A SURVEY ON TEXT SUMMARIZATION USING OPTIMIZATION ALGORITHM

Raj Kumar V.S  
M.E Computer Science,  
Kumaraguru College of Technology,  
Coimbatore,  
India.  
rajkumar.rbd@gmail.com

Chandrakala.D  
Associate Professor,  
Department of Computer Science and Engineering,  
Kumaraguru College of Technology,  
Coimbatore,  
India.  
chandrakala.d.cse@kct.ac.in

ABSTRACT:

Text Summarization is one of the fastest growing research areas because of availability of huge amount of information in an electronic form. Text summarization is the process of condensing a large amount of information in one or more documents into a concise form of summary which contain informative textual units that are relevant to the user. To generate effective summary and organization of information, optimization algorithms such as particle swarm optimization, artificial bee colony algorithms, genetic algorithms and ant colony optimization are used. These optimization algorithms can be applied in both single document summarization and multiple documents summarization. In this paper, a survey on text summarization using optimization algorithms has been presented.

Keywords: Differential Evolution Algorithm, Genetic algorithm, Latent Semantic Analysis, Modified Particle Swarm Optimization Algorithm

1. Introduction

Today’s world is all about information, most of it present in online. The World Wide Web contains billions of documents and is growing at a high rate. Tools that provide timely access to various sources are necessary in order to mitigate the information overload people are facing. These concerns have sparked interest in the development of automatic summarization systems. Automatic document summarization is the process of automatically creating a concise version of a text containing the main content of the original document.
[1] Generally, automatic document summarization methods are divided into two categories of abstractive and extractive. Abstractive method can be described as “reading and understanding the text for recognizing its content which is then compiled into a brief text”. Extractive method provides a summary which consists of units of text taken from the source and presented exactly. In fact, majority of researches have been emphasized on summary extraction, which selects pieces such as sentences, keywords or even paragraph from the source to generate a summary.

Based on the amount of documents to be summarized, the summary can be a single-document or a multi-document. Single-document summarization can condense only one document into a shorter representation, whereas multi-document summarization condenses a set of documents into a summary. It is impossible to summarize a large number of documents, without using automatic summarization and another problem faced is redundancy in the information presented by the multiple documents.

[3] Text summarization is the process of generating summary which conveys information present in the one or more documents and the summary is less than the original documents. Text summarization also produces summary that reduce the redundancy in the text thereby preserving the information. In short summarization must satisfy the following three important optimization properties:

(1) Summaries should contain factual text units that are relevant to the user.

(2) Summaries should not contain multiple textual units that convey the similar information.

(3) Summaries are bounded in length.

[1] The classical optimization techniques such as Lagrangian Relaxation, branch and bound and dynamic programming techniques are available for generating an effective summary that satisfy the optimization properties and improve the performance of text summarization. These classical optimization methods face difficulty in handling dimensionality problem in integer-programming and also computation time increases rapidly with the size of a problem. Therefore, artificial intelligence
techniques such as particle swarm optimization algorithms, genetic algorithms, artificial bee colony and ant colony optimization algorithms which are population based searching techniques and also search for the global or near global optimal solution for large-scale integer-programming problems. The artificial intelligence techniques are used to improve the relevancy score and computational cost of the text summarization process.

2. Literature Survey

Rasim M. Alguliyev et al. [1] proposed an optimization-based unsupervised approach to automatic document summarization. Boolean programming problem model was used for text summarization. This model generally attempt to optimize three properties, namely, (1) relevance: summary should contain factual text units that were relevant to the user; (2) redundancy: summary should not contain numerous textual units that convey the same information; and (3) length: length of the summary has to be within a limit. It is applicable for both single- and multi-document summarization tasks. In both tasks, the document is separated into sentences in preprocessing. In this method, salient sentences are selected from the documents to accumulate into a summary. Eventually, by threading all the selected sentences in their order of appearance in the actual documents, the summary is generated.

The author proposed a differential evolution (DE) algorithm which is established on the framework of genetic algorithms for solving the optimization problem. DE algorithm is used to improve the performance of the summarization process. This method is compared with existing summarization methods such as iRANK, TranSumm and Biased LexRank on an open DUC2005 and DUC2007 data sets, it was reported that this method improves the summarization results significantly. The author demonstrated that the summarization results depend on the similarity measure. The genetic algorithm consumes more computational time for generating the summaries in case of presence of multiple documents with more sentences. And also genetic algorithm suffers from local convergence problem.

