An Algorithm for Summarization of Paragraph Up to One Third with the Help of Cue Words Comparison

Noopur Srivastava
Department of Computer Science
Shri RamSwaroop Memorial University
Barabank-225003, UP, India

Bineet Kumar Gupta
Department of Computer Science
Shri RamSwaroop Memorial University
Barabanki-225003, UP, India

***Abstract*—In the fast growing information era utility of technology are more precise than completing the assignment manually. The digital information technology creates a knowledge-based society with high-tech global economy which spreads over and influence the corporate and service sector to operate in more efficient and convenient way. Here an attempt was made on Extract Technology based on research. In this technology data could be refined and sourced with certainty and relevance. The application of artificial intelligence matched with the theories of machine learning would prove to be very effective. Sometime summarization of paragraph required rather than page or pages. So, Auto Summarization Model is an agnostic content summarization technology that automatically parses news, information, documents and many more into relevant and contextually accurate abbreviated summaries. This is a concept to convert a whole paragraph into one third. The Auto summarization technology reads a document, much better way than manually prepared, where, keywords and key phrases accurately weighted as they are found in the document, text or web page.**

***Keyword*—Data Mining; Data Warehouse; Artificial Intelligence**

I. Introduction

In present paper an attempt is made to introduce the essential research area of the data mining algorithm implementation and suggest important line on the basis of 'cue words'. Where, Auto Summarization uses a patented set of core algorithms to extract keywords and key phrases from any text-based document [24, 7]. In essence a machine learned method for reading or summarizing any text written in an electronic text format [27]. On the basis of cue word analysis one can select important lines from one paragraph [10, 12].

Auto Summarization is exceptionally good at content summarization incorporating its patented technology to summarize text, e-mail and html content into weighted lists of structuring of Web data and solve the problem about effectiveness in retrieval accordingly [7, 10].

Auto Summarization is exceptionally good in summarizing the text i.e. important part of the paragraph automatically without changing meaning of the paragraph. It will summarize text, emails, news, speech, etc into weighted lists of keywords and key phrases extracting the primary contextual sentence highlight of how the keyword / key phrase has been used [15, 29]. Uniquely positioned for web services, Auto Summarization is immediately capable of consuming documents of any length and subject matter, distilling the precise, contextual meaning of the content into keyword and key phrase summary formats. Extractor's unique patented technology delivers precise content summaries of any subject domain without retraining and without human intervention [29].

Auto Summarization is extremely effective for objectively distilling a document down to its key concepts providing users with highly focused keywords and key phrases including contextual examples of exactly how the keywords / key phrases have been used in the document - an extraction [20].

In contrast, a synopsis is a subjective interpretation of a document providing the end user with a high level statement of what that person believes the author intended the reader to comprehend. Such as an abstract Subjective is an important note - to date automated processes for generating a synopsis has not been perfected - and why they remain a human based process [12, 15].

II. The Aspects of Auto Summarization

Not just information but contextually accurate, relevant information is a critical tool for the success of business today. Rather than working through traditional, time consuming, and interactive search engine processes, incorporating Auto Summarization into Enterprise systems, empowers corporate information with relevant and meaningful representations meeting the needs of today's social workforce [7, 12]. Simple demonstration of auto summarization shows the summarization of notes from paragraph as shown in Fig. 1.

