IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 2 Issue 4, April 2015.
www.ijiset.com
ISSN 2348 – 7968

Automatic Text Summarization Excerption System Using Fuzzy Values Based on Title and Number of Useful Sentences

Pooja Mehta **Khushal Chheda** **Aashish Kukreja**

(poojamehta.20893@gmail.com) (khushal.chheda061092@gmail.com)(aashishkukreja143@gmail.com)

poojamehta.20893@gmail.com

Department of Computer Engineering
Shah & Anchor Kutchhi Engineering College, Chembur

Abstract

An experimental approach is proposed to address the problem of improving content selection in automatic text summarization by using combination of frequency allocation and fuzzy values. The pre-processing step consists of stopword elimination and stemming. In first step, the system removes the stop words, parses the text and assigns a frequency for each word in the text. The second step is to extract the important keywords in the text by stemming. Each sentence is ranked depending on the existence of the keywords in it, the relation between the sentence and the title by computing fuzzy values. The third step is to extract the sentences with the highest rank. The forth step is the filtering stage. This step reduces the amount of the selected sentences in the summary in order to produce a qualitative summary.

Keywords:Text Summarization, frequent words, stemming, fuzzy values, extractive summary.

1. Introduction

Text summarization [1]has become an important and timely tool for assisting and interpreting text information in today's fast-growing information age. It is very difficult for human beings to manually summarize large documents of text. There is an abundance of text material available on the internet. However, usually the Internet provides more information than is needed. Therefore, a twofold problem is encountered: searching for relevant documents through an overwhelming number of documents available, and absorbing a large quantity of relevant information. The goal of text summarization is condensing the source text into a shorter version preserving its information content and overall meaning.

A summary [2] can be employed in an indicative way as a pointer to some parts of the original document, or in an informative way to cover all relevant information of the text. In both cases the most important advantage of using a summary is its reduced reading time. A good summary system should reflect the diverse topics of the document while keeping redundancy to a minimum. Microsoft Word's AutoSummarize function is a simple example of text summarization.

Text Summarization methods can be classified into extractive and abstractive summarization. An extractive summarization method consists of two types of summarization often addressed in the literature: key phrase extraction, where the goal is to select individual words or phrases to "tag" a document, and document summarization, where the goal is to select important sentences as a whole from the original document and concatenating them into shorter form. The importance of sentences is decided based on statistical and linguistic features of sentences.

An Abstractive summarization [3][4] attempts to develop an understanding of the main concepts in a document and then express those concepts in clear natural language. It uses linguistic methods to examine and interpret the text and then to find the new concepts and expressions to best describe it by generating a new shorter text that conveys the most important information from the original text document. This paper focuses on extractive text summarization methods.

Our project is developed using an extractive method to solve the problem with the idea of extracting qualitative sentences from the original document. The main contribution of the proposed system is the design of extracting qualitative sentences by using frequency allocation for each stemmed word and then computing fuzzy values assigning significance to the title and set of number of useful and useless sentences.

2. The Proposed System Architecture

The overall proposed algorithm is represented in fig1, where all the steps are depicted in sequential manner. The system is divided into 3 major parts, an input text

ISSN 2348 – 7968

document, a summarizer, and a summarized text document as output. The summarizer algorithm is further divided into the three parts- the text pre-processing module, frequent terms generation module along with semantically similar terms and sentence filtering module for summarization.

