© 2019 IJSRST | Volume 6 | Issue 1 | Print ISSN: 2395-6011 | Online ISSN: 2395-602X
Themed Section: Science and Technology
DOI : 10.32628/IJSRST196192

# A Survey on Various Techniques for Multi-Document Summarization

**Apurva Sawwalakhe, Nikita Wanjari, Shreya Paliwal, Shubhangi Katare, Vidhya Malve**
BE, Department of Computer Science and Engineering, Shrimati Rajshree Mulak College of Engineering, Nagpur, Maharashtra, India

## ABSTRACT

Natural language processing gives Text Summarization which is the most well-known application for data pressure. Content rundown is a procedure of creating a synopsis by decreasing the span of unique report and relating critical data of unique record. There is emerging a need to give top notch synopsis in less time on the grounds that in present time, the development of information increments massively on World Wide Web or on client's desktops so Multi-Document outline is the best apparatus for making rundown in less time. This paper introduces a study of existing techniques with the curiosities highlighting the need of astute Multi-Document summarizer.

**Keywords :** Multi-Document Summarization; Clustering Based; Extractive and Abstractive approach; Ranked Based; LDA Based; Natural Language Processing

## I. INTRODUCTION

Natural language processing (NLP) is a field of software engineering, manmade brainpower and machine learning with the associations amongst PCs and human dialect. The utilization of World Wide Web and many sources like Google, Yahoo! surfing likewise increments because of this the issue of over-burdening data additionally increments. There is immense measure of information accessible in organized and unstructured shape and it is hard to peruse all information or data. It is a need to get data inside less time. Subsequently we require a framework that consequently recovers and condense the archives according to client require in time confine. Report Summarizer is one of the practical answers for this issue. Summarizer is an apparatus which serves a helpful and proficient method for getting data. Summarizer is a procedure to separate the vital substance from the archives. When all is said in done, the outlines are characterized in two ways. They are Single Document Summarization and Multiple Document Summarization. The synopsis which is extricated and made from single archive is called as Single Document Summarization though Multiple Document Summarization is a programmed procedure for the extraction and making of data from different content reports.

The fundamental point of outline is to make synopsis which gives least excess, most extreme significance and coreferent question of same subject of rundown. In straightforward words, outline ought to cover all the significant parts of unique report without immateriality while keeping up relationship between the sentences of synopsis. Along these lines, Extractive rundown and Abstractive outline approach is utilized. Extractive rundown works by selecting existing words, expressions or number of sentences from the first content to shape outline. It picks the most important sentences or watchwords from the reports while it additionally keeps up the low repetition in the synopsis. Abstractive outline

technique which produces a rundown that is nearer to what a human may make. Essentially this sort of synopsis may contain words not expressly introduce in the first record arrange. It gives reflection of unique record shape in less words. This review covers Cluster Based approach,

LDA Based approach and Ranking Based approach. The fundamental point of Multi-archive synopsis has been likewise explained. The rest of the paper is introduced as takes after. Segment II depicts related work in the field of multi archive synopsis utilizing Cluster Based approach, LDA Based approach and Ranking Based approach, Section III presents last conclusion.

## II. RELATED WORK

Multi-Document Summarization is a programmed method intended to separate and make the data from various content archives about a similar theme. The multi-record outline is an exceptionally complex assignment to make a synopsis. It is a system where one synopsis should be converged from many archives. There are number of issues in multi record rundown that are not the same as single archive outline. It requires higher pressure. The present execution incorporates advancement of an extractive and abstractive technique. A 10% outline might be adequate for one report however in the event that we require it for various records then it is hard to get a synopsis from link prepare. In most if the examination, the analyst takes a shot at section extraction or sentence extraction in light of the fact that the gathering of watchwords contains a low measure of data though passage or sentences can cover the specific idea of record. There are loads of techniques which speak to multi-record outline, however in this paper we principally concentrate on Cluster based, LDA based approach and Ranking based approach of multi-report rundown.

