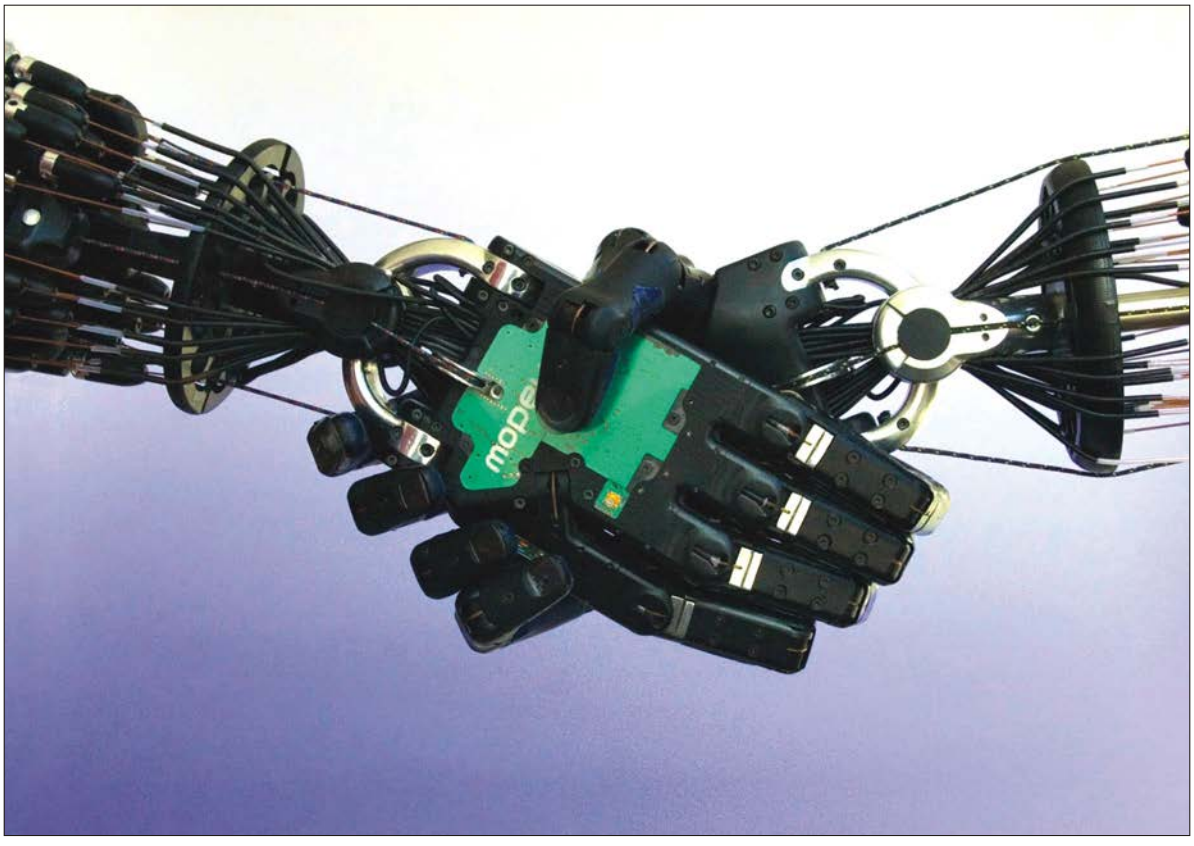

Journal of Artificial Intelligence

ISSN 1994-5450

science alert

ANSInet
an open access publisher
http://ansinet.com

Journal of Artificial Intelligence 8 (1): 1-9, 2015
ISSN 1994-5450 / DOI: 10.3923/jai.2015.1.9

Automatic Multi-Document Arabic Text Summarization Using Clustering and Keyphrase Extraction

Hamzah Noori Fejer and Nazlia Omar
Center for AI Technology, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia (UKM), Bangi, Selangor, Malaysia