Maolong Xi, Jun Sun, Wenbo Xu [2] proposed an improved quantum-behaved particle swarm optimization (QPSO) algorithm with weighted mean best position that works according to fitness values of the particles. The author states that the improved
QPSO has faster local convergence speed, results in a balanced global and local searching of the algorithm, and thus generating good performance. In this, the author compared weighted QPSO (WQPSO) algorithm with QPSO and standard PSO and also tested on several benchmark functions such as Sphere, Rosenbrock, Rastrigrin, Griewank and De Jong’s functions to show the superiority of WQPSO. An efficient parameter control method is required to enhance the performance of QPSO.

S.A.Babar, Pallavi D.Patil [3] focuses on improving the performance of text summarization. Extraction based text summarization involves selecting sentences with high relevance from the document based on sentence and word features and put them together to generate summary which is modeled using Fuzzy Inference System. Based on the level of the importance of the sentences in the document, the summary of the document was created. In this, the author proposed the Fuzzy logic Extraction approach for text summarization and the semantic approach of text summarization using Latent Semantic Analysis. This method improves the quality of summary by imposing the Latent Semantic Analysis into the sentence features extracted fuzzy logic system for extracting the semantic relations among the concepts in the original text. This method has a limitation in single document summarization, which is to be extended for multi document summarization with a large datasets and domain specific data.

You Ouyang, Wenjie Li, Sujian Li, Qin Lu [4] suggested a regression models to the sentence ranking problem in query-focused multi-document summarization. Regression models were implemented using SVR (Support Vector Regression). SVR is a regression type of Support Vector Machines which is capable of building state-of-the-art optimum approximation functions. The learned function is used to predict the importance of the sentences in the test data. The learned regression functions works better than the linear combinations containing manually assigned weight. The author conducted series of evaluations on the DUC data sets efficiency and the robustness of the proposed approaches and also compared with classification models and ranking models to show that the regression models are consistently preferable. This approach was impractical to precisely annotate sentence importance manually.
J. Tang, L. Yao, D. Chen [5] suggested a multi-topic based query-oriented summarization. Query-oriented summarization (QS) tries to extract a summary for a given query which is a common task in many text mining applications. The author conducts a thorough investigation on the problem of multi-topic based query-oriented summarization and proposed a probabilistic approach to solve the problem. A statistical topic model is suggested by the author is used to discover multiple topics in a document collection.

The author suggested two strategies for integrating the query information into the topic model. The first strategy, the query information is directly integrated into the generative process of the topic model. Thus the model estimates a mixture of a query-specific topic distribution and a document-specific topic distribution. Another strategy is to use a regularization form to constrain the topic model by using the query information. This approach has a limitation in capturing the dependencies between document contents and the query to obtain query oriented topic distribution of the document cluster.

Rasim Alguliev, Ramiz Aliguliyev [6] suggested an evolutionary algorithm for extractive text summarization. In this, the author demonstrated an extractive text summarization method which is based on creating the summary by extraction and clustering sentences from the original document. Similarity measure plays an increasingly important role in document clustering. New criterion functions for sentence clustering have been offered for this purpose. The author developed a discrete differential evolution algorithm to optimize the criterion functions. The extractive summarization systems are generally based on techniques for sentence extraction which also aims to cover the set of sentences that are most important for the overall understanding of a given document. The author conducted an experiment to show that the suggested approach can improve the performance compared to state-of-the-art summarization approaches. This approach contains a drawback of more information loss.

Rasim M. Alguliev et al. [7] suggested a maximum coverage and minimum redundant text summarization model. A query-oriented summary is used to present the information that is more relevant to the given queries, whereas a generic summary gives an overall theme of the document’s content. In this, the authors focused on the unsupervised generic text
summarization that generates a summary by extracting key textual units from the given document collection.

In particular, the authors model the text summarization as an Integer Linear Programming (ILP) problem. The advantage of this model is that it can directly discover key sentences from the given document(s) and cover the main content of the original document(s). This model also guarantees that in the summary cannot have redundancy of sentences that convey the same information. The authors conducted an experiment to show that combination of the NGD-based and cosine similarity measures conducts to better result than their use separately. If the size of the problem is limited, sometimes this model does not obtain the exact solution within a practical time by means of the branch-and-bound method.

S.Prabha et al. [8] proposed a Context-Based Similarity Analysis for Document Summarization. The authors suggested a context sensitive document indexing model based on the Bernoulli model of randomness which is used for document summarization process. The Bernoulli model of randomness is used to find the probability of the co-occurrences of two terms in a document and use the classical semantic information theory to quantify the information presents in the co-occurrences of these two terms. The lexical association between terms is used to provide a context sensitive weight to the document terms. The sentence similarity matrix is computed by using the context sensitive indexing weights and as a result, the informative sentences are placed on the top of the summary to make a positive impact on the quality of the summary.

