IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 6, June 2016

Clustering Sentence Level Text using Hierarchical FRECCA Algorithm

Mujawar Nilofar Shabbir[1], **Prof. Amrit Priyadarshi** [2]

PG Scholar, Department of Information Technology, DGOI, FOE, Bhigwan, Savitribai Phule Pune University, India[1]

Assistant Professor, Department of Computer Engg., DGOI, FOE, Bhigwan, Savitribai Phule Pune University, India[2]

Abstract: Text Processing is essential to organize data or to extract needful information from a heap of available Big Data. Sentence clustering is one of the processes used in Text mining task. Text document may contain Hierarchical structure which relate to more than one theme at a same time. Hence Hierarchical Fuzzy Clustering Algorithm can be used for clustering such text data. The paper presents a novel Hierarchical Fuzzy Relational Eigenvector Centrality-based Clustering (HFRECC) Algorithm which is extension of FRECCA Algorithm. It solves the problems like complexity, sensitivity and changeability of clusters and is useful for natural language document (NLP) and operates in Expectation-Maximization Framework and is capable to identify overlapping clusters. The algorithm uses graph representation of data and works on relational data provided viz., data in pairwise similarities among data objects.

Keywords: HFRECCA, FRECCA, Page-rank, Sentence clustering, Expectation-Maximization.

I. INTRODUCTION

Clustering organizes the documents to improve browsing and information retrieval. Clustering sentence is important area of text processing and text mining which helps in knowledge discovery. Clustering, groups similar data objects together and helps to discover the hidden similarity, main concepts and summarizes a large amount of text into groups. Most of the documents contain inter-related themes or terms and many sentences are inter-related in some degree to these. In many Natural Language Processing (NLP) tasks Clustering algorithms are used. They are popular and effective to use and discover similar groups of linguistic objects. In generally clustering algorithms can be classified into two categories as hard clustering and soft (fuzzy) clustering. In hard clustering, each data element is included in a single cluster group. Perhaps in soft clustering, data elements belong to more than single cluster. The work here deals with the ability to capture such Fuzzy relationships and scope of sentence clustering to solve it. It also solves the problem of content overlapping. Clustering documents in Information Retrieval is achieved by many clustering algorithms, but clustering text at sentence level is efficiently performed using Fuzzy Algorithm. The assumption of measuring similarity within documents or sentences based on word co-occurrences lead to many of the sentence similarity measures. These measures are not sufficient to represent sentences in common metric space; hence the topic of interest is Fuzzy Relational Clustering, based on pairwise similarity on relational data inputs. The first Fuzzy clustering algorithm which proved to be successful is Relational Fuzzy c-Means (RFCM) Algorithm. It operates on relational input data which requires relation to be Euclidean.

The paper represents use of a Fuzzy Relational Clustering Algorithm. The use of graph representation is done here to represent data objects in which nodes are used to represent objects and the weighted edges for denoting the similarity between objects. For each node cluster membership values are assigned which shows the degree to which object represented by that node belong to each respective clusters and the probability of object being generated by that component is shown by mixing coefficient. This can be achieved by application of PageRank to each cluster and computing the PageRank score of an object as likelihood in some of clusters. To determine the model parameters (i.e., mixing coefficient and membership values), the use of Expectation-Maximization (E-M) Framework is used. The result can be applied in any domain were the relationship between objects is expressed in terms of the pairwise similarity.

The paper proposes the Hierarchical Fuzzy Relational Eigenvector-Centrality based Clustering Sentence Level Text (HFRECCS), an extension of FRECCA used for clustering of sentence level text. The paper is structured in following way.

II. RELATED WORK

In sentence level text clustering [1], the common technique used is based on the statistical analysis of either word or phrase. Fuzzy relational clustering algorithm is used in text clustering since sentences may be related to more than a single theme. Fuzzy relational algorithm can deal with patterns belonging to more than one cluster. PageRank algorithm and Expectation-Maximization are used to measure graph centrality and to contract a complete fuzzy relational clustering algorithm. The algorithm is able to identify overlapping clusters hence used in various texts mining processing.

The paper [2] proposed a statistical homogeneous attribute which is quantifying and implements clustering, SIMFINDER, which organizes minute pieces of text from

DOI 10.17148/IJARCCE.2016.56138

one or multiple documents into tight clusters. The approach implements linguistic features and complex clustering algorithm to construct sets of related sentences. Here homogeneous attributes is nothing but prediction based on shared words. The approach makes use of k-Mediods clustering techniques identifying meaningless data arriving from unstructured text and identifies only high quality clusters. The vector space model in IR is successful, as it is able to identify much of the semantic content in document-level text. As the documents that are mostly semantically related and are likely to contain many words which are in common, which are based on word co-occurrence as per the vector space model [3].This is valid only for document level text, the assumption can't hold good for sentence level texts or short text fragments. Hence to solve this problem at sentence level, a number of sentence similarity measures are proposed [4].

