IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-12, Dec 2018 E-ISSN: 2347-2693

# Extractive Incremental Multi-Document Summarization by Ranking Sentences Relevant to Key Phrase

**J.Tamilselvan[1*], A.Senthilrajan[2]**

[1]Dept. of Computer Science, K.S.Rangasamy College of Arts and Science (Autonomous), Tiruchengode, Tamilnadu, India
[2]Computer Centre, Alagappa University, Karaikudi, Tamilnadu, India

[*]*Corresponding Author: tamilselvan@ymail.com, agni_senthil@yahoo.com*

**Available online at: www.ijcseonline.org**

Accepted: 06/Dec/2018, Published: 31/Dec/2018

***Abstract*—** The summarization deal's with giving the concepts precisely. The multi-document summarization gives the extract of the multiple documents into summarized single document. Here we summarize the document individually by extracting the key phrase using the RAKE algorithm, which perform well on the single document and does not depend on the corpus. This enables the reader to find out the documents, which are highly related to the document by using the TextRank algorithm that ranks the sentence based on the key phrase selected from the single document and they can read the entire document without going through all. The work finds the summary from the given documents and those are ranked and the high ranked documents selected are then used as input to the documents at the next level. The information gained from the previous level (i.e. Summary from documents) are used as the input for the next phase, which will give more information.

***Keywords*—** Multi Document Summarization, Extraction, Sentence Ranking

## I. INTRODUCTION

In this digital era, the growth of information is expanding in exponential throughout the world. when we consider the judicial domain day by day lot of judgments were given it is hard for the judges, lawyers, law scholars and for practitioners to go through all the judgments, the main concept in this judicial system is the lawyers has to quote the judgments previously given to defend their client in the court of law. The solution to this problem is shortening the judgments without losing its sense. Many systems are proposed to link the judgments by the articles and sections referred by it here we prefer the system which connects the judgments semantically.Documents summarization is the automatic system that produces the gist of the text from the single document or multiple documents by holding the information, significance and the order of sentences in the original text. Cohn.T, et. al., States that, "Text summarization is the process of distilling the most important information from a source or sources to produce an abridged version for a particular user or task." [1].

Many algorithms are stated for Knowledge Acquisition but for some specific Domain different style of text handling method are adopted In this situation obtaining Knowledge sources by manual approaches are very tedious and tiresome, rather automatic summarization produces a Knowledge Source that require huge set of training data further the result won't be up to the standard.

In this paper, we present our approach to summarize every single document present in the collection into a separate document based on the key phrase collected from the document. The summarization involves different kinds of information can be taken into account to locate important content, at the corpus level, word level, document level and at the sentence level, the way such attributes interact is likely to depend on the context of specific cases.

We apply the TextRank algorithm to extract the key phrase from the document given as input, graph-based ranking model for contents extracted from natural language texts. Graph-based ranking algorithms are essentially a way of deciding the importance of a vertex within a graph, based on the global information recursively drawn from the entire graph. The task of a key phrase extraction is to automatically, identify the content in a text as a set of terms that best summarize the document. Such key phrase may constitute useful entries for building an automatic index for a document, can be used to classify a text, or may serve as concise summary for a given document. The system for automatic identification of important terms in a text can be used for the problem of terminology extraction, and construction of domain-specific summary.

Section I contains the Introduction of the extractive incremental multi-document summarization system by ranking the key phrase, the Section II comprises of the

 

related work which are carried out of N-Gram, Genetic Algorithm, etc., the Section III consists of the working procedure of the summarization system, Section IV includes the TextRank based sentence selection to get the summary by extractive summarization, Section V has experiment of the data acquired from Aquaint and the data constructed by web crawling, the Section VI includes the analysis of the result and the conclusion given at Section VII by checking with ROUGE to find similarity with the benchmark datasets.

## II. RELATED WORKS

The general term used in text document summarization is "bag-of-words" which calculates the number of occurrences of words and the combination of words that present in a document [2][3]. To enhance the "bag-of-words" representation, many works has been carried out instead of using a word, phrases are used to keep intact the information present in the document. This leads to a "feature vector representation". Replacing single word with the combination of two or more words from a document called "N-gram". If a single word used it is unigram, two words combination forms bigram and so on. The limit for the word selection fixed accordingly to the document, which we represent so that it will yield reasonable feature selection, the "N" is fixed to preserve the information present in the text document, D.Mlademic [4] uses N-grams analysis against single word from a document.

