|| Volume 5 || Issue 5 || May 2020 || ISSN (Online) 2456-0774

INTERNATIONAL JOURNAL OF ADVANCE SCIENTIFIC RESEARCH AND ENGINEERING TRENDS

SURVEY AND ANALYSIS ON AUTOMATIC TEXT SUMMARIZATION METHODS

Miss.Sushama Pawar [1], Dr.Sunil Rathod [2]

PG Students, Department of Computer Engineering, DYPSCOE, PUNE, INDIA [1]

Professor, Department of Computer Engineering, DYPSCOE, PUNE, INDIA [2]

pawarsushma23@gmail.com[1], sunil.rathod@dypic.in [2]

-- ***--

Abstract: The document summarization is becoming essential as lots of information getting generated every day. Instead of going through the entire text document, it is easy to understand the text document fast and easily by a relevant summary. Text summarization is the method of explicitly making a shorter version of one or more text documents. It is a significant method of detecting related material from huge text libraries or from the Internet. It is also essential to extract the information in such a way that the content should be of user's interest. Text summarization is conducted using two main methods extractive summarization and abstractive summarization. When method select sentences from word document and rank them on basis of their weight to generate summary then that method is called extractive summarization. Abstractive summarization method focuses on main concepts of the document and then expresses those concepts in natural language. Many techniques have been developed for summarization on the basis of these two methods. There are many methods those only work for specific language. Here we discuss various techniques based on abstractive and extractive text summarization methods and shortcomings of different methods.

Keywords: - *Text Summarization, extractive summary, information extraction*

--***--

I. INTRODUCTION

With increasing amount of data it becomes more and more difficult for users to derive material of interest, to search efficiently for specific content or to gain an overview of influential, important and relevant material. In today's information technology number of people is searching for informative data on web, but every time it is not possible that they could get all relevant data in single document, or on a single web page. They could get number of web pages as a search result [5]. This problem has given the new solution that is associated to data mining and machine learning which returns query specific information from large set of offline documents and represents as a single document to the user. So, automated summarization is an important area in Natural Language Processing (NLP) research. Automated summarization provides single document summarization and multi-document summarization [3].

1. MULTI-DOCUMENT MERGER:

The merging of data from multiple documents is called multi-document merger. Data is found in unstructured or structured form and many times we have to generate summary from multiple files in less time, so, multi-document merger technique is useful. Multi-document summarization generates information reports that are both concise and comprehensive. With different opinions being put together, every topic is described from multiple perspectives within a single document. The goal of a brief summary is to simplify information search and save the time by pointing to the most relevant information.

Text summarization is gaining much importance currently. One reason for this is, due to the rapid growth in material, requirement for involuntary text summarization has enlarged. It is very difficult for human beings to manually summarize big documents of text. There is a profusion of text material available on the internet. However, usually the Internet offers more material than is required. Therefore a problem of repetition is encountered: examining for similar kind of documents through a large amount of documents is very tedious task [3]. The aim of text summarization is to reduce the source text into a shorter form preserving its information content and overall meaning. If sentences in a text document were of equivalent significance, creating a summary would not be very effective. With different opinions being put together & outlined, every topic is seen and described from multiple perspectives within a single document. While the main aim of a brief summary is to simplify information search and cut the time by pointing to the most relevant source documents, multi-document summary should itself contain the required information, hence limiting the need for accessing original files to cases when refinement is required. In this study various techniques for sentence based extractive summarization has been encountered also various similarity measures and their comparisons.

