IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 3, No. 2, June 2014, pp. 73~78
ISSN: 2252-8938

Extractive Based Single Document Text Summarization Using Clustering Approach

Pankaj Bhole, A. J. Agrawal
Departement of Computer Science and Engineering, SRCOEM, Nagpur ,India

Article Info

Article history:

Received Dec 22, 2013
Revised Mar 22, 2014
Accepted Apr 28, 2014

Keyword:

K-mean clustering
Stemming
Term Frequency
Text summarization

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.

Corresponding Author:

Pankaj Bhole,
Departement of Computer Science and Engineering,
Shri Ramdeobaba College of Engineering and Manegment, Nagpur
Email: bholepankaj123@gmail.com

1. 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. Document summarization can generate a summary that contains the most important points of a document, which has been applied to many specific domains including biomedical (Ling et al., 2007), email threads summarization (Zajic et al.,2008) and patent document analysis (Tseng et al. 2007) [3]. This technology may also benefit text processing such as document classification (Shen et al. 2004) [4] and question answering (Demner-Fushman and Lin 2006) [5].

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.

Yong et al. [6] worked on developing an automatic text summarization system by combining both a statistical approach and a neural network. Mohamed Abdel Fattah & Fuji Ren [7] 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. [8] proposed a new technique to

optimize text summarization based on fuzzy logic by selecting a set of features namely sentence length, entence position, titles similarity, keywords similarity, sentence-to-sentence cohesion and occurrence of proper names [9].

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 will helps in better summary.

2. PROPOSED METHOD

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:

a. Pre-processing
b. Sentence clustering
c. Cluster ordering
d. Representative sentence selection
e. Summary generation

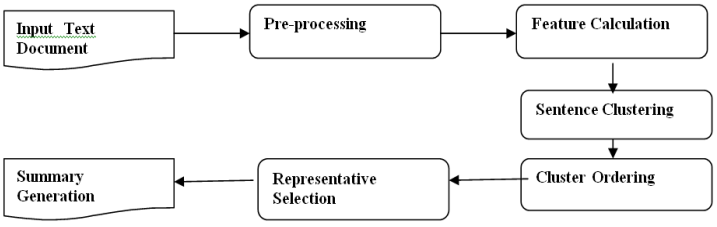


Figure 1. The framework of the proposed sentence clustering based summarization system

2.1. 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={ @, &, #, *, (,), !.", +, _ ,- ,…. }

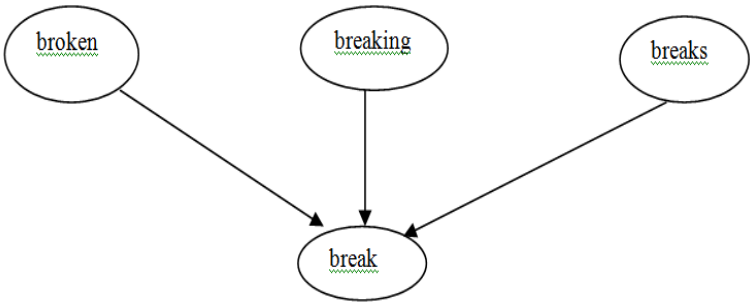


Figure 2. Example of stemming of different forms of word broken

2.2. 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) [1] (Murdock, 2006) [2]. 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 [10], [11]. 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⌢ S j|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.

2.3. 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)

2.4. 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.

2.5. 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

2.6. 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.

3. RESULTS

The Experimental result is applied on reuter 21578 newspaper corpus.