The "TextRank Graph based Model" by Michaleca et al., states that "In the web search technology used for link-structure analysis, citation analysis and social networks perform on the link available in the web page, based on the link and citation the importance of the web page is analyzed. The same concept is uded in the Natural language text documents, where the text documents will not have any reference to other document. Here, the word that combines with other words the higher priority for ranking".

Text Summarization approach has been conducted from the year of 1950 onwards [5]. Since then lot of research work carried on this summarization using word frequency with statistical approach, TF-IDF weighting approach [6]. We apply the graph based strategy to extract the key phrase from a document which does not need to rely on the corpus [7][8][9][10]. The TF-IDF approach using a probabilistic method based on the implicit assumption of TF-IDF classifier by Joachimes [11] proposed a new classifier PrTFIDF which optimizes the parameter selection.

Choi et al., [12] defines how "The hierarchical structure of categories takes control over text classification system". For multi-document summarization, selection and compression of multiple documents, selecting the sentences based on rank, ambiguity among the sentences are the major concerns in summarization.

In the earlier researches Statistical tools plays a vital role to make summaries because of its poor performance various optimization techniques such as Genetic Algorithm [13][14][15]. A Genetic Algorithm based single document summarizer uses sentences based on its features and tested with ROUGE, Particle Swarm Optimization [16][17] and Differential Evolution [18][19][20] are evolved in the recent years. Rautary et al., proposed a single document generic summarizer by comparing the Differential Evolution and Particle Swarm Optimization on document summarization. They also proposed summarization by using sentence features. Alguliev et al., [21] presents a text summarizer, which checks similarity metric to get the whole content and to restrict the summary size from multiple documents.

## III. INCREMENTAL MULTI-DOCUMENT SUMMARIZATION

Incremental Multi-Document Summarization (IMDS) is the process of selecting the documents based on the key-phrase and again some more key-phrases are extracted from the selected document then by using those key-phrases some more documents are added and the iteration goes on until it meets the threshold value. The entire process of IMDS is divided into four steps such as preprocessing, key-phrase extraction, documents selection and summary representation.

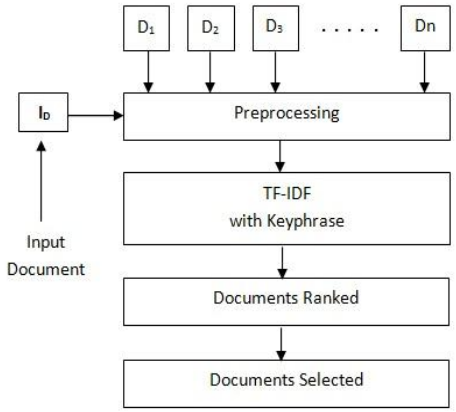


Fig. 1 - Documents Selection

The diagrammatic representation is given in figure – 1. A document Dinp given as input for key-phrase selection and multiple documents given as MD={D1,D2,D3,…Dn}, where each Di represents a single document. At first, the input document (Dinp) and the document collection (MD) for summarization are preprocessed and the key-phrases are extracted from the input document Dinp then documents are selected using the TF-IDF ranking method and finally the summarization is produced for the selected documents.

*Preprocessing*
*Input Representation*
*Documents Selection*
*Summarization*

 

*1 Preprocessing*

Stop Word Removing: This is the first step in the preprocessing, the general terms which are most commonly used words that supports the sentence formation but gives no special meaning are removed from the document.

Stemming: Removes the word endings so as to get the root word from different words like "fish" which is the root word for "fishing, Fisherires, fishes".

Terms Extraction: The unique words are extracted from the document and it is named as terms which referred as tokens where {t1,t2,t3,…,tn} are N number of tokens present in the document.

*2 Input Representation*

The collection of documents represented as {D1,D2,D3,…,Dn} are N number of documents and among these documents a document is given as input for extracting the key-phrase which termed as Dinp.

*3 Document Selection*

The key-phrase is the term served as to select the documents that are highly related with it again the obtained documents used as Input Document for the key-phrase extraction.

*4 Summarization*

The documents that are selected based on the key-phrase relation are summarized by defining the score for each sentence through optimization algorithm are extracted and the sentences with high score are extracted to form the extractive summary by checking the given threshold value.

## IV. TEXTRANK BASED IMDS

We propose the Incremental Multi-Document Summarization method to produce automatic summary of multiple documents which are highly related to the key-phrase, extracted using the TextRank algorithm [22]. The context of the text derived from the entire document rather than from the individual word, that probably does not give more impact about the document.