### 2.1 Cluster Based Approach

Center of Cluster Based technique gives bunching calculation which is more viable and it relies on upon centroid of the group. Grouping strategy fundamentally includes just three undertaking as pre-handling, bunching and synopsis era. The accompanying strategy must be done before giving contribution to the grouping technique by utilizing pre-preparing. Essentially, pre-handling steps separated into taking after focuses.

Tokenization: It breaks the content into isolated lexical words that are isolated by white space, comma, dash, spot and so on [3] Stop words expulsion: Stop words like an, about, all, and so on., or other area subordinate words that must be removed.[3] Stemming: It evacuates additions like "s", "ing" thus on from documents.[3]

After Pre-handling, bunching strategy is connected to produce the synopsis. A paper on information converging by Van Britsom et al. (2013) [1] proposed a strategy in view of utilization of NEWSUM Algorithm. It is a kind of grouping calculation where isolates an arrangement of record into subsets and afterward produces a rundown of coreferent writings. It contains three stages: theme distinguishing proof, change and synopsis by utilizing distinctive bunches. Outline utilizes sentence extraction and sentence deliberation. It is part the sources by their timestamps. It is partitioned into two sets as late articles and non-late articles. It depends on score of sentence means if data is more precise then it is included rundown. It speaks to higher outcome for vast rundown yet broad information blending issue emerges when boundless information is accessible to consolidate.

This paper is on multi-archive rundown utilizing sentence bunching by Virendra Kumar Gupta et al. (2012) [3] states that sentences from single record rundowns are bunched and top most sentences from

every group are utilized for making multi-report synopsis. The model contains the means as pre-preparing, commotion expulsion, tokenization, stop words, stemming, sentence part and highlight extraction. Include extraction includes taking after strides as-

***Precision:*** It is defined as the fraction of retrieved docs that are relevant given as

Relevant = P (relevant | retrieved) [9] Pn = m/N-n+1

***Recall:*** Fraction of relevant docs that are retrieved given as Retrieved = P (retrieved | relevant) [9]
Rn= m/n

***TFIDF:*** Formulae [9]

$$\text{TF (term, document)} = \frac{\text{Frequency of term}}{\text{No of Document}}$$

$$\text{Term Frequency} = \frac{n_j}{\Sigma_k n_k}$$

***IDF*** *(inverse document frequency):* It calculates whether the word is rare or common in all documents. IDF (term, document) is obtained by dividing total number of Documents by the number of documents containing that term and taking log of that.

IDF (term, document) = $\log \frac{\text{Total No of Document}}{\text{No of Doc containing term}}$

***TF-IDF:*** It is the multiple of the value of TF and IDF for a particular word. The value of TF-IDF increases with the number of occurrences within a doc and with rarity of the term across the corpus.

*TF-IDF=TF*IDF*

In the wake of playing out these means, critical sentences are removed from every bunch. What's more, for this, there is two sorts of sentence bunching utilized as syntactic likeness and semantic comparability. English National Corpus is utilized for computing the recurrence of words. It contains 100 million words. It gives best performing framework result on DUC 2002 dataset yet it is not chipped away at DUC 2005 or DUC 2006 dataset.

A paper on Extracting Summary from Documents Using K-Mean Clustering Algorithm by Manjula K. S. et al. (2013) [7] proposed K-MEAN calculation and MMR (Maximal Marginal Relevance) strategy which are utilized for question subordinate grouping of hubs in content report and discovering inquiry subordinate outline, relies on upon the record sentences and tries to apply confinement on the archive sentence to get the pertinence imperative sentence score by MMR known as nonexclusive rundown approach. Synopsis of report can be found by k-mean calculation. This technique used to prepare the dataset by utilizing a few groups and finds earlier in the datasets. These discovers similitude of every archive and make the synopsis of the report. In this work, n-gram which is subtype of co-event connection is utilized. These procedures the informational collection through certain number of groups and locate the earlier in the informational indexes however MMR relies on upon the report sentences, and tries to apply confinement on the archive sentence.