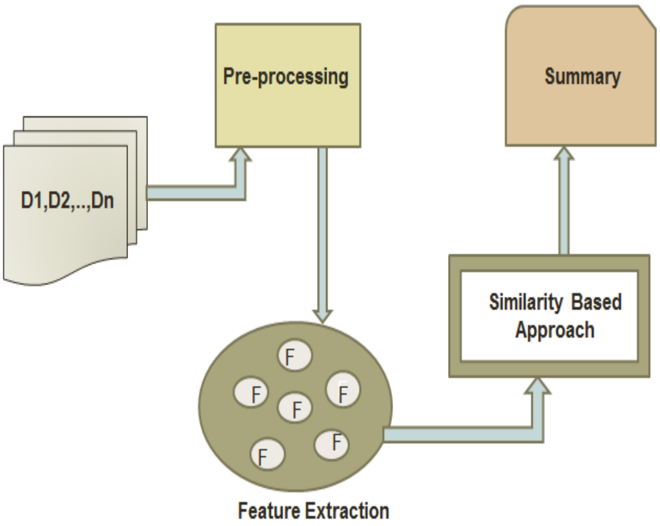

Figure 1 Generalized structure of document summarization

II. EXTRACTIVE TEXT SUMMARIZATION

Extractive summarizer aims at selecting the foremost relevant sentences within the document whereas maintaining a reduced redundancy within the outline. It is created by reusing portion (word, sentences etc.) of input text verbatim.

Example: Search engines typically generate extractive summaries from web pages.

A. Term Frequency-Inverse Document Frequency (TFIDF) approaches:

Bag-of-words model is made at sentence level, with the traditional term-frequency and sentence frequency algorithms, wherever sentence frequency is that the range of sentences within the document that have that term, words that occur frequently within the documents is also taken as the question words. Since these words represent the theme of the document, they manufacture generic summaries. Term frequency is typically zero or one for sentences [10].

A) Clustering based approach:

Documents area unit consist of mistreatment term frequency and inverse document frequency (TF-IDF) of various words. Term frequency in this context is that the average range of existences of similar kind of document over the cluster. The summarizer takes clustered documents as input. In each cluster the theme is portrayed by words with high ranking term frequency, inverse document frequency (TF-IDF) scores in this cluster. Sentence choice is dependent on the similarity of the sentences to the theme of the cluster [10][12].

B) Machine Learning Approach:

In a group of documents and their extractive summaries, the summarization algorithms are displayed as a classification problem: sentences area unit classified as outline sentences and non-summary sentences supported the options that they maintain. The classification likelihood is that learnt statistically from the obtained information, using Bayes' rule, there are also several machine learning techniques that can be used for document summarization [11].

Below figure shows the type of text summarization with its methodology used for summarization purpose.

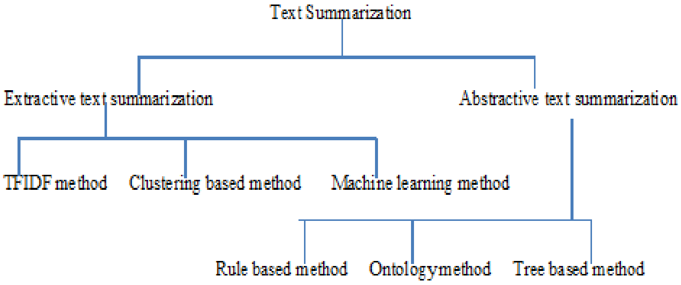

Figure 2 Classification of Text Summarization

III. ABSTRACTIVE TEXT SUMMARIZATION

Methods employ more powerful natural language processing techniques to interpret text and generate new summary text, as opposed to selecting the most representative existing excerpts to perform the summarization.

The information from source text is re-phrased. But it is harder to use because it provides allied problems such as semantic representations.

Example: Book Reviews-if we want a summary of book The Lord of The Rings then by using this method we can make summary from it.

A) Rule Based Method:

The rule based method[4]comprises of three steps:-- Firstly, the documents to be classified are represented in terms of their categories. The categories can be from various domains. Hence the first task is to sort these. The next thing is to form questions based on these categories amongst the various categories like attacks, disasters, health etc. taking the example of an attack category several questions can be figured out like:-What happened?, when did it happen?, who got affected ?, what were the consequences? etc. -Depending upon these questions, rules are generated. Here several verbs and nouns having similar meanings are determined and their positions are correctly identified.-The context selection module selects the best candidate amongst these.-Generation patterns are then used for the generation of summary sentences.

