Volume 2, Issue 2, February 2012 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

JC- Automatic Manifold Related Pages Reviewed by Jaccard's Coefficient

P.Sivakumar *
Department of CSE
KSR College of Engineering,
Namakkal, Tamilnadu, India

Dr. R.M.S Parvathi
Department of CSE
Sengunthar College of Engineering,
Namakkal, Tamilnadu, India

Abstract— **Nowadays, in the very fast development period of technology is moving, here work allocating and also gathering of message is very necessary. The person involves with a small amount of multiple extended text documents. In this paper, Various Pages Reviewed Method (VPSM) for automatic text reviewed concept. We give the new concept which is based on statistical features. In contrast to single document reviewed, the issues of compression, speediness, superfluous and passage opting are more decisive in multiple documents reviewed. For statement similarity, Jaccard's coefficient is to get better the value and feature of the reviewed. Similarity exists between our algorithms and dynamic time warping. The Jaccard's coefficient via our new concept property shows that it is valuable and efficient to improve the excellence of multiple documents summarizated.**

Keywords— **Multi-document reviewed, Jaccard's coefficient, sentence comparison, text mining**

I. INTRODUCTION

Reviewed manuscript is mostly a small account or additional only we can condition that it is a dividing up of the innovative set. Information reviewed is one of the divisions of information preprocessing [1]. Reviewed is also submitted to as classification or simplification [2]. It has been said that we hold extra quantity of information on our hands, pushing us to study large quantity of documents and removing important information from them. So to mix-up throughout such condition of interaction, search on automatic evaluation of formless text has engrossed much interest in current period.

More freely than single manuscript, now, more research work is going to launch performances for manual analyses of multiple documents [4]. Reviewed of a single document is silence uncomplicated as evaluated to multiple documents because in multiple documents reviewed, the difficulty of quickness, density and joblessness are more complicated [5]. Automated review of formless text significantly squeezes information content. up till now, the majority of the job 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 are some of the prominent approaches that have been used in the search for optimized text reviewed. The two concept "low sentence mining" and "the deep understand and generate" are commonly followed in automatic text reviewed research. There are two kinds of reviewed, linguistic and statistical. This paper gives a statistical approach to produce efficient review. More often than not, statistical summarizers do not make use of any linguistic information.

II .DIFFERENT TYPES OF DOCUMENTS REVIEWED

A Benefits

Following two points represent the situation of interaction in which different type of documents reviewed appears to be helpful [5]:

1. If there is a collection of different or dissimilar documents and yearn of a user is just to review the environment or situation enclosed in the entire group.
2. If there is a collection of strongly associated documents which are haul out from a more outsized miscellaneous assortment.

B.Related Work

The raking algorithm based on iterative graph, Luhn. H.P said about an concept of autonomous extractive summarization. They explained that unkindness of the language, the raking algorithm working efficiently.Their algorithm works efficiently. They did so by resources of assessment applied on single document summarization job. Those responsibilities were in Portuguese and as well as in English [3]. Luhn. H.P proposed a genetic algorithm using efficient model of many documents, their method briefly and contracts the of subjects and extra contents correspondingly.

During arrange to evaluate sentence, idea of every document, their relations and the middle plan of the group was examine which was created on Chinese proposal dictionary and quantity. On the foundation of sentence weight and as well as their importance from the connected documents, they find out the accurate judgment for extraction [4].

Jade Goldstein et. al. explain a novel clustering support text summarization method that utilize various sequence arrangement in order to improve the collection of sentence restricted by theme clusters [5]. By graph illustration for text, he proposed a novel approach for abbreviation similarity or resemblance and dissimilarity in a set of unrelated documents [10]. Via income of area independent approach and address the tribulations of fleetness, compression, superfluous, and way opting. Mainly, these methods were establish on swift, statistical dealing out, a metric for dropping extra and enlarge collection in the selected passages.

C Pre-processing

The common text preprocessing `segment` are generated by the Tokenization, punctuation and noisy words removal, and stemming.

