Asian Journal of Information Technology 15 (5): 933-939, 2016
ISSN: 1682-3915

Extractive Text Summarization System Using Fuzzy Clustering Algorithm for Mobile Devices

P. Vishnu Raja, K. Sangeetha and D. Deepa
Department of CSE, Kongu Engineering College, Perundurai, 638052 Erode, Tamil Nadu, India

Abstract: The digital world is full of tons and tons of electronic documents. This kind of large amount of information is cumbersome to read at a glance and also it takes more time to download the large content. In this paper, we propose sentence ranking measure to process the sentences and rank it with appropriate sentence rank score. Similar sentences are grouped into clusters. Based on the score, the sentences with highest score are considered as important sentences. The sentences with top scores from each cluster are retrieved for summary report. This summary report eases the user to quickly analyse the information. The experimental result proves the effectiveness of the system.

Key words: Fuzzy clustering, text summarization, clustering algorithm, clustering, information

INTRODUCTION

In recent decades, the growth of information is much higher due to the advancement of internet technologies. Inorder to search for a particular content, the common nature of the mobile users is to search for the summarized content. Unfortunately, some of the source information content may not have summary, in such case; the automatic summarization process helps the user to quickly have a glance at the summarized information.

In general, there are two different types of summarization. They are termed as extractive text summarization and abstractive text summarization. The process of Extractive text summarization is a simple task since, it extracts few sentences from the source and generates summary without any changes made with the words or sentence formation in the resulting summary. On the other hand, abstractive text summarization is considered to be more complicated process since it uses Natural Language Processing (NLP) techniques to generate a summary which would be more equivalent to the human generated summary. The resultant summary may have words that may not be present in the source information.

In this study, the sentence ranking measure processes each sentence and assigns a rank to each of those sentences. Accordingly, those sentences with similar content are grouped into clusters. Then their ranks are reversed such that the top scoring sentences are alone taken for summarized information content. The rest of this paper is organized as follows:

MATERIALS AND METHODS

There are different methods to perform automatic summarization.The semantic similarity measured in terms of word co-occurrence may be valid at the document level of clustering, but not suited for small sized text fragments such as sentences, since two sentences may be semantically related despite having few words in common. In order to solve this issue, sentence level similarity measures have been proposed. Zha (2002) proposed generic summarization and keyphrase extraction using mutual reinforcement principle and sentence clustering. This method explores the sentence link in the linear ordering of a document. The keyphrases and sentences are then ranked according to their salience scores and selected for inclusion in the top keyphrase list. The hierarchies of summaries are built for the documents at different levels of granularity. In general, clustering is well suited for automatic summarization. Okazaki *et al.* (2003) proposes Sentence extraction by spreading activation through sentence similarity. Most methods rely on statistical methods, disregarding relationships between extracted textual segments. The proposed method extracts a set of comprehensible sentences which enters on several key points to ensure sentence connectivity. It features a similarity network from documents with a lexical dictionary and spreading activation to rank sentences. Centroid Based Summarization (CBS) (Yuan and Sun, 2003) uses the centroids of the clusters to identify the sentences which are central to the topic of the entire cluster. Radev *et al.* (2004) proposed a centroid-based

Corresponding Author: P. Vishnu Raja, Department of CSE, Kongu Engineering College, Perundurai, 638052 Erode, Tamil Nadu, India

