IJournals: International Journal of Software & Hardware Research in Engineering
ISSN-2347-4890
Volume 4 Issue 11 November, 2016

TEXT SUMMARIZATION USING THEMANTIC HIERARCHY ALGORITHM

Aparna Khare
PG scholar, CSE department,
Truba Institute of Technology, Indore, MP.
Indore, MP.
aparnakhare.khare@gmail.com

Prof. Lalji Prasad
HOD, CSE Department,
Truba Institute of Technology,

lalji.prasad@trubainstitute.ac.in

ABSTRACT

In the past decade the special provision for retrieving specific data or required data from notes with huge amount of services like abstract of notes or paragraph, symbol or notes length etc.

In the advance age of technology there are number of application that has different facilities. But the main goal is design software containing all types of services including the abstraction feature. For better abstraction process we need a reliable and efficient approach by which one note automatically abstracted in their respective form according to user's choice like ½, 1/3 or 1/4.

This paper introduce the essential research challenges of the data mining algorithm implementation, analyses the abstraction feature including different other services. In our technology different approaches are included in single platform. In advance age of text summarization one more step is implemented with numerous number of services in one platform done through this research include the manipulation and rectify the data structure implementation on the basis of data mining and data warehouse concepts.

Finally, in this research paper, algorithm is implemented by "Themantic hierrarchy algorithm" which is more suitable approach can be used for text summarization in more advanced form.

General Terms

Text Summarization, Text Filter

Keywords

Data mining, data warehouse, artificial intelligence

INTRODUCTION

We all know that text summarizations are common now a days but basic need is abstraction of textual information. Every year we heard about many type of services included in text abstraction. But there is no provision is there by which user can select the abstraction filter level like half of text, one-third of text or one forth they want.

In the field of Natural Language Processing (NLP) the availability of online information has necessitated including intensive research in the area of automatic text summarization.

This survey intends to investigate some of the most relevant approaches in the area of summarization, giving special emphasis to empirical methods and extractive techniques. Some promising approaches that concentrate on specific details of the summarization problem are also discussed. Special attention is devoted to automatic evaluation of summarization systems, as future research on summarization is strongly dependent on progress in this area.

Summarization is a tuff task that requires understanding of the document content to determine the importance of the text. Lexical cohesion is a concept to identify connected portions of the text based on the relations between the words in the text.

The Lexical cohesive relations can be represented using lexical chains. Lexical chains are arrangement of semantically related words collection spread over the entire text. Lexical chains are used in variety of Natural Language Processing (NLP) and different types of Information Retrieval (IR) applications.

In current thesis, we propose a lexical chaining method based algorithm that is *"Themantic hierrarchy algorithm"* that includes the glossary relations in the chaining process. These relations are enable us to identify topically related concepts, for instance dormitory and student, and thereby enhance the identification of cohesive ties in the text.

The proposed algorithm consists of three stages. In the first stage, it detects the thematic hierarchy of a source text to decompose a source text into an appropriate number of textual units of approximately the same size. In the second stage, it adjusts each boundary between these textual units to identify a boundary sentence, indicating where a topic corresponding to a textual unit probably starts. It then selects a lead sentence that probably indicates the contents of subsequent parts in the same textual unit. In the last stage, it generates a structured summary of these sentences, thereby providing an outline of the thematic hierarchy of the source text.

We present methods and concepts that use the lexical chains to generate cluster of summaries by extracting sentences from the textual information. Headlines are generated by filtering the part of the sentences extracted, which is not matched with the meaning of the sentence. Headlines generated can be used in different real world application to float through the document collections in a digital library.

Document summarization is used in gaining demand with the drastic growth of online news sources. It requires identification of the different themes present in the collection to attain good compression and avoid duplicacy. In this thesis, we propose a special method to group the part of the text of a document collection into meaningful clusters or groups. Clustering enable us to extract the various concepts of the document collection or text collection. Sentences from clusters or different groups can then be extracted to generate a summarized form for the multi-document collection. Clusters or generated groups can also be used to design summaries with respect to a given query.

We designed a concept to compute the lexical chains for the given text and use them to extract the different portions of the document. Some specific tasks that are considered as follows:
Headline generation, and query-based summarization like use want to extract the task half, one third or one forth of original text. Our experimental evaluation shows that efficient summaries can be extracted for the above tasks.