Corresponding Author: Hamzah Noori Fejer, Center for AI Technology, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia (UKM), Bangi, Selangor, Malaysia

ABSTRACT

Automatic text summarization has become important due to the rapid growth of information texts since it is very difficult for human beings to manually summarize large documents of texts. A full understanding of the document is essential to form an ideal summary. However, achieving full understanding is either difficult or impossible for computers. Therefore, selecting important sentences from the original text and presenting these sentences as a summary present the most common techniques in automated text summarization. Arabic natural language processing lacks tools and resources which are essential to advance research in Arabic text summarization. In addition to the limited resources, there has been little attention and research done in this field. Arabic text summarization still suffer from low accuracy as they use simple summarization techniques. The aim of this research is to improve Arabic text summarization by using clustering and keyphrase extraction. This study proposes a combined clustering method to group Arabic documents into several clusters. Keyphrase extraction module is applied to extract important keyphrases from each cluster, which helps to identify the most important sentences and find similar sentences based on several similarity algorithms. These algorithms are applied to extract one sentence from a group of similar sentences while ignoring the other similar sentences. The Recall-Oriented Understudy for Gisting Evaluation (ROGUE) metrics were used for the evaluation. For the summarization dataset the corpus DUC2002 was used. This model achieved an accuracy of 43.4%. The experiments have proved that the proposed model has given better performance in comparison to other work.

Key words: Automatic text summarization, clustering, keyphrase extraction, similarity, ROUGE metrics, multi-document

INTRODUCTION

With summarization, one would simply summarize one article or document by creating a summary of the concepts that are most important. The summary would be in the correct order and coherent as well. This process takes a lot of effort and is time consuming, even for humans. A summary of the same article that is created by two different people would be different, based on what each person thinks is most important and includes this in their summary; they each have a unique perspective on the same article. This encouraged the requirement of an automated summarizer system that may carry out this process with less effort and time. This has created a large amount of research on this topic of automatic summarization, which was over 50 years ago (Luhn, 1958).

Generally, summarization of text is the task of creating a summary from an article or a collection of articles about the same subject by extracting the most important parts of the text and adding them in chronological order. However, automatic text summarization is the main task of generating a shorter version of a particular document or set of documents by machines. For instance, a document of text may be compressed version via the usage of applications.

A summary in general must provide the core concepts within the text that is input. Only main sentences must be available in the summary and the task of explaining the sentences depends on the method that is implemented for summarization. With the task of automatic summarization, there are generally two primary methods that are widely carried out. These approaches are extractive and abstractive. In the first approach, which is extractive based summarization, there is a limit in the extraction and only the important sentences are extracted and generated into a summary in the chronological order, in order for a coherent summary to be generated. The parts of text that are extracted vary, based on the summarizer used.

Numerous summarizers take sentences, as opposed to paragraphs, or other larger text units. Automatic text summarization is carried out often using extractive based approaches. However, abstractive based summarization deals with more tools that rely on language and Natural Language Generation (NLG) concepts. This type of summarization may incorporate terms that do not exist in the article. Abstractive summarization aims to copy methods used by humans, such as adding a concept that is available in the original article in a better and more comprehensive way. This type of summarization, however, is more complex to implement, although more effective. This work uses an extractive concept implemented on the proposed model.

Summarization has been investigated for over the past 50 years, with most studies focusing on the English language (Hirao *et al*., 2007). Studies on Arabic-language documents have been conducted at a much later period and remains far behind those studies on other languages. Research on automatic summarization of Arabic-language documents has started approximately 10 years ago (Conroy *et al*., 2006; Douzidia and Lapalme, 2004). Further studies on Arabic-language resources are required. Several studies have proposed relatively advanced methods for Arabic language summarization (El-Haj *et al*., 2011a, b; Giannakopoulos *et al*., 2008).