Sunita Sarkar et al. [9] presented a comparative analysis of K-means and particle swarm optimization algorithm for text clustering using Nepali WordNet. Here the author applied Particle Swarm Optimization (PSO) and hybrid PSO with K-means algorithms for clustering of text in Nepali language. By using inter cluster similarity and intra cluster similarity, experimental evaluation is performed. The author demonstrated an experiment to prove that hybrid PSO with K-means performs better than K-means and PSO algorithms, when documents are represented using WordNet.

symmetric non-negative matrix factorization (SNMF), sentence-level semantic analysis (SLSS) and mixture model. SLSS is able to capture the relationships between sentences in a semantic manner and SNMF can divide the sentences into groups for extraction. The similarity between documents is measured using the mixture language model and SNMF is used to factorize the similarity matrix and obtains the meaningful groups of sentences. The author selects the most informative sentences in each cluster by considering both internal and external information. The author ranks the selected sentences according to their relevance to users’ requests in a semantic way to increase the performance of multi-document summarization.

Chen Li et al. [11] focuses on the problem of using sentence compression techniques to improve multi-document summarization. The author proposed an innovative sentence compression model which is based on expanded constituent parse trees. The author utilizes a pipeline summarization framework where sentences are first compressed by the innovative sentence compression model to obtain top-n candidates and then generates the final summary by using a sentence selection module. The author evaluates innovative sentence compression method on the TAC 2008 and 2011 datasets using the standard ROUGE metric and human evaluation of the linguistic to show that the model yields significant performance gain in linguistic quality.

P.Sukumar, K.S.Gayathri [12] suggested semantic based sentence ordering approach for multi-document summarization and also addresses the semantic relationship among the sentences in the summary using WordNet synsets. The author presented an entailment model which infers the logical relationship among the sentences while arranging the sentences in the summary. The sentences are ranked by using graph method, where nodes represent the sentences and the edges represents the priority of one sentence over another sentence. The author demonstrated that the method provides high accuracy in providing efficient summary which significantly improves readability and understandability of the user when compared to statistical methods. To produce coherent summary, the casual relation measures must be taken into account in addition to symmetric and non-symmetric measures.

Mohamed Abdel Fattah [13] suggested a hybrid machine learning model
for multi-document summarization. The author improves content selection process in multi-document automatic text summarization by using statistical tools. The author constructed text summarizer models based on a naive-Bayes classifier, a maximum entropy model and a support vector machine and the three models are combined into a hybrid model which ranks the sentences in the order of importance to produce the final summary. The ROUGE score is used to measure effectiveness of the hybrid machine learning model in multi-document summarization. This approach has a limitation in extending to personalized single and multi-document summarization.

Ha Nguyen Thi Thu [14] suggested an optimization text summarization method which is based on Naïve Bayes algorithm and topic word for single syllable language. The processing of single syllable language is very complex for language processing problem because it’s very hard to determine word based on white space and all word segmentation tools not high accuracy. The author suggested an approach based on supervised learning method using Naïve Bayes with the goal of enhancing the quality of text summary, accuracy and to reduce the time for computing in single syllable language, through the selection of training data and training data by experts, and also the author use tagger tool for extracting topic words set in the training phase. The method is experimented with 320 Vietnamese texts to show that the method is really effective in solving the problems in processing the single syllable language. This approach has a limitation in extending to cross language summary for some single syllable languages and building automatic cross language text effectively.

Ramiz M. Aliguliyev [15] suggested a discrete particle swarm optimization algorithm which contains a mutation operation acquired from genetic algorithms to solve the clustering problem in multi-document summarization. The author proposed two weighted clustering methods, their three combinations and a new dissimilarity measure to optimize various aspects of intra-cluster similarity, inter-cluster dissimilarity and their combinations. The author demonstrated the good effectiveness of methods by experimenting the methods on the DUC2007 and DUC2005 data sets. The Normalized Google Distance (NGD) based measure is compared with the Euclidean distance to show that the NGD-
based dissimilarity measure outperforms the Euclidean distance.

3. Conclusion And Recommendations

This paper discusses about survey on text summarization using various optimization algorithms. The multi-document summarization contains a particular challenge in the document set which contains the diverse information that is either related or unrelated to the main topic. The multi-document summarization faces another challenge in which the information present in different documents overlaps with each other. Hence, an effective summarization method is required to analyze and extract the globally important information from the multiple documents and also to merge information stored in different documents. Hence this can be achieved by making use of modified particle swarm optimization algorithm (MPSO) to enhance the performance of text summarization. This will be considered as the future work for text summarization.

References


