International Journal of Future Generation Communication and Networking
Vol. 13, No. 4, (2020), pp. 4289–4294

Review of Recent Methods on Text Summarization and Dimensionality Reduction

Dipti Abhishek Bartakke --- Phd Student, Muit Lucknow
Dr Santosh Kumar --- Phd Guide, Muit Lucknow
Dr. Aparna Junnarkar---Phd Coguide, Muit Lucknow

Abstract.

Nowadays, the exponential growth of World Wide Web (WWW) leads the significant increase in online resources and hence the huge amount of data generated on Internet. Finding the relevant information from such enormous data is challenging tasks, thus the information retrieval (IR) becomes the more vital for searching the relevant data effectively. The Text Search Engines (TSE) returns the large number of pages which becomes very difficult task for end users to identify the relevant page. This process can be smoothening if documents are proved along with its short summary. The terminologies such as text search and text summarization (TS) are becoming the hot topics since from last decade for the researchers. The appropriate text summarization and dimensionality reduction of summarized text can leads to significant reduction in accessing time for the input requirements. TS the data mining process in which the original document is converted to the short version by fetching the key points from the document. The TS approach can be viewed as the Extractive Summarization and Abstractive Summarization. There are number of methods presented for the TS in literature, this paper presents the study on some recent works for TS. For TS, another approach called dimensionality reduction (DR) used to identify the relevant components from the document and remove the irrelevant text information from it. We present the review of some of the DR methods as well in this paper.

Keywords-: *Data mining, text search engine, text summarization, dimensionality reduction, information retrieval*

1 Introduction

In Human Summarization one has a tendency to abridge a single article by summing up the most vital thoughts and requesting to guarantee they are coherent. Notwithstanding for humans this undertaking would take a ton of work. Synopses created by two distinct individuals would normally be unique. Distinctive people may have an alternate view about what is imperative. This motivated the requirement for having an Automatic Summarizer that can play out the activity in less time and with the slightest exertion. This drove for the examination on Automatic Summarization to begin over 50 year's back [1].

Text Summarization when all is said in done is the way toward abridging a single article or an arrangement of related ones by summing up the most imperative occasions, ensuring the occasions succession is coherent by requesting them chronically. Then again automatic Text Summarization is the way toward delivering an abbreviated variant of a text by the utilization of computers [2]. The synopsis ought to pass on the key commitments of the text. In Automatic Summarization there are two primary methodologies that are extensively utilized Extractive and Abstractive. The primary strategy, the Extractive Summarization, removes up to a specific utmost the key sentences or passages from the text and requests them in a way that will create a coherent outline. The removed units contrast starting with one summarizer then onto the next. Most summarizers utilize sentences as opposed to bigger units, for example, sections [3] [4]. Extractive Summarization techniques are the attention strategy on Automatic Text Summarization. The other strategy, Abstractive Summarization, includes more language dependent tools and Natural Language Generation (NLG) innovation. Such summarizers can incorporate words not present in the first document [5]. The possibility of Abstractive Summarization

ISSN: 2233-7857 IJFGCN
Copyright ⓒ2020 SERSC

looks to emulate the Human Summarization strategies; however it is substantially harder to actualize. There have been different ways to deal with Automatic Text Summarization [6] [7]. Figure 1 and 2 shows the examples of single document and multi document text summarization respectively.