Table 1. Detail of Reuter21578 dataset

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

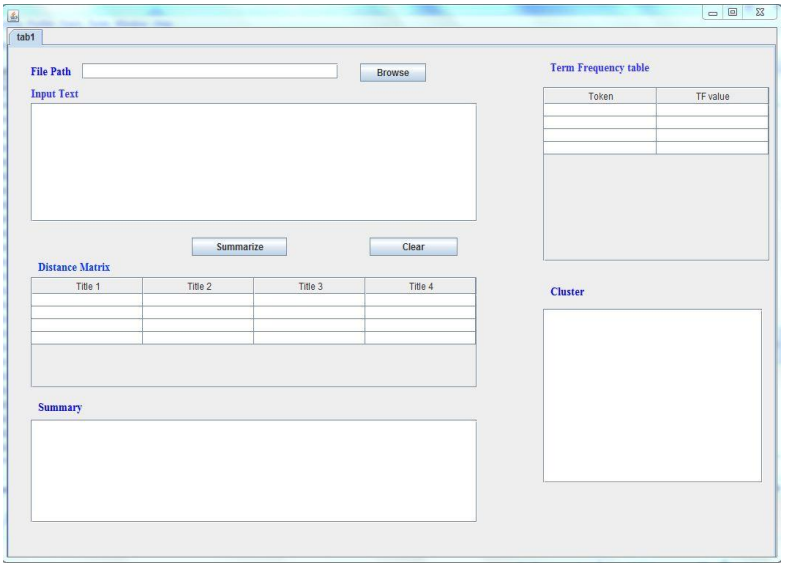
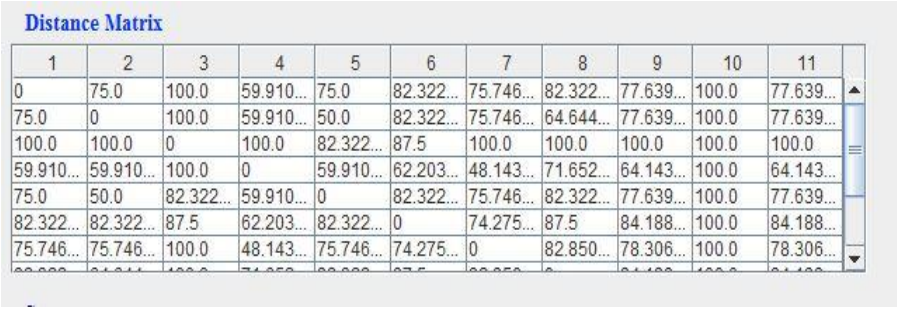


Figure 3. Main GUI of our application provide input text box, intermediate result and output text box

Term Frequency table

Token	TF value
advantage	0.7246376811594203
animal	0.7246376811594203
asia	1.4492753623188406
belong	1.4492753623188406
bengal	2.1739130434782608
black	0.7246376811594203
body	0.7246376811594203
cat	0.7246376811594203
central	0.7246376811594203
claw	0.7246376811594203
coat	1.4492753623188406
color	0.7246376811594203

Figure 4. The Term Frequency table of input text



Distance Matrix

1	2	3	4	5	6	7	8	9	10	11
0	75.0	100.0	59.910...	75.0	82.322...	75.746...	82.322...	77.639...	100.0	77.639...
75.0	0	100.0	59.910...	50.0	82.322...	75.746...	64.644...	77.639...	100.0	77.639...
100.0	100.0	0	100.0	82.322...	87.5	100.0	100.0	100.0	100.0	100.0
59.910...	59.910...	100.0	0	59.910...	62.203...	48.143...	71.652...	64.143...	100.0	64.143...
75.0	50.0	82.322...	59.910...	0	82.322...	75.746...	82.322...	77.639...	100.0	77.639...
82.322...	82.322...	87.5	62.203...	82.322...	0	74.275...	87.5	84.188...	100.0	84.188...
75.746...	75.746...	100.0	48.143...	75.746...	74.275...	0	82.850...	78.306...	100.0	78.306...

Figure 5. The Distance matrix of input text provide dissimilarity value of sentences

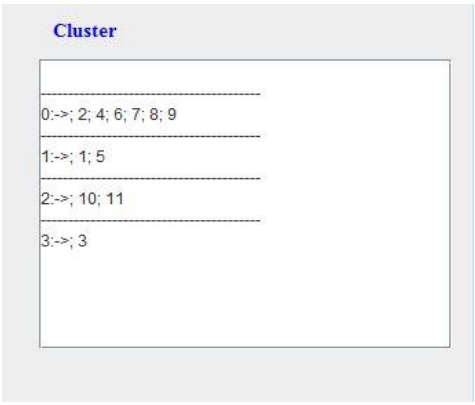


Figure 6. Clustering of sentences using Kmean algorithm

4 CONCLUSION AND FUTURE WORK

In this paper we have seen the how the Kmean clustering is applicable in summarization and how the cluster number is effective on qualitative summary Our work focuses on the design of a successful clustering based summarization and the related issues such as how to cluster sentences, how to order clusters and how to select representative sentences from the clusters. The better similarity measure will improve the clustering performance and this may improve the summarization performance. This summarization applied on news article or document for brief summary. If the sentence in input text increases then for better summary number of cluster should be increases.