B) Ontology Method:

In this method, domain ontology for news event is defined by the domain experts. Next phase is document processing phase. Meaningful terms from corpus are produces in this phase[7].The meaningful terms are classified by the classifier on basis of events of news. Membership degree associated with various events of domain ontology. Membership degree is generated by fuzzy inference. Limitations of this approach are it is time consuming because domain ontology has to be defined by domain experts. Advantage of this approach is it handles uncertain data.

C) Tree Based Method:

In this approach, the pre-processing is done of similar sentences using shallow parser [5]. After that we map those sentences to the predicate-argument structure. Different algorithms can be used for selecting the common phrase from the sentences such as Theme algorithm. The phrase conveying the same meaning is selected and also we add some information to it and will arrange in a particular order. At the end, FUF/SURGE language generator can be used for making the new summary sentences by combining and arranging the selected common phrase. Use of language generator increases the fluency of the language and also reduces the grammatical mistakes. This feature is the main strength of this method. The main problem with this method is that the context of the sentences does not get included while selection of common phrase and it is important part of the sentences even if it is not part of the common phrase.

IV. LITERATURE SURVEY

An improved method of automatic text summarization for web contents using lexical chain with semantic-related terms proposes an improved extractive text summarization method for documents by enhancing the conventional lexical chain method to produce better relevant information. Then, Author firstly investigated the approaches to extract sentences from the document(s) based on the distribution of lexical chains then built a transition probability distribution generator(TPDG) for n-gram keywords which learns the characteristics of the assigned keywords from the training data set. A new method of automatic keyword extraction also featured in the system based on the Markov chains process. Among the extracted n-gram keywords, only unigrams are selected to construct the lexical chain. [1]Top- K ensemble ranking algorithm is used to rank sentences TF-IDF is used to word count and word level feature extraction. In paper [2] author first extracted multiple candidate summaries by proposing several schemes for improving the upper-bound quality of the summaries. Then, proposed a new ensemble ranking method for ranking the candidate summaries by making use of bilingual features. Extensive experiments have been conducted on a benchmark dataset.[2]

Automatic text summarization within big data framework demonstrates how to process large data sets in parallel to address the volume problems associated with big data and generate summary using sentence ranking. TF-IDF is used for document feature extraction. MapReduce and Hadoop is used to process big data.[3]

Extractive document summarization based on hierarchical GRU proposes two stage structure 1) Key sentence extraction using Levenshtein distance formula 2) Recurrent neural network for summarization of documents. In extraction phase system conceives a hybrid sentence similarity measure by combining sentence vector and Levenshtein distance and integrates into graph model to extract key sentences. In the second phase it constructs GRU as basic block, and put the representation of entire document based on LDA as a feature to support summarization.[4]

Extractive algorithm of English text summarization for English teaching is based on semantic association rules. To summarize documents semantic association rule vectors is used. In this paper relative features are mined among English text phrases and sentences, the semantic relevance analysis and feature extraction of keywords in English abstracts are realized. [5]

Fairness of extractive text summarization is the first work that introduces the concept of fairness of text summarization algorithms. Author shows that while summarizing datasets having an associated sensitive attribute, one needs to verify the fairness of the summary. Especially, with the advent of neural network-based summarization algorithms (which involve super-wised learning), the question of fairness becomes even more critical. Author believe that this work will open up interesting research problems, e.g., on developing algorithms that will ensure some degree of fairness in the summaries. [6]

Automatic text summarization by local scoring and ranking for improving coherence approach provides automatic feature based extractive heading wise text summarizer to improve the coherence thereby improving the understand ability of the summary text. It summarizes the given input document using local scoring and local ranking that is it provides heading wise summary. Headings of a document give contextual information and permit visual scanning of the document to find the search contents. The outcomes of the experiment clearly show that heading wise summarizer provides better precision, recall and f-measure over the main summarizer, MS-word summarizer, free summarizer and Auto summarizer [7].

A paper on data merging by Van Britsom proposed a technique based on use of NEWSUM Algorithm. It is a type of clustering algorithm where divides a set of document into subsets and then generates a summary of correlated texts. It contains three phases: topic identification, transformation and summarization by using different clusters [8].

