Techniques of Big Data Text Summarization

Anish Mathew Kuriakose, V.Umadevi,

Abstract**: *In the web industry 4.0, big data are playing a key role in the organizations of their digital transformation journey. Data has also changed intensely in recent years, in volume, variety and velocity. Its fast development attributed to the extensive digitization of business progressions globally. In simple term, data has turn into the new currency of business and its further quick growth will be key transformation and growth of organizations globally. Vast amount of online information, available in healthcare, social media websites, e-commerce web pages, e-books, legal domain, e-news, etc. has made text processing a vital extent of research. The paper starts with introduction about the evolution of web industry 4.0 and digitalization. Followed by the introduction, the paper discusses about big data and text summarization techniques. Further, it describes the literature of text mining that have taken in the recent years. The main objective of this paper is to discuss about the big data text summarization issues and challenges. The paper starts with general introduction of big data and text mining and text summarization. Further it describes recent advances in big data text summarization, and then delve into extraction and abstraction-based text summarization. Finally, the paper concludes with some future research directions.***

Keywords* : *Web 4.0, Big Data, Text Mining, Text Summarization.

I. INTRODUCTION

In recent times the researchers and industrialists have been talking about web 4.0, with the connected forthcoming predictions. Web 4.0 is the term where web technologies are automated with influence of Artificial Intelligence (AI) and Machine Learning (ML). Nowadays, in the next phase of web technologies, big data, internet of things, cloud computing and augmented reality are having a main role in generating the huge amount of unpredictable data. It is alleged that web 4.0 will be a stage of development in which each one will have an advanced sense of self thinking and will talk increasingly more with new interfaces as artificial intelligent machines. Correspondingly an alluring tragic vision of the web in the future, with a more prominent control of the data that will influence the computerized world as well as the truth around us.

Web 4.0 is a revolutionary thought in advancement and there is no accurate description of how it would be. Web 4.0 is otherwise called harmonious web. The fantasy behind of the advantageous web is connection among people and machines in beneficial interaction. It will be conceivable to manufacture all the more dominant interfaces. For instance,

mind-controlled interfaces utilizing web 4.0 can be build. In basic words, machines would be smarter on scrutinizing the substance of the web and respond through a medium. The medium will be accomplishing and selecting what to execute first to stack the sites quick with extraordinary quality. Also build performance along with the quality and fabricate additionally effective interfaces [1]. Web 4.0 will be the perused compose execution-simultaneousness web. It accomplishes a minimum amount of support in online systems that convey worldwide straight forwardness, administration, dissemination, interest, and coordinated the effort into key networks. For example, it builds networks among the industry, political, social and other networks [2]. Web 4.0 or webOS will be a middleware in which will begin working like a operational framework [3]. The webOS will be parallel to the human cerebrum and suggests a gigantic snare of exceptionally shrewd connections [4]. In spite of the fact that there is no accurate thought regarding web 4.0 and its advancements, yet clearly the web is pushing toward utilizing man-made consciousness to progress toward becoming as a smart web.

In Web 4.0, the connection among people and machines is synergetic. Machines are outfitted with ground-breaking equipment and programming to dissect with existing substance and settle on the most ideal choices. For instance, if a client scans for a goal in one of the web crawlers or on a site identified with lodging or flight booking, Web 4.0 innovations enable the site to recommend all or a couple of the best alternatives. Furthermore, the ongoing improvement of a portable robot application known as "advanced companion of movement" will assist vacationers with planning and deal with any movement issues [5].

II. BIG DATA

The prominence of the Internet and the approach of the Web 4.0 advances have changed the substance of the web from distributer to client made substance [6]. Such presence has helped with distributing substance without the necessities of programming. Today, intriguing subjects, reviews, and conclusions from Web 4.0 and online networking can be effortlessly reachable via the global network by means of the Internet continuously. In addition, the increase and variety of online networking have given broad chances and difficulties to business analysts and research professionals.