The task of clustering within summarization of documents may be critical for choosing and retrieving related sentences and removing any redundancies. Even though approaches that include clustering have become widely used in text summarization, there has been a lack of research for clustering methods in summarization of documents in the Arabic language. In this work, focus is made on hierarchical and partitional clustering methods, namely, single linkage, complete linkage and k-means clustering because it is suitable with cluster summarization (Liu and Croft, 2004). This study propose a cluster-based summarization approach to group similar texts as well as employs a keyphrase extraction technique by an unsupervised machine learning algorithm to recognize sentences that include keyphrases and to summarize original text documents.

MATERIALS AND METHODS

This section describes the proposed framework for multi-document Arabic text summarization. A pre-processing text, clustering algorithms has been applied to group documents into many clusters. Keyphrase extraction is used to extract the important Keyphrases from each cluster, which helps to identify the most important sentences. Eliminate redundancy technique is applied to extract one sentence from a group of similar sentences and ignoring the other. Figure 1 shows the several phases in the overall architecture of our system, which are outlined as follows.

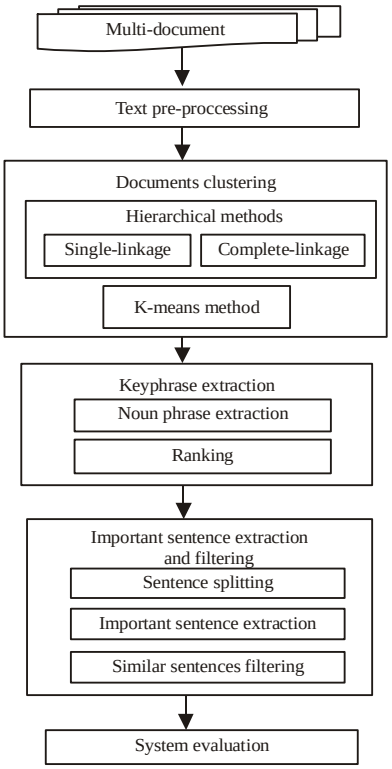

Fig. 1: Overall architecture of the multi-document arabic text summarization

Input documents: Input multi-documents will be used to extract texts from articles on different subjects, such as education, politics and sports. These documents contain texts of different sizes. The C# language is used to code the different stages of the proposed summarization model.

Text pre-processing: This phase involve four steps, namely, tokenizing, eliminating stop words, stemming and text representation and term weighting.

Clustering phase: Clustering is classifying data into many clusters. The elements within a cluster are similar to one another, but are dissimilar to elements in another cluster. Clustering decreases the amount of information by categorizing and grouping similar data. Document clustering involves unsupervised document organization, automatic topic extraction and fast information retrieval. Clustering is performed to group these results automatically into a list of meaningful categories.

Hierarchical clustering is the most commonly used clustering method. Hierarchical clustering method constructs the clusters by recursively partitioning the instances in either a top-down or bottom-up fashion. Single linkage clustering and Complete linkage clustering are used in this study.

Single linkage clustering (Nearest Neighbor Technique) is one of the simplest agglomerative clustering techniques. The distance between two groups is defined as that of the closest pair of individuals. One individual is considered from each group. Single linkage clustering is considered the best hierarchical method in terms of theoretical soundness. Several efficient methods are devised to apply the single linkage method on large data sets.

Complete linkage clustering (Furthest Neighbor Clustering) is the reverse of single linkage clustering in the sense that the distance between groups is defined as the distance between the most distant pair of individuals that are selected from each group. The definition of cluster membership in this method is much stricter than that in single linkage clustering and the large, straggly clusters in the latter method are replaced by many small, tightly bound clusters of equal diameter in this method. In graph theoretical terms, complete linkage clustering corresponds to the identification of maximally complete sub-graphs at a threshold similarity. Willett (1982) defined complete linkage clustering as the most effective method for document clustering despite its requirement for the greatest computational resources.

