Volume 4, Issue 12, December 2014 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

A Survey on Sentence-Level Text Clustering

[1]V. Arun Raj Kumar, [2]Dr. G.Arumugam
[1]MCA, MPhil., Dept. of Computer Science, M.K. University, Madurai, Tamil Nadu, India
[2]Senior Professor & Head, Dept. of Computer Science, M.K. University, Madurai, Tamil Nadu, India

Abstract: ***This paper surveys various results related to clustering. It explains about the problems in clustering in sentence level text and the solutions to overcome these problems. Nowadays, a web search engine often returns thousands of sentences in response to a query, making it difficult for users to browse or to find relevant information. To overcome these issues and to get high quality information, different clustering methods and algorithms are used. The search engine uses clustering methods to automatically group the retrieved sentences into a list of meaningful categories. Then the similarity between the sentences present in each cluster is found by applying some clustering algorithms. The sentence, which is most similar to the user query, is found as a relevant information or output of a user query. Then the algorithm gives high quality information.***

Keywords: EM Algorithm, Renovate Algorithm, Affinity Propagation, Novel Fuzzy Relational Clustering Algorithm.

I. INTRODUCTION

In a document, single statement is likely to be related to more than one theme or topic present in the same document. Hence a user gets relevant information with more than one statement having the same meaning instead of a single statement. So, it requires more time to retrieve all the statements and also large space is required to store all the statements having single meaning in a database. It is not necessary for a user to use all the statements that represent a single meaning. To overcome this situation, Sentence Clustering is introduced to measure the similarity among the number of statements having the same meaning. It plays an important role in text processing activities and it helps to avoid problem of content overlap. Here the sentences, which are related to each other, are grouped into clusters. By clustering the sentences, the system would intuitively expect at least one of the clusters to be closely related to the user query terms, but other clusters may contain information which is not related to the query and may be unknown to the user, and in such a case the system would have successfully mined new information. By applying any clustering method or clustering algorithm in that group of statements, user can get a single statement as relevant information, which is most similar to the user query.

II. VARIOUS CLUSTERING METHODOLOGIES

a) Expectation Maximization Clustering Algorithm.
b) RENOVATE Algorithm.
c) Affinity Propagation
d) Novel Fuzzy Relational Clustering Algorithm.

We discuss these methodologies in the subsequent sections.

a) Expectation Maximization Clustering Algorithm

Nowadays in web there are large amount of information available. It will be easy for user to search particular information by entering a keyword but it is difficult to find out useful, relevant, and significant information. Text summarization is an important tool for supporting and interpreting the data when a vast amount of data is available. Text summarization is the process of compressing the information without changing the original meaning. This is done by using the Expectation Maximization Clustering Algorithm. The focal point of summarization is Natural Language Processing. It will remove the uncertainty of the given sentence and will get an accurate summary maintaining the originality. The interpretation of the text and the available sentences are grouped in NLP. It is a preprocessing activity and the main aim is achieved by using the following two steps:

1) Implementation of Expectation Maximization Clustering algorithm.
2) Application of query dependent summarization by removing ambiguity.

For understanding the meaning of the text, the word net dictionary is used. Each sentence from the input text is considered as a single node and each node is compared with every other node by using the word net dictionary to calculate the weight. Then each sentence with the corresponding weight is represented in the form of document graph. The Expectation Maximization Clustering Algorithm **[2]** is used before summarization to generate an effective summary.

