Global Journal of Advanced Engineering Technologies and Sciences

SUMMARIZATION OF TEXT USING FUZZY RELATIONAL CLUSTERING

Deepa D, Vishnu Raja.P
Department of Computer Science and Engineering
Kongu Engineering College
Erode, India
deepa@kongu.ac.in

Abstract

Information overload is a major problem in the modern digital world. It is difficult to retrieve the relevant content from the billions of documents. Moreover, the mobile devices have restricted memory, display screen and processing power. The mobile users prefer to analyse the summarized report, if it is relevant to their requirement, the user may observe it deeper. However, it is difficult to manually summarize the large documents of text. Sentence level clustering can be employed to perform automatic summarization task. Most of the sentence similarity measures do not represent sentences in common metric space. Hence, the conventional fuzzy clustering approaches based on prototypes are not appropriate for sentence level clustering. The Expectation Maximization (EM) framework is used to perform lexrank analysis on relational input data iteratively. The proposed work computes the centroid sentence of each cluster to automatically generate the summary which in turn reduces the manual processing. The integration of automatic summarization process to mobile devices is deployed by using Extensible Markup Language (XML). Since XML is originally designed to support large scale electronic publishing of documents, it is a flexible interface for mobile devices inorder to relieve the reading burden of the users. Experimental evaluation on the famous quotations dataset shows that the fuzzy relational clustering algorithm is capable of extracting the semantically related important sentences for generating the summarized content of original source.

Keywords—Fuzzy relational clustering, mobile devices, lexrank, expectation maximization.

I. INTRODUCTION

In the modern digital world, the volume of data continues to grow exponentially, after the commencement of internet and its applications. The origination of online technology has made the sharing of information a rapid progression. The mobile users might want to quickly gather the information instead of reading the vast content since the mobile devices have restricted hardware such as memory, display screen and battery power. In such case,the process of automatic summarization aids the user to quickly analyze the essential information in the content. Automatic summarization task aims to obtain the information source, extract the important content from the source and then reveal that most important content to the user. Automatic text summarization is being used in many real-time applications such as to summarize news to Wireless Application Protocol format (WAP-format) for mobile phones or handheld devices such as Personal Digital Assistants (PDA), to present compressed descriptions of the search results as in case of search engines and also to search in foreign languages and obtain an automatically translated summary of the automatically summarized text and so on.

Automatic Summarization can be performed in two ways. They are extractive summarization and abstractive summarization. The extractive method selects a subset of existing words, phrases or sentences in the original text to form the summary. The abstractive method builds an internal semantic representation and then use natural language processing techniques to create a summary closer to what a human might generate. This kind of summary might contain words that are not explicitly present in the original document. However, this method is quite harder to develop, so extractive methods are focused rather than abstraction. The general approach to generate a summary is to perform clustering, among which sentence clustering plays an important role. Sentence level clustering is applicable in various types of text processing activities. There is a common dispute that incorporating sentence clustering into extractive multi-document summarization and single document summarization helps to avoid the problem of content overlap which leads to better coverage.

Based upon the cluster patterns there are two different classifications, hard clustering and soft clustering. The cluster pattern belongs to single cluster in case of hard clustering while soft clustering or fuzzy clustering algorithms allow patterns to belong to all clusters with differing degrees of membership. The concept of fuzzy relationships leads to an increase in the breadth and scope of problems to which sentence clustering can be applied. Hard clustering methods are based on classical set theory, and require that an object either does or does not belong to a cluster. Hard clustering refers to partitioning the data into a specified number of mutually exclusive subsets. Fuzz clustering methods allow the objects to belong to several clusters simultaneously, with different degrees of membership. In many situations, fuzzy clustering is more natural than hard clustering since objects on the boundaries between several classes are not forced to fully belong to one of the classes, but they are assigned to membership degrees between 0 and 1 indicating their partial membership. Relational Fuzzy c-Means (RFCM) algorithm is considered to be the initial successful fuzzy clustering model. The Euclidean requirement for RFCM is considered to be restrictive and various alternatives have been proposed. The Spectral clustering approaches employs the data points which are mapped into the space defined by eigenvectors associated with affinity matrix. The fuzzy relational clustering algorithm adopts the graph representation, instead of density model, in which nodes represent objects and weighted edges represent the similarity between objects.

The rest of this report is organized as follows: Section 2 discusses the related work on automatic summarization and sentence similarity measures. The proposed work to perform Extractive Text Summarization in mobile devices using fuzzy relational clustering algorithm is discussed in Section 3. In Section 4, the experimental evaluation for the sample dataset is discussed. Finally, additional discussions, conclusion and future work are presented in Section 5.