Generally, the graph based ranking algorithm for natural language text documents have

1. Add key texts to the vertices
2. Connect the vertices that are related with one another
3. Execute step 1 and step 2 until the convergence
4. Using the final score, vertices are sorted for selection/Ranking the key-phrase.

Based on the TextRank model proposed by Michalcea, 2004, we use the TextRank algorithm for (i) Extraction of key-phrases from the text documents that represents the whole document (ii) Deriving the most important sentences from the text document, These above work carried on a single document, here we work with mutli-documents to get summarization of individual documents in respect with the key-phrases generated by TextRank for that we apply (iii) the TF-IDF for page ranking to get the near related text document to the given points (i) and (ii), which are used to build the extractive summary by identifying the sentences.

The steps involved in IMDS is given below

Step 1: A document is given as input, which has the contents needed to extract.

Step 2: The given document in step 1 is preprocessed by using step 3

Step 3: The documents are preprocessed by removing the stop words that are commonly used and produces no or less meaning such as the articles 'a', 'the', etc., Stemming the words to its base form to get the accurate word and word tokenization based on the limit.

Step 4: Calculates the inter word similarity by using the degree of word, the word that supports by other words that gives more sense.

Step 5: Calculate the TF-IDF from the multiple documents where MD={D1,D2,D3,…,Dn} each Di represents the individual documents in the collection.

Step 6: Select the most relevant document that matches the key-phrase.

Step 7: The resultant documents derived from the step 6 are again taken as input and the steps from 4 starts its process again until it reaches the fixed limit.

Step 8: Merge the documents that matches the criteria.

## V. EXPERIMENT

By using the proposed IMDS experiments conducted for the document collection made by us manually. The IMDS results are compared with the benchmark DUC datasets. The IMDS method are implemented using the Python scripts. The summarization result obtained are compared with ROUGE tools using its score.

1. Dataset

The multiple documents collected for summarization are the judgments copy from the year of 2016 and 2017 around 2250 copies were collected from that after the noise removal (converting from PDF to text format) 2000 documents were selected for experiment. The same experiment is carried out to the benchmark datasets DUC AQUAINT 2006 and DUC AQUAINT 2007 datasets. Table describes the contents present in the DUC dataset and our own dataset.

Table 1: Description about the Data sets

| Data set | DUC 2006 | DUC2007 | JUDIS |
|---|---|---|---|
| Document Numbers | 1200 | 1100 | 1600 |
| Average sentence per doc. | 32 | 35 | 30 |
| Max. No. of sentence per doc. | 81 | 126 | 137 |
| Min. No of sentence per doc. | 6 | 11 | 24 |

 

| Data source | AQUAINT | AQUAINT | JUDIS |
|---|---|---|---|
| Summary length | 300 | 300 | 300 |
| Summary level | 4 | 4 | 4 |

## VI. RESULT ANALYSIS

The automatic summarization using N-Gram verified using statistical methods [23][[24]. The result from the proposed method IMDS, is verified through the ROUGE [25] package used to check the summaries produced by the IMDS for the data sets DUC 2006, DUC2007 and our self created data set JUDIS. The ROUGE model will analyze the document summary with reference to the manually created summary. The ROUGE - 1 model verifies the unigram, (single word), The ROUGE – 2 model verifies the bigram (two words) and the ROUGE – N model clocks for N-gram based on the number of words present in the sentence. The recall, precision and F – measure scores are calculated for the benchmark data sets and the data set created by us (JUDIS) are compared. They are closely nearer to the benchmark datasets.

## VII. CONCLUSION

In the paper are proposed a method called IMDS "International multi document summarization". It produces multiple summarizations from the multi – document collection. It produces different levels of summarization. The level 1 summary will be more closer to the key – phrase generated for the Input document, Level 2 will be next closer to the key – phrase and goes on. The summarizes with different levels are compared with the benchmark data sets DVC 2006 and 2007 along with JUDIS data set. From the observations mode it is verified that the system produces summaries for data sets created and the DVC data sets are verified with ROUGE package.

## REFERENCES


[1] Cohn. T and Lapata M, "Sentence compression as tree transduction. J", Artif. Int. Res. 34(1): 637-674, 2009.

[2] D. Koller and M. Sahami, "Hierarchically classifying documents using very few words", Proceedings of the 14th International Conference on Machine Learning, 1998.