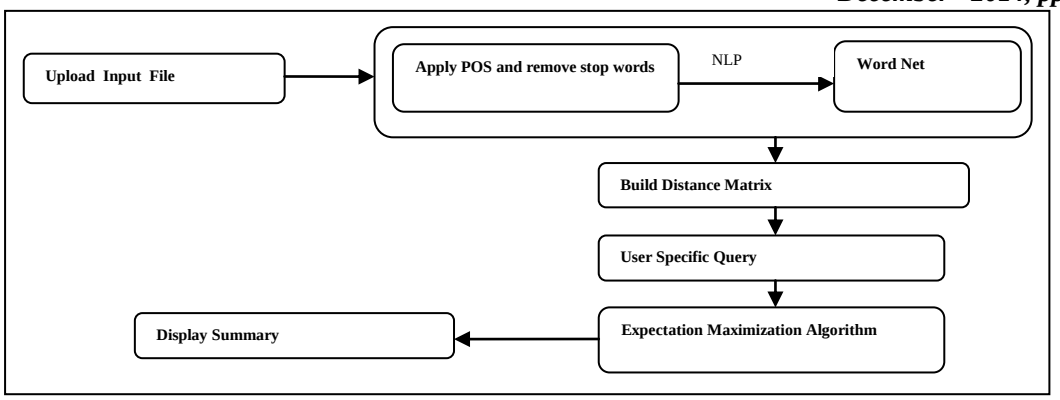

Fig: 1.1 Phases of Text Analysis

Phases of NLP Parser Engine

The phases of NLP parser engine are important task where it will chunk the text in the form of a graph by understanding the meaning of the text and the process is described as follows.

Split Sentence

The input data are split into separate sentences by the new line character and converted into an array of paragraphs by using split method. This can be done by treating each of the characters '.', '!', '?' as separator rather than definite end-of-sentence markers.

Tokenization

The input text is separated into tokens. Punctuation marks, spaces and word terminators are the word breaking characters.

Part Of Speech tagger

POS tagger is applied for grammatical semantics. Part-of-speech tagging is the process, which is applied after tokenization. The input to a tagging algorithm is a string of words of a natural language sentence. The output is a single best POS tag for each word. Part-of-speech tags are NN: Noun DT: Determiner VBN: Verb, past participle etc..

Chunker

Text chunking is dividing the text into parts of words and forming groups like verbgroup and noun group.

Parser

It generates the parse tree for a given sentence. Parsing is converting an input sentence into a hierarchical structure that corresponds to the units of meaning in the sentence.

A Distance Matrix is built to find closely related words. After building a distance matrix the user can fire a query. Then clustering is done by using the EM algorithm.The EM algorithm is an iterative procedure that consists of two alternating steps: (i) an expectation step and (ii) a maximization step. The expectation step is with a high opinion to the unknown basic variables, using the current estimates of the parameters and conditioned upon the observations. The Maximization step provides new estimates of the parameters. At each iteration, the estimated parameters provide an increase in the maximum-likelihood (ML) function until a maximum is achieved. The clustering algorithm plays an important role in space and time consumption. Then finally document graph is constructed to find the sentence with maximum weight and it is considered in summarization.

Merits and Demerits of Expectation Maximization clustering Algorithm

The main advantages of the EM algorithm are its simplicity and ease of implementation. Implementing the EM algorithm does not usually require heavy preparatory analytical work. It is easy to program, either it reduces to very simple re-estimation formulae or it is possible to use standard code to perform the Expectation step.

The main disadvantage of the EM algorithm is its slow linear convergence (junction). The hurrying schemes are used but they generally require some preparatory analytical work and this increases the complexity of the implementation.

To avoid the increase in the complexity of an implementation,E.Priya, C.Nivetha, A.Rajalakshmi introduced the system called **the fast recovery system** (i.e.,) **clustering data using labels [3]**. It assigns labels to each cluster and makes the implementation very easy. In this system, one more problem like preparation of analytical work need not be prepared. It also automatically updates the cluster by assigning the input data to the cluster that is most similar to the cluster. For example, one of the applications like the **disease prognostication** is done by using this system and it shows as follows:

b) RENOVATE Algorithm

The survey is made for use of clustering data using labels for making fast recovery. This system uses a new algorithm called RENOVATE algorithm **[3]** in a particular application like clustering the patient record for mentioning the disease

