Single Document Text Summarization Using Clustering Approach Implementing for News Article

Pankaj Bhole[#1], Dr. A.J. Agrawal[*2]

[#1] *M.tech Scholar, Department of Computer Science, Shri Ramdeobaba College of Engineering & Management, Nagpur, India*
[#2] *Associate Professor Department of Computer Science, Shri Ramdeobaba College of Engineering & Management, Nagpur, India*

Abstract— **Text summarization is an old challenge in text mining but in dire need of researcher's attention in the areas of computational intelligence, machine learning and natural language processing. We extract a set of features from each sentence that helps identify its importance in the document. Every time reading full text is time consuming. Clustering approach is useful to decide which type of data present in document. In this paper we introduce the concept of k-mean clustering for natural language processing of text for word matching and in order to extract meaningful information from large set of offline documents, data mining document clustering algorithm are adopted.**

Keywords— **Natural Language Processing, Stemming,, Clustering**.

I. INTRODUCTION

With the rapid growing popularity of the Internet and a variety of information services, obtaining the desired information within a short amount of time becomes a serious problem in the information age. Automatic text summarization provides an effective means to access the exponentially increased collection of information. This technology may also benefit text processing such as document classification (Shen et al. 2004)[1] and question answering (Demner-Fushman and Lin 2006)[2].

Automated text summarization focused two main ideas have emerged to deal with this task; the first was how a summarizer has to treat a huge quantity of data and the second, how it may be possible to produce a human quality summary. Depending on the nature of text representation in the summary, summary can be categorized as an abstract and an extract. An extract is a summary consisting of a number of salient text units selected from the input. An abstract is a summary, which represents the subject matter of the article with the text units, which are generated by reformulating the salient units selected from the input. An abstract may contain some text units, which are not present in to the input text. In general, the task of document summarization covers generic summarization and query-oriented summarization. The query-oriented method generates summaries of documents according to given queries or topics, and the generic method summarizes the overall sense of the document without any additional information.

Traditional documents clustering algorithms use the full-text in the documents to generate feature vectors. Such methods often produce unsatisfactory results because there is much noisy information in documents. The varying-length problem of the documents is also a significant negative factor affecting the performance. This technique retrieves important sentence emphasize on high information richness in the sentence as well as high information retrieval. These multiple factors help to maximize coverage of each sentence by taking into account the sentence relatedness to all other document sentence.

These related maximum sentence generated scores are clustered to generate the summary of the document. Thus we use k-mean clustering to these maximum sentences of the document and find the relation to extract clusters with most relevant sets in the document, these helps to find the summary of the document. The main purpose of k-mean clustering algorithm is to generate pre define length of summary having maximum informative sentences. In this paper we present the approach for automatic text summarization by extraction of sentences from the Reuters-21578 corpus which include newspaper articles and used clustering approach for extraction summary. Work done for Text Summarization is given in the section (II). Section (III) provided our methodology for Text Summarization, Section (IV) provide the result of our text summarization system.

1.1 Motivation

The motivation of natural language based text summarization system on newspaper come from news based application for mobile. Every person wants to be globalized with knowledge and information. Most of the user read news on mobile application. But the news always very large and descriptive. In modern world everyone wants fast and full information, so in this case reading complete news time consuming.

So for fasten and important news we can provide text summarization system that will analysis text information and generate short, optimal, knowledge based summary to end user. This will help us to save time and form better summary.

II. RELATED WORK

The text summarization has drawn attention primarily after the information explosion on the Internet, the first work has been done as early as in the 1950s (Luhn, 1958)[3]. Extractive

summarization selects sentence s from documents to form summaries directly without any sort of paraphrase. Many automated techniques for text summarization exist. Among the most successful of these techniques are those that are based on the position of terms in the source document and those that are based on Text Retrieval. Lead summaries are based on the idea that the leading terms (i.e., the first terms) in the source document are the most important.

Filatova and Hatzivassiloglou (2004)[4] modeled extractive document summarization as a maximum coverage problem that aims at covering as many conceptual units as possible by selecting some sentences. The format of summaries is another criterion to differentiate text-summarization approaches. Usually, a summary can be an extract or an abstract. In fact, a majority of researches have been focused on summary extraction, which selects salient pieces (keywords, sentences or paragraphs) from the source to yield a summary. In extractive document summarization, finding an optimal summary can be viewed as a combinatorial optimization problem which is NP-hard to solve. There are a few papers exploring an optimization approach to document summarization. The potential of optimization based document summarization models has not been well explored to date. This is partially due to the difficulty to formulate the criteria used for objective assessment. As far as we know, the idea of optimizing summarization was mentioned in Filatova and Hatzivassiloglou (2004)[4].