The output of single-or complete-link clustering is a number of clusters (K-value). The k-means algorithm is easy to apply, because a number of clusters are determined from hierarchical clustering. Then, k-means clustering is applied.

K-means clustering: K-means is the most famous clustering method, because of its easy implementation and rapid convergence. Nevertheless, this method is significantly influenced by the selection of the initial solution. In this method, t iterations are achieved on a sample size of m items that are each characterized by n attributes. t: iteration k: number of clusters n: attributes m: item.

A lot of works have been proposed in order to improve the efficiency of the original k-means (Kanungo *et al*., 2002; Jain and Dubes, 1988) defines that cluster analysis is the process of classifying elements into sets of similar objects based on a similarity/distance measure. Nonetheless, k-means are applied in a wide number of different fields including text mining, information retrieval and machine learning of neural network, pattern recognition, classification analysis, artificial intelligence, image processing and machine vision.

As such, the k-means approach is selected instead of its fast variants for comparative purposes throughout the experiments conducted in this research. Moreover, as k-means is adopted in the algorithm proposed in this study, fast variants of k-means can be used to improve its speed K-means is a method that has been widely used for partitional clustering with a linear time complexity (Cimiano *et al*., 2005). As stated by Hartigan (1975), the k-means algorithm argues that the mean of the documents assigned to that cluster represents. each of the k-clusters and as a result the k-means technique is largely regarded as the centroid of that cluster. The benefit of k-means clustering are simple and flexible easy to understand and can be easily to implemented.

The disadvantage of k-means clustering method is clusters number (k-value) must be determined at the beginning, but after applying k-means algorithm, number of clusters may not be the same as the value specified in advance (Vora and Oza, 2013). For this reason we proposed consolidation method combining hierarchical clustering to get the exact number of clusters (topics) and k-means clustering. Then, keyphrases are extracted from each cluster (each topic will be summarized, separately).

Keyphrase extraction: Keyphrases are important words/phrases that reflect the subject of the text. We extract keyphrases from each cluster during this phase. Keyphrase extraction comprises the following steps:

Noun phrase extraction: In this phase, all possible noun phrases of one, two, three, or four consecutive words that appear in a given document are generated as n-gram terms. Keyphrase often corresponds to frequently occurring noun phrase in a text. This extraction algorithm considers only noun phrases as candidate keyphrases.

Ranking: For each noun phrase, some set of features are extracted. The following is used for ranking the candidate keyphrase:

- **Sentence count:** This feature refers to the total number of sentences in which members of the keyphrase can be found. We normalize this feature using the total number of sentences in the document
- **Term frequency:** Frequency refers to the number of occurrences of the candidate phrase. Frequency is calculated using an aggressive stemming algorithm (iterated Light stemmer). Frequency is normalized by the number of noun occurrences in the document
- **First occurrence in text:** This feature refers to the first occurrence of a word in the text. This feature is normalized by the total number of sentences in the document
- **Last occurrence in text:** This feature refers to the last occurrence of a word in the text. This feature is normalized by the total number of sentences in the document
- **C-value:** This feature is normalized for potential keyphrases by using the length of the phrase and the frequency, with which it occurs as a sub-string of other phrases

Important sentences extraction and filtering: This phase extracts the most important sentences from each cluster. This phase involves the following steps:

Sentence splitting: In this step, each document is splitting into several sentences using delimiters (e.g., full stop, question mark and exclamation mark).

Important sentences extraction: Obviously, every sentence acquired from a text is not important for text summarization.

We assume that sentences containing keyphrases are only important. Therefore, sentences set can be filtered to provide better results by removing sentences without contain any keyphrases. The score of sentence yields:

$$\text{Score (sentence)} = \frac{\text{Total \#of words in every keyphrases in the sentence}}{\text{Sentencs length in words}} \quad (1)$$

Furthermore, it is assumed that the sentence is considered as important, if the is greater than a threshold otherwise it is excluded from the sentences set.

