IJCSMS International Journal of Computer Science & Management Studies, Vol. 12, Issue 03, Sept 2012
ISSN (Online): 2231 –5268
www.ijcsms.com

A Cluster Based Keyword Filtration Approach for Web Document Summarization

Kirti Bhatia[1], Dr Rajendar Chhillar[2]

[1]**M. Tech. Student, DCSA, MDU, Rohtak, Haryana, India**
bhatia.kirti.it@gmail.com

[2]**Professor, DCSA, MDU, Rohtak, Haryana, India**
chhillar02@gmail.com

Abstract

Summarization, an extremely important technique in Data Mining is an automatic learning technique aimed to extract the most valuable information from a large size document or the articles. The goal is to create the summary of the document, but substantially different from each other. Text Document summarization refers to the summarization of text documents based upon their content. The proposed work is about to extract a summary from a large document, webpage or the email respective the occurrence of words, heading and their frequency.

Keyword: *Summarization, Mining, Document, Feature* Selection, Clustering.

1. Introduction

Text Summarization is a data reduction process. The use text summarization allows a user to guess a sense of the content of a full-text, or to know its information content, without reading all sentences within the full text. The reduction in the amount of data has the advantage of increasing scale as follows:

1) Allow users to find relevant full-text sources more quickly.
2) Assimilating only essential information from many texts with reduced effort.
3) The Text Summarization is process used to extract the main part from the whole document. It helps a user to derive the Conclusion or the main abstract from the Document.
4) It is never easy for a user to read complete documents having thousands of Sentences because of this he require such an approach in which we retrieve the abstracted code from the document.

Web Summarization

It the process of automatic extraction of relevant information given a list of topics from different web sites- is of great utility for a variety of application and in particular for automatic indexing and categorization in order to facilitate the production and accessibility of new multimedia contents.

To give an example, Let's Consider a news reporter that's needs to have some info delivered to his PDA or cell phone for writing an article a news and does not want to waste time analyzing a great amount of information sources: a summarization system could helpfully produce a short summary of the gathered retrieved information, and in addition, such a summary could be easily managed and accessed using light mobile devices.

Summarization is well known research field in the artificial intelligence community. Most of the proposed summarization technique are query independent and follow one of the following two approaches: they simply extract relevant part of the text, considering the document as a sequence of unstructured set of text blocks, or they employ Natural Language Processing techniques. The former approach ignores the structural information of documents, while latter is more computationally expensive for large data- sets and sensitive to the different writing style.

Text Mining

Text Mining is an essential theme in data mining and is all about looking for patterns in a text. Nowadays most of the information related to government, business, industry and other institutions are available electronically in the form of text databases. Text databases such as news articles, digital libraries, research papers, e-mail messages, and web pages comprise substantial portion of the available information. Text databases are usually semi structured in nature. Due to abundance of text information the field of Information Retrieval and Text Mining are highly inter-related. Their common

applications are on-line library catalogue systems, on-line document management systems, recommender systems, Web Search Engines etc. Document clustering is an integral text mining task performed on the extracted keywords, tags or semantic information.