of the patient. The RENOVATE algorithm clusters the data based on everyday update and produces the best result for the doctor to understand the state of affairs of the patient. The renovate algorithm clusters the record on the core of the labels and provides the result. And also gives how the clustering process is carried to place them in labeled order. The algorithm described in this system performs clustering by calculating the outliers and similarities and also have the advantage of clustering the large set of data with minimal time consuming and clustering in labeled order.The application uses the following steps to do the clustering on a set of data by implementing the RENOVATE algorithm in the patient test details to find out the disease of the patient.Details of patients in hospital are taken as an input and they are stored in the database of hospital management. These data are analyzed in order to find out by which disease among the five diseases like pneumonia, diabetes, cancer, tuberculosis and malaria the patient is suffering from. Hence to detect the disease of the patient, the comparison takes place on the test results and symptoms of the patient to predict the disease of the patient. Based on the output, it provided suggested treatment for the patients. In **[13]**YuvrajKumbharey, SuweshShukla, SushilChaturvedi, proposed that the implementation of Renovate algorithm is as follows. The implementation is done by representing the data objects in a graph. Each node has different data objects. The nodes are divided into clusters. When a node is created, the state of a node is considered as an 'undecided' state. At first this node is to start a timer and transmits a "HELLO" message to the cluster-head. If a cluster-head receives this HELLO message, it replies immediately with a triggered HELLO message. After that, when the node receives this answer, it will change the undecided state into the "member" state. But when no message is received, then it makes itself as a cluster-head, but only when it has bi-directional link to one or more neighbors. Otherwise, when it has no link to any other node, it stays in the "undecided" state and retransmits the HELLO message. Cluster-heads are changed infrequently and it has not only the information's about the members of its cluster in the table, but it also maintains a cluster adjacency table that contains information about the neighboring clusters. In this table the gateway through which the neighbor cluster can be reached saved, and also the ID of the cluster-head.

Disadvantages:

The demerits of this system are, when any node has no link to any other node, it stays in the "undecided" state and repeats the procedure by sending the message again. The repeated procedure takes more time to implement the algorithm. To avoid this situation, a data point that is representative of itself (exemplars) is found by randomly choosing an initial subset of data points and then iteratively refining it. B.J. Frey and D. Dueck**[1]** proposed a method for measuring the similarity between the data points. This method is called as "affinity propagation".

c) Affinity Propagation

Clustering data based on the similarity measure is somewhat a typical step in scientific data analysis. To make clustering in an easy way of using a common approach is to use data to learn a set of centers by using the sum of squared errors between data points and their nearest centers. If the centers are selected from real data points, they are called "exemplars".In a network, viewing each data point as a node and a method was developed that recursively transmits real-valued messages along the edges of the network until a good set of exemplars and corresponding clusters appear. The Messages are updated on the basis of simple formulas and the energy function is chosen appropriately for minima. The magnitude of each message reflects the current affinity that one data point has for choosing another data point as its exemplar at any point in time and hence the method is called as **"affinity propagation" [1]**.

Affinity propagation:

An algorithm that identifies exemplars among data points and forms clusters of data points around these exemplars. This operation is performed simultaneously considering all data points as potential exemplars and exchanging messages between data points until a good set of exemplars and clusters emerges.Affinity propagation takes as input a collection of real-valued similarities between data points, where the similarity is s(p,i) indicates how well the data point with index i is suited to be the exemplar for data point p. If the squared error need to be minimized, then each similarity is set to a negative squared error (Euclidean distance): For points xp and xi, s(p,i) =square(−||xp− xi||).There are two kinds of messages exchanged between data points. Sometimes messages are combined to decide which points are exemplars and the points other than exemplars in which exemplars it belongs to. The "responsibility" r(p,i), sent from data point p to candidate exemplar point i. The "availability" a(p,i), sent from the candidate exemplar point i to point p.The availabilities are initialized to zero: a(p,i) = 0. Then, the responsibilities are computed using the following rule, the max value of (the summation of the availability and the similarity value) is subtracted with the similarity value and then the resultant value is considered as a responsibility value.The evidence gathers from data points by using following availability updates as to whether each candidate exemplar would make a good exemplar.The availability is found by calculating the minimum value of(0, the summation of the responsibility value of (i,i) and the max value of 0,the responsibility value of(p,i)).

Drawbacks of Affinity Propagation:

Affinity propagation has many advantages over other methods, but one disadvantage (for some applications) is that the number of clusters cannot be pre-specified. After each iteration, it is often the case that the resulting number of clusters starts high (all data points) and shrinks down to the number of clusters that finally occur. This algorithm also uses N*N comparisons of data and it leads more time to execute the algorithm.

To overcome this situation (i.e.,) to avoid N*N comparisons Andrew Skabar, and KhaledAbdalgader proposed an algorithm called the **Novel Fuzzy Relational Clustering Algorithm[4]**. This algorithm operates only in a relational data

which is got from the execution of user query. It has the capability of deciding the number of clusters automatically. The comparison takes place only on the fuzzy data. So the implementation time of this algorithm is lesser than the previous method and the algorithm is discussed in the next section.

d) Novel Fuzzy Relational Clustering Algorithm

The Novel Fuzzy Relational Clustering Algorithm operates on fuzzy relational data within the cluster. The Novel information is discovered from the set of documents by executing the user query. The input data is represented in a graph and each node represents the data object. The operation of the graphical representation of an input data is in an Expectation-Maximization framework in which the graph centrality of an object in the graph is taken as likelihood. After applying the algorithm in sentence clustering tasks the results demonstrate that it is capable of identifying overlapping clusters of semantically related sentences.

