A Survey on A hybrid method for query based automatic summarization system for text

Sannidhya Leuva[1], Assistant Prof. Pooja Jardosh[2]

[1] Student, M.E (CE), SOCET, GTU
Ahmedabad, Gujarat, India
sidleuva91@gmail.com

[2] Assistant Professor, M.E (CE/IT), SOCET, GTU
Ahmedabad, Gujarat, India

Abstract

Text summarization is the part of Information Retrieval system which comes under the area of Text Mining. This is the most popular application for information compression. Text summarization is a process of generating a summary by reducing the size of original document and concern important information of original document. There is arising a need to provide high quality summary in less time because in present time, the growth of data increases hugely on World Wide Web or on user's desktops so Multi-Document summarization is the best tool for making summary in less time. This paper presents a survey of existing techniques with the trinket highlighting the need of intelligent Multi-Document summarizer.

Keywords: *Text Summarization, extractive summary, generic, query based, automatic, single document, multi-document*

1. Introduction

Every person stores their data in mainly text format. At every place like government offices, financial company's data are being stored in mostly text format. Infect survey also says that the most of data (about 85%) are stored in text format by human. So text mining has large scope to get good and find better solutions. It is slightly complex and fuzzy work as it needs to be deal with unstructured data.

The process of text mining is the extraction of non-trivial and interesting data from the unstructured text. Text mining makes the use of different search techniques, but the difference between searching and text mining is that search method needs a user to know what he or she is looking for, whereas text mining attempts to find information in a pattern which is not known before [1].

Text Summarization comes under the area of information retrieval. It condenses the source text into a shorter version preserving its information content and overall meaning. It is very difficult for human beings to manually summarize large documents of text [2].

• Text summarization is having two main approaches:

Extractive Approach:

Extractive summarization uses statistical and linguistic features to determine the important features and fuse them into a shorter version. Extractive approach point outs the most important text from several documents and then fuse together and produce a summary

Extractive summaries do not focus on the understanding of text. It extracts the most important part based on statistical and linguistic features such as cue words, location, and word frequency

Abstractive approaches:

Abstractive-summarization understands the whole document and then generates the summary. While abstractive approach understands the source text and outputs the precise and concise summary by using linguistic methods and compression techniques [3]

Single document vs. Multi document summaries

Single document vs. Multi document summaries to generate a single output that summarizes the important points across multiple documents is more difficult. Since the documents are related by a common topic, they likely contain similar content.

Query based vs. Generic summaries

Automatic text summarization systems often produce generic summaries that show the most important points of a given text. However in the online search and retrieval context, a summarization system has access to the query entered by the user and should modify its output to suit the user's information need.

Mono lingual vs. Cross lingual

Cross lingual summarization require summarizing the documents when they are available in different languages. That time require a machine translation to generate the final summary in any one of the languages.

Text vs. Audio and Video

Important clips from audio and video can be extracted and summarized to give the list of the whole audio or video.

2. TEXT SUMMARIZATION

There are several ways in which one can characterize different approaches to text summarization. We present two possible classifications of text summarization systems can be found from literature but not all. The first classification is based on the goal of text summarization, follows [7]. The second proposed in [8], is based on characteristic of text summarization. The third and the last classification summarized by [9] are based on the level of processing and the kind of information

2.1 Goal of Text Summarization

Usually describe in terms of certain key features which relate to the concepts of intent, focus, and coverage. **Intent** describes the potential use of the summary. It can be

Classified into three types:

Indicative: Indicative summaries, provide just enough information to judge the relevance of the full text, use to alert the user as to what the source is about and decide to continue read the full source.

Informative: Informative or substantive summaries serve as acting for the full documents, keeping all important details

Evaluation: Evaluative summaries express the point of view of the author on a given topic.

Focus refers to the scope of the summary, either generic or user directed.

A generic summary is based on the main concept of a document, while directed summary is based on the topic of interest by the recipient of the summary.

Coverage indicates the summary is based on a single document or multiple documents.