LITRATURE SERVEY

Yihong Gong and Xin Liu have printed the concept of using LSA in text summarization in 2002 [1]. They, super impressed by the latent linguistics compartmentalization, applied in the singular value decomposition (SVD) to generic text summarization. The method or functionality starts with creation of a term by sentences matrix A = [A1, A2, A3 ..., A_n] with every column vector Ai, explaining the weighted term-frequency vector of sentence I within the document into different account. If there are collections of complete of m terms and n sentences within the document, then we'll have an m × n matrix A for the same document. Since each word doesn't unremarkably find in every sentence, the matrix A is thin.

Another paper is "Automated Text Summarization in SUMMARIST", designed by Eduard Hovy and Chin-Yew Lin ,Information Sciences Institute of the University of Southern California.

The target of a text summarizer is to provide an abstract of any textual document (or collection of documents) submitted there. The amount of sophistication of an abstract will vary from the easier list of isolated keywords that indicate the main content of the textual information, through an inventory of independent single sentence or multiple sentences that along with categorization of the main content, to a coherent, totally planned and generated text that compresses the textual information. The lot of subtle an abstract, the lot of effort it typically takes to provide.

Another paper is ―An Automatic Text Summarization Using Lexica Cohesion And Correlation Of Sentences. A.R.Kulkarni Computer Science & Engineering Department, Walchand Institute of Technology, Sholapur. Text report is that the method of designing a abstract version of original document. This condensed version should have unique content of the initial document. And analysis is being done since a few years to get coherent and indicative summaries using completely different techniques. Per (Jones, 1993) the text report is represented as 2 step method
i) Building a supply illustration from the initial document.
ii) Generating outline from the supply illustration.

Berzilay & Elhada [6] gave an improved formula that constructs all doable interpretations of the supply text using lexical chains. It‛s an economical methodology for text report as lexical chains determine and capture vital ideas of the document while not going into deep linguistics analyses.

Another paper is ―COMPENDIUM: A text summarization system for generating abstracts of research papers, By Dr. Elena Lloret,Prof. Dr. M Teresa Romá-Ferri. This article analyses the appropriateness of a text report system, COMPENDIUM, for generating abstracts of medicine papers. 2 approaches are suggested: Associate in Nursing extractive (COMPENDIUME), that solely selects and extracts the foremost relevant sentences of the documents, an abstractive-oriented one (COMPENDIUME– A), so facing conjointly the challenge of theoretical report. This novel strategy combines extractive info, with some items of knowledge of the article that are antecedently compressed or amalgamated.

Another paper is ―Extractive Text Summarization, By Namita Mittal, Basant Agarwal, Himanshu Mantri, Rahul Kumar Goyal and Manoj Kumar Jain‖ Text summarization also helps in reducing the length of a text whereas conserving its info content. In this paper, a text report approach is projected based mostly on removal of redundant sentences. Initially, every sentence from original text (input) is scored supported what quantity redundant the sentence is and at what extent that sentence is in a position to hide different sentences by itself. This approach is best effective on the documents that square measure extremely redundant and contain repetitive opinions regarding a topic.

Another paper is ―AUTOMATIC TEXTS SUMMARIZATION: CURRENT STATE OF THE ART By Nabil ALAMI, Mohammed MEKNASSI ― to facilitate the task of reading and looking info, it became necessary to realize a manner to scale back the size of documents while not affecting the content. The answer is in Automatic text account system, it allows, from an input text to result another smaller and additional condensed while not losing relevant data and that meaning sent by the original text.

PROPOSED METHODOLOGY

Thematic Hierarchy Detection

In the first stage, the proposed detect the thematic hierarchy of a text based on lexical cohesion measured by term repetitions. The output of this stage is a cluster of lists consisting of thematic boundary candidate sections (TBCS). The lists correspond individually to every layer of the hierarchy and are composed of TBCSs that separate the source text into thematic textual units of approximately the same size.

First, the algorithm calculates a cohesion score at fixed-width intervals in a source text. It is calculated based on the lexical similarity of two adjacent fixed-width windows (which are eight times larger than the interval width) set at a specific point by the following formula:

$$c(b_l, b_r) = \frac{\sum_t w_{t,b_l} w_{t,b_r}}{\sqrt{\sum_t w_{t,b_l}^2 \sum_t w_{t,b_r}^2}}$$

where bl and br are the textual block in the left and right windows, respectively, and wt;bl is the frequency of term[1] t for bl, and wt;br is the frequency t for br . Hereafter, the point between the left and right windows is referred to as the reference point of a cohesion score.