A novel technique for efficient text document summarization as a service by Anusha Banalkotkar represents the different techniques that explain as the main two fundamental techniques are identified to automatically summarize texts i.e. abstractive summarization and extractive summarization. Complex summarization technique (cohesive, readable, intelligible, multi-disciplinary approaches, machine learning) all are coming under this paper.[9]

Multi-document summarization using sentence clustering by Virendra Kumar Gupta states that sentences from single document summaries are clustered and top most sentences from each cluster are used for creating multi-document summary. The model contains the steps as pre-processing, noise removal, tokenization, stop words, stemming, sentence splitting and feature extraction. After performing these steps, important sentences are extracted from each cluster.[10]

The method reinforcement ranking on the Semantic Link Network can be applied to any structural text and the provision of various summarization services such as automatically generating the Mind Map of scientific paper, slides for a given paper, and extended abstract for a long scientific paper or book to give readers a quick impression of the core content.[12]

There are different types of text summarization techniques but here focused on two main content-based types of summaries: generic summaries and query-based summaries. If the system does not depend on the document subject and the user does not have any previous understanding of the text, all the information will be in the same level of importance. In such system we can say it is a generic summarization system. Differently, in a query-based summarization, before the summarization process starts, the user has to determine the topic of the original text in a query form. The user asks for special information in form of a query and the system only extracts that information from the source text and presents it as a summary.[13] [14] The proposed approach is based on the semantic information of the extracts in a text. So, different parameters like formats, positions of different units in the text are not taken into account. But in few cases, there are dominating numbers of named entities in a text. In those cases, hybridization of the proposed approach with some specific rules regarding Named Entity Recognition should give more effective results.

The paper proposes a novel system called PPSGen to generate presentation slides from academic papers. Author trained a sentence scoring model based on SVR and uses the ILP method to align and extract key phrases and sentences for generating the slides. Experimental results show that proposed method can generate much better slides than traditional methods. In this paper, Author only considers one typical style of slides that beginners usually use. [16]

In this paper, author examined how to use data merging techniques to summarize a set of co-referent documents that has been clustered while using soft computing techniques. The main focus of this paper lies on the fβ-optimal merge function which that uses the weighted harmonic mean to find a balance between precision and recall. The global precision and recall measures mentioned are defined by means of a triangular norm receiving local precision and recall values as an input, in order to generate a multi-set of key concepts that can use to generate summarizations.[17]

To overcome the low-frequency and misinterpretation problems for text mining pattern discovery technique is used. The proposed technique uses two processes, pattern deploying and pattern evolving, to refine the discovered patterns in text documents. The experimental results show that the proposed model outperforms not only other pure data mining-based methods and the concept-based model, but also term-based state-of-the-art models, such as BM25 and SVM-based models.[18]

Proposed algorithm relies on WordNet which is theoretically domain independent, and also author have used Wikipedia for some of the words that do not exist in the WordNet. For summarization, author aimed to use more cohesion clues than other lexical chain based summarization algorithms. Evaluated results were competitive with other summarization algorithms and achieved good results. Using co-occurrence of lexical chain members, our algorithm tries to build the bond between subject terms and the object terms in the text.[19]

The technique discussed in this paper is considered to be a pioneering attempt in the field of NLP (Natural Language Processing). The technique involves an information extractor and a slide generator, which combines certain NLP methods such as segmentation, chunking, summarization etc, with certain special linguistic features of the text such as the ontology of the words, noun phrases found, semantic links, sentence centrality etc., In order to aid the language processing task, two tools can be utilized namely, MontyLingua which helps in chunking and Doddle helps in creating an ontology for the input text represented as an OWL (Ontology Web Language) file.[20]

In this paper, author proposed the algorithm PASCAL which introduces a novel optimization of the well-known algorithm Apriori. This optimization is based on a new strategy called pattern counting inference thatrelies on the concept of key patterns. System shows that the support of frequent non-key patterns can be inferred from frequent key patterns with- out accessing the database. Experiments comparing PASCAL to In this paper, author proposed the algorithm PASCAL which introduces a novel optimization of the well-known algorithm Apriori. This optimization is based on a new strategy called pattern counting inference that the three algorithms Apriori, Close and Max-Miner, show that PASCAL is among the most efficient algorithms for mining frequent patterns.[21]