2. Literature Survey

Since text documents are high-dimensional structures, pre-processing and dimensionality reduction is another critical issue to be addressed in Summarization of text documents. This issue has been dealt with various techniques in literature. The popular dimension reduction methods include Independent Component Analysis (ICA), Latent Semantic Indexing (LSI), and a feature selection technique based on Document Frequency (DF)[15]. When tested with traditional K-means method, more efficient and accurate results are obtained with ICA and LSI as compared to DF. Some good representation techniques (Traditional word representation, N-gram representation) are also combined with dimension reduction methods to achieve better results. An extended version of Probabilistic Latent Semantic Analysis (PLSI) has been implemented jointly with Multi-nomial Mixture Modals (MM) [13] on standard datasets like Reuters-21578, WebKB, and 20Newsgroup which resulted in significant improvement in the Summarization solutions in a reduced concept space. Concept space document Summarization using LSI (Latent Semantic Indexing) has also been used in [13] to produce better results in combination with fuzzy c-means algorithm. Some other pre-processing techniques for feature selection include Term Variance (TV), Term Strength (TS), Term Contribution (TC)[14] Variants of K-Means have been largely implemented for document Summarization to improve efficiency and accuracy. Some of them include Euclidean K-Means, Spherical K-means [11] and Bisection K-means [17]. The quality of partitioning Summarization algorithms is highly dependent on the initial centroids. A new partitioning algorithm has been recently proposed [14] and has been tested efficiently on two English and two Chinese text datasets. This new technique generates initial centroids during the dynamic Summarization process by finding document which has closer neighbor but shares less neighbors with the partitioned clusters. The standard K-Means algorithm is quite sensitive to the selection of initial centroids and can generate a local optimal solution. Optimization techniques can be hybridized with traditional Summarization approaches to generate efficient global solutions. Hybrid techniques have been widely used in document Summarization literature by judiciously combining two or more techniques so as to exploit the strong points of all the combined algorithms [2]. Meta-heuristics, optimization techniques and model based Summarization form an important component of hybrid Summarization techniques used in literature. Harmony K-means Algorithm (HKA) is a hybridization of K-means and Harmony Search (HS) Optimization method that has been used for document Summarization. Harmony Search algorithm is utilized for global optimization. K-means algorithm has been used for better tuning of the algorithm to improve the speed of convergence of HKA. Incorporation of synthetic prototypes into the Spherical K-means (spk-means) procedure for document Summarization has been explored in [3] for discovering and describing topics in a collection of text documents. The synthetic prototype is a novel type of cluster representative which is computed in two steps: 1) a reference prototype is constructed for the cluster and then 2) feature selection is applied on it. This synthetic prototype favors the representation of the objects of the dominant class in a cluster (the class to which the majority of the cluster objects belong). The generic spk-means iterative procedure has been modified by incorporating synthetic prototypes. This leads to a novel, effective and quite simple Summarization method called k-synthetic prototypes (k-sp). This technique is a novel parameter-less method for discovering the topics from standard collections (Reuters-21578, TDT2 English corpus and AFP Spanish collection) and at the same time it attaches suitable descriptions to the discovered topics. A Simple Agglomerative Hierarchical K-Means Summarization (SAHKC) algorithm and a modified version of VSM known as Multiple-Feature VSM (MFVSM) are used to used improve the efficiency of Web document Summarization in terms of run time and accuracy [17]. This algorithm uses a new initialization method based on finding a set of medians extracted from a dimension with maximum variances. This proposed algorithm turned out to be better than traditional techniques in terms of various evaluation metrics like accuracy, response time. This technique could also be adopted by various search engines.

3. Proposed Work

3.1 Preprocessing

The documents to be clustered are in an unstructured format therefore some pre-processing steps need to be performed before the actual Summarization begins. The pre-processing includes Tokenization, Stemming of document words, and Stop word removal. Tokenization means tagging of words where each token refers to a word in the document.

Stemming involves conversion of various forms of a word to the base word. E.g. 'computing' and 'computed' will be stemmed to the base word 'compute'. Similarly 'sarcastically' is stemmed to the word 'sarcasm'. The Porter's Algorithm is the most popular stemming technique for English Language documents. Snowball is a popular tool using this stemming algorithm.

Stop word removal: Stop words are the words present in documents which do not contribute in differentiating a collection of documents hence, are removed from the documents. These are basically articles, prepositions, and pronouns. Standard stop lists are available but they can be modified depending upon the kind of dataset to be clustered.

3.2 Feature Selection

Documents need to be represented in a suitable form for Summarization. The most common representation includes the Vector Space Model (VSM) **[36]** which treats documents as a bag-of-words and uses words as a measure to find out similarity between documents. In this model, each document D_i is located as a point in a m-dimensional vector space, $Di = (w_{i1}, w_{i2}, \ldots, w_{im})$, $i = 1, \ldots, n$, where the dimension is the same as the number of terms in the document collection. Each component of such a vector reflects a term within the given document. The value of each component depends on the degree of relationship between its associated term and the respective document.

3.3 Summarization Algorithm

The Summarization algorithm generates clusters based on similarity measure and data representation model. Numerous Summarization algorithms have been implemented in document.