[3] K. Lang, "Newsweeder: Learning to filter news", Proceedings of the 12th International Conference on Machine Learning, 331-339, 1995.

[4] D. Mladenic, "Machine Learning on non-homogeneous distributed text data", Ph.D. thesis, University of Ljubjjana, Slovenia, 1998.

[5] Luhn HP, "The automatic creation of literature abstracts", IBM Journal of Research and Development, 159-165, 1958.

[6] Vanderwende L, Suzuki H, Brockett, C and Nenkova A, "Beyond sumbasic: Task-focused summarization with sentence simplification and lexical expansion", Information processing and Management 43(6), 1606-1618, 2007.

[7] Canhasi E and Kononenko I, "Weighted archetypal analysis of the multi element graph for query focused multi-document summarization", Expert systems with Applications 41(2), 535-543, 2014.

[8] Ferreira R, de Souza Cabral L, Freitas F, Lins R D, de Franca Silva G, et al, "A multi document summarization system based on statistics and linguistic treatment", Expert systems with Applications 41(13), 5780-5787, 2014.

[9] Glavas G and Sanjeder J, "Event graphs for information retrieval and multi-document summarization", Expert systems with Applications 41(15), 6904-6916, 2014.

[10] Zhao L, Wu L and Huang X, "Using query expansion in graph-based approach for query focused multi-document summarization", Information Processing & Management, 45(1), 35-41, 2009.

[11] T. Joachims, "A Probablistic analysis of the Rocchio algorithm with TF-IDF for text categorization", International Conference on Machine Learning, 1997.

[12] B. Choi and X. Peng, "Dynamic and hierarchical classification of web pages", Online Information Review, 28(2), 139-147, 2004.

[13] M.A.Fattah, F.Ren, "GA,MR, FFNN, PNN and GMM based models for automatic text summarization", Comput. Speech Lang, 23 (1), 126-144, 2009.

[14] M.D. Gordon, "Probabilistic and genetic algorithms for document retrieval", Commun. ACM 31 (10), 1208-1218, 1988.

[15] Y.X. He, D.X. Liu, D.H. Ji, H.Yang, C.Teng, "Msbga: A multi-document summarization system based on genetic algorithm", Machine learning and Cybernetics, 2006 International Conference on IEEE, August, PP. 2659-2664, 2006.

[16] M.S.Binwahlan, N.Salim, L.Suanmali, "Swarm based text summarization", Computer Science and information Technology-Spring Conference, IACSITSC'09, International Association of, IEEE, April, PP.145-150, 2009.

[17] R.Rautray, R.C.Balabantaray, A. Bhardwaj, "Document summarization using sentence features", Int.J. Inf. Retrieval Res. (IJIRR) 5(1), 36-47, 2015.

[18] R.M. Aliguliyev, "A new sentence similarity measure and sentence based extractive technique for automatic text summarization", Expert Syst. Appl.36 (4), 7764-7772, 2009.

[19] R.M.Alguliev, R.M.Aliguliyev, N.R.Isazade, "CDDS: Constraint – driven document summarization models", Expert Syst. Appl. 40 (2), 458 – 465, 2013.

[20] R.M.Alguliev, R.M.Aliguliyev, C.A. Mehdiyev, "Sentence selection for generic document summarization using an adaptive differential evolution algorithm", Swarm Evolutionary Comput. 1(4), 213-222, 2011.

[21] R.Rautray, R.C.Balabantaray, "Comparative study of DE and PSO over document summarization", Intelligent Computing, Communication and Devices, Springer India, PP. 1-5, 2015.

[22] R.M.Alguliev, R.M.Aliguliyev, M.S. Hajirahimova, C.A. Mehdiyev, "MCMR: Maximum coverage and minimum redundant text summarization model", Expert Syst. Appl. 38 (12), 14514 – 14522, 2011.

[23] S.L. Patil, K.P.Adhiya, "Textual Similarity Detection from Sentence", International Journal of Computer Sciences and Engineering, Sep, PP.835-839, 2018.

[24] B. Batra, S. Sethi, A.Dixit, "Improved Text Summarization Method for Summarizing Product Reviews", International Journal of Computer Sciences and Engineering, Sep, PP.113-122, 2018.

[25] C.Y. Lin, E.Hovy, "Automatic evaluation of summaries using n-gram co-occurrence statistics", Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology – Volume 1, Association for Computational Linguistics, May, PP.71-78, 2003.


 