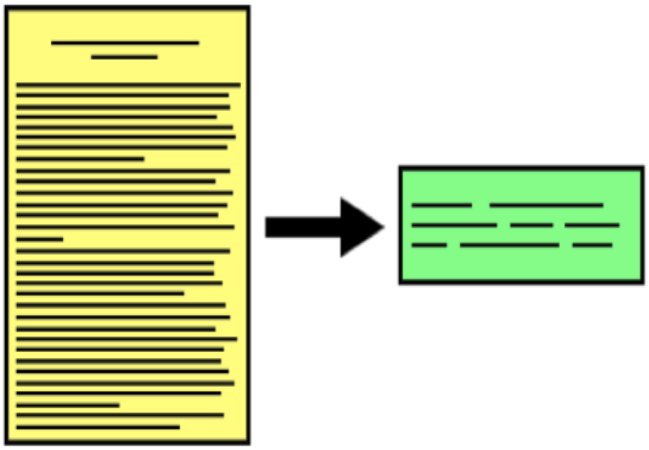

Fig.1. Single-Document TS

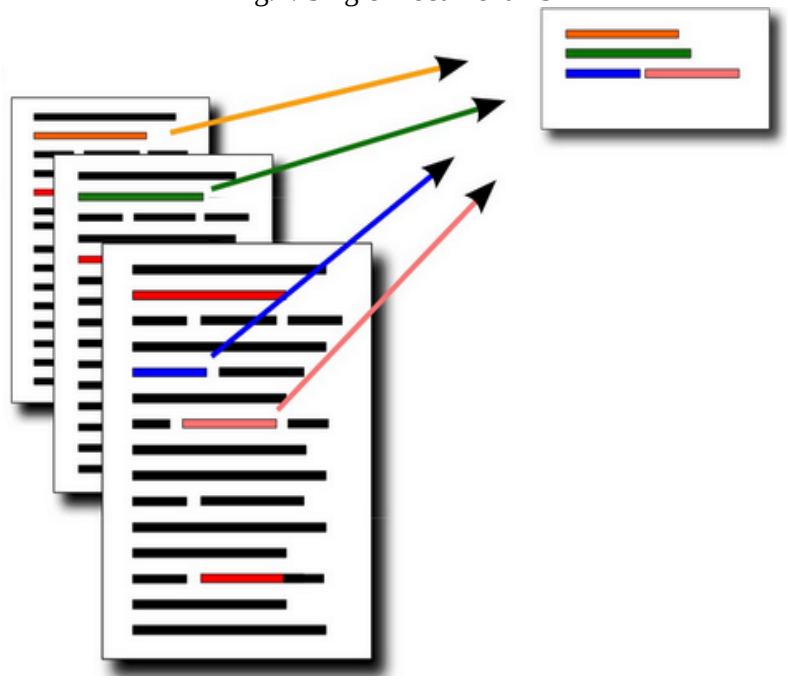

Fig.2. Multi-document TS

Text summery and dimensionality reduction of unique text is one part of data mining that extremely important to help pursuers effectively find essential information from a long text by shortening the length and substance of the first text [8]. Building an effective automatic text synopsis framework require various elements.

- First, a text synopsis framework must be extricated the most essential information from unique text.
- Second, this framework must be extremely viable and sets aside a short opportunity to indicate results.
- Third, cost to construct this framework isn't excessively.

Due to three issues above, analysts dependably endeavour to discover successful answers for assemble an arrangement of automatic TS and DR. For English text, there are numerous automatic TS techniques displayed in ongoing pasts. In this paper, we present the review of such methods those are reported very recently. After the review, we study the research gaps and challenges of current methods. In section II, the review of TS and DR methods presented. In section III, the research gaps identified and future directions discussed. Finally the conclusion presented in section IV.

ISSN: 2233-7857 IJFGCN
Copyright ⓒ2020 SERSC

2 Literature survey

In [1] author explained “A semantic approach for text clustering using Word Net and lexical chain” To beat this issue, presenting semantic data from ontology, for example, Word Net has been generally used to enhance the nature of content clustering. Regardless, there still exist a couple of challenges, for instance, comparable word and polysemy, high dimensionality, removing center semantics from text, and naming appropriate portrayal for the created clusters. In this paper, we report our undertaking towards fusing Word Net with lexical chains to alleviate these issues. The proposed approach mishandles ontology hierarchical structure and relations to give a more exact evaluation of the comparability between terms for word sense disambiguation. In addition, we familiarize lexical chains to evacuating a game plan of semantically related words from texts. This can speak to the semantic substance of the texts. Our joined way can perceive the subject of files in light of the disambiguated center highlights removed, and in parallel scale back the measurements of highlight space.

