Automated Multiple Related Documents Summarization via Jaccard's Coefficient

Huda Yasin
Department of Computer Science,
University of Karachi, University Road, Karachi, 75270, Pakistan

Mohsin Mohammad Yasin
Department of Computer Science,
National University of Computer and Emerging Sciences, Super Highway, Karachi, Pakistan

Farah Mohammad Yasin
Department of Bio Medical Engineering, SSUET, University Road, Karachi, 75300, Pakistan

ABSTRACT
Today, in the hasty advancement epoch of technology, allotting and gathering of information are imperative. Readers enthrall with an undersized edition of copious prolonged text documents. In this paper, we represent our approach which we used in our Automated Text Summarization System known as MDSS (*Multiple Documents Summarization System*). We elucidate a new fangled approach which is based on statistical (rather than semantic) factors. In contrast to single document summarization, the issues of compression, speediness, superfluous and passage opting are more decisive in multiple documents summarization. For sentence comparison, Jaccard's coefficient is used to improve the worth and quality of the summarization. Resemblance exists between our algorithms and dynamic time warping. Our experimental domino effects indicate that it is useful and effectual to enhance the quality of multiple documents summarization via Jaccard's coefficient. Our system MDSS is implemented in Java (jdk 1.6).

General Terms: Text mining, text summarization

Keywords: Multi-document summarization, Jaccard's coefficient, sentence comparison, text mining

1. INTRODUCTION
Summarized document is basically a short version or more merely we can state that it is a subset of the original set. Data summarization is one of the segments of data preprocessing [1]. Summarization is also referred to as characterization or generalization [2]. The study on automated summarization has been initiated 40 years before [3]. It has been said that we hold surplus amount of information on our hands, shoving us to read great number of documents and extracting germane information from them. So to muddle through such state of affairs, investigation on automated summarization of unstructured text has engrossed much attention in recent times. More willingly than single document, now, more research work is going to establish techniques for automated summarization of manifold documents [4]. Summarization of a single document is quiet simple as compared to manifold documents because in multiple documents summarization, the intricacy of swiftness, compression and redundancy are more convoluted [5].

Automated summarization of unstructured text drastically squeezes information content. Hitherto, most of the work has been done in English and other European language. Nevertheless many other languages seem to be appears swiftly emerging in this field.

Neural network [6], regression models [7] and decision trees [8] are some of the prominent approaches that have been used in the search for optimized text summarization.

The two approaches 'shallow sentence extraction' and 'the deep understand and generate' are generally followed in automatic text summarization research [9]. In exploiting summarization, many modern information retrieval applications need summarization systems which scale up to huge volumes of unhampered text. Those multiple documents which cover analogous information are a general issue which gets up in some application [10]. For instance, multiple stories which covers same incident.

For content withdrawal of multiple documents summarization, large ranges of techniques are present. By the degree of dependence on domain, these techniques show a discrepancy from each other [10].

There are different approaches which facilitates us in managing different problems which occur in data summarization e.g. intrinsically coherent nature of the clustering [11]. Direct management of free style unstructured data can be done by various approaches of data mining. Memory based reasoning is one of the approaches of nearest neighbor modes. This technique can operate on free style unstructured data [12].

Clustering, coverage, anti superfluous, rationality, summary uniformity criteria, identification of source inconsistencies and effectual user interfaces are some of the features which are involved in multiple documents summarization [5].

There are two categories of summarizers, linguistic and statistical. This paper bestows a statistical approach to engender effectual summary. More often than not, statistical summarizers do not make use of any linguistic information.

This paper is structured as follows. In section 2, we discuss the benefits and related research work on multi-documents summarization. Section 3 symbolizes the system architecture of MDSS. In section 4, we discuss the methods and their results.

2. MULTI-DOCUMENTS SUMMARIZATION
2.1 Benefits
Following two points represent the state of affairs in which multiple documents summarization seems to be constructive [5]:

1. If there is an assortment of divergent or unlike documents and yearn of a user is just to review the backdrop or milieu enclosed in the entire assortment.

