Review of Text Reduction Algorithms and Text Reduction using Sentence Vectorization

Sneh Garg
C-DAC Mohali
India

Sunil Chhillar
C-DAC Mohali
India

ABSTRACT

The reduced text of a document is the collection of sentences that contains the important sentences containing keywords of the document. The authentic keywords extraction is the primary target for any text reduction algorithm. The presented survey shows the primary algorithm used for document summarization based on keywords. Also, the work presents a novel approach for keywords identification and in turn text reduction based on words histogram, the no. of sentences containing the words and knowledge corpus. The text summary is extracted using the sentence vectorization process. The sentence vectorization gives the sentences that have at least one of the key words in the sentence from the entire document. The algorithm works fine for the textual matter in the document in MS Notepad format. Factual information that is normally covered under double inverted comas is also given due attention in text summary.

Keywords

Text Reduction, Text summary, Sentence Vectorization, Word Histogram, Reduction algorithm, Synonyms

1. INTRODUCTION

Text summary is the reduced text from source document by extraction or generation. As a human point of view, a text summary is the human perception based on his text understanding. However, from the computational point of view, the text summary is based on logical quantification text features like keywords weightage, sentence ranking and phrase ranking etc.

Most of the text summarization is done using the same information obtained from the same document rather than the concept analysis/exploring of the document. Most of summarization methods extract keywords for document only that are written in document method. However, the study shows that the importance of synonyms and relevant terms of keywords is ignored most of the time during text summarization. Therefore, comprehensive gain in summarization is required for a quality text summary of the document considering the effects of synonyms as well.

The primary contents of a summary include the principal information in least amount of words or sentence or factual statements. Identification of main content from rest information is the major challenging job in summarization.

Basically, summary is the important information from original text and that is not more than half or one third of the original text. The target is to extract the useful information from a document in less space. All sentences in a document do not contribute in generating the text summary and are only language supportive. Therefore, the sentences may be given weightage and may be made part of the text summary depending upon the programmable or controlled weightage parameter. It is the key point for the reduction of text. As text and documents are growing day by day so it's a tedious task to extract useful information from such a large text and documents data base. Text reduction algorithm is an efficient way to get important information or brief summary of the whole document.

2. RELATED WORK

A knowledge based feature set is extricated using the Sentence coverage weight (SCW), corpus coverage weight (CCW) and term coverage weight (TCW). The text summary is computed using the above features. The performance evaluation is analyzed using precision and recall parameters. [1]

Rhetorical Structure Theory (RST) is proposed for text summarization using an analytic frame work. This framework considers text structure at the clause level. Extraction of text's rhetorical structure and relations among sentences are calculated. Less important segments are removed. [2]

Sentences are prioritized using their connective strength (CS) values. K-mixture probabilistic model is used to establish term weights in a statistical sense, and further identifies the term relationships to derive the connective strength (CS) of nouns. [3]

The sentences are assigned some feature for the summary called ranking sentences and then select the best ones. The first step in summarization is by extraction is the identification of important features. To improve the quality and extract important feature for each sentence HMM tagger [4] is used. Ranked sentences are collected by identifying important features and summary is generated. [4]

The ATS is the identification of most important sentences from the given text by identifying the prime features of the sentences properly. A Conditional Random Field (CRF) based ATS can be used to identify and extract the correct features. [5]

Existing summarization approaches inherently assign more weights to the important sentences, whose extractive summary approach predicts the summary sentences that are important as well as readable to the target audience with good accuracy. Neural network technique is used for summary extraction of science and social subjects in the educational text. [6]

Synonyms based approach is used when the text summary is not target oriented or is very less i.e. less than 5% of the whole document. It is imperative to note that the length of text summary may be in between 30-40% of the whole document. [7].

The comprehension of summary is increased by considering many aspects of source text. At the time of summary some key points must be remembered. Summary should be based on main idea as well as it should also cover sub ideas of source text. Important information should be covered in summary. Summary includes restricting of sentences [9].

Computer generated summary helps a lot in analyzing clinical data. Summaries save a lot of time taken in comparison of patient history. Quality of generated summary depends upon the quality of records of patient. for accurate and high quality summary records or input data should be proper. Summary generation in medical field is highly useful and beneficial [10]. Dynamic system is implemented by extracting features from the opinions obtained from web given by customers about the product .From extracted features opinion summary is built which helps buyers in making decision for the product. Features are extracted in two phases. In first part of speech tagging is done by attaching to a word its relative part of speech. In second phase domain specific features are extracted [11].

