Volume 5, Issue 8, August 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

An Extractive Technique for Automatic Text Summarization Using Extended Boolean Model

Shivani Khurana, Chhavi Rana
CSE Department, MDU Rohtak,
Haryana, India

Abstract— Text summarization is the method of compressing the source document into a shorter version that conveys the important information of original document and its overall meaning is preserved. It is not easy for human beings to manually summarize large documents of text. In this paper, we are proposing a domain independent and single document technique for text summarization which is using "Extended Boolean model" of Information Retrieval.

Keywords—Extended Boolean Model, Inverse Document Frequency, Term Frequency, Text Summarization

I. INTRODUCTION

Text summarization has become a necessary tool for extracting the relevant information from a huge collection of text documents available on the internet. It is not easy for human beings to manually summarize large documents of text [1]. Therefore an automatic text summarizer is required which deals with searching for relevant documents through large number of documents available on internet and electronic media and then extracting the useful information while preserving its overall meaning [2]. A good summarizer should include all the important topics while keeping minimum redundancy. The three important aspects which should be considered during text summarization are [3]:

- Summary process may be result of single document or multiple documents.
- Important information should be preserved.
- Length of summary should be short.

A summary can be expressed in an indicative as well as informative manner [3], [4]. A pointer is pointed to some important parts of the original document in indicative manner whereas all relevant information of text is covered in the informative manner. The advantage of using both the methods is reduced reading time. There are two types of text summarization methods: Extractive and Abstractive. In an extractive method [5], important sentences and paragraph are extracted from the original document and then they are concatenated into the shorter form. Statistical and linguistic features of sentences are used to decide the importance of sentences. Sentences are extracted on the basis of word/phrase frequency, location or cue words. Most frequent sentences are considered as most important sentences. Extractive methods are simple and easy to implement. An abstractive method [6] deals with understanding of the important concepts of a document and then expressing those concepts in natural language. Linguistic methods are used to examine and interpret that text which conveys the most important information of the original document.

The two steps of extractive text summarization [7] are: Pre Processing step and Processing step. Pre Processing step includes: a) Sentence separation on the basis of separator "dot" at the end of the sentence. b) Stop-word elimination-Elimination of common words with no meaning which do not add important information to the task. c) Stemming-Process of reducing the words to their stem, base or root form. In relevance step, relevance of the sentences is decided on the basis of weights assigned to the sentences. Top ranked sentences are included in the summary document.

Summarization systems [8] are classified as domain-dependent and domain-independent systems. In domain dependent systems, important information depends on the knowledge of texts in that domain. These systems can handle documents from a particular domain only whereas domain independent systems can handle documents from any domain. Construction of abstracts can be done from single documents as well as from multiple documents. In case of multiple document summarization, summary is produced after considering important points in all the source documents and the source documents can be in single language or in different languages.

II. LITERATURE SURVEY

Text Summarization technique presented in [9] is a statistical method based on Salton's Vector Space model. This tool is domain independent and single document summarization tool. The summarizer initially breaks the entire document on the basis of separators. In the second step, unnecessary words known as stop words are removed from the document. The third step deals with removing those words which have same meaning by a method known as stemming. After that weight of words is calculated on the basis of term-frequency and inverse document frequency. In the next step, similarity between the sentences is calculated by using the formula of cosine similarity based on Salton's Vector Space model. In the last step sentences are ranked based on the similarity and highest ranked sentences are included in the summary document. This tool provides the summary of lengthy text documents in about 10 lines.

© 2015, IJARCSSE All Rights Reserved

A multiple document query based document summarizer based similarity of sentences and word frequency is presented in [10]. The technique presented by them removes redundancy based on grouping similar sentences and word frequency. The technique presented consists of pre-processing step and processing step. In the pre-processing step, the query is processed and summarizer collects required documents.

After pre-processing step, following steps are followed to obtain the summary:

- Similarity of sentences present in documents with user query is calculated.
- Grouping of sentences based on their similarity values.
- Calculating sentence score using word frequency and sentence location feature.
- Selection of best scored sentences from each group and putting it in summary.
- Reducing the summary document to 100 words.