summarization of multiple documents (Radev *et al*., 2004). This method utilizes Term Frequency Inverse Document Frequency (TF-IDF) value to calculate the centroid value. The centroid is generated by using the first document in the cluster. Then the new documents are processed using their TF-IDF values to compare with the centroid value. MEAD extraction algorithm is used to perform sentence ranking. Erkan and Radev (2004) proposes LexRank as a Graph-based Lexical Centrality as Salience in Text Summarization. LexRank is proposed for computing sentence importance based on the concept of eigenvector centrality in a graph representation of sentences. In this model, a connectivity matrix based on intra-sentence cosine similarity is used as the adjacency matrix of the graph representation of sentences. Von Luxburg *et al*. (2007) proposes a model for improving text categorization using the importance of sentences (Ko *et al*., 2004). The importance of each sentence is measured by two methods. First, the sentences which are more similar to the title, have higher weights. In the next method, we first measure the importance of terms by TF, IDF and x^2 statistics values. Then we assign the higher importance to the sentence with more important terms. Finally, the importance of a sentence is calculated by combination of two methods. Corsini *et al*. (2005) proposed a new fuzzy relational clustering algorithm based on the fuzzy C-means algorithm. This method is known as Any Relation Clustering Algorithm (ARCA) which remains to be stable without requiring any particular restrictions on the square binary relations. The ARCA represents a cluster in terms of a representative of the mutual relationships of the objects which belongs to the cluster with a high membership value. The ARCA represents a cluster in terms of a representative of the mutual relationships of the objects which belong to the cluster with a high membership value.

Li *et al*. (2006) proposed a sentence similarity measure based on semantic nets and corpus statistics. The standard Euclidean measure is applied to determine the distance between data objects. The semantic similarity of two sentences is calculated using information from a structured lexical database and from corpus statistics. The algorithm depends on semantic information and word order information implied in the sentences with significant correlation to human intuition and so it can be adaptable to different domains.

Aliguliyev proposed a new sentence similarity measure and sentence based extractive technique for automatic text summarization. This method consists of two steps. Initially sentences are clustered and then on each cluster representative sentences are defined. The discrete differential evolution algorithm is proposed to optimize the objective functions. The inter sentence word to word similarities are derived either from distributional information such as corpus based measures or semantic information represented using external sources such as WordNet. These sentence similarity measures are not based on representing sentences in a common metric space, hence the conventional fuzzy clustering approaches based on prototypes are not applicable to sentence clustering.

Geweniger *et al*. (2010) proposed median fuzzy c-means (MFCM) for clustering dissimilarity data. MFCM is a combination of fuzzy c-means and median c-means. This method is a median variant of FCM as a prototype based non-hierarchical cluster algorithm. MFCM takes the dissimilarities between data points into account but retain the concept of prototype based clustering. MFCM is designed to handle relational data but it tends to stuck in local minima depending on initialization. Garcia *et al*. (2006) proposed a context aware summarization system to adapt text for mobile devices. This method is based on ontologies which generates summaries from text according to the profile of the user. Ontologies are used to identify the textual data which is relevant to the profile and the context. Summaries are generated for each combination of profile and context. The context is determined using spatial and temporal localization. This method reduces the time for knowledge acquisition and minimizes the problem of information overload.

Liu *et al*. (2013) proposed XML query oriented text summarization for mobile devices. Most of mobile and interactive multimedia devices have limited hardware such as processing, battery power, memory and display screen. Hence, it is essential to compress an XML document collection to a brief summary before it is delivered to the user. Query oriented XML text summarization aims to provide readable summarized content to relieve the burden of users instead of reading the whole content.

Skabar and Abdalgader (2013) proposed clustering sentence level text using a novel fuzzy relational clustering algorithm. It initially randomly assigns the fuzzy cluster membership value to each sentence. The Page Rank score analysis over the Expectation and Maximization framework using relational data iteratively to group the similar sentences. This method is even though processing over multiple iterations, it is quite faster in processing. There are enormous amount of information sources available in digital form all over the world, which in turn, gives way to the problem of information overload.

Due to limited hardware facilities, the mobile users are in need of summarized information to quickly search the relevant information that matches with the constraints of the mobile devices.