In [2] creator clarified "Extractive text rundown framework to help data extraction from full text in systematic review improvement" Objectives: Extracting information from generation reports is a standard technique in systematic review (SR) progression. Regardless, the extraction methodology still depends too much on manual effort which is direct, costly, and subject to human slip-up. In this examination, we developed a substance rundown system went for overhauling proficiency and reducing blunders in the customary information extraction process

In [3] creator clarified "Assessment of Automatic Text Summarizations Based on Human Summaries" the goal of this paper is to look at outlines made by different customized text synopsis strategies and those delivered by people. To achieve this end, we finished two courses of action of investigations: in the primary, we utilized normally made extractive outlines; in the second one, physically delivered rundowns gotten by a couple of English instructors were used. Our customized plots were procured using Fuzzy method and Vector approach.

Using Rouge assessment system, we looked at about the physically delivered outlines and the normally made ones. Rouge evaluation of made synopses showed the prevalence of outlines made by individuals over the thus produced rundowns. On the other hand, the examination between the made rundowns showed that synopses conveyed by Fuzzy methodology were generously more commendable and legitimate contrasted and outlines made by Vector approach.

In [4] creator clarified "Vertex Cover Algorithm Based Multi-Document Summarization Using Information Content of Sentences" In later, the need for the time of multi-record synopsis has grabbed a great deal of consideration among examiners. Multi-document synopsis systems center around delivering the compacted kind of the records which keeps up the applicable highlights of the genuine documents. Generally, text synopsis systems use the sentence extraction strategy where the striking sentences in the different reports are picked and displayed as a layout. In our proposed structure, we have developed a sentence extraction based multi-document rundown system using the standard of vertex cover algorithm which therefore picks imperative sentences that cover the primary thoughts of the information documents. This edge work speaks to the documents as a weighted undirected graph with sentences as the vertices and the similarity between the sentences as the edge weight between the looking at vertices in the outline. The preliminary outcome on the DUC 2002 data set demonstrates the practicality of the proposed procedure in document rundown.

In [5] creator clarified "Enhancing Performance of Text Summarization" gigantic data is available on the web; it is difficult to get the data brisk and by and large capably. There are such an extensive number of text materials open on the web, with a particular true objective to expel the most essential data from it, we require a better than average system. Text rundown strategy deals with the weight of expansive document into a shorter adaptation of text. Text synopses pick the hugest bit of text and make smart rundowns that express the essential purpose behind the given record. The extraction based text rundown incorporates picking sentences of high congruity (rank) from the record in light of word and sentence highlights and set up them together to make the layout. This is exhibited using Fuzzy Inference Framework. The rundown of the document is made upon the level of the hugeness of the sentences in the record. This paper centers around the Fuzzy rationale Extraction approach for the text rundown and the semantic approach of text outline using Latent Semantic Investigation.

In [6] creator clarified "Space Independent Framework for Automatic Text Summarization" Because of the exponential advancement of documents on the web, customers require all the vital information in one place with no issue. This incited the advancement of Automatic text Summarization. Consequently, different systems have been proposed by examiners yet no technique can manage all zones of text documents. A couple of methods which work for News space may flop in Medicinal area to give capable results. In this paper, we proposed a zone self-governing framework for Automatic Text Summarization. The Procedure at first requests the source text and from that point onward, it applies the individual grouping's optimal course of action of rules or weights or strategy. The major ideal position of the structure is that it might be appropriate for both extractive and abstractive text synopsis.

In [7] author explained “Enhanced continuous and discrete multi objective particle swarm optimization for text summarization” Analyzing down this enormous substance would be helpful in basic leadership for different partners. So text summarization frameworks turn out to be significantly in breaking down this tremendous substance. The summaries are produced in view of critical features utilizing multi-target approaches where adequate writing isn't accessible. Significant confinements of content outline frameworks are adaptability and execution.