Manuscript published on November 30, 2019.
* Correspondence Author

Anish Mathew Kuriakose*, Ph.D Research Scholar, Department of Computer Science, Jairams Arts and Science College Karur affiliated to Bharathidasan University Tiruchirappalli

DrV.Umadevi,Director, Department of Computer Science,Jairams Arts and Science College Karur affiliated to Bharathidasan University Tiruchirappalli

In excess of a billion of individuals around the globe are utilizing internet-based life platforms that create overpowering unstructured information in moderately short timescales. The colossal measure of information created by clients is the consequence of the reconciliation of their experience subtleties and day by day exercises in such stage. This huge measure of produced information alluded to as "big data" which has been seriously investigated as of late. There are around 5, 92,000 articles have been indexed for the keyword “big data”. This shows the recent research awareness that has been evolving on the big data. The huge information from the tremendous measure of the dataset gathered in either organized, semi-organized or potentially unstructured organization. These organizations have been explored in different areas, for example, medicinal services, space science, astrology, social web, and geoscience [7]. Web based life substance, for example, tweets, remarks, posts, and reviews, have added to the making of huge information broadly from either stage suppliers or various sites [8][9]. The development of enormous information from online life has achieved another haste of enthusiasm into the field of man-made brainpower and information investigation. Breaking down internet-based life information which utilizing different customary data mining and AI systems is as yet a functioning space of research.

The most interesting part of the web 4.0 is “data deluge”. As it is discussed above, the enormous amount of data generating by the universe are having lot of hidden information. The most common data that is being used today are “textual data”. This paper addresses the role of text mining and big data text summarization.

III. TEXT MINING

The size of data is increasing and growing at a forceful and rapid rate due to evolution of web 4.0 and big data. Subsequently the information is impressively colossal all the corporate firms, foundations and associations who store the facts and data in the digitalized platform. Presently this enormous measure of information is put away and traded through web as advanced libraries and literary data like online journals and other internet-based life stages. Subsequently it is hard to remove data by utilizing conventional information mining methods since they are not ready to deal with printed information (hard copy) viably. When there is massive measure of information then all the substantial and essential data are required, Text mining is applied. Text Mining (TM) is likewise called as data text mining or text information retrieval. TM analyzes the text data which is in regular language in fact and find the lexical examples which are recognized to remove insignificant data. Pattern extraction from content records of content reports are the key objectives of content mining system improvement. Here both the unstructured information and semi organized information can be utilized for content mining. The phases for TM are follows:

Phase – 1: Transform unstructured information into organized information by gathering information from sources like natural content, website pages and information records

Phase – 2: Pre-preparing and purging tasks are executed to recognize and expel peculiarities

Phase – 3: Cleansing procedure uncovers the genuine quintessence of content which is accessible and is executed to dispose of word stemming is a procedure of perceiving a specific word root and information ordering perceive the examples from the organized information.

Phase – 4: Examines the structures and utilizes the content mining strategies.

Phase – 5: In Processing system cleaning and organizing of information is done and after the Text mining strategies like classification and clustering algorithms are applied to orchestrate to content archives.

Phase – 6: Extracts and roots out the helpful data from the content.

Computerized libraries, Web mining, medicate revelation, Clustering, Social media, recognition of connections among way of life and conditions of wellbeing, business knowledge are a portion of the uses of content mining.

[10] Health-care experts produce inexhaustible literary data in their everyday clinical practice and this data is put away in numerous different sources and, by and large, in printed structure. The extraction of bits of knowledge from all the accumulated data, is one of the significant difficulties in computational prescription due to unstructured data and lack in standardization. In this regard, TM collects various strategies to get important bits of knowledge from unstructured printed information, so it has prompted be particularly significant in medication. The point of this paper is thusly to give a broad audit of existing systems and assets to perform TM undertakings in medication. In this audit, in excess of 90 significant research studies have been broke down, portraying the most significant functional applications, expressed assets, instruments, and open difficulties of TM in medication.

The text mining techniques are as follows: information retrieval, information extraction, text summarization, clustering and categorization. This research paper focuses on the text summarization.

IV. TEXT SUMMARIZATION

A summary can be characterized as a text that is delivered from at least one text or more texts [11], that contains a substantial portion of the data in unique text document, which is no longer than half of the exclusive text [11]. As indicated by added reference, text summarization is the way toward refining the most significant data from a source to deliver abbreviated rendition for a specific client and errand [12].

