ISSN: 0976-3031

Available Online at http://www.recentscientific.com

CODEN: IJRSFP (USA)

International Journal of Recent Scientific Research
Vol. 9, Issue, 3(H), pp. 25182-25184, March, 2018

International Journal of Recent Scientific Research

DOI: 10.24327/IJRSR

Research Article

CLUSTERING FOR TEXT SUMMARIZATION

Ranjeet Ramesh Pawar and Mithun Vishnu Mhatre

Bharati Vidyapeeth Institute of Technology, Navi Mumbai

DOI: http://dx.doi.org/10.24327/ijrsr.2018.0903.1807

ARTICLE INFO

Article History:

Received 15th December, 2017
Received in revised form 25th January, 2018
Accepted 23rd February, 2018
Published online 28th March, 2018

Key Words:

Text Mining, Clustering, Summarization, Dynamic partitioning, Sentence clustering.

ABSTRACT

Text summarization is a process of reducing the size of a text while preserving its information content. This paper proposes a sentences clustering based summarization approach. With the abundance of documents through World Wide Web and corporate document management systems, the dynamic partitioning of texts into previously unseen categories ranks top on the priority list for all business intelligence systems. This increase in both the volume and the variety of data requires advances in methodology to automatically understand, process, and summarize the data. Data Clustering can be considered the most important *unsupervised learning* problem.

INTRODUCTION

Efficient and effective text document categorization is the need of every system, so that the documents can be grouped into the desired and required categories. Text mining [1] - [13] includes basically two different techniques, text classification and text clustering. The text classification [1] - [7] techniques and methods require prior knowledge and information as input and make the classifier learnable by providing the required processing with some additional cost, whereas text clustering [3], [13] provides a base to cluster the documents using unsupervised learning, where no prior information is available for this.

Text classification plays an important, appealing and vital role to legitimately categorize the text documents using supervised learning techniques. To reduce the large volume of extracted feature sets from the text documents, they must be compressed or made low-dimensional, so that their use and meaning are not lost and the system performance is increased greatly.

Data clustering has been used for the following three main purposes.

1. Underlying structure: to gain insight into data, generate hypotheses, detect anomalies, and identify salient features.
2. Natural classification: to identify the degree of similarity among forms or organisms (phylogenetic relationship).
3. Compression: as a method for organizing the data and summarizing it through cluster prototypes.

Typically text mining tasks include information extraction and text clustering [7], text classification, information retrieval and text summarization. Researchers have put a lot of interest on mining data that are in the form of structured format where they assume that the information to be mined is already in the form of a relational database [5]. Unfortunately some research papers, e-books and news articles are in the form of unstructured format which is not easy to apply data mining or knowledge discovery directly. It needs other processing techniques to allow data mining to be applied. A large amount of data available on Internet is in the form of unstructured long text documents. This information is being read and analyzed by many different people for different purposes like knowledge discovery, for decision-making and knowledge management through text mining. Text mining as the automatic discovery of new, previously unknown useful knowledge from unstructured text starts by extracting information (facts and events) from textual documents and then enables traditional Data Mining and data analysis methods to be applied. Document clustering (also known as text clustering) is one of the most important text mining methods that are developed to help users effectively navigate, summarize, and organize text documents [6].

Corresponding author:* **Ranjeet Ramesh Pawar
Bharati vidyapeeth Institute of Technology, Navi Mumbai

Clustering Algorithms

Clustering algorithms may be classified as listed below

1. Flat clustering (Creates a set of clusters without any explicit structure that would relate clusters to each other; It's also called exclusive clustering)
2. Hierarchical clustering (Creates a hierarchy of clusters)
3. Hard clustering (Assigns each document/object as a member of exactly one cluster)
4. Soft clustering (Distribute the document/object over all clusters)

The System

A summarization system consists of reduction of a text document to generate a new form which conveys the key meaning of the contained text. Due to the problem of information overload, access to sound and Correctly-developed summaries is necessary. Text summarization is the most challenging task in information retrieval. Data reduction helps a user to find required information quickly without wasting time and effort in reading the whole document collection. This paper presents a combined approach to document and sentence clustering as an extractive technique of summarization.

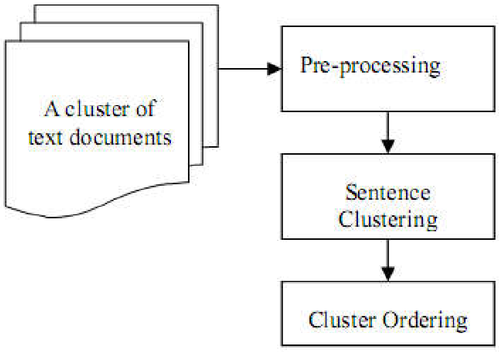
2. Features for Text Summarization.

Ageneric summary of a document has to reflect its key points. We need here statistical features which give different information about the relevance of sentences for the summary. If these features are sufficiently relevant for the SDStask, one can expect that

They assign high scores to summary sentences in a document but rank them differently. We argue that there exists an optimal combination of these features which gives better results than the performance of the best feature. These features constitute the input of

The ML algorithm we developed here[18] defined different sentence feature she considered important for a generic summary and grouped them into seven categories: Indicator phrases (such ascue-words oracronyms), Frequency and title keywords, location as well as sentence length cut-off heuristics and the number of semantic links between a sentence and its neighbours. These features have partially or completely been used in the state of the art since then.[13,15,9].