Yong et al. [5] worked on developing an automatic text summarization system by combining both a statistical approach and a neural network. Mohamed Abdel Fattah & Fuji Ren [6] applied a model based on a genetic algorithm (GA) and mathematical regression (MR) in order to obtain a suitable combination of feature weights to summarize one hundred English articles. Hamid et al. [7] proposed a new technique to optimize text summarization based on fuzzy logic by selecting a set of features namely sentence length, sentence position, titles similarity, keywords similarity, sentence-to-sentence cohesion and occurrence of proper names. The summarization approach discussed by Jing and McKeown [9] is based on statistical methods. Initially the most important sentences are extracted from the source text. The extracted sentences are then joined together by analyzing their discourse structure and modifying them as required.

Clustering-based approaches were explored in recent years. For example, Qazvinian and Radev [8] applied hierarchical agglomerative clustering algorithm to obtain sentence clusters, and then developed two strategies to extract sentences from the clusters to build a summary. One was to extract the first sentence in the order it appeared in the original documents from the largest to the smallest cluster, then the second ones and so on, until the summary length limit is reached. Wan and Yang [10], on the other hand, proposed a clustering-based HITS model which formalized the sentence-cluster relationships as the authority-hub relationships in the HITS algorithm. Finally sentences which had high authority scores were selected to form a summary. Besides, Wang et al. [10] proposed a language model to simultaneously cluster and summarize documents. A flaw of the clustering-based approaches is that clustering and ranking are independent of each other and thus they cannot share the information that is useful for both, e.g. the spectral information of sentence similarity matrix. A new approach that can really couple clustering and ranking together is required in order to improve the performance of each other.

III. PROPOSED WORK

Automatic Text Summarization important for several tasks, such as in search engine which provide shorter information as result. Assuming that the summarization task is to find the subset of sentences in text which in some way represents main content of source text, then arises a natural question: ‘what are the properties of text that should be represented or retained in a summary’. A summary will be considered good, if the Summary represents the whole content of the document. Motivated from Text Summarization, we have used decided to use this approach for information extraction. This is very difficult to do abstractive summarization because of very large text and their interdependence between sentences, difficult to make abstractive summary. We have proposed Text Summarization methodology as follows.

In this section, we describe in detail the various components of the framework of the our methodology

The major components are:

- Pre-processing
- Sentence clustering
- Cluster ordering
- Representative sentence selection
- Summary generation

3.1 Pre-Processing

We provide the input in the form of text document. This text contains many unnecessary text data and symbols. So that text will not give any optimal solution. For efficient and important summary we need to remove the unnecessary data.

Therefore pre-processing is the necessary and first step of application. In pre-processing we apply Stop Word Removal, Stop Symbol Removal, White space removal, and Stemming to make root form of word in preprocess text. Here we use the Word Net Library for efficient stemming. If there are different words but same root form the it count as single word instead of counting individually.

Stop Words={that, in, this, so, we, is , are , had, have, because, …}

Stop Symbol={ @, &, #, *, (,), !.”, +, _ ,- ,…. }

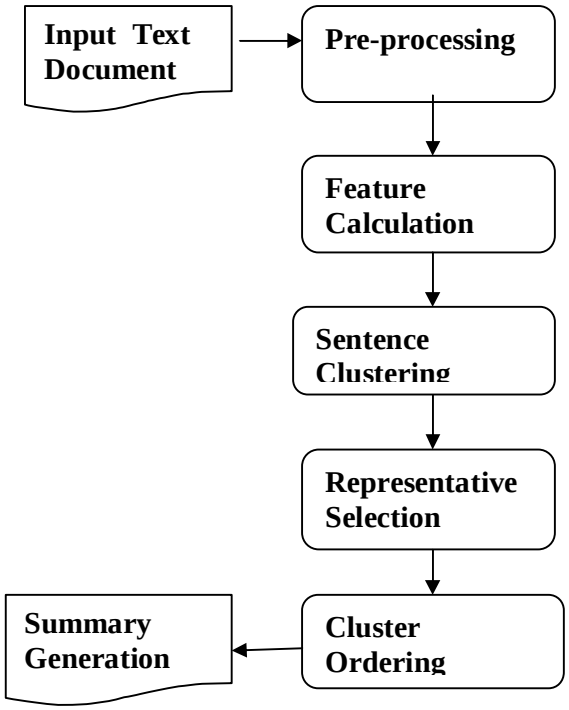

Fig 1:The framework of the proposed sentence clustering based summarization system