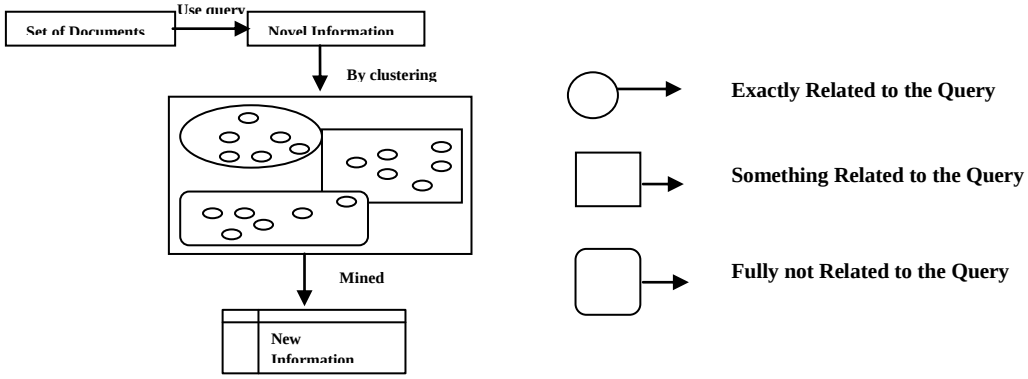

Fig: 3 Fuzzy Relational Clustering System

From the collection of retrieved data, some amount of data only related to the user Query and whether the remaining data is related to the query or not is unknown to the user. The relationship between those data should be considered as a fuzzy relationship and it is found out by using the Relational Fuzzy Clustering Algorithm.

RFCM operates on relational data input; it still requires that the relation expressed by this data be Euclidean. The squared Euclidean distance between points in the space matches those in the dissimilarity relation. RFCM is considered as restrictive and hence ARCA (Any Relational Clustering Algorithm) is used. This algorithm is attribute based representation. The input used in this algorithm also a relational data. An object is represented by a vector of its relationships with other objects in the data set.Afterwards K-Means algorithm is used. It is a prototype based algorithm and involves calculation of the Euclidean distance between the pairs of data objects in the cluster. This process is considered as a minimization step in which cluster means and covariance values are updated.Then K-Medoids algorithm is used to find out the fuzzy relationship of data. The calculations performed here are only based on the pairwise relations but not on the distances.

Novel information:

Information received by the five senses (sight, sound, taste, touch, and smell) are new, different, and abnormal. Any novel information has the potential to be aggressive and it triggers an automatic response. Novel information is calculated in contrast to unnecessary information, which is familiar and apparently nonthreatening. For example: Healthy person, depressed patient etc.A N**ovel Fuzzy Relational Clustering Algorithm[4]** uses the data with a graph representation. Each node in a graph represents an object and the similarity between the objects is represented by using the weighted edges. Cluster membership values for each node represent the degree to which the object is represented by that node. By applying the PageRank algorithm **[4]** to each cluster, and inferring the Page-Rank score of an object within some clusters as a likelihood, the system can then use the Expectation- Maximization (EM) framework [4] to determine the model parameters (i.e., cluster membership values and mixing coefficients). It provides a fuzzy relational clustering algorithm which can be applied to any domain in which the relationship between objects is expressed in terms of pairwise similarities.

PageRank assigns to every node in a directed graph a numerical score between 0 and 1 known as its PageRank score (PR).

To deal with weighted undirected edges the PageRank algorithm can easily be modified by using the following calculation,

$PR(V_i) = (1-d) + d \times \sum j=1 \text{ to } N \ (w_{ji} x \ PR(V_j) \ / \ \sum k=1 \text{ to } N \ w_{jk})$. The algorithm uses Expectation Maximization to optimize the parameters used here. The system assumes that the similarities between objects are stored in a similarity matrix $S = \{s_{ij}\}$, where s_{ij} is the similarity between objects i and j. The process of finding similarity is described below.