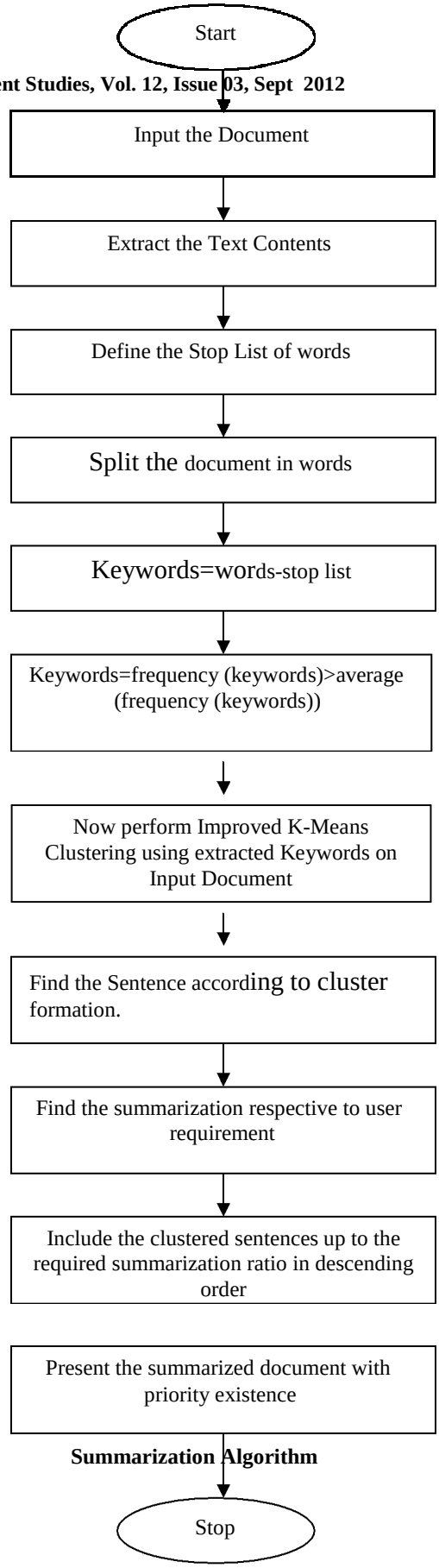

Fig. 1 Summarization Algorithm

4. Result Validation:

This is post Summarization technique in which the quality of the final result is validated. There are numerous evaluation measures to validate the cluster quality. The validity criteria can be categorized in two ways:

4.1 External Criteria:

Measures performance by matching Summarization structure to some a priori knowledge e.g. degree of correspondence between the cluster numbers obtained from a Summarization algorithm and category labels assigned a priori.

4.2 Internal Criteria:

Assess the fit between the structure and the data by making use of the data alone. It basically deals with the typical objective function of Summarization analysis of maximizing the intra-cluster similarity and minimizing the inter-cluster similarity [2]. e.g. the degree to which a partition obtained from a Summarization algorithm is justified by the given proximity matrix.

5. Analysis and Results

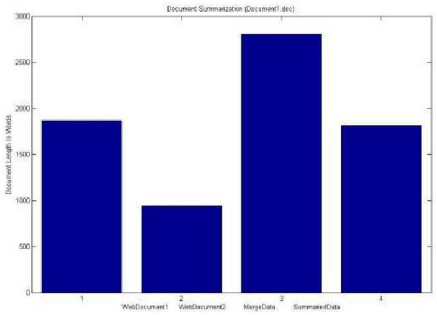

Fig. 2 Summary Document 1

Here figure 4.5 shows the results driven from the summarization process for the document1. The summarization process is based on two web documents extracted from the web based on similar kind of articles or the news. The size of webdocument1 is 1864 words and for webdocument2 the size is 943 words. We merge these documents and the size becomes 2806 words. After implementing the summarization process the result document size is 1811 words. Here we get the summarization ratio about 64.54%.

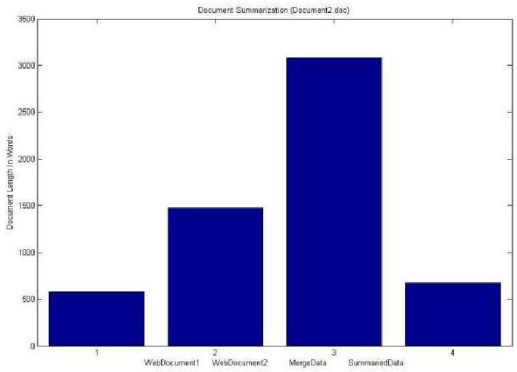

Fig. 3 Summary Document 2

Here figure 4.6 shows the results driven from the summarization process for the document1. The summarization process is based on two web documents extracted from the web based on similar kind of articles or the news. The size of webdocument1 is 574 words and for webdocument2 the size is 1474 words. We merge these documents and the size becomes 2058 words. After implementing the summarization process the result document size is 670 words. Here we get the summarization ratio about 21.73%.