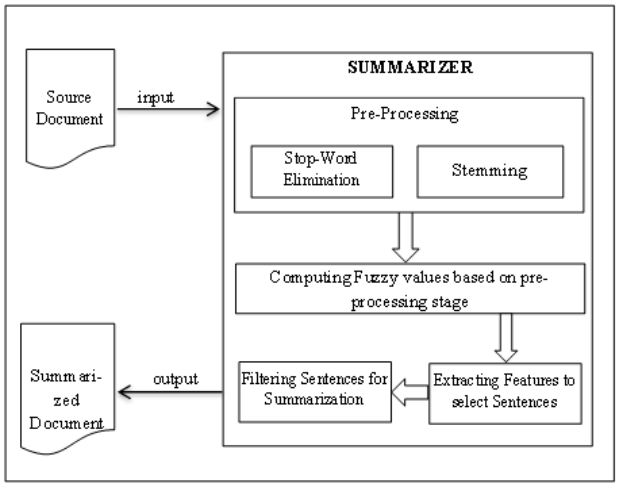

Fig. 1 Proposed System Model

The overall methodology of semantic similarity basessingle document summarization can be expressed in terms of an algorithm. The algorithm takes two input parameters – the input text document and number of frequent terms. As the output it generates a summarized text document along with the stemmed keywords and fuzzy values.

The model consists of following stages:

2.1 Preprocessing

The preprocessing is a primary step to load the text into the proposed system and involves cleaning the noisy text containing grammatical and typographical errors. The major problem in text summarization is that the size of the document is not well known. Thus each word in the documents wasrepresented by the terms in the vector space model, which causes the number of dimensions to be too high for the text summarization algorithms. This preprocessing method plays a vital role in reducing the number of dimensions passed to the text summarization process. In this paper, the followed pre-processing methods were applied, namely, Removal of Stop Words, Punctuation Removal, Removal of Extra White Spaces, Word Stemming, Key Phrase Identification, Sentence Segmentation and Tokenization.

The major Preprocessing steps are:

i. *Stop Word Elimination:*The procedure is to create a filter for those words that are unimportant and which appears frequently in the document and provides less meaning in text processing. Using the stop list has the advantage of reducing the size of the candidate keywords.

ii. *Stemming:*Stemming is a pre-processing technique that is widely applicable in many text mining applications, such as, Information Retrieval (IR), Topic detection, database search system, and linguistic applications. It is a process of converting the variant words of conflation group to correct root word. For example, the following variant words belong to the conflation class “accept”.

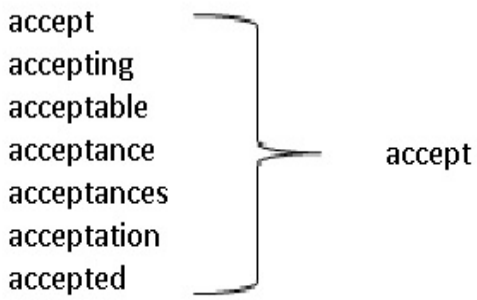

Fig. 2An example of Stemmed Keyword

Mostly, the stemming algorithms are rule based and use the logical approach for removal or sometimes replacement of inflectional and derivational suffixes. It stems the input word, so that all the inflected forms of a word are conflated to the root word. Many rule based algorithms transform these inflected forms of word to a stem rather than root word. It helps to reduce the size and complexity of the data in the document collections and in turn it is useful to improve the performance of IR.

Frequently, the performance of a keyword extraction system will be improved if term groups such as these are conflated into a single term. This may be done by removal of the various suffixes *-ed, -ing, -ion, -able,-ance* to leave the single term “accept”. In addition, the suffix stripping process will reduce the number of terms in the system, and hence reduce the size and complexity of the data in the system, which is always advantageous.

ISSN 2348 – 7968

iii. *Frequency Estimation:*After the stop words are eliminated and the candidate keywords are scanned to a common root (stemming), all the words including the stemmed keywords are assigned a frequency based on the total number of occurrences in the input text document. For example if the word "technical" is occurred 57 times including the stemmed words then its frequency is57.

2.2 Conversion to Fuzzy Values

Fuzzy logic system design usually implicates selecting fuzzy rules and membership function. The selection of fuzzy rules and membership functions directly affect the performance of the fuzzy logic system.

In this paper, we are proposing title based weighing and ranking, we had to convert the frequency counts to fuzzy values based on the title of the document. The title words will be given more importance, i.e., their fuzzy value will be 1. Now this value will be averaged with the normal fuzzy value of the word based on its occurrence in the remaining part of the document.