This paper is on Context Sensitive Text Summarization Using K Means Clustering Algorithm by Harshal J. Jain et al. (2012) [12] speaks to K-MEAN calculation. K-mean bunching is utilized to gathering all the comparative arrangement of records together and partition the report into k-group where to discover k centroids for every bunch. These centroids are not orchestrated appropriately so it gives diverse outcome. Along these lines, we put it legitimately to gather the closest centroid. Hence we rehash this progression until the finish of collection to the whole report. After this we need to re-compute k new centroid by considering the focal point of past stride bunches. These k new centroids create the new informational index purpose of closest new centroid. Here circle is produced and k-centroids change their place well-ordered until any

progressions are happened. It discovers question subordinate rundown. Adequacy and time utilization is the principle issues in this approach.

This paper is on Word Sequence Models for Single Text Summarization by Rene Arnulfo Garcia-Hernandez et al. (2009) [13] proposed the Extractive synopsis procedure which gives an outline to the client for comparable content archives. In this paper, here likewise utilizes the n-gram (non-linguistic) which comprises of arrangement of n words inside a specific separation in the content and continuously show up in the content. N-gram is utilized as parts of a vector space demonstrate in deciding the extractive content rundown. At the point when succession of a few words is utilized then their probabilities are evaluated from a CORPUS which comprises of set of archives. At the last, the probabilities are consolidated to get from the earlier likelihood of most plausible understanding. In this work, n-gram is utilized as a component of a sentence in an unsupervised learning technique. This strategy is utilized for bunching the comparative sentences and structures the groups where most illustrative sentences are decided for producing the outline. The calculation characterized as takes after-

- Pre-Processing: First, kill stop words, evacuate clamor and afterward apply stemming process on it.
- Term choice must be taken what size of n-grams as highlight is to be utilized to speak to the sentences. The recurrence edge was 2 for MFS demonstrate.
- Term weighting-choice must be taken that how every components are ascertained.
- Sentence grouping choose the contribution for the k-mean calculation.
- Sentence choice after completing k-mean calculation; pick the closest sentence to every centroid for producing the outline. It gives a rundown to the client for comparative content reports. It is important to discover from the earlier method for deciding the best gram estimate for content synopsis what is not clear how to do.

### 2.2 Ranking Based Approach

Ranking Based Approach ordinarily gives the higher positioned sentences into the rundown. Positioning calculations removes the rank sentences and unions the every single rank sentence and create the outline. Fundamentally, it applies positioning calculation; removes rank sentences and create a rundown.

This paper on SRRank: Leveraging Semantic Roles for Extractive Multi-Document Summarization by Su Yan and Xiaojun Wan (2014) [19] clarify a methodology that it positions sentences by utilizing SR-Rank calculation on Extractive content synopsis. SR-Rank calculation is a sort of diagram based calculation. Firstly, appoint the sentences and get the semantic parts, and afterward apply a novel SR-Rank calculation. SR-Rank calculation at the same time positions the sentences and semantic parts; it removes the most essential sentences from an archive. A diagram based SR-Rank calculation rank all sentences hubs with the assistance of different sorts of hubs in the heterogeneous chart. Here three sorts of charts are clarified as diagram group, chart output and fundamental chart. So in this paper, three sorts of charts are created as SR-Rank, SR-Rank-traverse and SR-Rank-group. Exploratory outcomes are given on two DUC datasets which demonstrates that SR-Rank calculation outperforms couple of baselines and semantic part data is approved which is exceptionally useful for multi-archive synopsis.

Another paper Document Summarization Method in light of Heterogeneous Graph by Yang Wei (2012) [20] clarifies the Ranking calculation that applies on heterogeneous chart. Existing procedure for the most part uses measurable and etymological data to extricate the most critical sentences from numerous reports where they can't give the relationship between various granularities (i.e., word, sentence,

and theme). The technique in this paper very connected by developing a chart which reflect relationship between various granularity hubs which have distinctive size. At that point apply positioning calculation to ascertain score of hubs lastly most noteworthy score of sentences will be chosen in the record for creating synopsis. By utilizing DUC2001 and DUC 2002, it shows the great exploratory outcome.