At hand there are two distinct gatherings of text content summary. Inductive and Informative. Demonstrative synopsis gives the principle thought of the content to the client. The length of this synopsis is around 5 percent of the given content. The enlightening outline framework gives brief data of the primary content. The length of the informative summary is around 20 % of the given content [13]. Moreover, based on the size of source document, summarization strategies can be arranged, where, the source may be multiple or individual document. In single record summary, just one report is accommodated for generating the summary.

It is a basic and most punctual methodology for text summarization. Multi report summarization is additionally significant piece of summarization. More than one data sources are accommodated synopsis age. Many online grouping frameworks like news were enlivened from multi report outline. Nonetheless, errand of various report rundown method is more troublesome and complex than single archive procedures. The genuine point isn't just to expel excess and identify right content for outline yet in addition to give curiosity and guaranteeing that last rundown ought to be intellectual and complete in itself, it was a challenge for the authors to think about every one of the records and relate the outline [14].

Text summarization is referred as, gathering and creating brief portrayal of records with unique content [12]. At first the raw content is considered for preprocessing and handling activities in the text summarization process. In pre-preparing, three techniques are smeared i.e., tokenization, stemming, and word evacuation strategies are applied [12]. At taking care of phase of content outline, age of dictionary records happens. The presentation of programmed content synopsis was impacted by pace of appearance of words or expressions over the most recent couple of years. Later to build accuracy of results some more techniques were conceded into training with the standards system of content mining. Various reports can apply content rundown strategies simultaneously. The subject of the proceedings relies upon the quality and kind of classifiers. Meticulous content is made from number of records in Summarization. It is a rarity indeed conceivable to embody immense literary document. [15]. Likewise, in focuses utilized in for looking at all the records can't be perused. They essentially abridge reports and make up the synopsis of record from significant focuses.

In this big data era, most of the traditional research have being conducting in big data. Due to the huge amount of text data generation it is difficult to summarize the huge text document. To solve this, big data analytics are used to summarize the huge document, this procedure is called as “Big Data Text Summarization”. To know the depth of the articles published in the two areas text summarization and big data text summarization, the number of article published are collected based on those keywords. Figure 1 depicts the bar chart of the comparison of text summarization and big data text summarization publications. It is clearly indicated that big data text summarization are used in the recent 3 years. This represents the trend of big data in text analytics.

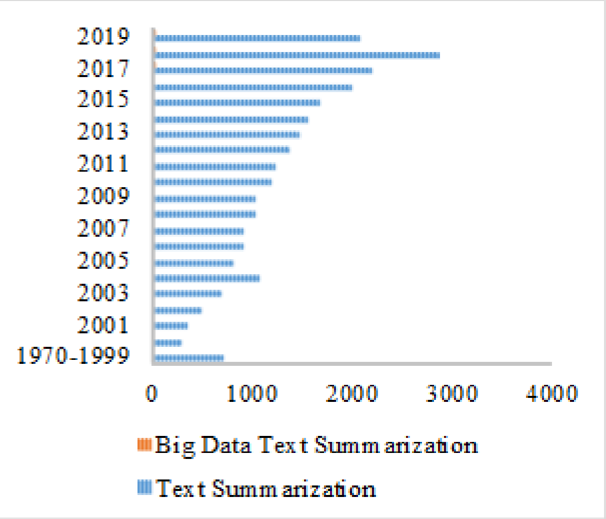

Fig. 1.Article comparison of text summarization and big data text summarization

Extractive and abstractive are two simple procedures for summarizing the structured or unstructured text data. In extractive strategy, a short form of content is framed. This is finished by the proper determination of significant sentences, and section and so forth from a printed report. In abstractive technique, the whole record is comprehended. In the wake of understanding the record, regular language is utilized to express the idea of the whole report. Semantic strategies are utilized to portray the whole record. Likewise, it helps in conveying noticeable quality of unique archive by making another content. Content outline makes it simple to sort the archive as indicated by client and furthermore checks on the off chance that it ought to be considered for additional data.