2.3 Extraction of Sentences

Extraction of sentences is based on the fuzzy values computed. Here, we calculate the gain of the sentences based on parameters like the number of useful sentences, the keywords in the sentence, etc. Then all document sentences are ranked in a descending order according to their scores.

A set of highest score sentences are extracted as document summary. Therefore, we extracted the appropriate number of sentences as selected by user according to nearest to 1 fuzzy value and highest frequency. After calculation, we return a set of sentences (sentence numbers) to the user to be displayed as a summary. Finally, the summary sentences are arranged in the original order.

2.4 Post Processing

The system makes filtering on the generated summary to reduce the number of the sentences, and to give more compressed summary. The system at first removes redundant sentences; second, the system removes the sentence that has a similar context to another one which has similar words more than 65%. For example in case of technical papers, this is necessary becauseauthors often repeat the idea using the same or similar sentences in both introduction and conclusion sections.

The similarity value is calculated as the vector similarity between two sentences represented as vectors. That is, the more common words in two sentences, the more similar they are. If the similarity value of two sentences is greater than a threshold, we eliminate one whose rank based on the features is lower that of the other. For the threshold value, we used 65% in the current implementation.

3. Result Analysis

In this section, we would focus on analysing the test results that we got after performing the test cases as shown in Table 1. We have performed testing on popular text formats, different text sizes and also on different summary lengths.

3.1Analysing the results of a small document in .txt format:

We have tested two .txt documents of two different sizes (3KB & 6KB), each at two different summary lengths (30%, 60%).The same document is summarized for 30% and 60% summary lengths. It is observed that at 30% summary length, the summary provides an outline to the subject matter and hence presenting the idea, but there are a few segments that are not included along with quite a few minute details. The document is summarized in a short way, where some points may not be covered.

Although, when the given source document is tested at 60% summary length, the results show an example as to how a document can be precisely reduced and framed in a 60% bracket. The details mentioned in the summary refer to the essential subject matter of the document that requires the most attention. The summary of 60% length and above becomes much detailed.

3.2 Analysing the results of a large document in .doc format:

After testing the summary of a document of type .doc size 43KB, which was in a 50% bracket, not only is the main and essential subject matter covered, but the rest of the minute, yet important details have been included. This provides a greater platform for scrutinized reference of the document, henceforth increasing the reliability of the summary.

ISSN 2348 – 7968

When made a comparison between the 30% summary and 50% summary of this document, the 30% summary included the essentials of the subject matter, framing the document, but at the same time leaving out some important details along with a few minute details. One can very well comprehend the theme of the document with this type of summary.

3.3 Analysing the results of a document in .rtf format:

Similarly 15% summary of the .rtf format provides in brief the detailing of the document, hence making it possible to analyse the document with very less data. But at the same time it lacks the capacity to include all the aspects of the document. This type of summary helps in overviewing the document.

The 40% summary of the .rtf format includes a lot more important aspects and details of the document. Considering the performance with respect to speed of generating the summary .rtf format also requires sufficient amount of time.

3.4 Analysing the cross references in the documents:

Documents of various genres were tested. Genres like informative document, biography, etc. were tested and the cross references were found for most sentences in the documents. However, some sentences were not cross referenced like the ones starting with "*he*" and "*she*". We can say that currently the software successfully cross references most sentences and has a passing probability of about 70%.

In Table 1, the accuracy of summary is scored on the basis of performance, speed of generating summary, its relevance, the size of the document, the format of the document and the length of summary.

Table 1. Comparative Analysis

Type	Size	Summary length	Score
.txt	3KB	30%	5
.txt	3KB	50%	5
.txt	6KB	30%	4
.rtf	4KB	15%	3
.rtf	4KB	40%	4
.rtf	7KB	15%	2
.doc	43KB	30%	3
.doc	43KB	50%	4

If the result is scored 5, then it signifies that the summary fulfils its objective whereas if score is 1 then the result is not satisfactory.