Redundancy elimination: In this phase, all sentences are taken in each cluster. Similarity measures are used to identify similar sentences. Then, similar sentences are coherently grouped. The well-known measures of distance between patterns are majorly focused in this study. Consequently, the Cosine Similarity (CS) and Jaccard Coefficient (JC) are computed as similarity or distance measures (Cimiano *et al*., 2005; Huang, 2008). Finally, one sentence is selected by ignoring others. The selected sentence becomes the most important sentence based on the calculated score. The CS and JC are defined as follows:

Cosine similarity: Cosine similarity is one of the most famous similarity measures. The cosine similarity of two sentences, namely, $\vec{t_a}$ and $\vec{t_b}$ is computed as follows:

$$SIM_C\left(\vec{t_a}, \vec{t_b}\right) = \frac{\vec{t_a}.\vec{t_b}}{|\vec{t_a}|*|\vec{t_b}|} \quad (2)$$

where, $\vec{t_a}$ and $\vec{t_b}$ are the m-dimensional vectors over the term set T = {t1, tm}. Each dimension represents the positive weight of a term in the document.

Jaccard coefficient: The Jaccard coefficient, also known as the Tanimoto coefficient, is used to measure the similarity in the intersection that is divided by the union of the objects. For text sentences, the Jaccard coefficient compares the sum weight of shared terms and the sum weight of non-shared terms that are presented in either of the two sentences. The Jaccard coefficient is formally defined as follows:

$$SIM_j\left(\vec{t_a}, \vec{t_b}\right) = \frac{\vec{t_a}.\vec{t_b}}{|\vec{t_a}|^2 + |\vec{t_b}|^2 - \vec{t_a}.\vec{t_b}} \quad (3)$$

The Jaccard coefficient equals to 1 when $\vec{t_a} = \vec{t_b}$ and equals to 0 when $\vec{t_a}$ and $\vec{t_b}$ are disjointed. The corresponding distance measure is computed as $D_j = 1\text{-}SIM_j$.

Evaluation: Evaluating the accuracy of a generated summary is difficult, because an optimal summary remains undefined (Fiszman *et al*., 2009). Achieving system evaluation may help address this problem.

In our work, ROUGE-N is used to calculate the scores of a candidate summary based on the n-gram overlap between candidate and reference summaries (Lin, 2004). ROUGE-N consists of many metrics, such as ROUGE-1, ROUGE-2, ROUGE-3 and so on, with each corresponding to the size of the n-grams used in the evaluation. In this study N is 1-4. ROUGE-N scores are computed as follows:

Thus, the recall (R) measure calculates the proportion of n-grams from reference summaries that occur in a candidate summary, the precision (P) calculates the proportion of n-grams from a candidate summary that occur in reference summaries and F-score combines recall and precision into one metric. The following formulas compute, recall, precision and F-score, as:

$$\text{Recall} - N = \frac{|G_{ref} \cap G_{can}|}{|G_{ref}|} \quad (4)$$

$$\text{Precision} - N = \frac{|G_{ref} \cap G_{can}|}{|G_{can}|} \quad (5)$$

$$ROUGE - NF - score = \frac{2 * ROUGE - Nrecall * ROUGE - Nprecision}{ROUGE - Nrecall + ROUGE - Nprecision} \quad (6)$$

where, G_{ref} includes the grams of reference summary and G_{can} includes the grams of candidate summary.

RESULTS AND DISCUSSION

Data description: For the document collection, machine translation was used to automatically create an Arabic multi-document summary corpus. The machine translation was used to translate an English data set into Arabic. The output of the translation process is an Arabic data set of related news articles. The data set used in the machine translation process was the DUC-2002 data set provided freely by the National Institute of Standards and Technology (NIST) through DUC2002. DUC-2002 is an English data set that contains 567 articles, 17,340 sentence and 199,423 word in addition to 1,111 gold-standard (model) summaries. National Institute of Standards and Technology produced 60 reference sets.