The approach presented in [11] is based on clustering of multiple documents for producing cluster wise summaries based on feature profile oriented sentence extraction strategy. Clustering algorithm is used for grouping of related documents in the same cluster. Feature profile is generated based on word weight, sentence position, sentence length, sentence centrality, proper nouns in the sentence and numerical data in the sentence. Sentence score is calculated for each sentence based on the feature profile. Individual word weight is calculated using Term Synonym Frequency-Inverse Sentence Frequency (TSF-ISF) Sentences are extracted from each cluster using different compression rates and ranked in order of importance based on sentence score. After this cluster wise summary is generated after arranging sentences in chronological order as in original documents. The output is a concise cluster wise summary which provides the condensed information of the input documents.

Various issues related to information retrieval system with vector space model are discussed in [12]. The technique is implemented using MATLAB on Cranfield data collection of aerodynamics domain. Stop-word elimination step is also performed to obtain precise summary of the document. Pre processing step is performed using PHP script to obtain the compressed version of the document text. A set of queries had been created from the title of documents and these queries are input to the tool. Term-frequency matrix and query matrix are generated. The term-frequency matrix is used to get the term weights considering tf-idf scheme. Query matrix is divided by their corresponding Euclidean lengths to obtain the normalized weights. The final result is obtained by ordering the weights in a result matrix in decreasing order of their weights.

In [13] a combined approach is presented for document and sentence clustering as an extractive technique of summarization. The input to this summarizer is user selected collection of documents & query. A list of maps is maintained where each term from document collection is stored in a map with its number of occurrences. All the synonyms of the words present in document collection is stored in a map. WordNet dictionary is used to find synonyms. Query modification technique is used in which a query is split into tokens and finds the synonym for each token. Query is strengthened by appending most frequently occurred words from corpus to the query. After this sentence score is calculated using some features such as noun feature, Cue Phrase Feature, Sentence Length Feature, Numerical data Feature, Sentence Position, Sentence centrality (similarity with other sentences), Upper Case word feature, Sentence similarity with user query, Term frequency & Inverse Document Frequency. After this document is clustered using cosine similarity. Then sentences are clustered in every document cluster based on similarity measures. In the next step, score for each sentence cluster group is calculated and sentence cluster are sorted in reverse order of group score. Best scored sentences from each cluster is selected and added to summary.

Multiple text documents sentence clustering technique is presented in [14]. Three important factors considered in this technique are: (1) clustering sentences (2) cluster ordering (3) selection of representative sentences from the clusters. Similarity histogram based incremental clustering method is used for sentence clustering. The clustering approach used is fully unsupervised & is an incremental dynamic method of building the sentence clusters. Then the importance of a cluster is measured based on the number of important words it contains. Top n clusters are selected after ordering the clusters in decreasing order of their importance. Best scored sentence is selected from each cluster and included in the summary. This process of selection is continued until a predefined summary size is reached.

A novel technique using neural networks for summarizing news articles is presented in [15]. Three important steps of this process are neural network training, feature fusion, and sentence selection. The input to this system is either real or binary vectors. In the first step training of neural network is done to recognize the selection of the summary sentences in terms of their importance by learning the relevant features of sentences that should be included in the summary of the article. Neural network is trained on a corpus of articles. After this neural network is modified to generalize and combine the relevant features present in summary sentences. After this feature fusion step is performed in which trends and relationships among the features that are inherent in majority of sentences are discovered. In this step uncommon features are eliminated and common features are collapsed. Importance of various features is discovered by the network

III. RESEARCH METHODOLOGY

This paper proposes an approach to single document text summarization system. The technique is based on Extended Boolean model given by Gerard Salton, Edward A. Fox, and Harry Wu. Extended Boolean model [16] was introduced to overcome the drawbacks of standard Boolean model. This model combines the simplicity of Standard Boolean Model and efficiency of Vector Space Model. In this model document and queries are represented as vectors which make partial matching possible as in Vector Space Model. Similarity between queries and documents is ranked on the basis of term weights which makes a document relevant even if some of the terms of queries are matched in the document.

