constraints. Integer linear optimization is employed for phrase selection and merging to achieve global optimal solution for summary. Results shows that using TAC 2011 data set the proposed approach outperforms and generates good summary result.

In ENRICH FRAMEWORK FORMULTI-DOCUMENT SUMMARIZATION USING TEXT FEATURESAND FUZZY [9], proposed a Collection of documents increases rapidly and finding the important information becomes complex task. Multiple document summarizations are the approach to get the assured summary of document set. The methods which are already proposed for summarization uses two or three features of text to find the importance of combined sentences. In this paper, multiple extracted features of the text are considered by applying NLP protocol. Extracted features are classified on the basis of fuzzy logic to get the best document summary. Key features are preprocessing, feature scoring and fuzzy logic with inference engine.

## III. PROPOSED SYSTEM ARCHITECTURE

In the proposed approach, multiple documents are summarized according to the given topic information. The set of documents are in big dataset is available for searching information. User input topic is matched with the multiple documents set and data is passed as filtered data. The preprocessing is invoked after filtering of data in which sentence segmentation, tokenization is performed. Preprocessed sentences are divided into a cluster which is done on the basis of semantic similarity. We proposed DBSCAN algorithm for clustering which works on density component. WordNet is powerful open source java based library which provides synonym for words. Redundant sentences are removed for getting effective summary. Clustered results are then modeled using Hidden Markov model and summarization process is completed. Finally by applying custom grammar rules to the words meaningful sentences are built.

### Preprocessing:

Initially, the relevant documents to the topic are parsed to select all sentences. These documents are considered as a filtered data and passed for preprocessing. Here it follows sequence of steps.

From Mapreduce framework the filtered data which is nothing but relevant to topic. Multiple documents are split and globally passed to each mapper function. Mapper has the responsibility to create a key-value pair.

### Stop Words Removal:

These are the words from sentences which do not interpret any meaning. They are the connector of words which joins words and create a meaningful sentence. Stop words occurs frequently in the database records, articles and web pages which are insignificant. When document summary is generated with extraction method these are eliminated from the process of weighting and ranking which makes summary effective. Stop words are generally single character, common-two character and three-character words also they repeat frequently in sentences.

### Stemming:

Stemming is applied in preprocessing step which converts words to their respective root form. For example, “stemming” is converted to “stem”. Stemmer is a program which generates a morphological root of the word. Common stem terms are generally having same meanings.

### Clustering using DBSCAN:

Clustering helps to create separate blocks of data which essentially reduces the problem of information overhead. It also enables the faster access of data as it is divided.DBSCAN is the density based algorithm which can generate any number of clusters and also for the distribution of spatial data. DBSCAN is advantageous as it does not require number of classes to be formed in advance.

The preprocessed data is having different words which are needs to be sorted according ti their similarity. In the proposed word we are going to use WordNet Open Source Java library for measuring similarity of the words and removing redundant words. The map functions works on the words and builds key-value pair and perform DBSCAN algorithm to initial cluster. A combiner function identifies the same key value pairs and passes to reduce process. Reduce process does hierarchical merging of the initial clusters generated by map and passes to summarization step.

### Summarization using Hidden Markov Model:

Clustered data is sorted and fully contains unique words from sentences which are passed from Reduce function in clustering. Finally Hidden Markov model is applied for getting abstract summary of the documents.

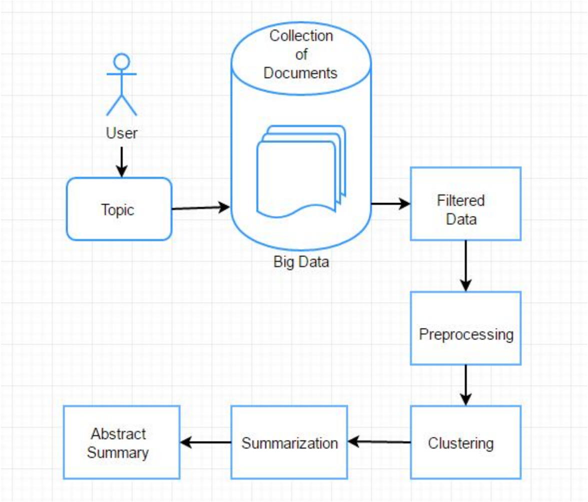

Figure 1. Block Diagram of Proposed System Architecture

## III. CONCLUSION

As Summarization of multiple documents require a proper analysis of data and correct selection of words from documents so in this paper, a systematic approach is proposed for summarization using Mapreduce framework. DBSCAN for clustering is used which clusters the similar words in block of clusters. Hidden Markov model is used for final summary generation which helps in effective topic modelling. Using Mapreduce framework summarization will system will provide faster and effective summary.Open source WordNet API is used for getting semantic similarity of words. Processing on distributed data is well handled by Mapreduce programming which results faster summary generation of multiple documents.