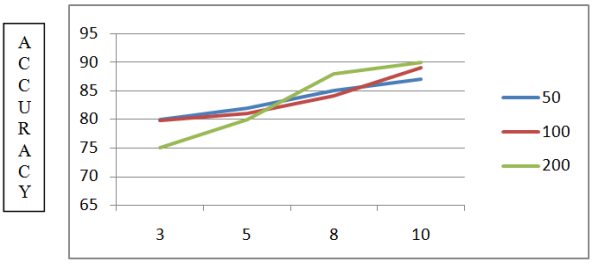


Figure 7. The relation between number of cluster to number of sentences

REFERENCES

[1] Satoshi, Chikashi Nobata., Satoshi, Sekine., Murata, Masaki., Uchimoto, Kiyotaka., Utiyama, Masao., & Isahara, Hitoshi. *Keihanna human info-communication. Sentence extraction system assembling multiple evidence*. In Proceedings 2nd NTCIR workshop, pp. 319–324, 2001.
[2] Murdock, Vanessa Graham. Aspects of sentence retrieval. Ph.D. thesis, University of Massachusetts, Amherst. 2006.
[3] Tseng, Y., Lin, C., & Lin, Y. Text mining techniques for patent analysis. *Information Processing & Management*, vol 43(5), pp 1216–1247, 2007.
[4] Shen, D., Chen, Z., Yang, Q., Zeng, H., Zhang, B., Lu, Y., et al. Web-page classification through summarization. In Proceedings of the 27th annual international ACM SIGIR conference on research and development in information retrieval. ACM, pp. 249, 2004.
[5] Demner-Fushman, D., & Lin, J. *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, pp. 848, 2006.
[6] Yong, S.P., Ahmad I.Z. Abidin and Chen, Y.Y. *A Neural Based Text Summarization System*, 6th International Conference of DATA MINING, pp. 45-50, 2005 .
[7] Mohamed Abdel Fattah and Fuji Ren. Automatic Text Summarization, *International Journal of Computer Science*, No.1, pp.25-28, 2008.
[8] Hamid Khosravi, Esfandiar Eslami, Farshad Kyoomarsi and Pooya Khosravyan Dehkordy. Optimizing Text Summarization Based on Fuzzy Logic", *Springer-Verlag Computer and Information Science*, SCI 131, pp. 121-130, 2008.
[9] Mohammed Salem Binwahlan, Naomie Salim and Ladda Suanmali. 'Swarm Based Features Selection for Text Summarization', *International Journal of Computer Science and Network Security*, Vol. 9, No. 1, pp. 175-179, 2009.
[10] G. Erkan and D. R. Radev. LexRank: Graph-based centrality as salience in text summarization. *Journal of Artificial Intelligence Research (JAIR)*, 2004.
[11] X. Wan. Using only cross-document relationships for both generic and topic-focused multi-document summarizations. InformationRetrieval. Vol 11: 25–49, 2008.

BIOGRAPHIES OF AUTHORS

Pankaj K Bhole: received Bachelor of Engineering Degree in Information Technology from Amravati University, and Master of Technology degree in Computer Science & Engineering from Shri Ramdeobaba College of Engineering & Management Nagpur, India in 2012and 2014 respectively. His research area is Natural Language Processing. He is having 11 months of teaching experience.

Avinash J. Agrawal: received Bachelor of Engineering Degree in Computer Technology from Nagpur University, India and Master of Technology degree in Computer Technology from National Institute of Technology, Raipur, India in 1998 and 2005 respectively. He received Ph.D. from Visvesvaraya National Institute of Technology, Nagpur, India in 2013. His research area is Natural Language Processing and Databases. He is having 15 years of teaching experience. Presently he is Assistant Professor in Shri Ramdeobaba College of Engineering & Management Nagpur, India He is the author of seven research papers in International and National Journal, Conferences.