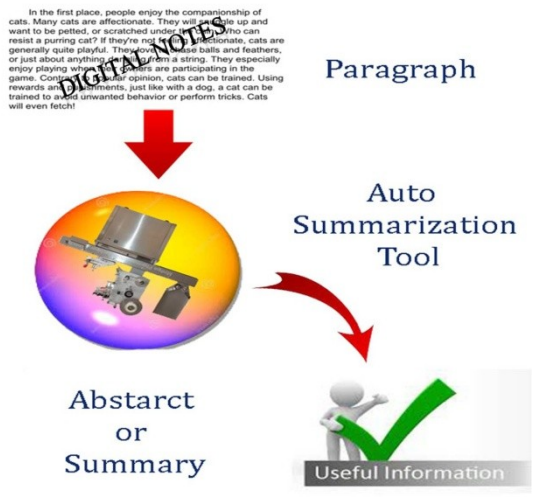

Fig. 1 Process of Auto Summarization

A. Artificial Intelligence Approach

It is based on fuzzy logic in which artificial intelligence tool also works. This technology is based on fuzzy logic and artificial intelligence [20, 9, 16]. Here artificial intelligence help to maintain the discipline in calculation of percentage through which one signify the actual subject percentage, while, fuzzy logic is to calculate the percentage of frequency.

With combination of both approaches that are fuzzy logic and artificial intelligence a unique formula prepared for the word, which help the user to identify the words of paragraph more accurately without wasting the time and money [20, 9].

In present research, summarization of cue words and maintaining frequency of each word and sentence by comparing it with cue word collections and stop word collections belongs under Artificial Intelligence approach.

B. Data Warehouse Approach

It is a central repository of data which is created by integrating data from one or more disparate sources [20]. Data warehouses store current as well as historical data and are used for creating trending reports for senior management reporting such as annual and quarterly comparisons [14, 25].

By this approach, proposed model collect the cue words means from the high frequency word store in respective subject at the time of adding paragraph of new subject. This process is going on for future reference when these words will match by auto arranger for final distribution of single note into different subject's collection [20]. Fig. 2 shows the collection or summarization of cue words and stop words, which maintain their frequency after comparison with each word, worked as Data Warehouse.

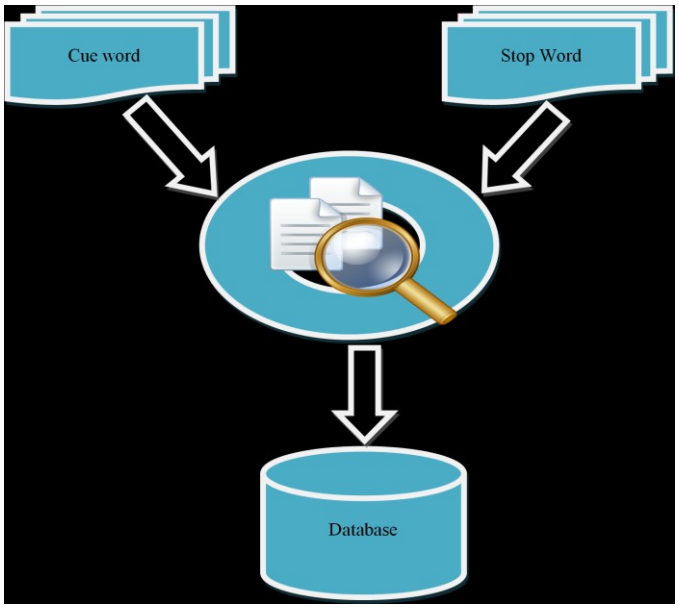

Fig. 2 Frequency of words in database

C. Data Mining Approach

Data mining is very helpful for extracting word or sentence after comparing each word from data warehouse. In present approach, comparison of cue words with new word in paragraph and make frequency table, is worked as Data Mining Fig. 3 [2, 28, 13]. Here the role of Data mining is to match the frequency of cue words to find out how many times the paragraph word is coming in the compared text. By this approach this model works more efficiently [5].