Automatic text summary's similarity measure is done by Latent Dirichlet Allocation (LDR) method . In LDA documents are represented randomly over topic . Each topic is attributed by the distribution of words. In this words,sentences,corpus ,documents are represented by vectors.[12]

3. ALGORITHMS REVIEW

A very common disadvantage in all discussed algorithms is that the synonyms of the keywords are not covered in text summary. Also in many algorithms, factual information is also ignored in text summary steps. Further, in text summary, some sentences are repeated due to different attributes/properties of keywords. Speedy processing of document is another important factor in text reduction algorithms performance. These disadvantages are taken care of in proposed algorithm as discussed in next section.

The literature studied for text reduction presents different algorithms based on different approaches. The main approach includes k-mixture probabilistic approach, Rhetorical Structure Theory (RST), feature terms based method for improving text summarization with supervised POS tagging, Hidden Markov Tagging (HMM) tagging method for ranking sentences and Conditional Random Field (CRF) based automated text summarization. A brief summary of the above discussed approaches are summarized in below table:

Sr. No.	Brief Summary
1	Single Document Text Summarization by Knowledge-Corpus **Brief Summary:** • Use of Knowledge corpus. • Three statistical measuring metrics. • Sentence Coverage weight, Knowledge coverage weight, term coverage weight.
2	Automatic Text Summarization Based On Rhetorical Structure Theory **Brief Summary:** • Use analytical framework RST at clausal level. • Extracts Rhetorical structure of text and relationship. • Use NLG model to produce summary.
3	A hybrid approach to automatic text summarization **Brief Summary:** • Use K-mixture probalistic approach to establish term weights. • Use connective strength (CS) of noun for term relations. • Sentences are ranked and extracted according to CS.
4	A Feature Terms based Method for Improving Text Summarization with Supervised POS Tagging **Brief Summary:** • Use Extraction approach. • Feature term identification. • HMM tagging method for ranking sentences.
5	CRF Based Feature Extraction Applied for Supervised Automatic Text Summarization **Brief Summary:** • Conditional random field based ATS. • Use trainable supervised method • Identify correct features. • Segment sentences based on selected features.
6	A Survey of Extractive and Abstractive Automatic Text summarization Techniques **Brief Summary:** Investigate techniques and methods used by researchers for automatic text summarization
7	Feature Based Summarization of Customers' Reviews of Online Products **Brief Summary:** • Opinion summarization of product • Features are extracted in two phase • Part of speech tagging is done • Domain specific feature are extraced

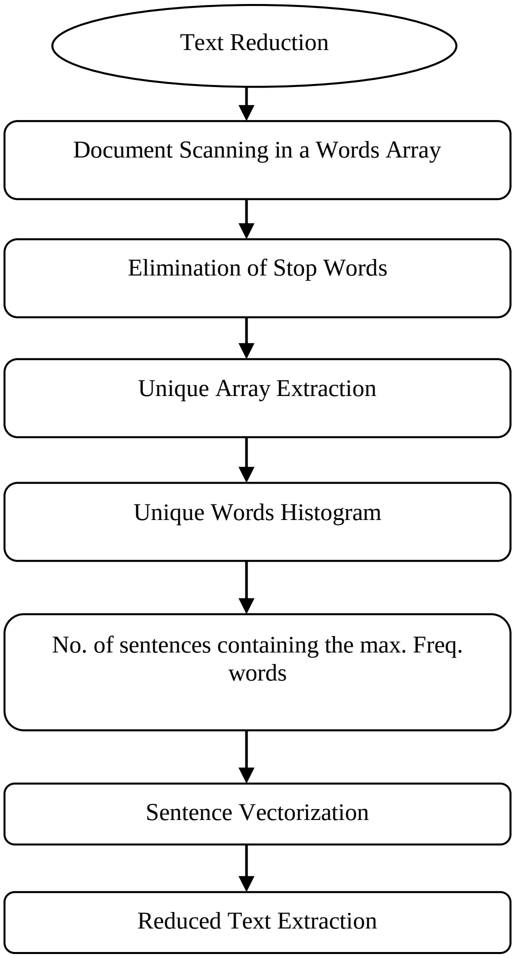
Sr. No.	Brief Summary
8	Quality of written summary texts: An analysis in the context of gender and school variables **Brief Summary:** • Summary should have main idea • Summary should have sub ideas. • Restructuring of sentences.
9	Data-to-text summarization of patient records: Using computer-generated summaries to access patient histories **Brief Summary:** • Summary save a lot of time. • Quality depends upon the quality of input. • Highly beneficial in medical field.
10	Feature Based Summarization of Customers' Reviews of Online Products **Brief Summary:** • Extract features from opinions. • Two phases for extraction. • Part of speech tagging. • Domain specific features.
11	The Similarity Measure Based on LDA for Automatic Summarization **Brief Summary:** • Similarity measure is done. • LDA method is used. • Documents are distributed randomly over topic. • Vector representation.