This paper presents an innovative pattern enhanced topic model for information filtering including user interest modeling and document relevance ranking. The proposed MPBTM model generates pattern enhanced topic representations to model user's interests across multiple topics. In the filtering stage, the MPBTM selects maximum matched patterns, instead of using all discovered patterns, for estimating the relevance of incoming documents. The proposed approach incorporates the semantic structure from topic modeling and the specificity as well as the statistical significance from the most representative patterns. The proposed model has been evaluated by using the RCV1 and TREC collections for the task of information filtering. In comparison with the state-of-the-art models, the proposed model demonstrates excellent strength on document modelling and relevance ranking.[22]

In this paper, author proposed a systematic framework for frequent pattern-based classification and gives theoretical answers to several critical questions raised by this framework. Author stated that the proposed method is able to overcome two kinds of over fitting problems and shown to be scalable. A strategy for setting min_sup is also suggested. In addition, author proposed a feature selection algorithm to select discriminative frequent patterns. Experimental studies demonstrate that significant improvement is achieved in classification accuracy using the frequent pattern-based classification framework. The framework is also applicable to more complex patterns, including sequences and graphs.[23]

Author has presented and evaluated the Max-Miner algorithm for mining maximal frequent item sets from large databases. Max- Miner applies several new techniques for reducing the space of item sets considered through superset-frequency based pruning. The result is orders of magnitude in performance improvements over Apriori-like algorithms when frequent item sets are long, and more modest though still substantial improvements when frequent item sets are short. Max-Miner is also easily made to incorporate additional constraints on the set of frequent item sets identified. Incorporating these constraints into the search is the only way to achieve tractable completeness at low supports on complex data-sets. [24]

In this paper author presented a brief overview of the current status and future directions of frequent pattern mining. Overview provides a rough outline of the recent work and gives a general view of the field. In general, author feels that as a young research field in data mining, frequent pattern mining has achieved tremendous progress and claimed a good set of applications. However, in-depth research is still needed on several critical issues so that the field may have its long lasting and deep impact in data mining applications.[25]

Author presented and evaluated CHARM, an efficient algorithm for mining closed frequent itemsets. CHARM simultaneously explores both the itemset space and tidset space using the new IT-tree framework, which allows it to use a novel search method that skips many levels to quickly identify the closed frequent itemsets, instead of having to enumerate many non-closed subsets. We utilized a new vertical format based on diffsets, i.e., storing the differences in the tids as the computation progresses. An extensive set of experiments confirms that CHARM can provide orders of magnitude improvement over existing methods for mining closed itemsets.[26]

Author proposed LDA-based document models for ad-hoc retrieval, and evaluated the method using several TREC collections. Based on the experimental results, author made following conclusions. Firstly, experiments performed in the language modeling framework, including combination with the relevance model, have demonstrated that the LDA-based document model consistently outperforms the cluster-based approach, and the performance of LBDM is close to the Relevance Model, which incorporates pseudo-feedback information. Secondly, it shows that the estimation of the LDA model on IR tasks is feasible with suitable parameters based on the analysis of the algorithm complexity and empirical parameter selections.[27]

Author stated that problem with association rule mining is the redundancy existing in the extracted association rules which greatly impacts the effective use of the extracted rules in solving real world problems. A satisfactory solution to the problem should be one that can maximally remove redundancy but does not damage the inference capacity of and the belief in the extracted rules. Moreover, an appropriate criterion to define a boundary between redundancy and non-redundancy is desirable. In this paper, author proposed a concise representation of association rules called Reliable basis was presented which can ensure the removal of the maximal amount of redundancy without reducing the inference capacity of the remaining extracted rules. Moreover, author proposed to use the certainty factor as the criterion to measure the strength of the discovered association rules.[28]