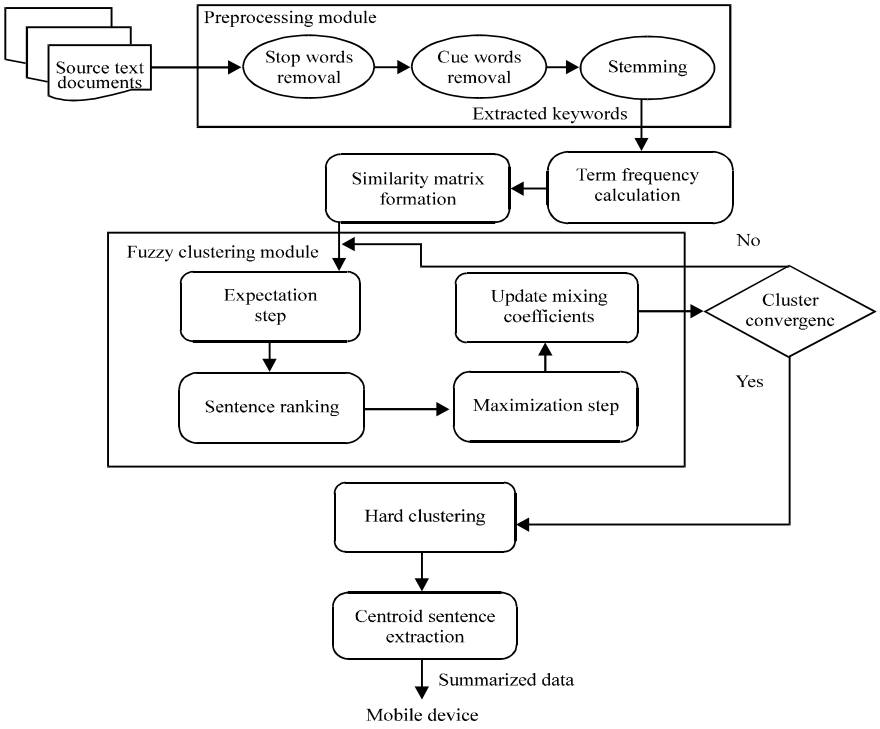

Fig. 1: Overview of the automatic summarization process

Proposed work: The proposed automatic summarizer is based on the lexrank measure by Erkan and Radev (2004) andfuzzy relational clustering algorithm proposed by Skabar and Abdalgader (2013). The modified page rank proposed by Skabar and Abdalgader (2013) is replaced by lexrank measure of Erkan and Radev (2004). Unlike the original PageRank method, the similarity graph for sentences is undirectedsince cosine similarity is a symmetric relation. However, this does not make any differencein the computation of the stationary distribution. The lexrank measure can be further enhanced by including the cluster membership parameter. Hence, in this study, we propose a combination of fuzzy relational clustering algorithm and modified lexrank measure.

Fuzzy clustering algorithm: The enhanced fuzzy clustering algorithm uses themodifedLexRank score of the sentence within each cluster as a measure of its centrality to that cluster. The proposed algorithm consists of five stages while passing through the summarization system (Fig. 1).

Algorithm: Modified LexRank based Fuzzy Clustering (MLFC) algorithm:

Input:
$S=\{s_{ij}|i=1,2,\ldots,N,j=1,2,\ldots,N\}$
//Pairwise Similarity matrix
C= Number of expected clusters
Output:
$p_i^m i=1,2,\ldots,N,$
$m=1,2,\ldots,C$//Cluster membership values
Summarized content
Method:

- Initialize random cluster membership value for all sentences
- Calculate normalized cluster membership value
- Calculation of Sentence Rank
- Grouping of similar sentences into clusters
- Generate summary with centroid sentence of each cluster

Preprocessing: The initial task in preprocessing module is to separate the keywords from the stopwords and other insignificant words to effectively calculate the term frequency. The preprocessing consists of following tasks:

- Stopwords removal
- Cue words removal
- Stemming

The words with no useful information such as preposition, pronoun, article, alphamerics, numerals and so on. These words are called as stop words. A connective expression is used for linking spans of

discourse and signals semantic relations in a text. For instance, some of the sample cue words are thus, concludes, however and so onare also eliminated since they have no specific meaning. Stemming is the process of converting the words to their root words For example consider a word stabilize which is derived from the root word stable. These words are not removed but converted into original root word.