V. TEXT SUMMARIZATION TECHNIQUES

According to Sherry et. al [14], Automatic Text Summarization is a difficult issue and it increasing a lot of immensity significance nowadays. The programmed text summary implies a consequently summarized production when an information is given. Despite the fact that the exploration on the Automatic Text Summarization has been initiated in the year 1950 in Research Laboratories of IBM [15], in the ongoing years the field of Text Summarization has encountered an exponential development as a result of the Internet. It is extremely hard to physically outline the enormous records of the content in light of the fact that there is a huge measure of data on the Internet. Then again, Internet is a rich library that gives more data than is requested. Thusly, it is significant for looking through pertinent report through an enormous number of records accessible. The objective of content synopsis is abbreviate the source content into a lessened variant saving its data substance and in general significance.

[16] Document outline gives an instrument to quicker understanding the gathering of content archives and has various genuine applications. Semantic closeness and grouping can be used productively for creating viable outline of enormous content accumulations.

Abridging enormous volume of content is a difficult and tedious issue especially while considering the semantic closeness calculation in rundown process. Rundown of content accumulation includes concentrated content handling and calculations to create the synopsis. MapReduce is demonstrated condition of workmanship innovation for dealing with Big Data. In this paper, a novel structure dependent on MapReduce innovation is proposed for abridging huge content accumulation. The proposed system is planned utilizing semantic closeness based bunching and theme displaying utilizing Latent Dirichlet Allocation (LDA) for outlining the huge content accumulation over MapReduce structure. The outline undertaking is performed in four phases and gives a particular execution of various reports rundown. The exhibited method is assessed as far as versatility and different content outline parameters specifically, pressure proportion, maintenance proportion, ROUGE and Pyramid score are likewise estimated. The upsides of MapReduce system are plainly noticeable from the investigations and it is additionally shown that MapReduce gives a quicker usage of condensing enormous content accumulations and is an incredible asset in Big Text Data examination.

A. Abstractive summarization

Abstractive Summarization is a tactic for novel stating portraying the substance of the content which requires substantial hardware from characteristic language preparing, including punctuations and vocabularies for parsing and age. Deliberation methodologies give modern synopses and adjust well to high pressure rates while extraction methodologies are anything but difficult to adjust bigger sources even though the subsequent outlines might be disjointed.

Kiani et. Al. [17] had discussed the most import abstractive summarization methods in their article. The abstractive summarization techniques are as follows: Query Based and Generic Summarization, Bayesian Classifier, Hidden Markov Model, Neural Networks Based Text Summarization and Fuzzy Logic Based Text Summarization.

[18] With the experience of the web there is immense increment in the capacity of data. Practically all the data exists in advanced structure which lessens heaps of the desk work and expands simplicity of capacity. Looking through applicable data in gathering of archives is a monotonous errand. The arrangement comes in picture for this issue is programmed content rundown. In this paper, unique rundown age of enormous measure of reports for huge information is proposed which will consider client contribution as theme. The proposed framework is planned utilizing DBSCAN calculation which works with MapReduce structure for grouping and Hidden Markov Model for rundown. The synopsis procedure is performed in three primary stages and gives a secluded execution of enormous number of records for outline. The proposed strategy pursues preprocessing step in which records are filtered with similitude and different AI system are applied. The consequence of applying grouping improves the summarizer framework to gather careful words instead of duplicating excess words. Subject based dynamic synopsis from enormous information is testing task especially when there are various archives with same or distinctive substance. Hadoop with its programming systems can give better methods for producing outline and it additionally improves the multifaceted nature of rundown procedure utilizing dispersed registering.

B. Extractive Summarization

Extractive Summarization is a technique for deciding notable content units (regularly sentences) by taking a gander at the content unit's lexical and factual importance or by coordinating phrasal examples [19] Extractive Summaries are detailed by removing key content portions, for example, sentences or entries from the content, in view of measurable investigation of individual or blended surface level highlights [20].