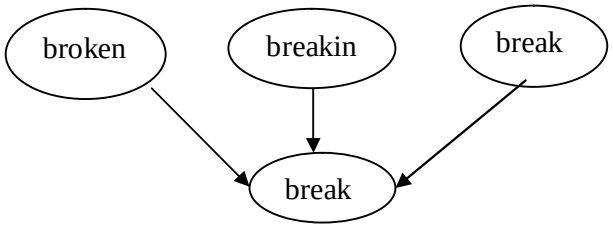

Fig. 2: Example of stemming of different forms of word brake

3.2 Some Feature Calculation:

For efficient summarization, it is necessary to calculate some efficient feature for optimizing the clustering and summary of text.

A) Term Frequency:

The hypothesis assumed by this approach is that if there are ‘‘more specific words’’ in a given sentence, then the sentence is relatively more important. The target words are usually nouns except for temporal or adverbial nouns (Satoshi et al., 2001)[11] (Murdock,2006)[12]. This algorithm performs a comparison between the term frequencies (TF) in a document

$$TF(W)=\frac{NUMBER\ OF\ W\ IN\ DOCUMENT}{TOTAL\ NUMBER\ OF\ TERMS\ IN\ DOCUMENT}$$

B) Cosine Similarity:

Cosine similarity is a popular sentence-to-sentence similarity metric used in many clustering and summarization tasks[13][14]. Sentences are represented by a vector of weights while computing cosine similarity. But, the feature vector corresponding to a sentence becomes too sparse because sentences are too short in size compared to the input collection of sentences. Sometimes it may happen that two sentences sharing only one higher frequent word show high cosine similarity value.

$$Sim(S_i, S_j)=(2*|S_i \frown S_j|) / (|S_i|+|S_j|)$$

Where S i and S j are any two sentences belonging to the input collection of sentences.
The numerator |Si⌢ Sj| represents number of matching words between two sentences and
| Si| is the length of the i-th sentence, where length of a sentence =number of words in the sentence.

3.3 Sentence Clustering

Sentence clustering is the important component of the clustering based summarization system because sub-topics or multiple themes in the input document set should properly be identified to find the similarities and dissimilarities across the documents.

Clustering of sentences provide grouping the sentence which provide similar information. Sentence clustering is the important component of the clustering based summarization system because sub-topics or multiple themes in the input document set should properly be identified to find the similarities and dissimilarities across the documents. Clustering should be tight and not generate redundancy of sentences in inter-cluster and intra-cluster.

Here K-Mean is suitable for this type of clustering. It makes classification of vector on distant measure. We are calculating distance matrix from the cosine similarity matrix.

Dist(s1,s2)=1-Cosine(s1, s2)

3.4 Cluster Ordering

Since our sentence-clustering algorithm is fully supervised and it assume prior knowledge about the number of clusters to be formed, it is crucial to decide which cluster would contribute the representative first to the summary. Instead of considering the count of sentences in a cluster as the cluster importance, we measure the importance of a cluster based on the number of important words it contains.

3.5 Representative Sentence Selection

Selecting most informative sentences from cluster need ranking algorithm to give the sentences. After ranking sentences in the cluster based on its scores, the sentence with highest score is selected as the representative sentence

3.6 Summary Generation

We select one sentence from the topmost cluster first and then continue selecting the sentences from the subsequent clusters in ordered list until a given summary length is reached.

4 EXPERIMENTAL RESULT

IR researchers have developed evaluation measures specifically designed to evaluate efficiency in summary. Most of these measures combine number of sentences and amount of information. The result of text summarization with clustering is calculated on the basis of number of cluster and number of informative sentences.

Here we are using unsupervised data but supervised clustering method. We pre-define number of cluster for grouping of sentences. We used reuter 21578 newspaper corpus for testing and experimental purpose. The details of reuter 21578 is in table 1.

Table 1. Detail of reuter 21578 dataset

Number of files	21
Document in each Files	Nearly 1000
Total	21578

Preprocessing is most important step during summarization it remove the unwanted word that will only increase the value of that word but not the information about text.

After the preprocessing the text will be more compressed and useful. The preprocess text likely to be 80% of input text. Those preprocess text will used for further any feature calculation or text operation.

Table 2: Number of words at different preprocessing step

Input text (number of words)	After stop word removal	After stop Symbol removal	Preprocess text
123	117	90	70%
279	250	213	82%

Clustering is one of the effectiveness factor of text summarization. It eliminate similar type of sentences by clustering them in single cluster and selecting most informative senteces among them. If any news article is repeating same thing repeatadly using same set of words then clustering method select one or two sentences from semilar sentences according selective criteria and add that sentences in summary.