Experimental results: Five clustering methods were examined: k-means with cosine algorithm (KC), k-means with Jaccard coefficient algorithm (KJ), Single Linkage (SL), Complete Linkage (CL) and combined Clustering Method (CM). A keyphrase extraction module was then applied to extract important keyphrases from each cluster, identify the most important sentences and find similar sentences based on several similarity algorithms to eliminate redundancy. For test this model, three different sets of experiments were carried out in this study. The first experiment aimed to summarize three documents that are classified under the educational category, the second experiment was conducted to summarize five documents that belong to the political category and the last experiment was carried out to summarize eight documents that are classified under the two categories (educational and political). Table 1 shows the results of these experiments, respectively.

All experiments were designed with the number of clusters (k value). This number was specified by the user as two and is randomly initiated in the k-means methods. The number of clusters in the hierarchical methods is based on the distance/similarity between the documents was between 0.03-0.05. The threshold value chosen for the experiments was 0.05. The ROUGE-1 (n-gram = 1) were considered for the evaluation of this study. According to these results, the clustering by using a Combined Method (CM) was better and more effective than the k-means methods and the hierarchical methods. Based on the results, the best overall recall, precision and ROUGE-1, which were achieved in the second experiment. Figure 2 shows the compression of this model's summaries with the summaries of El-Haj (2012), which is the most recent related work. The recall, precision and ROUGE-1 scores of the El-Haj system were 39.52, 38.49 and 38.99%, respectively, whereas those of this model were 45.23, 41.80 and 43.45%, respectively, as obtained from the second experiment. The experiments proved that this model achieved better results than the El-Haj system.

Table 1: Overall ROUGE-1 evaluation of the first, second and third experiment

Measure methods	CM	KC	KJ	SL	CL
	(%)				
First experiment					
Recall	45.5	40.5	40.5	45.2	46.6
Precision	37.5	38.8	38.4	34.6	33.5
ROUGE-1	40.7	39.6	39.4	39.6	39.0
Second experiment					
Recall	45.2	32.3	42.3	37.0	43.4
Precision	41.8	56.6	39.6	48.6	40.0
ROUGE-1	43.5	41.0	40.9	42.0	41.7
Third experiment					
Recall	41.7	36.5	35.3	37.0	46.6
Precision	44.2	46.8	47.2	42.9	45.0
ROUGE-1	42.9	41.0	40.4	39.8	39.2

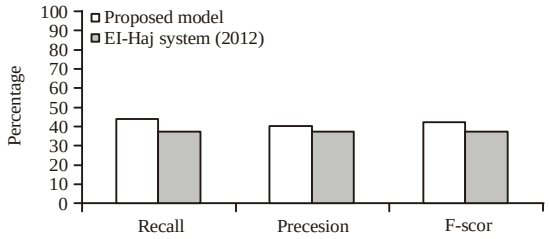

Fig. 2: ROUGE-1: Comparison the proposed model with El-Haj (2012)

CONCLUSION

In this study, a new automatic Arabic multi-document text summarization model is presented and discusses the structure of the proposed frameworks for multi-document Arabic text summarization. The most important steps in this research are document clustering and keyphrase extraction. We also describe important sentence extraction and filtering. Similarity algorithms, namely, cosine similarity and the Jaccard coefficient, are used to choose one sentence from each set of similar sentences and ignoring the rest (eliminate redundancy). The sentences are used to represent the text summary. The major contribution which can be advocated by the current study is a new model for automatic Arabic multi-document summarization, which is represented by using combined clustering method and keyphrase extraction. The combined clustering is a new approach of clustering which combines between hierarchical clustering methods and k-means clustering. The results of experiments proved the proposed model gives better performance in comparison to other works.