© 2015, IJARCSSE All Rights Reserved

A document is represented as a vector in which each i dimension corresponds to a separate term associated with the document.

Given Kx is a term associated with document dj. The weight of term Kx is measured by its normalized Term frequency which can be defined as:

$$W_{x,j}=f_{x,j}* \frac{Idf_x}{\max_i Idf_i} \quad (1)$$

where Idf_x is inverse document frequency.

The weight vector associated with document dj can be represented as:

$$V_{dj} = [W_{1,j} , W_{2,j} , . . . , W_{i,j}]$$

Using P-norms that is including p distances, where $1 \leq p \leq \infty$, the similarity between query and document can be represented in the following manner:

$$[q_{or} = K_1 \vee^p K_2 \vee^p .... \vee^p K_t] \text{ and } [q_{and} = K_1 \wedge^p K_2 \wedge^p .... \wedge^p K_t]$$

are generalized conjunctive and generalized disjunctive queries respectively.

$$sim(q_{or} , dj) = \sqrt[p]{\frac{W_1^p + W_2^p + ... + W_t^p}{t}} \quad (2)$$

$$sim (q_{and} , dj) = 1 - \sqrt[p]{\frac{(1 - W_1)^p + (1 - W_2)^p + ... + (1 - W_t)^p}{t}} \quad (3)$$

Equation 2 represents similarity between conjunctive query q_{or} and document dj . Equation 3 represents similarity between disjunctive query q_{and} and document dj.

The representation of the method used is shown in Fig. 1

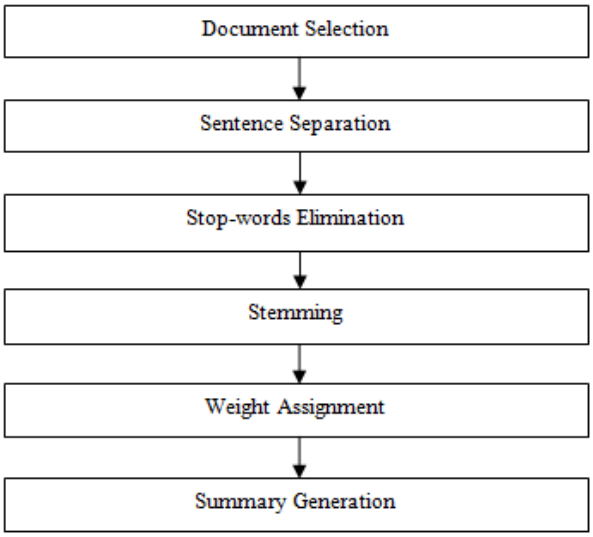

Fig. 1 Steps in automatic text summarization

A. Algorithm

1. Input to the summarizer is a text document of any domain.
2. The summarizer initially breaks the entire document into sentences on the basis of separator. The separator used is full stop (.).
3. After this stop words or unnecessary words which are not relevant in the document are removed from the documents.
4. In the next step, stemming is performed to remove the redundancy from the document by removing the words which are of same meaning from the document. Words are reduced to their stem or root words.
5. Using the stemming mechanism, occurrence of a word is calculated. This is known as term-frequency of any term associated with the document.
6. Using equation (1) weight of each word occurred in a sentence is calculated.
7. Based on weight of each and every word, total weight of a sentence is calculated using the equation (2). Equation (3) can be used in case of multiple document text summarization.
8. After calculating the weight for each sentence, sentences are ranked based on their weights. Highest weight sentence is ranked number 1 followed by other sentences.
9. The final step is producing summary of the document in which highest ranked sentence is included in the summary

IV. TESTING

The summarizer is tested on various domains and result of three domains is included in this paper. Table I shows various domain areas, total number of words and number of unique words on which our text summarizer is tested.

© 2015, IJARCSSE All Rights Reserved

TABLE I VARIOUS DOMAINS ON WHICH SUMMARIZER IS TESTED

Domain Number	**Domain Area**	**Total Number of Words**	**Number Of Unique Words**
1	Software Engineering	404	321
2	Networking	350	295
3	Economics	319	278

A. Result For Domain 1:

Fig. 2 shows Frequency Count of each unique term in respective document. Table II shows summary of the document and frequency count of significant words which are included in the summary.