In order to determine the importance of a term in a document, term frequency is calculated. The Term frequency is given by the Eq. 1:

$$TF(t) = \frac{n_j}{\sum n_k} \quad (1)$$

where, n_j represents the number of occurrences of $term_j$ in the document and n_k represents the total number of words in the document k.

Similarity matrix formation: The similarity matrix is used to reveal the relationship between each sentence in the original document. Most often cosine similarity is used to generate the similarity matrix. Erkan and Radev (2004) proposes a modified cosine similarity measure which is more effective than the standard cosine similarity measure. Consider two sentences x_i and y_i, their term frequency be tf, their modified cosine similarity measure is given by the following Eq. 2:

$$Cosine(x_1, y_1) = \frac{\sum_{wex,y} tf_{w,x} tf_{w,y} (isf)^2}{\sqrt{\sum_{xi \in x} (tf_{xi,x} isf)^2 \; x \sum_{yi \in y} (tf_{yi,y} isf_{yi})^2}} \quad (2)$$

Where:

$tf_{w,x}$ = The number of occurrences of word w in sentence x

n = Represents the total number of sentences

x_i and y_i = The i^{th}sentence of the cluster x and cluster y respectively and isf is the inverse sentence frequency since clustering is focused on sentences rather than documents

The cosine similarity between the sentence x and sentence y is non-negative and bounded between [0, 1]. The value of the exact match is 1 and also the matrix is symmetric.

Clustering module: Each sentence in the original document is randomly initialized to a cluster membership value as proposed by Skaber. Even though hard clustering algorithms such as spectral clustering algorithm can be adopted for cluster membership initialization, the initialization does not affect the final cluster membership value. The cluster membership value is normalized such that the sum of the objects contributes to unity over all clusters. The mixing coefficients are initialized with appropriate value such that the priors for all clusters are equal. The Expectation step calculates the sentence ranking score of each object in each cluster using the weighted affinity matrix obtained by scaling the similarity matrix. The sentence rank is computed using modified lexrank as shown in Eq. 3:

$$l(x) = \frac{d}{n} \times (1-d) \times \frac{mat_{ij}}{degree_{ij}}$$

Where:

d = The damping factor which is added for absorbing the errors due to convergence

l(x) = The modified lexrank score measure to depict the importance of sentence x

mat_{ij} = The weighted affinity matrix which is the product of similarity matrix and the cluster membership matrix and

$degree_{ij}$ = The corresponding highest cluster membership degree of the sentence x

The maximization step involves updating mixing coefficients based on the cluster membership values estimated in the Expectation step.

Hard clustering: The fuzzy clusters generated in the above module is converted into hard clustering for generating summary. Based on the highest cluster membership value of the sentence and the number of links to each of the fuzzy cluster, the sentences can be grouped into the corresponding hard cluster. This is analogous to the Gaussian mixture model case in which an object predominantly belongs to one Gaussian mixture component.

Summarization: In each of the converted hard cluster, the sentence which is having highest score is considered to be the centroid sentence. The centroid sentence is the most important sentence, interrelated to the original information source. Hence, from each cluster the centroid sentence is extracted to generate the summarized content. The summarized content is deployed using XML and java Application Programming Interface (API) to support the mobile environment.

RESULTS AND DISCUSSION

The C50 dataset is used for summarization task. It consists of 50 folders each with 50 text documents. Consider a sample of three sentences as follows:

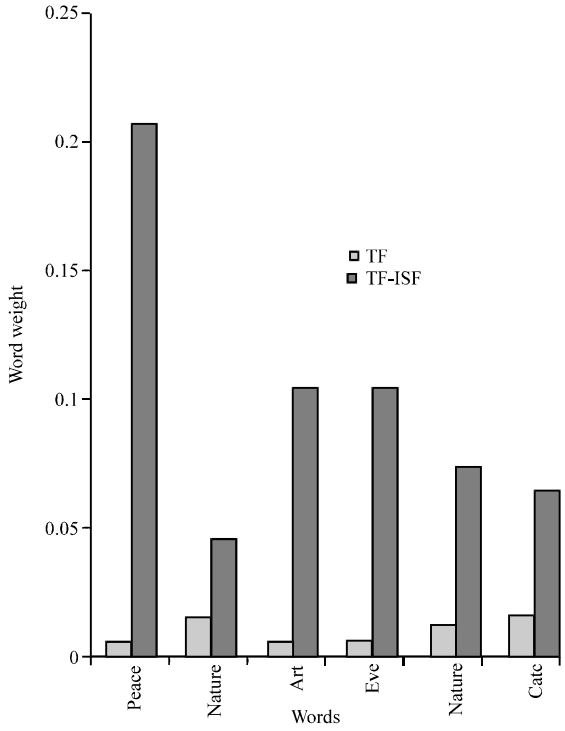

Fig. 2: Word weight measure analysis

- S1: Peace is everywhere
- S2: Nature is an art of God
- S3: Nature is a catchment of Peace

The initial step proceeds with pre-processing. The pre-processing process removes the insignificant words. The stopwords such as is, an, of and cuewords such as everywhere are removed. After removing them, the term frequency is calculated for remaining keywords which ever shown as follows:

- S1: Peace is everywhere
- S2: Nature is an art of God
- S3: Nature is a catchment of Peace

The word weight calculated for the above sample sentences are shown in Fig. 2. This Fig. 2 compares the word weight of each word calculated using TF and TF-ISF. The result shows that word weight of sample words are improved in TF-ISF measure. The cluster membership is assigned using random initialization because randomness is suppressed by the use of damping factor in sentence rank calculation. The cluster membership matrix order and similarity matrix order are N×C and N×N, i.e., 3×2 and 3×3 respectively, where N is the number of sentences and C is the number of clusters where C must be always less than N.

The term frequency is calculated using cosine similarity measure in equation 3. Then the weighted affinity matrix is the product of above calculated similarity matrix and cluster membership matrix. The cluster membership is assigned using random initialization or k-means approach or else the other kind of methods because randomness is suppressed by the use of damping factor in sentence rank calculation. The cluster membership matrix order and similarity matrix order are N×C and N×N respectively, where N is the number of sentences and C is the number of clusters where C must be always less than N. Here, cluster membership matrix and similarity matrix will be 3×2 and 3×3 matrices respectively and so the weighted affinity matrix will be 3×3 matrix since it is their product matrix.

The number of clusters is taken here as 2 since there are only three sentences available. In case of actual application, this option is left out to the user to decide the number of sentences.

Accordingly, each sentence with highest sentence rank from each of 2 clusters is taken. Thus there will be at most 2 sentences in the final generated summarized content. The equation 3 is applied for term frequency calculation using Term-Matrix Generator (TMG) in Matlab for each significant words in each sentence. With support of WordNet Database, the more weightage is given to words peace and nature. So there will be 2 clusters namely cluster 1 related to peace and cluster 2 related to nature. The weighted affinity matrix is as shown in equation. This may vary since the initial assignment of cluster membership is randomly designed:

$$C_{ij} = \begin{pmatrix} 0.613 & 0.383 & 0.260 \\ 0.380 & 0.960 & 0.328 \\ 0.26 & 0.32 & 0.448 \end{pmatrix}$$

Here, C_{ij} is the cluster membership matrix where i is the number of rows and j is the number of columns. Then the sentence ranking is estimated using equation 4 for each sentence. Let the ranking score for Sentence 1, Sentence 2 and Sentence 3 will be 0.28,0.30 and 0.98. The cluster 1 has Sentence 1 and cluster 2 has Sentence 2. While the Sentence 3 is linked to both, thus forming a overlapping structure of fuzzy cluster membership nature as shown in Fig. 3.