Author proposed an algorithm for recommending scientific articles to users based on both content and other users' ratings. Experimental analysis showed that this approach works well relative to traditional matrix factorization methods and makes good predictions on completely unrated articles. Further, algorithm provides interpretable user profiles. Such profiles could be useful in real-world recommender systems. For example, if a particular user recognizes her profile as representing different topics, she can choose to “hide” some topics when seeking recommendations about a subject.[29]

Author proposed latent Dirichlet allocation (LDA), a generative probabilistic model for collections of discrete data such as text corpora. LDA is a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying set of topics. Each topic is, in turn, modeled as an infinite mixture over an underlying set of topic probabilities. In the context of text modeling, the topic probabilities provide an explicit representation of a document. Author presented efficient approximate inference techniques based on variational methods and an EM algorithm for empirical Bayes parameter estimation.[30]

Limitations:

In paper [1] effectiveness and time consumption are the main issues. Average precision measure is between 4.8 to 6.0 percent on the cost of maximum time complexity because thesystem first extract sentences from the document(s) based on the distribution of lexical chains then built a transition probability distribution generator (TPDG) for n-gram keywords which learns the characteristics of the assigned keywords from the training data set which takes maximum amount of time.

In paper [2] the system is designed for multiple language document summarizations but the accuracy of summarization is not up to the mark, as stated in paper [2] the accuracy for summarization is 60 percent with complex execution framework.

The system demonstrated in [3] is designed only for big data framework. Author used MapReduceframework to minimize data mining time and MapReduceframework is designed to work only with big data so system will not work on other data storage frameworks [3].

In paper [12] author used reinforcement ranking on the Semantic Link Network of various representation units only within scientific paper for summarization, but to make system work on other domain documents like news articles or sports system dataset needs to be trained.

Automatic Text Summarization Based on Fuzzy Logic [19] is able to deal with imprecise linguistic information and can model nonlinear functions of arbitrary complexity the main advantage of this method is it does not need lots of data to train but the process of designing fuzzy rules, which have to cover all the relationships among the parameters, is quite time consuming. The system relies on WordNet and Wikipedia for document term extraction and only focused on single document summarization.

Heading wise summarizer [7] can be performed on single as well as multi-document generic summarization and uses Singular Value Decomposition (SVD) to find outinterrelation between two documents which has less time complexity. Considering minimum time complexity framework can extract redundant sentences under different headings. And it only work on document paragraph with headings.

Hierarchical GRU [4] combines the traditional RNN with Levenshtein Distance formula gives accurate results. The main advantage of this framework is that it shows better results with noisy datasets. Framework contains RNN which increases the calculation complexity of the framework.

DSDR [6] selects the most representative sentences that can best reconstruct the entire document that is it measures the relationship between the textual units using linearcombinations and reconstructions, and generates the summaryby minimizing the reconstruction error and can generate less redundant sentences. System has high accuracy rate on the cost of maximum time complexity.

Latent Semantic Analysis [29] is capable of assuring decent results, much better than plain vector space model. It works well on dataset with diverse topics.LSA can handle Synonymy problems to some extent (depends on dataset though). Since it only involves decomposing term document matrix, it is faster, compared to other dimensionality reduction models. Since it is a distributional model, so not an efficient representation, when compared against state-of-the-art methods (say deep neural networks).

In above literature survey we found that all summarization frameworks are unique in their own way with respect to document processing, algorithms and final outputs. To overcome the limitations discussed above for existing systems, we suggest following methods.

i) Clustering with cosine similarity algorithm for sentence extraction.

Previously we analyzed some limitations of existing systems one of them was single domain summarization that is algorithm only works on specific documents like scientific, sports, news document to avoid this we are using cosine similarity algorithm which gives very good sentence extraction result regardless of the type of document or size of the document.While extracting sentences we will treat a heading as a general sentence so the system will perform on documents with or without heading.

i) The NEWSUM algorithm for generating clusters.

To increase accuracy we are using clustering so that we can avoid unrelated documents, on top of that both algorithms have minimum time complexity which will help to minimize overall system execution time.

ii) Position score algorithm to rank the sentences.