ACKNOWLEDGMENT

Our sincere thanks to Universiti Kebangsaan Malaysia (UKM), which have supported this study. Also, we would like to acknowledge and thank the General Directorate of Education in Qadisiyah, Ministry of Education, Republic of Iraq, for contributing in this study.

REFERENCES

Cimiano, P., A. Hotho and S. Staab, 2005. Learning concept hierarchies from text corpora using formal concept analysis. J. Artif. Intell. Res., 24: 305-339.

Conroy, J.M., J.D. Schlesinger, D.P. O'Leary and J. Goldstein, 2006. Back to basics: CLASSY 2006. Proceedings of the 6th Document Understanding Conferences, November 2006, New York.

Douzidia, F.S. and G. Lapalme, 2004. Lakhas, an arabic summarising system. Proceedings of the 4th Document Understanding Conferences, May 2004, Rochester, pp: 128-135.

El-Haj, M., U. Kruschwitz and C. Fox, 2011a. Multi-document arabic text summarisation. Proceedings of the 3rd Computer Science and Electronic Engineering Conference, July 13-14, 2011, Colchester, UK., pp: 40-44.

El-Haj, M., U. Kruschwitz and C. Fox, 2011b. University of essex at the TAC 2011 multilingual summarisation pilot. Proceedings of the Text Analysis Conference, November 14-15, 2011, Pilot, Maryland, USA.

El-Haj, M., 2012. Multi-document arabic text summarisation. Ph.D. Thesis, University of Essex, UK.

Fiszman, M., D. Demner-Fushman, H. Kilicoglu and T.C. Rindflesch, 2009. Automatic summarization of MEDLINE citations for evidence-based medical treatment: A topic-oriented evaluation. J. Biomed. Inform., 42: 801-813.

Giannakopoulos, G., V. Karkaletsis, G. Vouros and P. Stamatopoulos, 2008. Summarization system evaluation revisited: N-gram graphs. ACM Trans. Speech Language Process., Vol. 5. 10.1145/1410358.1410359.

Hartigan, J.A., 1975. Clustering Algorithms. Books on Demand, New York, USA., ISBN-13: 9780608300498, Pages: 365.

Hirao, T., M. Okumura, N. Yasuda and H. Isozaki, 2007. Supervised automatic evaluation for summarization with voted regression model. Inform. Process. Manage., 43: 1521-1535.

Huang, A., 2008. Similarity measures for text document clustering. Proceedings of the New Zealand Computer Science Research Student Conference, April 14-17, 2008, Christchurch, New Zealand, pp: 49-56.

Jain, A.K. and R.C. Dubes, 1988. Algorithms for Clustering Data. Prentice Hall Inc., Englewood Cliffs, USA., ISBN: 0-13-022278-X, Pages: 320.

Kanungo, T., D.M. Mount, N.S. Netanyahu, C.D. Piatko, R.S. Angela and Y. Wu, 2002. An efficient k-means clustering algorithm: Analysis and implementation. IEEE Trans. Pattern Anal. Mach. Intell., 24: 881-892.

Lin, C.Y., 2004. A package for automatic evaluation of summaries. Proceedings of the Workshop on Text Summarization Branches Out, July 25-26, 2004, Barcelona, Spain, pp: 25-26.

Liu, X. and W.B. Croft, 2004. Cluster-based retrieval using language models. Proceedings of the 27th ACM SIGIR Conference on Research and Development in Information Retrieval, July 25-29, 2004, Sheffield, UK., pp: 186-193.

Luhn, H.P., 1958. The automatic creation of literature abstracts. IBM J. Res. Dev., 2: 159-165.

Vora, P. and B. Oza, 2013. A survey on k-mean clustering and particle swarm optimization. Int. J. Sci. Mod. Eng., 1: 24-26.

Willett, P., 1982. A comparison of some hierarchical agglomerative clustering algorithms for structure-property correlation. Analytica Chim. Acta, 136: 29-37.