A paper on A Novel Relational Learning-to-Rank Approach for Topic-Focused Multi-Document Summarization by Yadong Zhu et al. (2013) [21] gives Optimization calculation and R-LTR (Learning-to-rank) approach. Social R-LTR structure is utilized instead of conventional R-LTR in an exquisite way which maintains a strategic distance from differing qualities issue. Assorted qualities are a testing issue in extractive outline technique. The positioning capacity particularly characterize as the blend of ran sentences from records and for this which is connected first then misfortune capacity is connected on Plackett-Luce demonstrate which gives positioning method on client sentences. Stochastic slope drop is then used to direct the learning procedure, and the synopsis is created by anticipating insatiable determination system. Quantitative and subjective approach can be given by trial comes about on TAC 2008 AND TAC 2009 which gives condition of-workmanship techniques. To suit the learning technique which will use on other kind of dataset past the conventional report.

Another paper on Learning to Rank for Query-centered Multi-Document Summarization by Chao Shen, Tao Li (2011) [22] investigate how to utilize positioning SVM to set up the element weight for question centered multi-archive outline. As abstractive outline gives not very much coordinated sentences from the archives and human produced synopsis is abstractive so consequently positioning SVM is pertinent here. To begin with, gauge the sentence-to - sentence relationship by considering likelihood of sentence from the archives. Second, cost touchy misfortune capacity is made inferred preparing information less delicate in the positioning SVM's goal work. Trial result shows viable aftereffect of proposed strategy.

### 2.3 LDA Based Approach

Latent Dirichlet Allocation (LDA), has been as of late presented for producing corpus subjects [22], and connected to sentence based multi-record synopsis strategy. It is not impulse to gauge points are of equivalent significance or pertinence gathering of sentence or centrality subjects. A portion of the points can contain diverse subject and unimportant so for this LDA is utilized for theme demonstrate.

The paper Mixture of Topic Model for Multi-archive Summarization by Liu Na (2014) [15] in light of Titled-LDA calculation which models title and substance of reports then blends them by hilter kilter technique. Here blend weights for points to be resolved. Theme show represent a thought how records can be displayed as likelihood conveyances over words in an archive. Titled-LDA isolated into three errands: First, dissemination of theme is done over the subject which is tested from a Dirichlet conveyance. Second, a solitary point is chosen by this appropriation for every word in the record. At long last, every word is tested from a polynomial dispersion over words which are characterized in inspected point. What's more, get the title data and the substance data in suitable way which is useful in execution of Summarization. The exploratory outcomes indicate great come about by proposing another calculation contrasted with other calculation on DUC 2002 CORPUS.

The paper Latent Dirichlet Allocation and Singular Value Decomposition in light of Multi-Document Summarization by Rachit Arora et al. (2008) [16] proposed LDA-SVD (Latent Dirichlet Allocation and

Singular Value Decomposition) Multi-Document Summarization calculation. As multi-record rundown covers distinctive occasions from the sentences in the reports and LDA separate that archives into various themes or occasions. Yet, here orthogonal vector is required to lessen regular data substance and it gives relationship of sentences. SVM is utilized to get the orthogonal representations of vectors and furthermore can speaks to as sentence orthogonal. LDA finds diverse subjects in the archives though SVD finds the sentences which are best speak to these points. At last, assess the calculations on DUC 2002 CORPUS multi-report synopsis errands utilizing the ROUGE evaluator to assess the rundowns. This calculation gives better outcomes for ROUGE-1 review measures in examination of DUC 2002.In this. LDA-SVD Multi-Document rundown calculation is superior to GISTEXTER and WSRSE.