Finally we are using position score algorithm to rank the extracted sentences which will help to maximize the accuracy rate of the system. Here we are expecting the accuracy around 75 to 80 percent.

As time complexity can be extracted from the different modules as follows:

Module 1: Pre-processing of Input multiple documents-if we are preprocessing on the input documents then we extract the number of words or sentences by removing stop words. So, time complexity for this module becomes O(n).

Module 2: Feature extraction of multiple documents-after removing stop words n stemming on sentences then we extracted feature from all documents. So, complexity for this module becomes O(n).

Module 3: Similarity Based Approach-after extracting feature from all documents then compare two or more sentences with each other that they are related to each other or not. So, complexity becomes O(n).

The proposed work focuses on using different aspects of text mining to come up with an efficient approach that will aid the document creator with the draft format of the contents essentially conveying the important concepts of the text by surpassing the accuracy rate of existing frameworks with minimum time

V. CONCLUSION AND FUTURE WORK

The data can be retrieved by using the background knowledge for generalization. Now a day the growth of data increases in structured or unstructured form and we want a summary from that data in less time. To overcome the drawback of previous model we propose a new system. The work is under implementation and focuses on using different aspects of text mining to come up with an efficient approach that will aid the document creator with the draft format of the contents essentially conveying the important concepts of the text. So, multi-document merger summarization is used. It reduces our time and gives efficient output.

REFERENCES

[1] HtetMyet Lynn 1 , Chang Choi 2 , Pankoo Kim “An improved method of automatic text summarization for web contents using lexical chain with semantic-related terms”, Springer-Verlag Berlin Heidelberg 2017

[2] Xiaojun Wan 1 , FuliLuo 2 , Xue Sun Songfang Huang3 , Jin-ge Yao “Cross-language document summarization via extraction and ranking of multiple summaries” Springer- Verlag London 2018

[3]Andrew Mackey and Israel Cuevas “AUTOMATIC TEXT SUMMARIZATION WITHIN BIG DATA FRAMEWORKS”, ACM 2018

[4] Yong Zhang, Jinzhi Liao, Jiyuyang Tang “Extractive Document Summarization based on hierarchical GRU”, International Conference on Robots & Intelligent System IEEE 2018

[5] Lili Wan “Extractive Algorithm of English Text Summarization for English Teaching” IEEE 2018

[6]AnuragShandilya, KripabandhuGhosh, SaptarshiGhosh “Fairness of Extractive Text Summarization”, ACM 2018

[7] P.Krishnaveni, Dr.S. R. Balasundaram “Automatic Text Summarization by Local Scoring and Ranking for Improving Coherence”, Proceedings of the IEEE 2017 International Conference on Computing Methodologies and Communication

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

[9] 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.

[10] Gupta, V. K., &Siddiqui, T. J. (2012, December). “Multi-document summarization using sentence clustering”, In Intelligent Human Computer Interaction (IHCI), 2012 4th International Conference on (pp. 1-5). IEEE.

[11] Min-Yuh Day Department of Information Management Tamkang University New Taipei City,

Taiwan myday@mail.tku.edu.tw Chao-Yu Chen Department of Information Management Tamkang University New Taipei City, Taiwan susan.cy.chen@gmail.tw "Artificial Intelligence for Automatic Text Summarization",2018 IEEE International Conference on Information Reuse and Integration for Data Science

[12] Xiaoping SunandHaiZhuge*, Senior Member, IEEE Laboratory of Cyber-Physical-Social Intelligence, Guangzhou University, China Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, University of Chinese Academy of Sciences, Chinese Academy of Sciences, China System Analytics Research Institute, Aston University, UK "Summarization of Scientific Paper through Reinforcement Ranking on Semantic Link Network" ,IEEE 2018

[13] Ahmad T. Al-Taani (PhD, MSc, BSc) Professor of Computer Science (Artificial Intelligence) Faculty of Information Technology and Computer Sciences Yarmouk University, Jordan. ahmadta@yu.edu.jo "Automatic Text Summarization Approaches" ,IEEE 2017