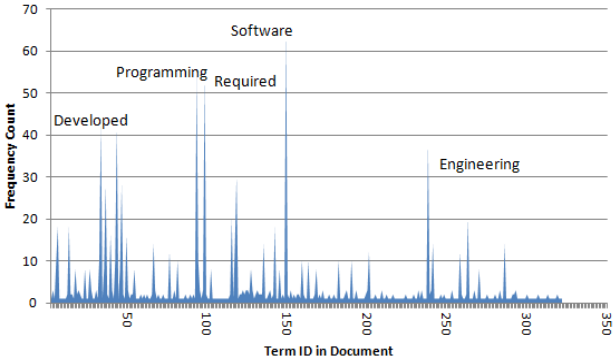

Fig. 2 Frequency count of each unique term in Domain 1

TABLE II SUMMARY AND FREQUENCY COUNT OF SIGNIFICANT TERMS IN INPUT DOCUMENT OF DOMAIN 1

Summary	**Software Engineering** includes the initial development or **programming** of software and its maintenance and updates, till **required** software product is **developed**, which satisfies the expected user requirements.
Words	**Frequency Count**
developed	42
programming	54
required	54
software	65
engineering	38

B. Result For Domain 2:

Fig. 3 shows Frequency Count of each unique term in respective document. Table III shows summary of the document and frequency count of significant words which are included in the summary.

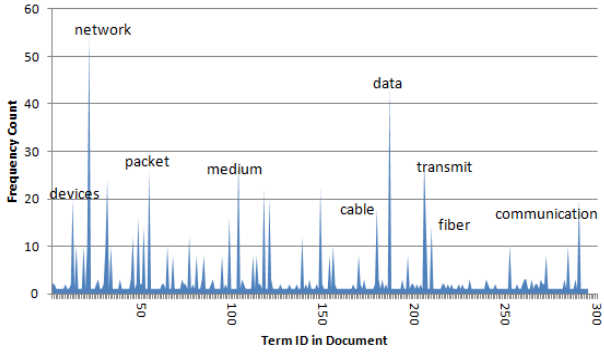

Fig. 3 Frequency count of each unique term in Domain 2

TABLE III VARIOUS DOMAINS ON WHICH SUMMARIZER IS TESTED

Summary	The transmission **medium** (often referred to in the literature as the physical **medium**) used to link **devices** to form a computer **network** to **transmit** information using **data packets** include electrical **cable** (ethernet, homepna, power line **communication**, g.hn), optical **fiber** (fiber-optic **communication**), and radio waves (wireless networking).

© 2015, IJARCSSE All Rights Reserved

Words	**Frequency Count**
network	55
devices	24
packet	26
medium	26
cable	18
data	44
transmit	28
fiber	14
communication	20

C. Result For Domain 3:

Fig. 4 shows Frequency Count of each unique term in respective document. Table IV shows summary of the document and frequency count of significant words which are included in the summary.

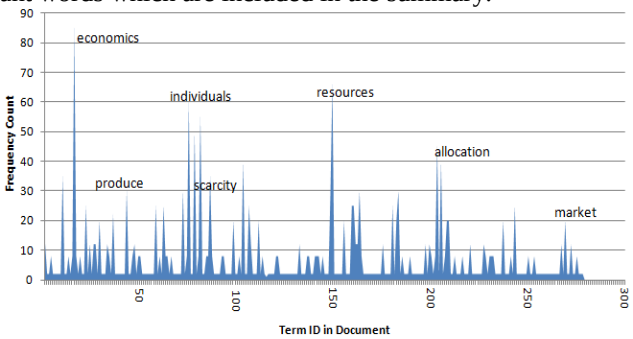

Fig. 4 Frequency count of each unique term in Domain 3

TABLE IV VARIOUS DOMAINS ON WHICH SUMMARIZER IS TESTED

Summary	Microeconomics and macroeconomics are intertwined; as **economics** help to gain understanding of **market**, economists can help nations and **individuals** make more informed decisions for **allocation** of **resources** and to **produce** sufficient **resources** to remove **scarcity**.
Words	**Frequency Count**
economics	85
produce	30
individuals	55
scarcity	35
resources	66
allocation	45
market	20