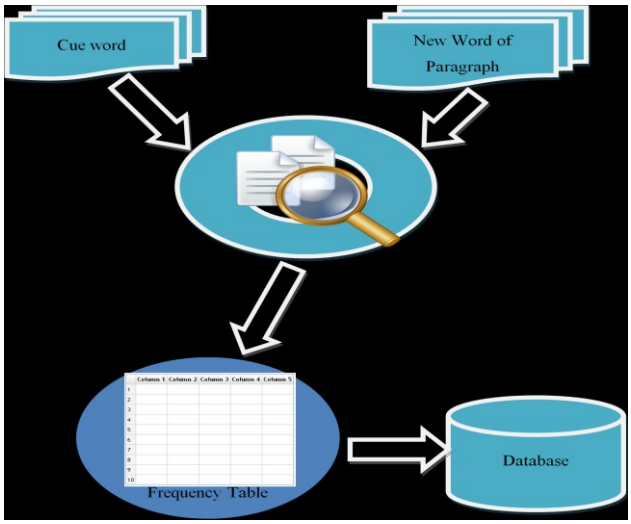

Fig. 3 Process of storing frequency table

D. Natural Processing Language (NLP)

A branch of Artificial Intelligence with analyzing understanding and generating the languages, which is used naturally in order to interface with computers in both written and spoken contexts using natural human languages instead of computer languages [17]. One of the challenges inherent in natural language processing is teaching computers to understand the way one can learn and use language. In the course of human communication, the meaning of the sentence depends on both the context in which it was communicated and each person understands the ambiguity in human languages [17, 4]. This sentence poses problems for software that must first be programmed to understand context and linguistic structures [6].

III. Proposed Technique

The proposed technique is based on NLP (Natural Language Processing) known as Gradual NLP algorithm. Automatic document summarization is an important research area in natural language processing (NLP). The technology of automatic document summarization is developing and may provide a solution to the information overload problem [17].

The process of summarization can be decomposed into three phases: analysis, transformation, and synthesis. The analysis phase analyzes the input text and selects a few salient features. The transformation phase transforms the results of the analysis into a summary representation. Finally, the synthesis phase takes the summary representation, and produces an appropriate summary corresponding to users' needs [2, 7]. In the overall process, compression rate, which is defined as the ratio between the length of the summary and that of the original, is an important factor that influences the quality of the summary. As the compression rate decreases, the summary will be more concise; however, more information is lost. While the compression rate increases, the summary will be larger; relatively, more insignificant information is contained. In fact, when the compression rate is 5–30%, the quality of the summary is acceptable as shown in Fig. 4.

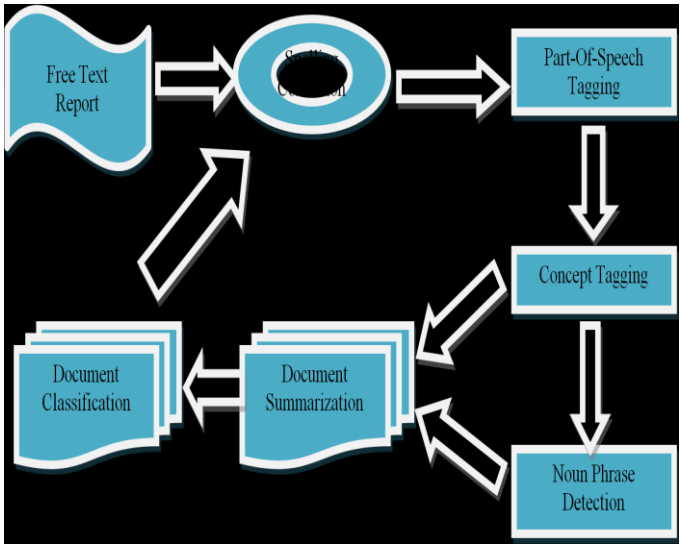

Fig. 4 Process NLP (Natural Language Processing)

A. Algorithm for Summarization

1) Select table word frequency (rs4), world list (rs5), stop word (rs3).

- Not stop word then frequency=frequency+1.
- Find total frequency and no. of words.
- Find average= total frequency/no. of words.
- If frequency used>average then store in rs5.

2) Open sentence (rs) and store each sentence with frequency 0.

3) Open cue word (rs6), basic id (rs7).