This paper Multi-archive Summarization in view of Hierarchical Topic Model by Hongyan Lill et al. (2011) [17] speaks to h-LDA (various leveled Latent Dirichlet Allocation) calculation presented for extractive multi-report synopsis technique. h-LDA calculation isolate into four stages as Pre-preparing of the informational index, Sentence weighting, Similarity Calculation and Summary sentence pressure. It speaks to profitable probabilistic model. This concentrates idle themes from various records and furthermore can arrange these subjects into a pecking order to increase semantic investigation. In the meantime sentence pressure innovation is utilized to exact the outlines. So by doing this, we get brief synopsis. Here TAC 2010 datasets are utilized for exploratory reason and furthermore ROUGE strategy is utilized for assessing the outcomes. It gives preferred outcomes over customary technique.

The paper on Topic-Sensitive Multi-record Summarization Algorithm Liu Na et al. (2014) [18] proposes Topic-Sensitive Multi-Document Summarization calculation. This calculation separates the theme into two classes as noteworthy subject and immaterial point. Huge point as LDA character of sentence strategy is utilized as a part of this proposed display for checking comparability between sentence theme. This approach highlights the advantages of insights attributes and collaborated with LDA point display. LDA highlight is utilized to ascertain sentence weight. This approach gives better outcome utilizing DUC 2002 CORPUS when contrasted with other condition of-craftsmanship calculations.

## III. PROPOSED SYSTEM

The concentration of our thought is on consolidating co-referent things. Co-referent things is an arrangement of reports identified with a similar point that one needs to condense which are prepared to be converged in the information consolidating issue. An archive is disintegrated into a multi-set of ideas. After decay of the reports into multi-set of ideas a weighted ideal union capacity is connected. The multi-set of ideas therefore acquired is considered as an arrangement of key ideas. For synopsis era an essential adjustment of the NEWSUM calculation is presented. It is an outline method that utilizations sentence extraction approach so as to produce rundowns.

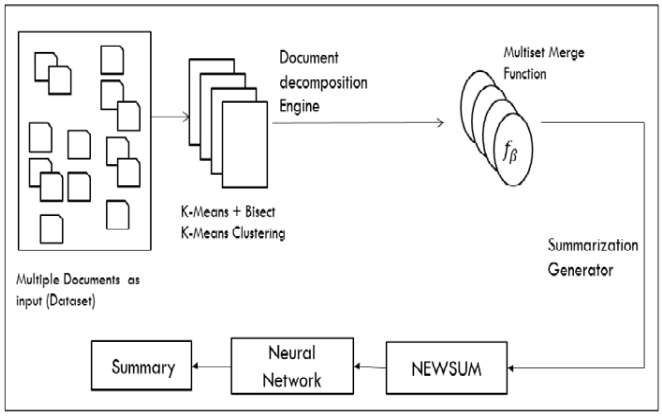


Fig. 1 System Architecture

## IV. CONCLUSIONS

In this paper, ideas of Multi-Document Summarization are assessed with various methodologies. This writing survey looks at the current pattern in rundown framework and regular dialect preparing is utilized to make the synopsis which depends on human communication and PC framework. All techniques utilized as a part of outline that gives corresponded data about the subject. There is affiliation found after outline of numerous archives. Around 22 papers have been talked about here and different techniques that is as of now exists that likewise depicted in this study. From over all study, obviously multi report synopsis is preferable method over single archive outline. Thus, anybody can get another heading for better recognition which will build another method for next age.

## V. REFERENCES

[1] Van Britsom, Daan, Antoon Bronselaer, and Guy De Tre. "Using data merging techniques for generating multi-document summarizations." in IEEE trans. On fuzzy systems, pp 1 -17, 2013.

[2] Bagalkotkar, A., Kandelwal, A., Pandey, S., &Kamath, S. S. (2013, August). A Novel Technique for Efficient Text Document Summarization as a Service.InAdvances in Computing and Communications (ICACC), 2013 Third International Conference on (pp. 50-53). IEEE.

[3] Gupta, V. K., &Siddiqui, T. J. (2012, December). Multi-document summarization using sentence clustering.In

[4] Intelligent Human Computer Interaction (IHCI), 2012 4th International Conference on (pp. 1-5).IEEE.