4. ALGORITHM

The proposed scheme starts with the document reading for words. An array of each word is generated and modified for its uniqueness. This means that each word appear only once. Primarily, the text summary is based on important key words that occur many times in the document and any factual information. It is observed that the document contains only few keywords and most of the text material is language supporting words and phrases. Therefore, before exercising for keywords extraction, common words, generally referred as stop words are eliminated using the string comparison method.

The filtered text array is exposed to keywords generation algorithm. The keywords are based on their frequency in the document and no. of sentences contacting the term. Further, the document title words are also considered in keyword category. Once a keyword vector set is derived, sentence vectorization process is performed.

In sentence vectorization, the keyword vector set is compared with each of the sentence in the document. The document that contains at least one of the keyword vector set entry, the sentence is put into the text summary array. The entire document is scanned using the sentence vectorization algorithm. Finally, the text summary is compiled by concatenating all the sentences obtained during the sentence vectorization process.


Text Reduction
Document Scanning in a Words Array
Elimination of Stop Words
Unique Array Extraction
Unique Words Histogram
No. of sentences containing the max. Freq. words
Sentence Vectorization
Reduced Text Extraction

5. RESULTS AND CONCLUSIONS

The proposed algorithm has been tested on number of documents in MS Notepad format documents. The text summary as obtained has found to be satisfactory. However, the algorithm works good on textual part, but the relevance of special character based information could not be accommodated in the text summary as the knowledge base cannot be generated using the special characters wordings. The summary length is of prime concern while deriving the text summary from the documents. The degree of text summary may be measured by taking the ratio of summary length to the document length i.e. words counts of summary and original document. Also, the factual text is made must part of the text summary so that no important information is missed out in the final text summary. The text summary may be bounded if the length of the text summary is prefixed. Further, the work is on using the keywords extraction based on synonyms so that the document could be scanned globally for the text summarization. The work may also be extended for tabular text material, for example comparison of different techniques or documents. If a text line contains more than one keyword, then it should be repeated for all the keywords in summary. This deteriorates the quality of summary due to repetitive texts.

6. REFERENCES

[1] Durga Bhavani Dasari, Dr. Venu gopala Rao. K., "Single Document Text Summarization by Knowledge-Corpus", 978-1-4799-1626-9/ 2013 IEEE.

[2] Li Chengcheng," Automatic Text Summarization Based On Rhetorical Structure Theory", 978-1-4244-7237-62010 IEEE.

[3] Te-Min Chang, Wen-Feng Hsiao," A hybrid approach to automatic text summarization", 978-1-4244-2358-3/2008 IEEE.

[4] Suneetha Manne, S. Sameen Fatima," A Feature Terms based Method for Improving Text Summarization with Supervised POS Tagging", International Journal of Computer Applications (0975 – 8887) Volume 47– No.23, June 2012.

[5] Nowshath K. Batcha, Normaziah A. Aziz," CRF Based Feature Extraction Applied for Supervised AutomaticText Summarization", Procedia Technology 11 (2013) 426 – 436.

[6] K. Nandhini, S.R. Balasundaram," Improving readability through extractive summarization for learners with reading difficulties", Egyptian Informatics Journal (2013) 14, 195–204.

[7] Alexander Yates, Oren Etzioni," Unsupervised Methods for Determining Object and Relation Synonyms on the Web", Journal of Artificial Intelligence Research 34 (2009) 255-296.

[8] Vipul Dalal, Dr.Latesh Malik,"A Survey of Extractive and Abstractive Automatic Text summarization Techniques", 978-1-4799-2560-5/2013 IEEE DOI 10.1109/ICETET.2013.31.

[9] Egitim Fakültesi, Mehmet Akif Ersoy," Quality of written summary texts: An analysis in the context of gender and school variables", 1877-0428 © 2010 Published by Elsevier Ltd.

[10] Donia Scott, Catalina Hallett, Rachel Fettiplace," Data-to-text summarisation of patient records: Using computer-generated summaries to access patient histories", D. Scott 156 et al. / Patient Education and Counseling 92 (2013) 153–159

[11] Kushal Bafna, Durga Toshniwal," Feature Based Summarization of Customers' Reviews of Online Products", 2013 The Authors. Published by Elsevier B.V.

[12] Tiedan Zhu, Kan Li," The Similarity Measure Based on LDA for Automatic Summarization", 2011 Published by Elsevier Ltd

7. AUTHOR'S PROFILE

Sneh Garg: The author is pursuing her M.Tech. (IT) thesis work in Text Mining from CDAC, Mohali, India.