2. If there is an assortment of closely associated documents which are haul out from a more outsized miscellaneous assortment.

2.2 Related Work

Based on an iterative graph based ranking algorithm, Rada Mihalcea and Paul Tarau explained an approach for language autonomous extractive summarization. They presented that in spite of the language, their algorithm works efficiently. They did so by means of appraisal applied on single document summarization task. Those tasks were in Portuguese and as well as in English [13].

Derong Liu et. al. proposed an efficient model for multiple documents summarization by means of genetic algorithm. Their model expands and contracts the coverage of subjects and superfluous contents respectively. In order to appraise sentences, theme of each document, their associations and the central idea of the collection was scrutinized which was founded on Chinese idea lexicon and corpus. On the basis of sentences weight and as well as their significance from the associated documents, they find out the correct sentences for withdrawal [14].

V. Finley Lacatusu et. al. elucidates a novel clustering based text summarization system that uses manifold sequence Alignment in order to enhance the arrangement of sentences contained by theme clusters [15].

Via graph representation for text, Inderjeet Mani and Eric Bloedorn proposed a new-fangled approach for summarizing likeness or resemblances and dissimilarities in a set of associated documents [10].

By means of domain autonomous approaches, Jade Goldstein et. al. [5] addressed the problems of swiftness, compression, superfluous, and passage opting. Principally, these techniques were established on swift, statistical dealing out, a metric for tumbling superfluous and expands miscellany in the opted passages.

3. SYSTEM ARCHITECTURE

Our system is divided in four main parts as shown in figure 1:

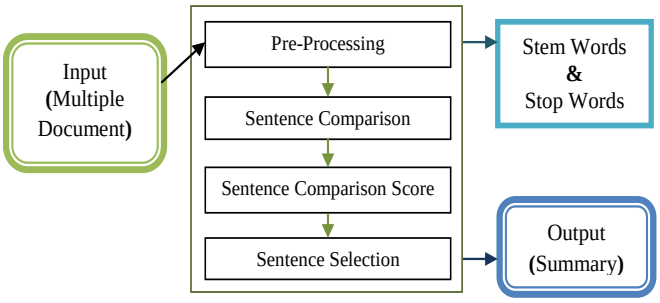

Figure 1. System Architecture of MDSS

3.1 Pre-processing

Tokenization, punctuation and noisy words removal, and stemming are assumed to be the general text preprocessing phases. The two foremost activities which are performed in this stage are:

- Stemming
- Removal of stop/ noisy words

These activities are considered to be the preliminary steps in summary generation to skim and scrutinize the documents. These steps are elucidated below:

3.1.1 Stemming

It is a procedure in which the word endings are cut off or more simply we can say that the words are abridged into their roots [16]. For instance, after applying stemming to words "challenged," "challenges," and "challenging," the corresponding root 'challenge' would be resulted. We have applied 'Paice Husk' algorithm for stemming.

3.1.2 Removal of Stop/ Noisy Words

Noisy words like is, an, the, or etc have no significance in unstructured text. We have detached such words in order to obtain optimized end result. In Japanese, noisy/stop words identification is based on grammatical information. As an exemplar, project search makes out whether the utterance is a noun or a verb, whereas the other dialects work with particular lists. We have used a list comprises of 521 stop words. Plus, this list of stop words is also available in [17].

3.2 Sentence Comparison

For multiple documents summarization, we have considered the 1st document as a base document i.e. its each sentence compares with each and every sentence of the rest of the documents. The similarity or association stuck between the sentences is premeditated by means of 'Jaccard's coefficient'. Jaccard's coefficient is utilized to measure the intersection of two sets as related to the entire set instigate by their union [2]. It is defined as:

$$\mathrm{Sim}(h_i, h_j) = \frac{\sum_{c=1}^{k} h_{ic}\ h_{jc}}{\sum_{c=1}^{k} h_{ic}^2 + \sum_{c=1}^{k} h_{jc}^2 - \sum_{c=1}^{k} h_{ic}\ h_{jc}}$$

where h_i and h_j represents words of a sentence of different documents.