V. CONCLUSION AND FUTURE SCOPE

We have concentrated on extractive summarization technique. We have tested this summarizer on various domains and results of three domains are included in this paper. We have obtained satisfactory results for these domains. But in some cases due to spread of important information across various sentences, relevant information is missing in summary of those documents. We have implemented this system using only "OR" queries for single document. In future this system can be extended for multiple documents using "AND" queries. For multiple documents, clustering technique can also be used in which similar sentences can be clustered in a group and then highest weighted sentence from each group can be included in the summary of the document. In future, we would also like to improve the system by adding sentence simplification technique for producing summary i.e. it can be used to simplify the sentences which are complex and very large. This approach can also be extended to multi lingual platform. We can also add paraphrasing technique to give abstractive feel to summary.

REFERENCES

[1] K. Jezek and J. Steinberger, "Automatic text summarization", Znalosti, pp.1-12, 2008.

[2] V. Gupta, G. S. Lehal, "A survey of text summarization extractive techniques", *Journal of Emerging Technologies in Web Intelligence*, vol. 2, no. 3, pp. 258-268, August 2010.

[3] A. N. Pimpalshende, "Overview of text summarization extractive techniques", *International Journal Of Engineering And Computer Science*, ISSN:2319-7242 vol. 2 issue 4, pp. 1205-1214, April 2013.

© 2015, IJARCSSE All Rights Reserved

[4] W. Fan, L. Wallace, S. Rich, and Z. Zhang, "Tapping into the Power of Text Mining", Communications of the ACM, vol. 49, issue 9, pp. 76-82, Sep. 2006.

[5] F. Kyoomarsi, H. Khosravi, E. Eslami and P. K. Dehkordy, "Optimizing text summarization based on fuzzy logic", in Proc. of 7th IEEE/ACIS International Conference on Computer and Information Science, IEEE, pp. 347-352, 2008.

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

[7] Vishal Gupta, G.Sl Lehal, "A Survey of Text Mining Techniques and Applications", *Journal of Emerging Technologies in Web Intelligence*, vol. 1, no. 1, pp. 60-76, Aug. 2009.

[8] Hongyan Jing, "Cut-and-Paste Text Summarization", Graduate School of Arts and Sciences, Columbia University, 2001

[9] http://sourceforge.net/projects/autosummarizationtoolusingjava/

[10] A. P. Siva kumar, Dr. P. Premchand and Dr. A. Govardhan, "Query- Based Summarizer Based on Similarity of Sentences and Word Frequency", *International Journal of Data Mining & Knowledge Management Process*, vol.1, no.3, May 2011.

[11] A. Kogilavani, Dr.P.Balasubramani, "Clustering and feature specific sentence extraction based summarization of multiple documents", *International journal of computer science & information Technology*, vol.2, no.4, Aug. 2010.

[12] A. B. Manwar, H. S. Mahalle, K. D. Chinchkhede, Dr. V. Chavan, "A vector space model for Information retrieval: a matlab approach", *Indian Journal of Computer Science and Engineering*, ISSN : 0976-5166 Vol. 3 No. 2, pp. 222-229, Apr-May 2012.

[13] A. R. Deshpande, Lobo L. M. R. J., "Text Summarization using Clustering Technique", *International Journal of Engineering Trends and Technology (IJETT)*, vol.4, Issue 8, ISSN: 2231-5381, pp. 3348-3351, Aug. 2013.

[14] K. Sarkar, "Sentence Clustering-based Summarization of Multiple Text Documents", *TECHNIA – International Journal of Computing Science and Communication Technologies*, vol. 2, no. 1, Jul. 2009.

[15] K. Kaikhah, "Automatic Text Summarization with Neural Networks", Second IEEE International Conference on Intelligent Systems, June 2004.

[16] https://en.wikipedia.org/wiki/Extended_Boolean_model

© 2015, IJARCSSE All Rights Reserved