[5] Ferreira, Rafael, Luciano de Souza Cabral, Rafael DueireLins, Gabriel Pereira e Silva, Fred Freitas, George DC Cavalcanti, Rinaldo Lima, Steven J. Simske, and Luciano Favaro. "Assessing sentence scoring techniques for extractive text summarization." Expert systems with applications 40, no. 14 (2013): 5755-5764.

[6] Guran, A., N. G. Bayazit, and E. Bekar. "Automatic summarization of Turkish documents using non-negative matrix factorization." In Innovations in Intelligent Systems and Applications (INISTA), 2011 International Symposium on, pp. 480-484.IEEE, 2011.

[7] ShashiShekhar “A WEBIR Crawling Framework for Retrieving Highly Relevant Web Documents: Evaluation Based on Rank Aggregation and Result Merging Algorithms” in Conf. on Computational Intelligence and Communication Systems, pp 83-88 ,2011.

[8] Manjula.K.S “Extracting Summary from Documents Using K-Mean Clustering Algorithm” in IEEE IJARCCE, pp 3242-3246, 2013.

[9] Gawali, Madhuri, MrunalBewoor, and SuhasPatil. "Review: Evaluating and Analyzer to Developing Optimized Text Summary Algorithm."

[10] P.Sukumar, K.S.Gayathri “Semantic based Sentence Ordering Approach for Multi-Document Summarization” in IEEE IJRTE, pp 71-76, 2014.

[11] JinqiangBian “Research On Multi-document Summarization Based On LDA Topic Model” in IEEE Conf. On Conference on Intelligent Human-Machine Systems and Cybernetics ,pp 113-116 , 2014

[12] Li, Yuhua, David McLean, Zuhair A. Bandar, James D. O'shea, and Keeley Crockett. "Sentence similarity based on semantic nets and

corpus statistics."Knowledge and Data Engineering, IEEE Transactions on 18, no. 8 (2006): 1138-1150.

[13] Harshad Jain et. al. "Context Sensitive Text Summarization Using K Means Clustering Algorithm" IJSCE, pp no 301-304, 2012.

[14] García-Hernández, René Arnulfo, and YuliaLedeneva. "Word Sequence Models for Single Text Summarization."In Advances in Computer-Human

[15] Ferreira, R., de Souza Cabral, L., Lins, R. D., Pereira e Silva, G., Freitas, F., Cavalcanti, G. D., ...&Favaro, L. (2013). Assessing sentence scoring techniques for extractive text summarization. Expert systems with applications, 40(14), 5755-5764.

[16] Liu Na et al."Mixture of Topic Model for Multi-document Summarization" In 2014 26th Chinese Control and Decision Conference (CCDC), IEEE, pp no 5168-5172.

[17] RachitArora et al. "Latent Dirichlet Allocation and Singular Value Decomposition based Multi-Document Summarization" In2008 Eighth IEEE International Conference on Data Mining, pp no 713-718.

[18] HongyanLill et al. "Multi-document Summarization based on Hierarchical Topic Model" HongyanLill, pp no 88-91.

[19] Liu, N., Tang, X. J., Lu, Y., Li, M. X., Wang, H. W., & Xiao, P. (2014, July). Topic-Sensitive Multi-document Summarization Algorithm. In Parallel Architectures, Algorithms and Programming (PAAP), 2014 Sixth International Symposium on (pp. 69-74). IEEE.

[20] Yan, Su, and Xiaojun Wan. "SRRank: Leveraging Semantic Roles for Extractive Multi-Document Summarization." Audio, Speech, and Language Processing, IEEE/ACM Transactions on 22, no. 12 (2014): 2048-2058.

[21] Yang Wei "Document Summarization Method based on Heterogeneous Graph" In 2012 9th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2012), pp no. 1285-1289, 2012.

[22] Zhu, Yadong, Yanyan Lan, Jiafeng Guo, Pan Du, and Xueqi Cheng. "A Novel Relational Learning-to-Rank Approach for Topic-Focused Multi-Document Summarization." In Data Mining (ICDM), 2013 IEEE 13th International Conference on, pp. 927-936. IEEE, 2013