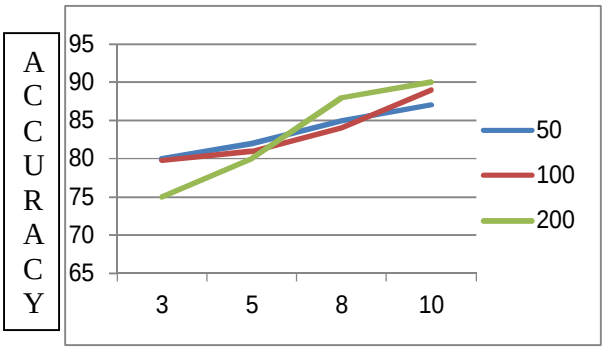

Fig 3: Graph shows relationship between number of sentence and number of cluster

If the number of sentences in text is increases then it is necessary to increase number of cluster for more descriptive summary. For example if sentence are 50 then cluster number should be near 7 and for 100 sentences cluster number should be 10 that give one third summaries.

5 CONCLUSION AND FUTURE SCOPE

Text summarization on newspaper article is helpful to generating short news from large article that give informative but short updates. The use of this application for mobiile news application which very helpful in small display device. This is time saving application. Without reading full news article and wastage of time we can get easily important detail about news. The result of this application will be enhanced by using spectral clustering and may side by side feature extration to provide high level clustering.

In Future work may continue in following directions:

- The system can be implemented for multiple news article.
- System may also consider the most high valuable news from all news article

References

[1] Shen, D., Chen, Z., Yang, Q., Zeng, H., Zhang, B., Lu, Y., et al (2004). Web-page classification through summarization. In Proceedings of the 27th annual international ACM SIGIR conference on research and development in information retrieval. ACM, p. 249

[2] Demner-Fushman, D., & Lin, J. (2006). Answer extraction, semantic clustering, and extractive summarization for clinical question answering. In Proceedings of the 21st international conference on computational linguistics and the 44th annual meeting of the association for computational linguistics. Association for computational linguistics (p. 848)

[3] Luhn, H. P. (1958). The automatic creation of literature abstracts. IBM Journal of Research and Development, 2(2), 159–165.

[4] Filatova, E., & Hatzivassiloglou, V. (2004). A formal model for information selection in multi-sentence text extraction. In Proceedings of the 20th international conference on computational linguistics (COLING'04), Geneva, Switzerland, August 23–27 (pp.397–403)

[5] Yong, S.P., Ahmad I.Z. Abidin and Chen, Y.Y. (2005). 'A Neural Based Text Summarization System', 6th International Conference of DATA MINING, pp.45-50.

[6] Mohamed Abdel Fattah and Fuji Ren (2008). 'Automatic Text Summarization', International Journal of Computer Science, Vol., No.1, pp.25-28
Hamid Khosravi, Esfandiar Eslami, Farshad Kyoomarsi and Pooya Khosravyan Dehkordy (2008). 'Optimizing Text Summarization Based on Fuzzy Logic", Springer-Verlag Computer and Information Science, SCI 131, pp.121-130.

[7] V. Qazvinian, D.R. Radev, Scientific paper summarization using citation summary networks, in: Proceedings of 22nd International Conference on Computational Linguistics, 2008, pp. 689–696

[8] H. Jing and K. McKeown. Cut and paste based text summarization. In Proc. of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics, pages 178--185, 2000

[9] D.D. Wang, S.H. Zhu, T. Li, Y. Chi, Y.H. Gong, Integrating clustering and multi-document summarization to improve document understanding, in: Proceedings of ACM 17th Conference on Information and Knowledge Management, 2008, pp. 1435–1436

[10] Satoshi, Chikashi Nobata., Satoshi, Sekine., Murata, Masaki., Uchimoto, Kiyotaka., Utiyama, Masao., & Isahara, Hitoshi. (2001). Keihanna human info-communication. Sentence extraction system assembling multiple evidence. In Proceedings 2nd NTCIR workshop (pp. 319–324).

[11] Murdock, Vanessa Graham. (2006). Aspects of sentence retrieval. Ph.D. thesis, University of Massachusetts, Amherst

[12] G. Erkan and D. R. Radev. LexRank: Graph-based centrality as salience in text summarization. Journal of Artificial Intelligence Research (JAIR), (2004).

[13] X. Wan: Using only cross-document relationships for both generic and topic-focused multi-document summarizations. InformationRetrieval (2008) 11:25–49 1997.