[21] Reviews are significant origins of data for some, significant basic leadership undertakings. Abridging the monstrous measure of audits, which are accessible nowadays on numerous substances and administrations, is basic to enable clients to even more likely overview the gradient about an element or an assistance and its angles (for example highlights of the element or the administration). [22] This article displays a novel viewpoint-based outline system that creates a theoretical from different surveys of a substance without the requirement for a handmade component scientific categorization or any preparation information. It create outlines utilizing Natural Language Generation by considering the significance of angles, just as the relationship between them. It model these data as a tree, called Aspect Hierarchy Tree, in which hubs demonstrate the significant perspectives and edges show the connection between them. It propose and explore three elective substance choice and organizing models for the programmed development of an Aspect Hierarchy Tree in their proposed rundown system [23]:

i) Rhetorical model, which catches the angles' significance and relationship by taking a gander at the manner in which individuals talk about and relate the perspectives when communicating supposition in their surveys.

ii) Conceptual model, which endeavors a sound judgment learning base (for example Concept net) to locate the reasonable relationship between perspectives.

iii) Hybrid model, which endeavors both the logical and applied data.

The author had claimed that their abstractive rundown system can possibly execute one of the proposed models relying upon the application or apply every one of the three models and let a client pick the yield, contingent upon his/her craving to utilize the calculated, expository or the two wellsprings of data [24]. Quantitative and subjective examination on the subsequent Aspect Hierarchy Tree of the three substance choice and organizing models for seven elements in three spaces demonstrates that the three models produce Aspect Hierarchy Tree that contrast in intriguing manners with regards to terms of both substance (for example chosen viewpoints to be incorporated into the outline) and structure (for example the connection between viewpoints).

VI. CONCLUSION

Big Data Text Summarization has been one of the active areas of research for about last two years. Even though very less amount of work has been done in this area when compared to the text summarization. Extraction-based big data summarization, automatic generation of abstractive summaries from text documents is comparatively a new area which is not explored in much detail and depth. As far as in healthcare, social media, e-commerce, etc. text is concerned, often the documents are quite long and different from text of other genres. The need for automatic summarization of text documents and other text processing has been felt for long but focus of computer scientists has come only very recently. In this survey, it have attempted to give an account of different text summarization techniques giving special emphasis on the legal text summarization. It started with general definition of text summarization with a brief account of the recent works in the domain so that an unfamiliar reader can readily relate to the techniques that are used in legal text summarization. Specifically, it covered some state-of-the-art summarization techniques, the datasets and metrics they used, their performances and a comparative study. The same treatment was then followed for legal text summarization with techniques based on different categories like features, graph, rhetorical roles and classification. The summarization of legal text has some domain-specific issues like the structure of the documents, different terminologies and evaluation criteria for the task. Even though sometimes scores achieved in the legal text summarization are comparable and competitive to those in general domain counterpart, they are not consistent across the datasets. Unless some more research is done, it is difficult to have a comparative analysis. Also, the summarization techniques it were able to find in the literature are only extraction-based, coming from single documents only. It is imperative to explore whether abstractive summarization is possible or not in the legal domain. Also, although it is a far cry, there is certainly a need to have automatic categorization of the similar court cases and their verdicts. Multi-document summarization from the similar cases can provide the legal practitioners a brief yet holistic view of a court-cases or a quick chronological account of important milestones in a single case. There are several areas along with a plenty of issues therein where information science community can explore as far as legal text summarization is concerned.