- If sentence contain number then freq= freq+1.
- If sentence contain cue word then freq=freq+1 for each word.
- If sentence contain wordlist then frequency=freq+1 for each word.
- If sentence not contain basic id then freq=freq+1 for each word.

4) Find total words in paragraph.

5) Find average length=total no. of words/no. in sentence

6) Find final score=score (average/length of sentence).*

7) Find total frequency=add all final score of the entire sentence.

8) Find average frequency= total frequency/no. of sentence.

9) If sentence frequency>avg frequency

- Extract that sentence.

10) Average frequency=total final score frequency/no. of sentence

11) Abstract summarized

B. Algorithm of summarization

Step 1: Implement Simple NLP algorithm in which first create 3 tables in the database with the names word frequency, word list, stop word, respectively. Then add a new paragraph in the summarization model. The words in the paragraph will be matched with the words available in the stop word. If the matched words are not stop word then the frequency will be increased by one. Then repeat the process for all the words in the paragraph and find the total frequency and count the number of words. In the last, find the average by using the formula given below:

Average = total frequency/no. of words.

Step 2: If the frequency used is greater than average then store and write the words in the word list table. Now, open the sentence from the paragraph that are added before and store in the newly created table sentence with zero frequency.

Now create a basic cue word table and store useful words. In "cue words" store the important words like noun, adjective and adverb. While, phrasing the sentence if found a number then increase the frequency by 1.

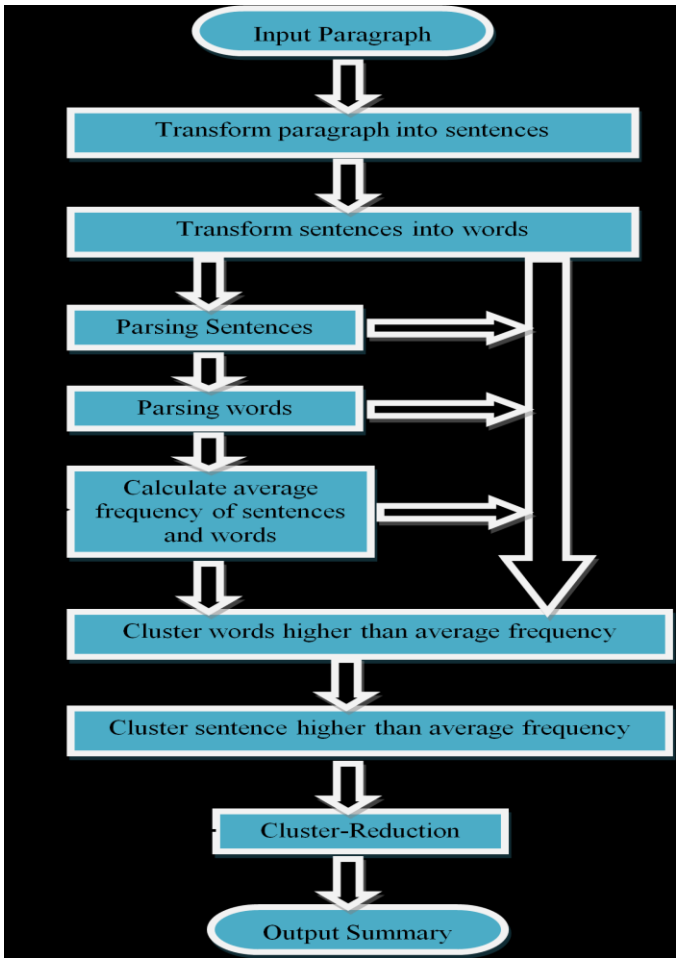

Fig. 5 Progressive representation of algorithm

If cue word is present in the sentence which is being phrased then increase the frequency by 1 of each word. If sentence contains wordlist then increase the frequency by 1 for each word. If basic is not found in the sentence, while phrasing then increases the frequency by 1.