II. RELATED WORK

There are different methods to perform automatic summarization. In general, clustering is well suited for automatic summarization. Aliguliyev (2009) proposed a new sentence similarity measure and sentence based extractive technique [1] 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. Corsini et al (2004) proposed a new fuzzy relational clustering algorithm [2] 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. 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. 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.

Tina Geweniger et al (2010) proposed median fuzzy c-means (MFCM) for clustering dissimilarity data [5]. 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. The algorithm is appropriate for many clustering tasks if non-metric data objects. have to be grouped in medical applications. Andrew Skaber and Khaled Abdalgader (2013) proposed clustering sentence level text using a novel fuzzy relational clustering algorithm [7]. The proposed algorithm is named as Fuzzy Relational Eigenvector Centrality-based Clustering Algorithm (FRECCA) since page rank centrality had been examined as a special case of eigenvector centrality. This method is applied on relational input data and operates in an Expectation Maximization framework in which the graph centrality of an object is interpreted as likelihood.

Centroid Based Summarization (CBS) uses the centroids of the clusters to identify the sentences which are central to the topic of the entire cluster. Radev et al (2003) proposed a centroid-based summarization of multiple documents [8]. 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. Summarization is performed by extracting important sentences based on the rank score. Erkan and Radev (2004) proposed a graph based lexical centrality as salience in text summarization [3]. The Extractive Text Summarization (ETS) approach relies on the concept of sentence salience to identify the important sentence in the document. LexRank approach is used to perform ranking to find the centrality of a sentence. The connectivity matrix based on intra-sentence cosine similarity is used as the adjacency matrix for graph representation of sentences.

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. Hongyuan Zha (2002) proposed generic summarization and keyphrase extraction [10] 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. Yuhua Li et al (2006) proposed a sentence similarity measure based on semantic nets and corpus statistics [9]. 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.

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. L.F.F Garcia et al (2007) proposed a context aware summarization system to adapt text for mobile devices [4]. 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. Dexi Liu et al (2013) proposed XML query oriented text summarization for mobile devices [6]. 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.

III. PROPOSED WORK

The proposed work generates summary automatically by performing sentence clustering at the initial stage and then ranking of those sentences is performed using Expectation Maximization (EM) algorithm along with LexRank analysis. The EM iteration alternates between performing an expectation (E) step, which creates a function for the expectation of the log-likelihood evaluated using the current estimate for the parameters, and maximization (M) step, which computes parameters maximizing the expected log-likelihood found on the E step. The convergence of cluster is assured in the EM algorithm since the algorithm is guaranteed to increase the likelihood in each iteration. While, the classical clustering algorithms assign each data object to exactly one cluster and forms a crisp partition of the given data, which may not be appropriate to certain applications, where the data object might be related to multiple clusters. Hence, fuzzy clustering allows the data object to belong to different cluster with different degrees of membership.

The architecture design to perform automatic summarization is shown in the Figure 4.1. The sentences are extracted from the input document and then the stopwords are removed from each sentence to calculate its term frequency in the preprocessing module. After preprocessing, the similarity matrix is formed between the sentences. Then the relevant sentences are clustered using EM algorithm. The semantically related fuzzy clusters are converted to hard cluster groups. Finally, the centroid sentence with highest lexrank score from each cluster is extracted to generate summarized content.