A summarization system consists of reduction of a text document to generate a new form which conveys the key meaning of the contained text. Due to the problem of information overload, access to sound and correctly developed summaries is necessary. Text summarization is the most challenging task in information retrieval. Data reduction helps a user to find required information quickly without wasting time and effort in reading the whole document collection. This paper presents a combined approach to document and sentence clustering as an extractive technique of summarization. Framework of document clustering summarization system.



Preprocessing

The preprocessing task primarily includes removal of stop words (prepositions, articles and other low content words), punctuation marks (except dots at the sentence boundary).

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. If sentences are grouped in to a predefined number of clusters, the clusters may not be coherent because some sentences may be forcibly assigned to some clusters although it should not be. The incoherent clusters may contain duplicate text units, which may lead to the selection of the redundant sentences in to summary. On the other hand, if the clusters are very tight, most of the clusters may be converted to singletons. Thus, we should have a clustering method, which ensures the coherency of the clusters and minimizes inter-cluster distance. For the sentence clustering, we adopt the similarity histogram based incremental clustering method presented in[10]. The clustering algorithm presented in [10] has been used for web document clustering. The method to sentence clustering. The sentence clustering task is not totally similar to the document clustering task because the sentences are short and less informative compared to documents. One of the important factors of any clustering technique is how to computes imilarity between two objects.

Representative Sentence Selection

Arbitrary (random) selection, longest candidate selection, sentence selection based on its similarity to the centroid of input document set. Sentence selection based on local and global importance. Ideally, the random selection can be thought to be a solution expecting that all the sentences in the cluster are perfectly similar to each other and any member of the cluster is sufficient to represent the cluster theme. But, practically, it does not happen so, because we traditionally use a similarity threshold to judge whether two sentences are similar or not. Hence, two similar sentences in clusters may share some dissimilar information. We observe that setting the similarity threshold to the highest possible value (i.e., 1) does not improve the summarization performance because it converts most of the clusters to single tons. The second solution i.e., longest candidate selection can be thought to be useful assuming that the sentences in the clusters are similar to each other and the longest sentence in the cluster can be the true representative sentence. The third method considers a sentence in a clusteras the representative sentence if it is closest to the common centroid. The common centroid is considered as the pseudo document consisting of a number of words whose weight is greater than a predefined threshold and it is the centroid of the cluster of input documents. We compute the weight of a word using the formula: log(1+tf), where tf (term frequency) = total number of times a term (word) occurs in the input collection of documents (Stop words are not taken into count). Closeness of a sentence to the centroid is measured by summing up the weights of centroid words appearing in the sentence.

CONCLUSION

As clustering is especially useful for organizing documents to improve retrieval and support browsing [3], we can use it to group research papers. Because we want to have meaningful cluster topics, we use multi word features in clustering process in such way that initial centroids are made of phrases. Another feature in our method is that we do not need to process

full text of paper, since it is a time consuming work, we only need a little important information of the paper which is enough to represent the paper's information. In automatic techniques for subject classification of research paper documents, a simple approach is to do a keyword based search for subject term or some of its synonyms in paper's title, keywords, and full text. On one hand, title and keywords provide only limited information which may lead to inaccurate decision and on the other hand, processing the whole text of a paper also takes a long time [9].

References

1. K.R. McKeown, R. Barzilay, D. Evans, V. Hatzivassiloglou, J.L.Klavans, A. Nenkova, C. Sable, B. Schiffman and S. Sigelman.
2. Tracking and summarizing news on a daily basis with Columbia's News Blaster. In Proceedings of Human Language Technology Conference (HLT 2002), (San Diego, CA, Mar. 2002).
3. D. R. Radev, H. Jing, M. Sty, D. Tam. Centroid-based summarization of multiple documents. *Journal of Information Process and Management.* 40(6): 919-938(2004).
4. D. R Radev., H. Jing, and M. Budzikowska. Centroid-based summarization of multiple documents: Sentence extraction, utility-based evaluation, and user studies. In ANLP/NAACL Workshop on Summarization, Seattle, April, (2000).
5. AminiM.R., GallinariP.: The Use of unlabeled data to improve supervised learning for text summarization. Proceedings of the 25th ACMSIGIR,105–112,(2002).
6. Aslam,J.A., Montague,M.: Models formet a search. In Proceedings of the 24th annual international ACMSIGIR conference on Research and development in information retrieval, (2001).
7. Caillet M., Pessiot J.F., AminiM.R., Gallinari P.: Unsupervised Learning with Term Clustering for the matic Segmentation of Texts Proceedings of RIAO,(2004).
8. Collins.M: Ranking algorithms for named-entity extraction: Boosting and the voted perceptron. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics(ACL-2002).
9. Dong Zhen-dong.: How Net[OL]: http://www.keenage.com
10. Barzilay, Elhadad, 1997. Using lexical chains for text summarization. Proceedings of the ACL'97/EACL'97.Workshop on Intelligent Scalable Text Summarization.

How to cite this article:

Ranjeet Ramesh Pawar and Mithun Vishnu Mhatre.2018, Clustering for text Summarization. *Int J Recent Sci Res.* 9(3), pp. 25182-25184. DOI: http://dx.doi.org/10.24327/ijrsr.2018.0903.1807

****