Are assumed to be the general text preprocessing phase. The two leading behavior which is achieved in this period are:

- Stemming
- Removal of stop/ noisy words

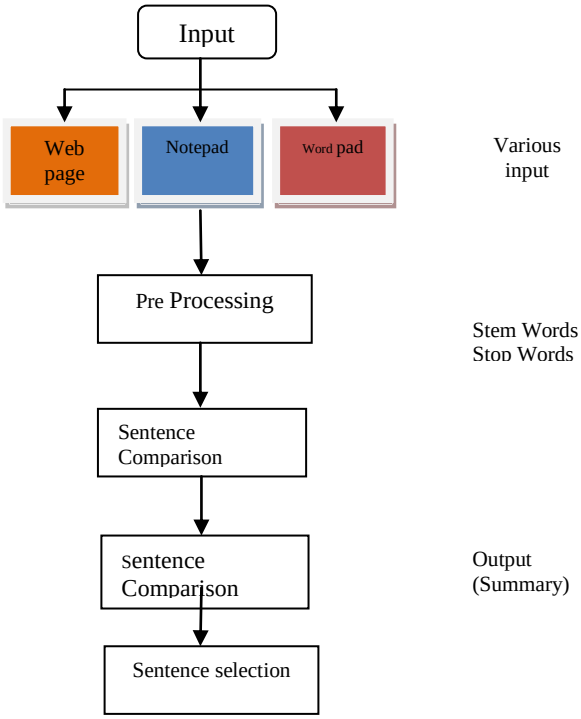
These behaviors are considered to be the introduction steps in hypothetical invention to glide and examine the documents.

III. SYSTEM ARCHITECTURE

Our method is alienated in four major divisions as shown in figure 1:

D Stemming

It is a method in which the word conclusion is cut off or further just we can state that the words are shortened into their roots [6]. For example, after submit an application stem to words "implemented," "implements," and "implementing," the matching origin "implement would be resulted. We have useful" Paice Husk algorithm for stemming.



E Removal of Stop/ Noisy Words

The noise words are captured similar to is, an, the, or etc have no meaning in unstructured text. We have separated such words in regulate to achieve optimized end result. In Japanese, noisy/stop words classification is based on grammatical in sequence. As an example, development look for create out whether the word is a noun or a verb, while the other dialect work with exacting directory. We have used a list contain of 521 stop words. in addition, this file of stop words is also obtainable here [7].

F Sentence Comparison

In support of several documents summarization, we have measured the 1st document as a stand document i.e. its each judgment contrast with each and every stretch 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 use to calculate the connection of two sets as related to the entire set activate by their combination [2]. It is define as:

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

Where and stand for words of a sentence of dissimilar documents.

G Sentence Comparison Score

VPRM stores the gain of each judgment in a vector. This score is purchase after the judgment between sentences and is utilized by the subsequent methods which are converse in detail in part 4:

- Make review via Jaccard's coefficient (both in ascending and descending order).
- Make review with Jaccard's coefficient (selecting sentence on the foundation of sentence weight).
- First, mine review of person documents and then using Jaccard's coefficient for evaluating judgments.

H . 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:

All paragraphs must be indented. All paragraphs must be justified, i.e. both left-justified and right-justified.

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

IV. METHODS AND RESULTS

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

*I.Generating Summary Using Jaccard's Coefficient (Ascending and Descending Order):*Sentences are extracting from multiple documents on the basis of similarity evaluation gain. Evaluation gain is set in the subsequent two instructions:

a) Ascending order

b) Descending order

Ascending order is used to compare the score. As a result taken form the review 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 similarity. Here consider two documents and we want the summary up to 50%, 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 50% summary completed.

Next, we set the similarity gain in descending order. Now, the review contains individuals 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 50% summary, this approach goes through the comparison surrounded by each and every sentence. From this approach, we have found the optimized abstract Text Font of Entire Document the entire document should be in Times New Roman or Times font. Type 3 fonts must not be used. Other font types may be used if needed for special purposes. Recommended font sizes are shown in Table 1.

*J. 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.

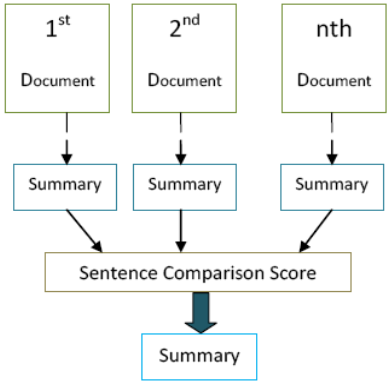


Figure 2. Generating Summaries from Summaries of Individual Document

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

V. CONCLUSION

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.

VI.FUTURE WORK

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.

VII. REFERENCES

[1] Rozita Jamili Oskouei, " Differential Internet Behavior's of Students from Gender Groups" International Journal of Computer Applications, Number 7 - Article 2, Year of Publication: 2010.

[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] Dragomir R. Radev, Hongyan Jing, Malgorzata Budzikowska,. "Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies".

[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.