The sample cluster membership for the above three sentences. The final summary consists of Sentence 3 since it is related to both clusters and ranks with top sentence score in both clusters. Then the sentence 2 is

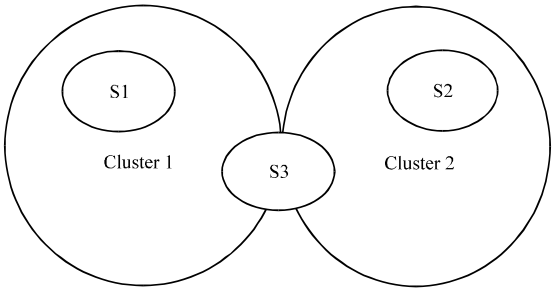

Fig. 3: Fuzzy cluster formation

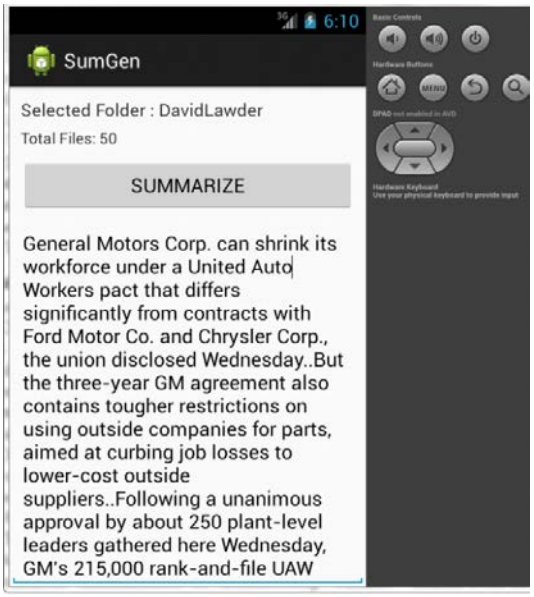

Fig. 4: Summarized output on AVD

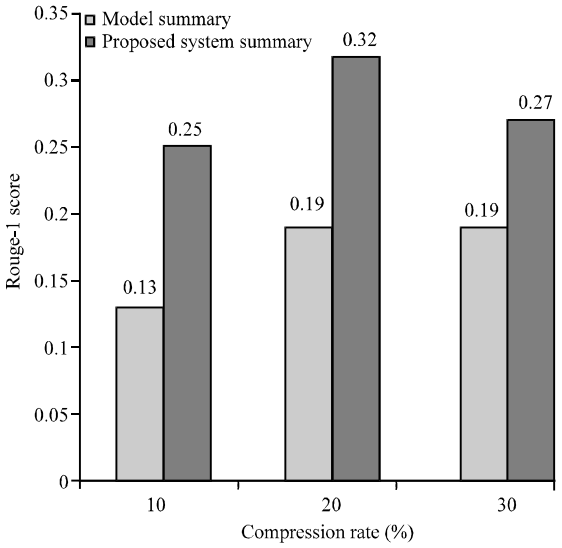

Fig. 5: ROUGE-1 evaluation measure

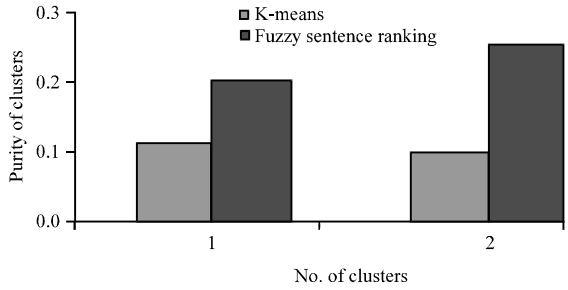

Fig. 6: Performance comparison between proposed algorithm and k-medoids based on Purity

added since it has the second top ranking score while compared to sentence 1. Similarly, the final summary for C50 dataset generated in the Android Virtual Device (AVD) is shown in Fig. 4.

ROUGE is a software package to evaluate the automatically generated summaries from the summarization system. It will compare the model summary and the system generated summary for evaluation by counting the number of matches. While comparing these scores, it shows that it is similar to that of manual summaries result.