4. Conclusion

In this paper, we have presented a fuzzy logic aided sentence extractive summarizer that can be as informative as the full text of a document with better information coverage. Here the significance is given to the title of the original document and the concept of useful and useless sentences, based on which fuzzy values are computed. The system is tested with the input of various formats and size of documents; and the results show that the summary that has been generated is in the context with the subject matter of the original document. The system rendered superior results in comparison to manual summarization excerption process. The system can give the most compressed summary with high quality. The main applications of this work are Collating Search Engine hits, Text-to-Speech for blind people, text compression and producing intelligent reports.

References

[1] Karel Jezek and Josef Steinberger, "Automatic Text summarization", Vaclav Snasel (Ed.): Znalosti 2008, pp.1-12, ISBN 978-80-227- 2827-0, FIIT STU Bratislava, Ustav Informality a softveroveho inzinierstva, 2008.

[2] Weiguo Fan, Linda Wallace, Stephanie Rich, and Zhongju Zhang, "Tapping into the Power of Text Mining", Journal of ACM, Blacksburg, 2005.

[3] G Erkan and Dragomir R. Radev, "LexRank: Graph-based Centrality as Salience in Text Summarization", Journal of Artificial Intelligence Research, Re-search, Vol. 22, pp. 457- 479 2004.

[4] Udo Hahn and Martin Romacker, "The SYNDIKATE text Knowledge base generator", Proceedings of the first International conference on Human language technology research, Association for Computational Linguistics, ACM, Morristown, NJ,USA , 2001.

[5] Farshad Kyoomarsi, Hamid Khosravi, Esfandiar Eslami and Pooya KhosravyanDehkordy, "Optimizing Text Summarization Based on Fuzzy Logic", In proceedings of Seventh IEEE/ACIS International Conference on Computer and

ISSN 2348 – 7968

Information Science, IEEE, University of Shahid Bahonar Kerman, UK, 347-352, 2008.

[6] G. Salton, C. Buckley, “Term-weighting approaches in automatic text retrieval” *Information Processing and Management* 24, 1988. 513-523. Reprinted in: Sparck-Jones, K.; Willet, P. (eds.) Readings in I. Retrieval.Morgan Kaufmann. Pp.323-328.1997.

[7] M.A. Fattah and Fuji Ren, “Automatic Text Summarization” In proceedings of World Academy of Science, Engineering and Technology Volume 27. Pp. 192-195. February 2008.

[8] Arman Kiani and M.R. Akbarzadeh, “Automatic Text Summarization Using: Hybrid Fuzzy GA-GP” In Proceedings of 2006 IEEE International Conference on FuzzySystems, Sheraton Vancouver Wall Center Hotel, Vancouver, BC, Canada. Pp.977-983.2006.

[9] S. Rucha, S.S.Apte, "Improvement of Text Summarization using Fuzzy Logic Based method", IOSR Journal of Computer Engineering, P:5-10, 2012.

[10] Ladda Suanamali, Naomic Salim, Mohammed Salem Binwahlan, "Fuzzy Logic Based Method for Improving Text Summarization", International Journal of Computer Science and Information Security, P: 150 -156, 2009.

[11] Kyoomarsi F,Khosravi H, Eslami E, and Davoudi M, "Extraction Based TextSummarization using Fuzzy Analysis", Iranian Journal of Fuzzy Systems, Vol. 7, No. 3,P: 15-32, 2010.

[12] Sayantani Gosh, Sudipta Roy, Samir K. Bandyopadhyay, “A Tutorial Review on Text Mining Algorithms", International journal of Advanced Research in Computer and Communication Engineering, 2012.

[13] D.Y.Sakhare, Dr.Raj Kumar, “Syntactic and Sentence Feature Based Hybrid Approach for TextSummarization”I.J. Information Technology and Computer Science, P: 38-46, 2014.

[14] Dipti, Sakhare, Raj Kumar, “Neural Network Based Approach to study the effect of Feature Selection on Document Summarization”, International Journal of Engineering and Technology, P:2585 – 2593, 2013.