REFERENCES

1. Hemnath (2010), “Web 4.0 - A New Web Technology”, http://website-quality.blogspot.com/2010/01/web-40-new-web-technology.html/.
2. Marcus, Cake (2008), “Web 1.0, Web 2.0, Web 3.0 and Web 4.0 explained”, http://www.marcuscake.com/economic-development/internet-evolution/.
3. Ron, Callari (2009), “Web 4.0, Trip Down the Rabbit Hole or Brave New World?”, http://www.zmogo.com/web/web-40trip-down-the-rabbit-hole-or-brave-new-world/.
4. Dan, Farber (2007), “From semantic Web (3.0) to the WebOS (4.0)”, http://www.zdnet.com/blog/btl/from-semantic-web-30-to-the-webos-40/4499/.
5. G. Soava “Development prospects of the tourism industry in the digital age”, Young Economists Journal / Revista Tinerilor Economişti, 12 (25) (2015), pp. 101-116
6. Alexander, B. (2006). “Web 2.0: A new wave of innovation for teaching and learning?” Educause Review, 41(2), 32.
7. Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., & Khan, S. U. (2015). Therise of“big data”on cloud computing: Review and open research issues.InformationSystems, 47,98–115.
8. Kwon, O., Lee, N., & Shin, B. (2014). Data quality management, data usage experienceand acquisition intention of big data analytics.International Journal of Information Management, 34(3), 387–394.
9. Lyu, K., & Kim, H. (2016). Sentiment analysis using word polarity of social media.WirelessPersonal Communications, 89(3), 941–958.
10. Carmen Luque, José M. Luna, Maria Luque and Sebastian Ventura, “An advanced review on text mining in medicine”, https://doi.org/10.1002/widm.1302
11. E. H. Hovy, Automated Text Summarization. The Oxford Handbook of Computational Linguistics, Chapter 32, pages 583-598. Oxford University Press, 2005.
12. I. Mani, D. House, G. Klein, The TIPSTER SUMMAC Text Summarization Evaluation. In Proceedings of EACL, 1999.
13. S.A. Babar, P. D. Patil, “Improving Performance of Text Summarization”, International Conference on Information and Communication Technologies, ICICT, 2014.
14. Sherry, P. Bhatia, “A Survey to Automatic Summarization Techniques”, International Journal of Engineering Research and General Science Volume 3, Issue 5, September-October 2015.
15. H. P. Luhn, "The automatic creation of literature abstracts”, IBM Journal of Research and Development, vol. 2, pp. 159-165, 1958.
16. N K Nagwani, “Summarizing large text collection using topic modeling and clustering based on MapReduce framework”, Journal of Big Data, 2015.
17. Kiani, F., and Tas, O. A survey on automatic text summarization. Pressacademia 5 (06 2017), 205–213.
18. Mr. Krushnadeo Tanaji Belerao and Dr. S. B. Chaudhari, “Summarization using Mapreduce Framework based Big Data and Hybrid Algorithm (HMM and DBSCAN)”, IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI-2017).
19. U. Hahn, & I. Mani, “The challenges of automatic summarization” IEEE-Computer, 33(11), 29–36, 2000.
20. V. Gupta, G.S. Lehal, “A survey of Text Summarization Extractive Techniques”, Journal of Emerging Technologies in Web Intelligence, Vol. 2, No.3, August 2010.
21. Shima Gerani, Giuseppe Carenini abd Raymond T. Ng, “Modeling content and structure for abstractive review summarization”, Published by Elsevier Ltd, 2016, http://dx.doi.org/10.1016/j.csl.2016.06.005
22. B. A. Mukhedkar, D. Sakhare, and R. Kumar. 2016. “Pragmatic analysis based document summarization” International Journal of Computer Science and Information Security, Vol. 14, no. 4, p. 145,.
23. Falguni N. Patel, Neha R. Soni. 2012.” Text mining: A Brief survey”, International Journal of Advanced Computer Research (ISSN (print): 2249-7277 ISSN (online):2277-7970) ,Volume-2 Number-4 Issue-6.
24. “Social media big data analytics: A survey”, Norjihan Abdul Ghani, Suraya Hamid, Ibrahim Abaker Targio Hashem and Ejaz Ahmed, Computers in Human Behavior, , https://doi.org/10.1016/j.chb.2018.08.039.

AUTHORS PROFILE

Anish Mathew Kuriakose is a Ph.D Research Scholar at Jairams Arts and Science College Karur affiliated to Bharathidasan University . He completed MCA and MBA from Bharathidasan University Tiruchirappalli. His areas of interest are Data mining, Mobile Communication and Big Data.

Dr. V.Umadevi M.Sc(CS &IT). M.Tech (IT)., M.Phil., PhD., D.Lit.Currently working as Director, Department of Computer Science, Jairams Arts and Science College, Karur.She is a reputed author of few journal publications and wrote three books.