In [8] writer clarified "Automatic Text Summarization of News Articles" the text Summarization has reliably been an area of dynamic eagerness for the scholarly community. Starting late, in spite of the way that couple of procedures have being made for the program text outline, capability is up 'til now a stress. Given the extension in size and number of records available on the web, a capable modified news summarizer is the need of awesome significance. In this paper, we propose a system of text outline which focuses on the issue of perceiving the most basic parts of the text and making cognizant rundowns. In our methodology, we don't require full semantic interpretation of the text; rather we make an outline using a model of subject development in the substance got from lexical chains. We show an upgraded and capable algorithm to deliver text synopsis using lexical chains and using the Word Net thesaurus.

In [9] author explained “Automatic Text Summarization Based on Multi-Agent Particle Swarm Optimization” Text summarization is the target extraction of a few sections of the text, for example, sentence and passage, as the report theoretical. In the event that there are documents with a lot of data, extractive text summarization would have emerged as an NP-complete issue. To take care of these issues, metaheuristic algorithms are utilized. In this paper, a technique in light of multi-agent molecule swarm enhancement approach is proposed to enhance the extractive text summarization. In this technique, every molecule will be overhauled with the status of multi-agent frameworks. The proposed technique is tried on DUC 2002 standard documents and broke down by ROUGE assessment programming.

In [10] author explained “A Systematic Study on Document Representation and Dimensionality Reduction for Text Clustering” In this examination, an orderly report is led to the utilization three Dimension Reduction Techniques (DRT) on three diverse record portrayal techniques with regards to the content grouping issue utilizing a few standard benchmark datasets. These three methods are connected on three Record portrayal strategies in light of the possibility of Vector Space Show, in particular word, term, and N-Gram portrayals.

In [11] author explained specific end goal to create document summaries that can meet reader necessities, this exploration builds up a programmed report summarization show that produces summaries based on singular prerequisites of reader. The report summarization issue is changed into a numerical issue by investigation of the quality components for the summaries and figuring of summaries quality records and constraints of quality factors. The proposed model can be utilized to appropriately summarize records by mulling over client prerequisites. Sooner rather than later, a web-based text summarization framework will be built up with a specific end goal to assess execution of the proposed methodology.

In [12] creator clarified "Enhancing text outline utilizing neuro-fuzzy approach" In the present propelled period, it transforms into a test for netizens to discover specific data on the web. Many web-based documents are recovered and it is hard to process all the recovered data. Modified text rundown is a technique that perceives the basic concentrations from all the related reports to make a compact synopsis. In this paper, we propose a text synopsis show in perspective of portrayal using neuro-fuzzy approach. The model can be set up to channel choice diagram sentences. We by then consider the

ISSN: 2233-7857 IJFGCN
Copyright ⓒ2020 SERSC

execution of our proposed show with the present procedures, which rely upon fuzzy rationale and neural network strategies. ANFIS showed upgraded results contrasted and the past systems with respect to ordinary exactness, review, and F-measure on the Document Understanding Conference (DUC) data corpus.

3 Recent method

This section presents the review of recent techniques presented for the automatic TS and DR approaches.

Table 1: Review of Recent Automatic TS Methods

Ref. No	Year	TS Methods	Dataset for evaluation
1	2015	DCS Method	DSC,ASG03,LMJ10 and CFS11
2	2016	Regex matching, Concept mapping, Supplement dictionary, Combined method.	N/A
3	2015	Fuzzy method and Vector approach	N/A
4	2015	NGD	DUC 2002
5	2015	Fuzzy summarization method	N/A
6	2015	Automatic Text Summarization	N/A
7	2018	Particle swarm optimization ,Swarm intelligence, Summarization	N/A
8	2017	Extractive Text Summarization, Lexical Chains, News Summarization, Natural Language Processing	lexical
9	2014	TF-IDF method	DUC 2002
10		Latent Semantic Indexing, Independent Component Analysis.	Classic3, NG, RD-256, RD-512, URCS.
11	2013	Automatic Text Summarization	N/A
12	2017	ANFIS	(DUC) 2002