## IV. ACKNOWLEDGMENT

I would like to give deep gratitude to all those who have supported me to prepare this paper. I am highly grateful to Prof. K T Belerao for his guidance and consistent supervision. Very special thanks goes to the Management, Principal, Head of department Dr.S.B.Chaudhari, Faculty of Trinity College of Engineering and Research, Pune, for all their support and kind co-operation for help in completion of this work.

## REFERENCES

[1] Pei-ying Zhang, Cun-he Li, Automatic text summarization based on sentences clustering and extraction, 2009 2nd IEEE International Conference on Computer Science and Information Technology, Beijing, 2009, pp.167-170.

[2] S.A. Babar, Pallavi D. Patil , Improving Performance of Text Summarization, Proceedings of the International Conference on Information and Communication Technologies, ICICT 2014, 3-5 December 2014 at Bolgatty Palace and Island Resort, Kochi, India.

[3] CHENGHUA DANG, XINJUN LUO, WordNet-based Document Summarization, 7th WSEAS Int. Conf. On APPLIED COMPUTER APPLIEDCOMPUTATIONAL SCIENCE (ACACOS '08), Hangzhou, China, April6-8, 2008.

[4] Dweepna Garg, Khushboo Trivedi, B.B.Panchal,A Comparative study of Clustering Algorithms using MapReduce in Hadoop, International Journal of Engineering Research Technology (IJERT) Vol. 2 Issues 10, October- 2013.

[5] Sandeep Sripada, Venu Gopal Kasturi, Gautam Kumar Parai, Multidocument extraction based Summarization, http://nlp.stanford.edu/ner/index.shtml.

[6] Xiufen Fu, Shanshan Hu and Yaguang Wang, Research of parallel DB-SCAN clustering algorithm based on MapReduce, International Journal of Database Theory and Application Vol.7, No.3 (2014), pp.41-48.

[7] Sameer Maskey, Julia Hirschberg, Summarizing Speech Without Text Using Hidden Markov Models, Proceeding NAACL-Short '06 Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers Pages 89-92.

[8] Lidong Bing, Piji Li, Yi Liao, Wai Lam, Weiwei Guo, Rebecca J. Passon-neau, Abstractive Multi-Document Summarization via Phrase Selection and Merging, Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, pages 15871597, Beijing, China, July26-31, 2015.

[9] SACHIN PATIL, RAHUL JOSHI, ENRICH FRAMEWORK FORMULTI-DOCUMENT SUMMARIZATION USING TEXT FEATURESAND FUZZY, Journal of Theoretical Applied Information Technology .6/30/2016, Vol. 88 Issue 3, p431-437. 7p.

[10] N.L.Drotz, F.G.P.Zetterberg,Wi-Fi Fingerprint Indoor Positioning Sys-tem using Probability Distribution Comparison, IEEE International conference on Acoustics,

Speech and Signal Processing, Kyoto, Japan, March2012:2301-2304.

[11] imisha Dheer, Chetan Kumar, Automatic Text Summarization: A Detailed Study, International Journal of Science and Research (IJSR) Index Copernicus Value (2013): 6.14.

[12] John Conroy, Dianne P. O'Leary, Text summarization via hidden markov models and pivoted QR matrix decomposition, Center for Computing Sciences Institute for Defense Analyses 17100 Science Drive Bowie, MD 20715.

[13] Arnold Overwijk, Generating Snippets for undirected information search, 9thTwente Student Conference on IT, Enschede, June 23th, 2008.

[14] Anjali R. Deshpande, Lobo L. M. R. J, Text Summarization using Clustering Technique, International Journal of Engineering Trends and Technology (IJETT) - Volume4 Issue8- August 2013.

[15] Arina Esmaeilpour1, Elnaz Bigdel, Fatemeh Cheraghchi, Bijan Raahemi, and Behrouz H. Far, Distributed Gaussian Mixture Model Summarization Using the MapReduce Framework, 29th Canadian Conference on Artificial Intelligence, Canadian AI 2016, Victoria, BC, Canada, May 31 -June 3, 2016.

[16] Yaminee S. Patil, M. B. Vaidya, K-means Clustering with MapReduce Technique, International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 11, November 2015.

[17] P.Sukumar, K.S.Gayathri, Semantic based Sentence Ordering Approach for Multi-Document Summarization, International Journal of Recent Technology and Engineering (IJRTE), Volume-3, Issue-2, May 2014.

[18] Daniel M. Dunlavy, John Conroy, Dianne P. OLeary, QCS: A system for querying, clustering and summarizing documents, Proceedings of HLTNAACL 2003, Demonstrations, pp. 11-12, Edmonton, May-June 2003.