The algorithm then detects thematic boundaries according to the minimal points of four-item moving average (arithmetic mean of four consecutive scores) of the cohesion score series. After that, it selects the textual area contributing the most to every minimal value and identifies it as a TBCS.

CONCLUSION

The document summarization problem is a very important problem due to its impact on the information retrieval methods as well as on the efficiency of the decision making processes, and particularly in the age of Big Data Analysis. Though a good kind of text summarization techniques and algorithms are developed there's a requirement for developing new approaches to supply precise and reliable document summaries that may tolerate variations in document characteristics. Due to this all we prefer to implement the *"Themantic hierrarchy algorithm"* that includes the detailed relations in the chaining process and extract the text on the basis of user's choice.

REFERENCES

[1] Vishal Gupta, "A Survey of Text Summarization Extractive Techniques", JOURNAL OF EMERGING TECHNOLOGIES IN WEB INTELLIGENCE, VOL. 2, NO. 3, AUGUST 2010.

[2] Josef Steinberger, "Using Latent Semantic Analysis in Text Summarization and Summary Evaluation", Department of Computer Science and Engineering, Univerzitní 22, CZ-306 14 Plzeň

[3] Eduard Hovy and Chin-Yew Lin, "Automated Text Summarization in SUMMARIST", Information Sciences Institute of the University of Southern California 4676 Admiralty Way Marina del Rey, CA 90292-6695, U.S.A

[4] Canasai Kruengkari and Chuleer at Jaruskulchai, "Generic Text Summarization Using Local and Global Properties of Sentences", Proceedings of the IEEE/WIC international Conference on Web Intelligence (WI'03),2003.

[5] Morris, J. and G. Hirst ―Lexical cohesion computed by thesaurus relations as an indicator of the structure of the text‖. In Computational Linguistics, 18(1):pp21-45. 1991.

[6] Barzilay, Regina and Michael Elhadad ―Using Lexical Chains for Text Summarization. in Proceedings of the Intelligent Scalable Text Summarization‖ Workshop.(ISTS'97), ACL Madrid, 1997.

[7] Dr. Elena Lloret,Prof. Dr. M Teresa Romá-Ferri, COMPENDIUM: ―A text summarization system for generating abstracts of research papers‖, November 2013

[8] Namita Mittal, Basant Agarwal, Himanshu Mantri, Rahul Kumar Goyal and Manoj Kumar Jain, "Extractive Text Summarization", INPRESSCO ,2014

[9] Nabil ALAMI, Mohammed MEKNASSI , "AUTOMATIC TEXTS SUMMARIZATION: CURRENT STATE OF THE ART" , Journal of Asian Scientific Research, 2015

[10] R. C. Ho, C. Han Yang, M. A. Horowitz, and D. L. Dill, "Architecture validation for processors," *in Proc. Int. Symp. Comput. Architecture*, 1995, pp. 404-413.

[11] D. Geist, M. Farkas, A. Landver, Y. Lichtenstein, S. Ur, and Y. Wolfsthal, "Coverage-directed test generation using

symbolic techniques." *in Proc. Int. Conf. on Formal Methods in Comput.-Aided Des.*, 1996, pp. 143-158

[12] D. Moundanos, J. A. Abraham, and Y. V. Hoskote, "Abstraction techniques for validation coverage analysis and test generation," *IEEE Trans. Comput.*, vol. 47, pp. 2-14, 1998.

[13] A. Kuehlmann, K. L. McMillan, and R. K. Brayton, "Probabilistic state space search," *in Digest of Technical Papers of Int. Conf. on Comput.- Aided Des.*, 1999, pp. 574-579.

[14] C. H. Yang and D. L. Dill, "Validation with guided search of the state space," *in Proc. Des. Automation Conf.*, 1998, pp. 599-604.

[15] I. Wagner, V. Bertacco, and T. Austin, "StressTest: an automatic approach to test generation via activity monitors," *in Proc. Des. Automation Conf.*, 2005, pp. 783-788.

[16] M. Li and M. S. Hsiao, "An ant colony optimization technique for abstraction-guided state justification," *in Proc. Int. Test Conf.*, 2009, Paper 16.2.

[17] T. Zhang, T. Lv, and X. Li, "An abstraction-guided simulation approach using Markov models for microprocessor verification," *Des., Automation & Test in Europe Conf. & Exhibition*, 2010, pp. 484-489.

[18] M. Mitzenmacher and E. Upfal. (2005). *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*.

[19] R. Motwani and P. Raghavan. (1995). *Randomized Algorithms*.