4 Research gap

From the above studies, the task of TS is used with various purposes based on application demands such as document categorization, information retrieval etc. However, many of such methods revealed the potential of TS as the feature selection method. Further the DR technique applied to reduce the features selected by TS approach to enhance the accuracy of machine learning based approaches in various applications. Some applications used the TS to enhance the terms weighting and hence the classification performance. Thus, TS methods are significantly studied since from last decade. There are several methods reported for the automatic TS from the documents, however still many research problems yet to resolve. The challenges for TS and DR methods are:

- Accurate abstraction from the large document is main research problem.
- The recent tools available for the automatic TS, but due to increasing volume of online data, it becomes difficult to generate quick and meaningful abstractions.
- There are different types of TS methods reported under the three main categories such as abstractive, extractive and hybrid recently presented, but the challenge is how to trust or evaluate the extract summary.
- The correctness and verification of extracted summary should be performed using the approach evaluation methods

The quicker and cost effective abstraction from large pool of multi-documents is another challenge.

5 Conclusion and future work

The Text Summarization is nothing but the process of abstracting unique content from one or more documents sources is now becomes the important task of day to day life. However, designing the TS method to extract the accurate, reliable, and less information from single document or multi-document is challenging research problem. Generally the TS methods reviewed in two or three categories, but in this paper, we have presented the random and recent review of automatic TS methods. Further, the DR method is generally used for the irrelevant features removal after the task of TS. We discussed the key challenges and research gaps identified from the recent study. For the future work, we would like to initiate the process of designing the generalized TS method based on parallel computing frameworks.

References

[1] Tingting Wei, Yonghe Lu, **"**A semantic approach for text clustering using Word Net and lexical chain", Expert System with Application 42(2015),2015.

[2] Duy Duc An Bui PhD[a.b.*], Guilherme Del Fiol MD, PhD[a], **"**Extractive text summarization system to aid data extraction from full text in systematic review development", Journal of Biomedical Informatics 64(2016), 2016.

[3] Farshad Kiyoumarsi , "Evaluation Of Automatic Text Summarizations Based On Human" Procedia - Social and Behavioral Sciences 192 (2015) 83 – 9 .

[4] Ansamma Johna, "Vertex Cover Algorithm Based Multi-Document Summarization Using Information Content of Sentences" Procedia Computer Science 46 (2015) 285 – 291.

[5] S.A.Babara, Pallavi D.Patil, "Improving Performance of Text Summarization" , Procedia Computer Science 46 (2015) 354 – 363.

[6] Yogesh Kumar Meenaa,*, Dinesh Gopalanib, "Domain Independent Framework for Automatic Text Summarization" , Procedia Computer Science 48 (2015) 722 – 727 ,2015.

[7] V. Priya1 • K. Umamaheswari[2] "Enhanced continuous and discrete multi objective particle swarm optimization for text summarization", Received: 12 February 2018/Revised: 20 March 2018/Accepted: 22 March 2018 Springer Science+Business Media, LLC, part of Springer Nature 2018.

[8] Prakhar Sethi1, Sameer Sonawane2 "Automatic Text Summarization of News Articles", 2017 International Conference on Big Data, IoT and Data Science (BID) Vishwakarma Institute of Technology, Pune, Dec 20-22, 2017

[9] Hamed Asgari, Sameer Sonawane2, "Automatic Text Summarization Based on Multi-Agent Particle Swarm Optimization", 978-1-4799-3351-8/14/$31.00 ©2014 IEEE

[10] Evangelos E. Milios, M. Mahdi Shafiei,"A Systematic Study on Document Representation and Dimensionality Reduction for Text Clustering" , A Systematic Study on Document Representation and Dimensionality Reduction for Text Clustering.

[11] Jiang-Liang Hou, Yong-Jhih Chen "Development and Application of Optimization Model for Customized Text Summarization", Proceedings of the 2013 IEEE 17th International Conference on Computer Supported Cooperative Work in Design,2013.

[12] Muhammad Azhari & Yogan Jaya Kumar, " Improving text summarization using neuro-fuzzy approach", DOI: 10.1080/24751839.2017.1364040 ,2017.

ISSN: 2233-7857 IJFGCN
Copyright ⓒ2020 SERSC