Find the total number of words in the paragraph and average length of the sentence by using the formula below:

Average length = total no. of words/no. in sentence

Step 3: Find the Final score by using given formula:

Final score = score* (average/length of sentence)

Step 4: After receiving the final score of the sentence find the total frequency by using the given formula:

Total frequency = add all final score of the entire sentence

Step 5: Get the average frequency by using the given formula below:

Average frequency = total frequency/no. of sentence

Step 6: Compare the frequency of the sentence and average frequency. If frequency of the sentence is greater than average frequency, then extract that sentence.

Calculate the Average frequency again by using following formula:

Average frequency = total final score frequency/no. of sentence

Step 7: Check Final score if it is greater than average frequency then extract those sentences and show the final abstracted or summarized paragraph as in result form in front of user.

IV. Conclusion

In present work it is clear that not only information but contextually accurate, relevant information is a critical tool for the success of business today. Being able to source relevant information in context of the subject matter gives organizations an ultimate competitive advantage rather than working through traditional, time consuming, and iterative search engine processes, thus, incorporating extractor into enterprise systems. This empowers corporate information with relevant and meaningful representations, meeting the needs of today's social workforce.

In this context, paragraph break down into one third where one third part is abstract or summary for that whole paragraph. This helps to generate or convert whole paragraph into one third with highly important part as in the form of extract without violating the meaning of paragraph. Sometimes, it is required to extract a paragraph rather than whole page. So, this model is very effective and efficient to extract a paragraph. In this research we summarize only a paragraph but in future aspect we can summarize whole document into one-third. It is very much helpful for those students or people who cannot tell or express their knowledge. So, with the help of this technique they express their extract content of particular subject Auto Summarization is responsible for summarize the textual information approximately one-third valuable information for further decision support system or management information system. It can also be used for fetching important headline from the news. So, this model is very effective in retrieving the correct information and reduces the time complexity of the user.

Acknowledgment

This study is a part of my M.Tech final year dissertation conducted at the faculty of computer Science and Technology,

Shri Ramswroop Memorial University, Lucknow Deva road,Uttar Pradesh,India.

References

[1] A. K. Choudhary, J. A. Harding, and M. K. Tiwari, "Data Mining in Manufacturing: A Review Based on the Kind of Knowledge", Journal of Intelligent Manufacturing 20(5), pp. 501-521, 2009.

[2] B. Gupta and md. Hussain,"Algorithm to Evaluate the Rank of Research Papers Using Citation Graph in CiiT", International Journal of Data Mining Knowledge Engineering, ISSN NO 0974, Paper ID 3033, 2012.

[3] B. Park and H. Kargupta,"Distributed Data Mining: Algorithms, Systems, and Applications", 2002.

[4] B. Z. Manaris,"Natural Language Processing: A Human-Computer Interaction Perspective", University of Southwestern Louisiana, Academic Press, New York, vol. 47, pp. 1-66, 1998.

[5] C. Yu, K. L. Liu, W. Meng, Z. Wu, and N. Rishe, "A Methodology to Retrieve Text Documents from Multiple Databases. Knowledge and Data Engineering", IEEE Transactions, 14(6), pp. 1347-1361, 2002.

[6] D. Lewis, "Natural language processing for information retrieval, ACM, 39 (1), pp. 92-101, 1996.

[7] D. Das and A. F. T. Martins, "A survey on Automatic Text Summarization," Language Technologies Institute, Carnegie Mellon University, 2007.

[8] D. P. Ballou and G. K. Tayi, "Enhancing Data Quality in Data Warehouse Environments",Communications of the ACM, 42(1), 73-78, 2009.

[9] E. Charniak and D. McDermott, "Introduction to artificial intelligence", Addison-Wesley Longman Publishing Co., Inc., Boston, MA, 1985.

[10] Eduard Hovy and Chin Yew Lin, "Automated text summarization in SUMMARIST", MIT Press, pp. 81–94, 1999.