The sample cluster membership for the above three sentences. The final summary consists of Sentence 3 since it is related to both clusters and ranks with top sentence score in both clusters. Then, the sentence 2 is added since it has the second top ranking score while compared to sentence 1. ROUGE is a software package to evaluate the automatically generated summaries from the summarization system. Figure 4 compares ROUGE-1 Score of model summary against proposed system summary for different CR. For evaluation purpose, ROUGE-1 Score is calculated for the summary of 10-30% compression rates. The result shows that by utilizing TF-ISF, at 20% CR the proposed system generated summary highly correlates with the model summary (Fig. 5).

The resultant score reveals that the proposed fuzzy clustering approach is better than the existing methods. When number of cluster increases, Purity is less in k-means when compared with proposed algorithm which is ~10% more than k-medoid as shown in comparison graph (Fig. 6).

The purity measure ensures how well the classes of objects are distributed within each cluster. Figure 5 reveals that the algorithm is more advantageous than the existing k-medoids algorithm. The existing algorithm lacks in depicting overlapping sentences. While the proposed method uses fuzzy logic for overlapping semantically related sentences as shown in Fig. 3.

CONCLUSION

The sentence level clustering is implemented using the fuzzy clustering approach. The algorithm is able to converge to an appropriate number of clusters, even though the number of initial clusters was set to a very high value. The proposed work computes lexrank for each sentence, since it is well suited for extractive multi-document summarization. The centroid sentence with highest score from each cluster is extracted and grouped together to form the summarized content of the original source. Finally, this automatic summarization process is deployed to the mobile devices. The future work objective is to extend this method to the development of hierarchical fuzzy clustering algorithm.

REFERENCES

Corsini, P., B. Lazzerini and F. Marcelloni, 2005. A new fuzzy relational clustering algorithm based on the fuzzy C-means algorithm. Soft Comput., 9: 439-447.

Erkan, G. and D.R. Radev, 2004. LexRank: Graph-based lexical centrality as salience in text summarization. J. Artificial Intell. Res., 22: 457-479.

Garcia, L.F.F., J.V. De Lima, S. Loh and J.P.M. De Oliveira, 2006. Using Ontological Modeling in a Context-Aware Summarization System to Adapt Text for Mobile Devices. In: Active Conceptual Modeling of Learning, Chen, P.P. and L.Y. Wong (Eds.). Springer, New York, pp: 144-154.

Geweniger, T., D. Zulke, B. Hammer and T. Villmann, 2010. Median fuzzy c-means for clustering dissimilarity data. Neurocomputing, 73: 1109-1116.

Ko, Y., J. Park and J. Seo, 2004. Improving text categorization using the importance of sentences. Inform. Process. Manage., 40: 65-79.

Li, Y., D. Mclean, Z.A. Bandar, J.D. O'Shea and K. Crockett, 2006. Sentence similarity based on semantic nets and corpus statistics. IEEE Trans. Knowl. Data Eng., 18: 1138-1150.

Liu, D., S. Wu, Y. Lan, G. Di, J. Peng, N. Xiong and A.V. Vasilakos, 2013. A query-oriented XML text summarization for mobile devices. Soft Comput., 17: 1585-1593.

Okazaki, N., Y. Matsuo, N. Matsumura and M. Ishizuka, 2003. Sentence extraction by spreading activation through sentence similarity. IEICE Trans. Inform. Syst., 86: 1686-1694.

Radev, D.R., H. Jing, M. Stys and D. Tam, 2004. Centroid-based summarization of multiple documents. Inform. Process. Manage., 40: 919-938.

Skabar, A. and K. Abdalgader, 2013. Clustering sentence-level text using a novel fuzzy relational clustering algorithm. IEEE Trans. Knowledge Data Eng., 25: 62-75.

Von Luxburg, U., 2007. A tutorial on spectral clustering. Stat. Comput., 17: 395-416.

Yuan, S.T. and J. Sun, 2003. Ontology-based task-oriented audio mining for Mobile B2E applications. ROC Technical Report, Fu Jen Catholic University, Taiwan.

Zha, H., 2002. Generic summarization and keyphrase extraction using mutual reinforcement principle and sentence clustering. Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 11-15, 2002, Tampere, Finland, pp: 113-120.