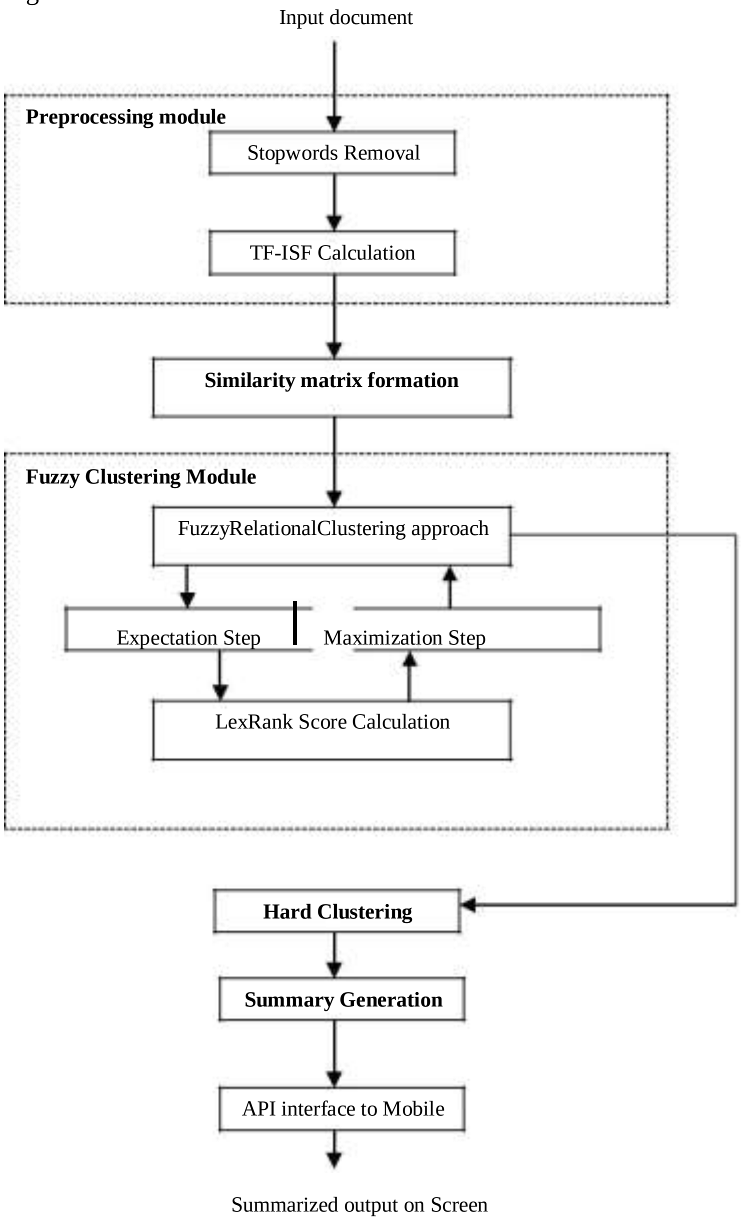

Figure 1. Overview of the Automatic Summarization process.

3.1 Algorithm

The enhanced fuzzy relational clustering algorithm uses the LexRank score of the sentence within each cluster as a measure of its centrality to that cluster. The proposed algorithm consists of four steps.

They are:

1. Initialization and Normalization
2. Expectation (E-step)
3. Maximization (M-step)
4. Summarization

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

Step 1: Initialization and
    Normalization for each
    sentence
    for each cluster
    Initialize random cluster
    membership value end for
    end for
    for each cluster
    Calculate normalized cluster membership
    value end for

Step 2: Expectation (E-step)
Repeat the following procedure until the convergence of cluster membership values occurs

for each
sentence
for each
cluster
for each other similar
sentence Calculate
weighted affinity matrix
end for
end for
Determine LexRank score of each sentence and assign to likelihood
for each
sentence
for each
cluster
Update new cluster membership
value end for
end for end for

Step 3: Maximization
(M-Step) for each cluster
Update mixing
coefficients end for