III. Related Works

Currently, most successful multi-document summarization systems follow the extractive summarization framework. These systems first rank all the sentences in the original document set and then select the most salient sentences to compose summaries for a good coverage of the concepts. For the purpose of creating more concise and fluent summaries

Authors [5] this proposed system on a graph-based method to summarize documents based on user's query. The proposed method includes two stages, the offline and online stages. In the offline stage, pre-processing tasks are performed. Give a document set which needs to be summarized, first, all stop words are removed from the sentences. Then, the documents are decomposed into a set of paragraphs. Each node of the graph represents a paragraph. An edge is added between two nodes if they are semantically related. If two nodes share common words, they are related. The similarity score between two nodes is calculated using the TF-IDF method. The similarity score is considered as the weight of the edge between two nodes. Finally, the nodes of the graph are clustered using the AHA approach (Davidson and Ravi2005) and nearest neighbor algorithm (Shekhar and Xiong2008) to reduce the processing time during the online stage.

At the online level, first, the similarity measure between each cluster and query is calculated using the okapi equation which is based on TF-IDF (Varadarajan and Hristidis 2006). Second, minimal clusters are identified. Minimal clusters are the clusters which are related to the input query and the weight of the edge between a cluster and the input query is non-zero.

These minimal clusters are shown in the result. For this purpose, the top-n clusters with the highest weight in relation to the input query are displayed.

Authors [6] the author introduces a algorithm that can summarize a document by extracting key text and attempting to modify this extraction using a thesaurus, they reduce a given body of text to a fraction of its size, maintaining coherence and semantics, they focus basically two types its extractive and abstractive first they apply Extractive summarization technique then improved further by replacing a few parts of it using an abstractive technique.

They used a text-ranking algorithm they use d WorldNet tool for abstract the generated summary its lexical database and also use the NLTK for access the database through the program.

Authors [7] This paper author survey about different types of method like Graph based, Cluster Based, Time Based and Term frequency - Inverse document frequency Based etc. The survey starts introducing Multi-document text Summarization (MDS) and then discusses various methods of MDS which fall under the Graph and Cluster Based methods. In this paper, they have analyze Graph and Cluster Based methods proposed by various researchers in the field and they sort out some of the problems in applied procedures and also pin out advantages, which would help future researchers working in the area, to get significant instruction for further analysis. Using this information one can generate new or even hybrid methods in Multi-document summarization.

Authors [8] the paper discussed they introduce Text summarization is the part of Information Retrieval system which comes under the area of Text Mining. A general format for storing data is text which is easy but unstructured. Text mining deals with the unstructured data and finds the interesting data. Text summary is important now a days for online library system that stores newspapers, books or/and magazine. Query based text summarization is process of generation of summary where each sentence in the summary is chosen as per the user given query. To generate a query Based text summary, sentence scoring is most important process at a whole. Statistical and linguistic approaches are followed for sentence scoring. Here to combine both and applying weighted average on each sentence scoring method will improve the results in comparison with simple average of those sentence scoring method. here the Sentence scoring can be done using statistical techniques and /or linguistic technique they introduce the hybrid method for the text summarization they method work based on the query, sentence scoring method sentence clustering, sentence ordering methods and they make proposed sentence scoring method=(a+b+c)3+d)/2 where a, word form similarity b. N-gram based similarity c. Word order similarity and for linguistic technique they use d. semantic similarity

Authors [9] in this paper the input text document are divide in the two parts 1) informative and 2) non informative and after summarizing and simplifying them individually. They use NLTK (Natural language tool kit) tagger to tag the words and get their parts of speech. The all nouns are simplified by WorldNet. In order to summarize and simplify non-informative sentences keyword selection approach used after both file combine together to obtain output file. They are used the Grammar rules for reduce the length of the non-informative sentences and for informative sentence it noun simplified via World Net.

Authors [10] this paper discusses the development of multi-document summarization for Indonesian documents by using hybrid abstractive-extractive summarization approach. Multi-document summarization is a technology that able to summarize multiple documents and present them in one summary. The method used in this research, hybrid abstractive-extractive summarization technique, that is the combination of WordNet based text summarization (abstractive technique) and title word based text summarization (extractive technique). After an experiment with LSA as the comparison method, this research method successfully generated well-compressed and readable summary with a fast processing time. in the methodology first they input the set of clustered document after the document should discuss the same topic and also have same category. they use the WordNet lexical database that contains word meanings and its semantic relations they did concatenation after the input set using the pre-processing remove unnecessary word and stop word and do tokenization and did future scoring for paragraph high score and after apply Feature ranking and extraction.