Initialization:

The system assumes here that the cluster membership values are initialized arbitrarily and standardized so that cluster membership for an object sums to unity over all clusters. Integrating coefficients are initialized such that the priority for all clusters are equal.

Expectation step:

The PageRank value is calculated for each object in each cluster with the affinity matrix weights w_{ij} obtained by scaling the similarities by their cluster membership values; i.e.,$W_{ij}^{m} = s_{ij} \times p_i^m \times p_j^m$

Where w_{ij}^{m} is the weight between objects i and j in cluster m, and p_i^m and p_j^m are the respective membership values of objects i and j to cluster m. The intuition behind this scaling is that an object's entitlement to contribute to the centrality score of some other objects depends on its similarity to another object and also on its degree of membership to the cluster. The obtained PageRank scores are treated as likelihood and used to calculate cluster membership values.

Maximization step:

In this maximization step there is no parameterized likelihood function and it involves only updating the mixing coefficients based on membership values calculated in the Expectation Step. The pseudo code is presented in Algorithm [4], where w_{ij}^{m}, s_{ij}, p_i^m, and p_j^m are defined as above, _m is the integrating coefficient for cluster m, PR_i^m is the PageRank score of an object i in cluster m, and l_i^m is the likelihood of an object i in cluster m.The FRECCA algorithm [4] is used to implement the above steps. This algorithm uses the input values as pairwise similarity values and the number of clusters. After applying the algorithmic concept by using these input values in the Expectation Maximization steps the cluster membership values are given as output.

The drawback of this system is that identifies only flat clusters. The future work may be taken as to extend these ideas to the development of a hierarchical fuzzy relational clustering algorithm.

III. CONCLUSION

Many users may not have a sense about the relationship between the data within a cluster. Even though they take care of the relationship between the data yet they do not consider it until the last data in a cluster. This type of system produces only the particular type of cluster namely flat cluster. When a user finds out the fuzzy relationship between the data they should consider until the last data present within a cluster. This could be done by using the hierarchical fuzzy relational clustering algorithm. The results of applying the algorithm to sentence clustering tasks demonstrate that the algorithm is capable of identifying overlapping clusters of semantically related sentences.

REFERENCES

[1] B.J. Frey and D. Dueck, "Clustering byPassing Messages between Data Points," Science,vol. 315, pp. 972-976, 2007.

[2] Ms. Meghana. N.Ingole, Mrs.M.S.Bewoor, Mr.S.H.Patil / International Journal ofEngineering**"**Text Summarization using ExpectationMaximizationClustering Algorithm**"**Vol. 2, Issue 4, July-August 2012, pp.168-171.

[3] E.Priya, C.Nivetha, A.Rajalakshmi, "Clustering data using labels for disease prognostication and making meticulous firmness for fast recovery" Vol.2, Issue 12, December 2013

[4] Andrew Skabar, Member, IEEE, and KhaledAbdalgader "Clustering Sentence-Level Text Using a Novel Fuzzy Relational ClusteringAlgorithm" Vol. 25, NO. 1, January 2013

[5] W.M. Rand, "Objective Criteria for the Evaluation of Clustering Methods," Am. Statistical Assoc. J.,vol. 66, no.338, pp. 846-850, 1971.

[6] A. Rosenberg and J. Hirschberg, "V- Measure: A Conditional Entropy-Based External Cluster Evaluation Measure," ProcConf.Empirical Methods inNatural Language Processing (EMNLP '07),pp. 410-420, 2007

[7] R.M. Aliguyev, "A New Sentence Similarity Measure and Sentence Based Extractive Technique for AutomaticText Summarization,"Expert Systemswith Applications, Vol. 36, pp. 7764 - 7772, 2009.

[8] G. Ball and D. Hall, "A Clustering Technique for Summarizing Multivariate Data," Behavioural Science, Vol. 12

[9] E.H. Ruspini, "A New Approach to Clustering," Information and Control, Vol. 15, pp. 22-32, 1969.

[10] E.H. Ruspini, "Numerical Methods for Fuzzy Clustering," Information Science, Vol. 2, pp. 319-350, 1970.

[11] YuvrajKumbharey, SuweshShukla, SushilChaturvedi, " Renovated Cluster Based Routing Protocol for MANET" Volume-3Number-1Issue-8