Step 4:
Summarization
for each cluster
for each sentence within
cluster Determine
centroid sentence end for
end for
```

After random initialization, the Expectation step (E-step) is followed by the Maximization step (M-step) are iterated until convergence. The E-step computes the cluster membership probabilities using lexrank. The M-step updates the mixing coefficients based on the cluster membership values. The Summarization step extracts the centroid sentence from each cluster.

3.2 Preprocessing

Inaccurate, incomplete and inconsistent data are common place properties of large real-world databases and data warehouses. Low quality data will lead to incorrect mining results. So, effective text mining process depends on preprocessing technique. The first step in preprocessing module is to identify the keywords. There are some words like a, an, the, is, of etc., which do not carry any useful information. These words are called as stop words. The stop words include articles, prepositions, tenses and so on. In order to determine the importance of a term in a document, term frequency is calculated. The Term frequency is given by the Equation 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.

3.3 Similarity matrix formation

After preprocessing, the similarity matrix formation is performed since it is the essential input to the fuzzy relational clustering algorithm. 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. Similarity between the sentences is calculated by modifying the cosine similarity with the inverse sentence frequency (ISF). The Inverse Sentence Frequency can be defined using the Equation 2.

$$ISF(t) = \log \frac{N}{n_i} \quad (2)$$

where N is the total number of sentences in the document collection and n_i is the number of sentences containing that particular term t. Consider two sentences x_i and y_i, their modified cosine similarity measure is given by the following Equation 3.

$$\mathrm{cosine}(x_i, y_i) = \frac{\sum_{w\in x,y} tf_{w,x}\; tf_{w,y}\; (isf)_2}{}$$

$$\sqrt{\sum_{x_i \in x} (tf_{x_i,x}\, isf_{x_i})_2}\ x \sqrt{\sum_{y_i \in y} (tf_{y_i,y}\, isf_{y_i})^2} \qquad (3)$$

where $tf_{w,x}$ is the number of occurrences of word w in sentence x, n represents the total number of sentences, x_i and y_i are 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.

3.4 Fuzzy cluster formation

Each sentence in the original document is randomly initialized to a cluster membership value. 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 LexRank score of each object in each cluster using the weighted affinity matrix obtained by scaling the similarity matrix. The lexrank is computed using cosine similarity as shown in Equation 4.

$$l(x) = \frac{d}{N} + (1-d) \sum_{y \in y_i} \frac{cosine(x, y)}{cosine(z, y)}\, x l(y) \qquad (4)$$

where d is the damping factor which is added for absorbing the errors due to convergence, x,y and z are the sentences that are linked in the fuzzy clusters. The maximization step involves updating mixing coefficients based on the cluster membership values estimated in the Expectation step.

3.5 Hard clustering

The fuzzy clusters generated using the fuzzy relational clustering is converted hard clustering. 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. In order to reveal the global importance of a sentence in a document, Global Page Rank is calculated using Equation 5.

$$GPR_i = \sum_{j=1}^{C} p_j\, PR_i^j \qquad (5)$$

where p_j is the mixing coefficient of the corresponding cluster and PR is the page rank score of sentence i in cluster j.

3.6 Summarization

In each cluster, the sentence which is having highest lexrank 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 retrieved to generate the summarized content. The summarized content is deployed using XML and java Application Programming Interface (API) with the help of alchemy interface to support the mobile environment.

IV. EXPERIMENTAL EVALUATION

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. ROUGE produces individual scores for 1,2,3 and 4-gram matching between the summaries. While comparing these scores, it shows that unigram score is similar to that of manual summaries result. The ROUGE-N measure can be defined using the Equation 6.

$$ROUGE-N = \frac{\sum_{S \in Summ_{ref}} \sum_{N-gram \in S} Count_{match}(N-gram)}{\sum_{S \in Summ_{ref}} \sum_{N-gram \in S} Count(N-gram)} \qquad (6)$$

where N is the length of the N-gram, $Count_{match}$(N-gram) is the maximum number of N-grams matches with

the comparative summaries and Count(N-gram) is the number of N-grams in the reference summaries. The following Table 1 shows the effectiveness of the proposed work.

Table 1. Average values of evaluation metrics.

System	ROUGE-1	ROUGE-2
SVM	0.44628	0.17018
Fuzzy Clustering	0.45645	0.17412
CRF	0.45512	0.17327

The resultant score reveals that the proposed fuzzy clustering approach is better than the existing methods.

V. CONCLUSION AND FUTURE WORK

The sentence level clustering is implemented using the fuzzy relational 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

I. Aliguyev R.M. (2009), 'A New Sentence Similarity Measure and Sentence Based Extractive Technique for Automatic Text Summarization,' Expert Systems with Applications, Vol. 36, pp. 7764-7772

II. Corsini P., Lazzerini F., and Marcelloni F. (2005) 'A New Fuzzy Relational Clustering Algorithm Based on the Fuzzy C-Means Algorithm,' Soft Computing, Vol. 9, pp. 439-447, 2005.

III. Corsini P., Lazzerini F., and Marcelloni F. (2005) 'A New Fuzzy Relational Clustering Algorithm Based on the Fuzzy C-Means Algorithm,' Soft Computing, Vol. 9, pp. 439-447, 2005.

IV. Erkan G. and Radev D.R. (2004), 'LexRank: Graph-Based Lexical Centrality as Salience in Text Summarization,' J. Artificial Intelligence Research, Vol. 22, pp. 457-479

V. Garcia L.F.F., Lima J.V., Loh S., Oliveira J.P.M. (2007), 'Using Ontogical Modelling in a Context Aware Summarization System to Adapt Text for Mobile Devices,' ACM LNCS 4512, pp. 144-154.

VI. Geweniger T., Zuhlke D., Hammer B. and Villmann T. (2010), 'Median Fuzzy C-Means for Clustering Dissimilarity Data,' Neurocomputing, Vol. 73, No. 7-9, pp. 1109-1116

VII. Liu D., Wu S., Lan Y., Di G., Peng J., Xiong N., Vasilakos A.V. (2013),' A query-oriented XML text summarization for mobile devices,' Soft Computing, pp.1585–1593.

VIII. Skaber A. and Abdalgader K. (2013), 'Clustering Sentence Level Text using a Novel Fuzzy Relational Clustering Algorithm,' IEEE Transactions on KNowledge and Data Engineering, Vol. 25, pp. 62-75

IX. Radev D.R., Jing H., Stys M. and Tam D. (2003), 'Centroid-Based Summarization of Multiple Documents,' Information Processing and Management, Vol. 40, pp. 919-938, 2004.

X. Yuhua Li, McLean D., Bandar Z.A., Shea J.D.O. and Crockett K. (2006), 'Sentence Similarity Based on Semantic Nets and Corpus Statistics,' IEEE Transactions on KNowledge and Data Engineering, Vol. 18, pp. 1138-1150.

XI. Zha H. (2002) 'Generic Summarization and Keyphrase Extraction Using Mutual Reinforcement Principle and Sentence Clustering,' Proc. 25th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 113-120.