Authors [11] in this paper the authors introduce the different method regarding to the extractive and abstractive text summarization. In The extractive summarization use the statistical and linguistic features to determine the important features and fuse them shorter version whereas the abstractive summarization understands the whole document and then the generates the summary. The Extractive summaries maintain the redundancy by extracting the relevant features from the document. there are different method for the extractive summarization like Team Frequency ,cluster based method, graph theory approach, machine learning approach, LSA approach, neural networking in text summarization, automatic text summarization on fuzzy logic, multi document extractive summarization, query based extractive summarization and other abstractive techniques is structure based approach and semantic based approach.

Authors [12] In this paper is cluster based approach similar document into cluster after then sentences from every document cluster are clustered into sentence clusters Best scoring sentences from sentence cluster are selected in to the final summary.

Here the find similarity between each sentence & query using cosine similarity measure

IV.CONCLUSION

Web is growing rapidly, but on the other hand the user’s capability to access Web content remains constant. Currently, Web personalization is the most promising approach to alleviate this problem and to provide users with tailored experiences. Web-based applications (ex, e-commerce sites, e-learning systems, etc.) improve their performance by addressing the individual needs and preferences of each user, increasing satisfaction of user. In this paper, we discussed Web personalization as one of the solutions to this problem, which makes use of Web usage mining. Summarizing, in this paper we explored the different faces of personalization.

V. REFERENCES

[1]. Recommender Systems Handbook Francesco Ricci · LiorRokach · BrachaShapira · Paul B. Kantor .Springer.

[2]. D. Das and A. F. Martins, "A survey on automatic text summarization," Literature Survey for the Language and Statistics II course at CMU, vol. 4, pp. 192-195, 2007.

[3]. Bridge, D., G¨oker, M., McGinty, L., Smyth, B.: Case-based recommender systems. TheKnowledge Engineering review 20(3), 315–320 (2006).

[4]. Intelligent Decision and Policy Making Support Systems edited by Da Ruan, Frank Hardeman, Klaas van der Meer

[5]. Yoon Ho Cho, Jae Kyeong Kim, SoungHie Kim. 2002.A personalized recommender system based on web usage mining and decision tree induction. Expert Systems with Applications 23, 329–342.

[6]. Feng Hsu Wanga, Hsiu-Mei Shao. 2004. Effective personalized recommendation based on time-framed navigation clustering and association mining. Expert Systems with Applications. 27, 365–377.

[7]. Baoyao Zhou, Siu Cheung Hui, Kuiyu Chang. 2005. A Formal Concept Analysis Approach for Web Usage Mining. Intelligent Information Processing II IFIP International Federation for Information Processing. 163, 437-441.

[8]. Mohamed KoutheaïrKhribi, Mohamed Jemni1 and OlfaNasraoui. 2009. Automatic Recommendations for E-Learning Personalization Based on Web Usage Mining Techniques and Information Retrieval. Educational Technology and Society, 12 (4), 30–42.

[9]. Harita Mehta, ShvetaKundra Bhatia, PunamBedi and Dixit, V., S. 2011. Collaborative Personalized Web Recommender System using Entropy based Similarity Measure. IJCSI International Journal of Computer Science. 8(6),3, 1694-0814.

[10]. Suneetha, K., and Usha Rani, M. 2012. Web Page Recommendation Approach Using Weighted Sequential Patterns and Markov Model. Global Journal of Computer Science and Technology. 12(9).

[11]. Qinjiao Mao, BoqinFeng, Shanliang Pan,. 2013. Modeling User Interests Using Topic Model. Journal of Theoretical and Applied Information Technology. 48(1).

[12]. R. Suguna., D. Sharmila. 2013. An Efficient Web Recommendation System using. Collaborative Filtering and Pattern Discovery Algorithms. International Journal of Computer Applications (0975 – 8887). Volume 70– No.3, May 2013. 37.