3.3 Sentence Comparison Score

MDSS stores the score of each sentence in a vector. This score is acquired after the comparison between sentences and is utilized by the following methods which are discussed in detail in Section 4:

- Generating summary using Jaccard's coefficient (both in ascending and descending order).
- Generating summary using Jaccard's coefficient (Opting sentence on the basis of sentence weight).
- First, extracting summary of individual documents and then using Jaccard's coefficient for comparing sentences.

3.4 Sentence Selection

For summarization, buffer stores the elected sentences. This selection process continues till the desired percentage for summarization.

In order to generate yearned percentage of summary, we have set a threshold. It is calculated as:

$$\text{Threshold} = ((\text{Total sentences of 1st document}) + (\text{Total sentences of 2nd document}) + \cdots + (\text{Total sentences of nth document})) \times (\text{desired percentage of summary})$$

4. METHODS, RESULTS AND DISCUSSION

We have applied Jaccard's coefficient in different ways with the aim to explore the optimized end result. These different techniques are explicated below:

4.1 Generating Summary Using Jaccard's Coefficient (Ascending and Descending Order):

Sentences are extracted from manifold documents on the basis of similarity comparison score. Evaluation score is arranged in the following two orders:

a) Ascending order
b) Descending order

First, we set the comparison score in ascending order. As a result, the summary consists of those sentences which have the minimum similarity score (may be zero). The thought behind this approach is that sometimes it may be possible that the score of an important sentence is minimum i.e. most of its content words do not match with the words of comparing sentence. From this approach, we do not found the efficient summary because some of the resulted sentences seem to be discrete or irrelevant and do not reflect the theme of the documents. We have also observed that the sentences in a summary are the initial sentences of the multiple documents. For example, consider two documents and we want the summary up to 25%, the first sentence of first document compares with all the sentences of the second document. This comparison contains many sentences with minimum score i.e. before comparing the second sentence of the first document with rest of the sentences- the 25% summary completed.

Second, we arranged the similarity score in descending order. Now, the summary contains those sentences that have the maximum similarity score. This approach gives an efficient result. The maximum similarity is may be sandwiched between the last sentence of the first document and any sentence of rest of the documents. We have noticed that even for 25% summary, this approach goes through the comparison surrounded by each and every sentence. From this approach, we have found the optimized summary.

4.2 Generating Summary using Jaccard's coefficient (Selecting Sentence on the Basis of Sentence Weight):

The basic idea behind this approach is same as that of the above. The variation comes on that point when we have two sentences of different documents (on the basis of maximum similarity score) and than we opt one of them on the basis of their sentence weight. Each sentence weight is calculated as:

$$SW = \frac{CW}{TW}$$

- SW: weight of a sentence
- CW: content words of a sentence
- TW: total words of a sentence

The sentence with maximum and minimum weight will be included and excluded respectively. This technique is efficient and engender useful summary but we have also identified that in this approach, the compression rate escalates i.e. for 25% summary, if we are getting 45 sentences (on the basis of threshold) from first approach than this approach gives us 23 sentences. Plus, most of these 23 sentences are the end sentences of those 45 sentences.

4.3 Generating Summary from Summaries of Individual Documents

This approach first generates summary of each individual document and than same similarity comparison (as discussed above) takes place between summaries of individual document. The architecture of this procedure is shown in figure 2.

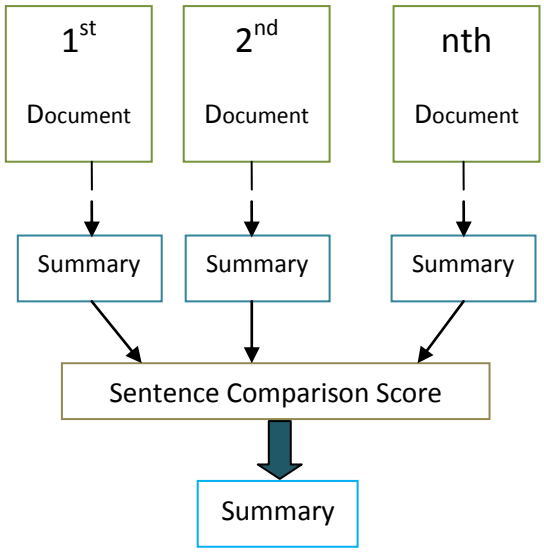

Figure 2. Generating Summaries from Summaries of Individual Document

We have perceived that the summary obtained from this approach is much similar to the summary obtained by means of Jaccard's coefficient in descending order but the time taken by this approach is more than any other approach.