[11] G. Kundu, "Adapting Text Instead of the Model: An Open Domain Approach", Conference on Computational Natural Language Learning, 2011.

[12] I. Mani, "Automatic Summarization", John Benjamin's Publishing Co. pp.1-22, 2001.

[13] I. Taksa,"Research and Trends in Data Mining Technologies and Applications". Information Retrieval, 11(2), pp.165-167, 2008.

[14] J. M. Pe´rez, R. Berlanga, Marı´a Jose´ Aramburu, and Torben Bach Pedersen," Integrating Data Warehouses with Web Data: A Survey", Ieee Transactions On Knowledge And Data Engineering, vol. 20, no. 7, 2008.

[15] K. Jezek, and J. Steinberger, "Automatic Text Summarization (the state of the art 2007 and new challenges)", Znalosti , pp. 1-12, 2008.

[16] L. Suanmali, N. Salim and M. S. Binwahlan, "Fuzzy Logic Based Method for Improving Text Summarization", International Journal of Computer Science and Information Security, Vol. 2, No. 1, 2009.

[17] Md. M. Haque, S. Pervin, and Z. Begum, "Literature Review of Automatic Single Document Text Summarization Using NLP", International Journal of Innovation and Applied Studies ISSN 2028-9324 Vol. 3 No. 3, pp. 857-865, 2013.

[18] M. J. Pe´rez, R. Berlanga, Marı´a Jose´ Aramburu, and B. T. Pedersen,"Integrating Data Warehouses with Web Data: A Survey", IEEE Transactions on Knowledge and Data Engineering, vol. 20, no. 7, 2008.

[19] M. Lease,"Natural Language Processing for Information Retrieval: the time is ripe (again)", ACM, 2007.

[20] N. Srivastava, B. K. Gupta and N. K. Tiwari, "An Approach to Develop a Framework to Enhance the Performance of Digital Notes Based on Auto Arranger", International Journal of Engineering Research and Development, Vol. 10, Issue 4, pp. 53-57, 2014.

[21] R. Arora, P. Pahwa and S. Bansal, "Alliance Rules of Data Warehouse Cleansing. IEEE, International Conference on Signal Processing Systems, Singapore", 743–747, 2009.

[22] R. Mihalcea, H. Liu, and H. Lieberman, "Natural processing language for natural processing programming", Springer-Verlag Berlin Heidelberg , pp. 319–330, 2006

[23] R. Studer, V.R. Benjamins and D. Fensel," Knowledge Engineering: Principles and Methods. Data & Knowledge Engineering", 25(1-2), 161-197, 1998.

[24] S. Gholamrezazadeh, Mohsen Amini Salehi, and Bahareh Gholamzadeh , "A Comprehensive Survey on Text Summarization Systems" , IEEE, 2009.

[25] S. Chaudhuri, and U. Dayal," An Overview of Data Warehousing and OLAP Technology", ACM SIGMOD record, 26(1), pp. 65-74, 1997.

[26] S. E. Madnick, Y. W. Lee, R. Y. Wang and H. Zhu, "Overview and Framework for Data and Information Quality Research",ACM Journal of Data and Information Quality, 1(1), 2, 2009.

[27] S. Suneetha, "Automatic Text Summarization: The Current State of the art," International Journal of Science and Advanced Technology (ISSN 2221-8386), vol. 1, no. 9, pp. 283-293, 2011.

[28] T. L. Daniel, "Discovering Knowledge in Data. An Introduction to Data Mining", John Wiley & Sons, Inc., 0-471-66657-2, 2005.

[29] V. Gupta and G. S. Lehal, "A Survey of Text Summarization Extractive Techniques", Journal Of Emerging Technologies In Web Intelligence, Vol. 2, No. 3, 2010.

[30] X. Li, "A Supervised Clustering and Classification Algorithm for Mining Data with Mixed Variables". IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans, 6(2), 376-406, 2006.