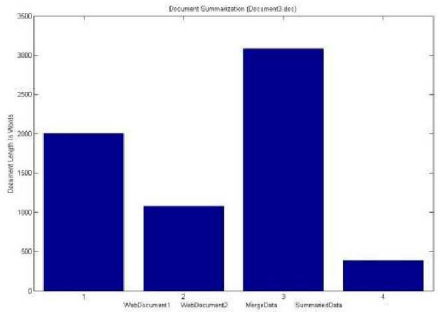

Fig. 4 Summary Document 3

Here figure 4.7 shows the results driven from the summarization process for the document1. The summarization process is based on two web documents extracted from the web based on similar kind of articles or the news. The size of webdocument1 is 2002 words and for webdocument2 the size is 1082 words. We merge these documents and the size becomes 3083 words. After implementing the summarization process the result document size

is 383 words. Here we get the summarization ratio about 12.42 %.

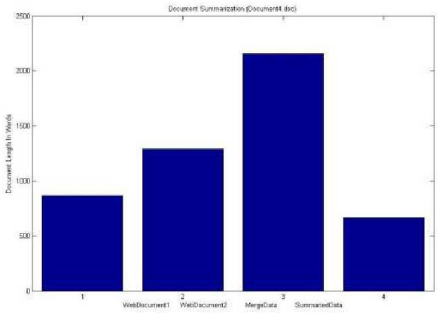

Fig. 5 Summary Document 4

Here figure 4.8 shows the results driven from the summarization process for the document1. The summarization process is based on two web documents extracted from the web based on similar kind of articles or the news. The size of webdocument1 is 866 words and for webdocument2 the size is 1289 words. We merge these documents and the size becomes 2154 words. After implementing the summarization process the result document size is 668 words. Here we get the summarization ratio about 31.01 %.

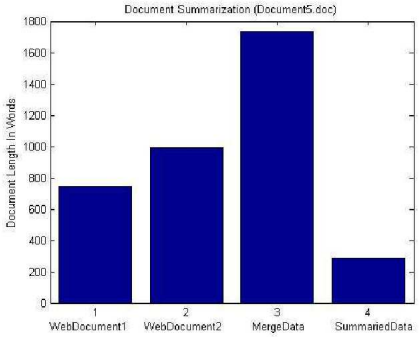

Fig. 6 Summary Document 5

Here figure 4.9 shows the results driven from the summarization process for the document1. The summarization process is based on two web documents extracted from the web based on similar kind of articles or the news. The size of webdocument1 is 743 words and for webdocument2 the size is 994 words. We merge these documents and the size becomes 1736 words. After implementing the summarization process the result document size is 285 words. Here we get the summarization ratio about 16.41 %.

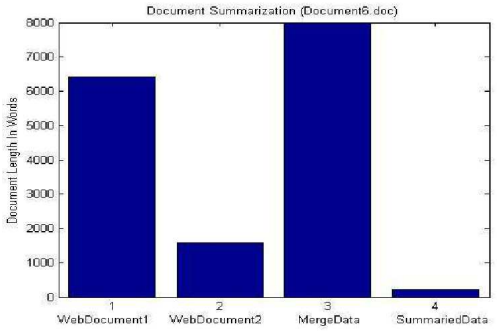

Fig.7 Summary Document 6

Here figure 4.10 shows the results driven from the summarization process for the document1. The summarization process is based on two web documents extracted from the web based on similar kind of articles or the news. The size of webdocument1 is 6410 words and for webdocument2 the size is 1578 words. We merge these documents and the size becomes 7987 words. After implementing the summarization process the result document size is 233 words. Here we get the summarization ratio about 4.67 %.