[14] AlokRanjan Pal Dept. of Computer Science and Engineering College of Engineering and Management, KolaghatKolaghat, India chhaandasik@gmail.com DigantaSaha Dept. of Computer Science and Engineering Jadavpur University Kolkata, India neruda0101@yahoo.com "An Approach to Automatic Text Summarization using WordNet", IEEE 2014

[15] Prakhar Sethi1, Sameer Sonawane2, Saumitra Khanwalker3, R. B. Keskar4 Department of Computer Science Engineering, Visvesvaraya National Institute of Technology, India 1 prakhar.sethi2@gmail.com, 2 sameer9311@gmail.com, 3 theapogee2011@gmail.com, 4rbkeskar@cse.vnit.ac.in"Automatic Text Summarization of News Articles", IEEE 2017

[16] Yue Hu and Xiaojun Wan "PPSGen: Learning-Based Presentation Slides Generation for Academic Papers" , IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 27, NO. 4, APRIL 2015

[17] Daan Van Britsom, AntoonBronselaer, and Guy De Tre "Using Data Merging Techniques for Generating Multidocument Summarizations" , IEEE TRANSACTIONS ON FUZZY SYSTEMS, VOL. 23, NO. 3, JUNE 2015

[18] NingZhong, Yuefeng Li, and Sheng-Tang Wu "Effective Pattern Discovery for Text Mining", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 24, NO. 1, JANUARY 2012

[19] Mohsen Pourvali and Mohammad SanieeAbadeh Department of Electrical & Computer Qazvin Branch Islamic Azad University Qazvin, Iran Department of Electrical and Computer Engineering at TarbiatModares University Tehran, Iran "Automated Text Summarization Base on Lexicales Chain and graph Using of WordNet and Wikipedia Knowledge Base" , IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 1, No 3, January 2012

[20] Daan Van Britsom, AntoonBronselaer, Guy De Tre´ Department of Telecommunications and Information Processing, Ghent University Sint-Pietersnieuwstraat 41, B-9000 Ghent, Belgium "Using data merging techniques for generating multi-document summarizations" ,IEEE TRANSACTIONS ON FUZZY SYSTEMS 2018

[21] Yang Gao,YueXu, Yuefengli, "Pattern-based Topics for Document Modeling in Information Filtering" in IEEE Transaction on Knoweledge and Data Engineering, vol.27,No.6,June 2015.

[22] Y. Bastide, R. Taouil, N. Pasquier, G. Stumme, and L. Lakhal, "Mining frequent patterns with counting inference," ACM SIGKDD Explorations Newslett., vol. 2, no. 2, pp. 66–75, 2000.

[23] H. Cheng, X. Yan, J. Han, and C.-W. Hsu, "Discriminative frequent pattern analysis for effective classification," in Proc. IEEE 23rd Int. Conf. Data Eng., 2007, pp. 716–725.

[24] R. J. BayardoJr, "Efficiently mining long patterns from databases," in Proc. ACM Sigmod Record, 1998, vol. 27, no. 2, pp. 85–93.

[25] J. Han, H. Cheng, D. Xin, and X. Yan, "Frequent pattern mining: Current status and future directions," Data Min. Knowledge. Discovery., vol. 15, no. 1, pp. 55–86, 2007.

[26] M. J. Zaki and C.-J. Hsiao, "CHARM: An efficient algorithm for closed itemset mining." in Proc. SDM, vol. 2, 2002, pp. 457–473.

[27] Y. Xu, Y. Li, and G. Shaw, "Reliable representations for association rules," Data Knowl. Eng., vol. 70, no. 6, pp. 555–575, 2011.

[28] X. Wei and W. B. Croft, "LDA-based document models for ad-hoc retrieval," in Proc. 29th Annu. Int. ACM SIGIR Conf. Res. Develop. Inform. Retrieval, 2006, pp. 178–185.

[29] C. Wang and D. M. Blei, "Collaborative topic modeling for recommending scientific articles," in Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., 2011, pp. 448–456.

[30] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," J. Mach. Learn. Res., vol. 3, pp. 993–1022, 2003.