5. CONCLUSION AND FUTURE WORK

In this paper, we proposed a method which is used in our multiple document summarization system. It is based on Jaccard's coefficient. We have presented three different algorithms. Our experimental consequence indicates that 'Generating summary using similarity score based on Jaccard's coefficient in descending order' gives the most optimized result. We compared our different summarization results with the manuals. We have analyzed that our system represents steady correlation with the human assessment outcome.

In future, we will broaden this paper to acquire more enhanced domino effects by using different text mining algorithms. In addition, we will apply fuzzy learning models for further enhanced estimation.

6. ACKNOWLEDGMENT

We are awfully thankful to Mr. Badar Sami for his unparalleled help and support with the evaluation.

7. REFERENCES

[1] Doru Tanasa, Brigitte Trousse, "Advanced Data Preprocessing for Intersites Web Usage Mining," IEEE Intelligent Systems, vol. 19, no. 2, pp. 59-65, Mar./Apr. 2004

[2] Margaret H. Dunham and S.Sridhar, 2006, Data Mining (Introductory and Advanced Topics). Pearson Education, chapter 1.

[3] Luhn. H.P. "The Automatic Creation of Literature Abstracts". IBM Journal of Research and Development, Vol. 2, No. 2, pp. 159-165, April 1958.

[4] Tsutomu HIRAO, Takahiro FUKUSIMA, Manabu OKUMURA, Chikashi NOBATA. "Corpus and Evaluation Measures for Multiple Documents Summarization with Multiple Sources".

[5] Jade Goldstein, Vibhu Mittal, Jaime Carbonell and Mark Kantrowitz., Multi-Document Summarization by Sentence Extraction.

[6] E. Qwiener, J.O. Pederson, and A.S.Weigned, "A neural network approach to topic spotting", in Proceedings of the fourth Annual Symposium on Document Analysis and Information Retrieval (SDAIR'95), 1995.

[7] Y.Yang and C.G.Chutte, "An example-based mapping method for text categorization and retrieval", ACM Transaction on Information Systems (TOIS), 12(3):252-277, 1994.

[8] Joachims, T., "Text Categorization with Support Vector Machines: Learning with Many Relevant Features", in European Conference on Machine Learning (ECML), 1998.

[9] Mani, I., Automatic Text Summarization. John Benjamins Publishing Company, (2000-01).

[10] Mani, I. and Bloedorn, E., Multi-document Summarization by Graph Search and Matching 1997.

[11] Witold Pedrycz, Knowledge based clustering from data to information granules.

[12] Michael J. A. Berry, Gordon S. Linoff, Data Mining Techniques (For marketing, sales, and CRM).

[13] Rada Mihalcea and Paul Tarau, A Language Independent Algorithm for Single and Multiple Document Summarization, University of North Texas

[14] Derong Liu, Yongcheng Wang, Chuanhan Liu, and Zhiqi Wang, Multiple Documents Summarization Based on Genetic Algorithm.

[15] V. Finley Lacatusu, Steven J. Maiorano and Sanda M. Harabagiu, Multi-Document Summarization using Multiple-Sequence Alignment, Human Language Technology Research Institute, Department of Computer Science, University of Texas at Dallas

[16] Huan Liu, Nitin Agarwal, Robert Grossman, 2009, Modeling and Data Mining in Blogosphere.

[17] Stop Words List Available at: http://www.lextek.com/manuals/onix/stopwords1.html and http://www.lextek.com/manuals/onix/stopwords2.html