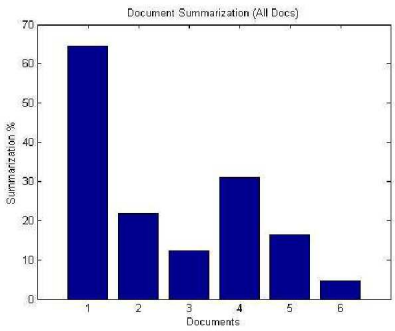

Fig. 8 Summarization (All Documents)

Here figure 4.11 shows the overall summarization driven from the system. The obtained summarization in all documents is

driven for the system. The results shows that based on the documents contents the summarization will vary. Here we get the the minimum summarization level of 65% and max up to 4%.

6. Conclusions

In this research work we have presented a statistical approach to summarize the documents. We have performed a comparative analysis of proposed approach with existing statistical approach. To perform the analysis we have performed the summarization on different size documents on different ratio 30%, 50% and 70%. The work is performed by creating a user friendly application in java and the analysis is performed in the form of graphs formed in matlab environment. All the results are showing that the proposed system is giving better summarization as compared to existing statistical approach.

Acknowledgments

I would like to express my deepest gratitude toward my guide Dr. Rajendar Chhillar, M.D University, Dept. of Computer Science & Applications Rohtak, Haryana, India for showing great interest in my thesis work, this work could not finished without his valuable comments and inspiring guidance.

References

[1].Jiawie Michelene Kamber Han, Data Mining: Concepts and techniques, Morgan Kauffman Publishers, Second Edition.

[2].Christopher Manning, Prabhakar Raghavan, Hinrich Schütze, An Introduction to Information Retrieval, Online Edition, Cambridge University Press, 2009.

[3].Anaya-Sánchez , Aurora Pons-Porrata, Rafael Berlanga-Llavori, “A document clustering algorithm for discovering and describing topics”, Henry, Pattern Recognition Letters 31, 2010.

[4].http://en.wikipedia.org/wiki/Soft_computing

[5].Selim, S. Z. And Ismail, M. A. 1984.” K means type algorithms: A generalized convergence theorem and characterization of local optimality”, IEEE Trans. Pattern Anal. Mach. Intel. 6, 81–87.

[6].Zhang, Ramakrishna, Livny, “BIRCH: An efficient Clustering Algorithm for very Large Databases”, SIGMOD 1996.

[7].Sudipto Guha, Rajeev Rastogi, Kyuseok Shim, “CURE: An efficient clustering algorithm or large databases”, SIGMOD 1996.

[8].George Karypis, Eui-Hong (Sam) and Han Vipin Kumar, “CHAMELEON: A Hierarchical Clustering Algorithm Using Dynamic Modeling”.

[9].Sudipto Guha, Rajeev Rastogi, Kyuseok Shim, “ROCK: A Robust Clustering Algorithm for Categorical Attributes”.

[10]. A.K.Jain, M.N.Murty, P.J.Flynn, ”Data Clustering: A Review”, ACM Computing Surveys, Vol. 31, No. 3, September 1999.

[11]. Hai-hui Wang, Wen-jie Zhao, “Data Clustering Based on Approach of Genetic Algorithm”, Chinese Control and Decision Conference (CCDC 2008)

[12]. Zhenya Zhang, Hongmei Cheng, Shuguang Zhang, Wanli Cheng, Qiansheng Fang, “Clustering Aggregation based on Genetic Algorithm for Documents Clustering”.IEEE 2008.

[13]. Naser El-Bathy, Ghassan Azar, Mohammed El-Bathy and Gordon Stein, “Intelligent Extended Clustering Genetic Algorithm”, IEEE.

[14]. R.J. Kuo, L.M. Lin, “Application of a hybrid of genetic algorithm and particle swarm optimization algorithm for order clustering”, Decision Support Systems 49 (2010).

[15]. Anirban Mukhopadhyay, Indrajit Saha, “Genetic Algorithm and Simulated Annealing based Approaches to Categorical Data Clustering”, 2008 IEEE Region 10 Colloquium and the Third ICIIS, Kharagpur, INDIA

[16]. Tahani Hussain, Sami J. Habib, “Optimization of Network Clustering and Hierarchy through Simulated Annealing”, IEEE, 2009

[17]. Michael Ng, “A Parallel Tabu Search Heuristic for Clustering Data Sets”, IEEE, Proceedings of the International Conference on Parallel Processing Workshops (ICPPW 2003).