The paper [5] proposed a novel measuring method which is applicable for sentence fragments and short sentences. Instead of traditional vector space model, this method calculates the similarity based on distance model.

The paper [6] proposed a method that is performed in the correlation with related attribute measure space. In this they are shown into low-dimensional semantics. Clustering technique used here is spectral with Low computation cost. This method is applicable when the number of clusters is small.

III. PROBLEM STATEMENT

Document clustering is well established and renowned but if one has to cluster sentence level text it may cause problem as sentence may contain words related to more than a single theme. The paper implements an application which is used to clustering sentence level text. The algorithm uses a graphical representation of data and works within a framework of Expectation-Maximization. HFRECCA is able to identify clusters which can overlap semantically.

IV. PROPOSED SYSTEM

The proposed Hierarchical Fuzzy Relational Clustering Algorithm can solve the problems of complexity, sensitivity and changeability of clusters. This algorithm is extended from FRECCA. The proposed system is based on FRECCA Algorithm.

Input: User Query

Output: Fuzzy Relation and related sentences

The algorithm operates in an Expectation-Maximization framework and uses graph representation of data.

Proposed architecture is designed considering following modules:

A) PageRank
B) EM algorithm
C) FRECCA
D) HFRECCA

The detailed description of the modules of proposed system is as:

A. Page Rank

Page Rank is used as a graph centrality measure and used to determine the importance of a particular node within that graph. Importance of node is also used as a measure of centrality. The algorithm assigns numerical score from 0 - 1 to every node in graph known as Page Rank Score. In graph node represents Sentence and weighted edges represent similarity value between sentences. To formulate the cluster and optimize the values of parameters PageRank can be used within the E-M Framework.

B. E-M algorithm

It tries to find the parameters of the probability distribution with maximum likelihood of data objects. It is an iterative method used to find maximum likelihood parameters of model. Parameter estimation is main role of this algorithm. The E-step involves the computation of cluster membership probabilities. The probabilities calculated from E-step are re-estimated with the parameters in M-step.

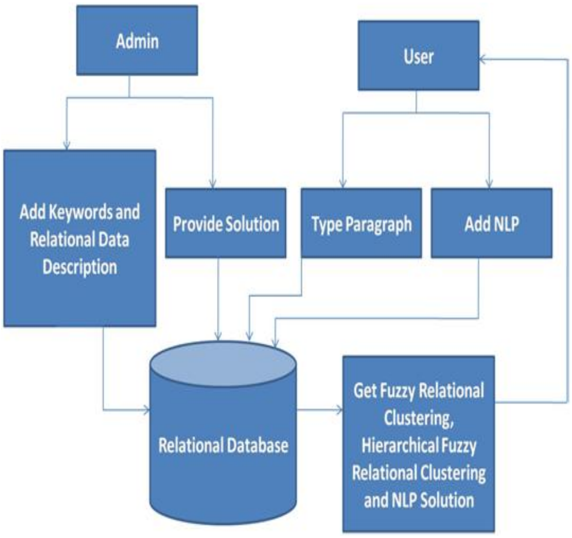

Admin
User
Add Keywords and Relational Data Description
Provide Solution
Type Paragraph
Add NLP
Relational Database
Get Fuzzy Relational Clustering, Hierarchical Fuzzy Relational Clustering and NLP Solution

Fig. 1 Proposed System Architecture

C. FRECCA

Andrew Skabar and Khaled Abdalgader proposed a novel fuzzy relational clustering algorithm called FRECCA [1]. A fuzzy relational clustering approach is used to cluster sentence level text. The clustering helps to indicate the strength of the association among the data objects and partition the data items into the clusters. The data points represent the membership values assigned for each and every cluster. Many existing clustering technique has difficulties in handling extreme outliers data objects but fuzzy clustering algorithm can handle these outliers and give them very small membership degree in surrounding clusters. FRECCA operates in three steps as:

Initialization: - : random initialization and normalization of cluster membership. Also the Mixing coefficient is initialized.

Expectation: **-** Page Rank values get calculated.

Maximization: - updating the mixing coefficients found in E step

DOI 10.17148/IJARCCE.2016.56138

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 6, June 2016

D. HFRECCA

Hierarchical clustering is used to partition the data into number of groups. Each group has a data that is similar in some sense to other [7]. An expectation–maximization (EM) algorithm is a repetitive process which depends on unobserved hidden variables. In the EM, the expectation (E) step creates a function to compute the cluster membership probabilities and maximization (M) step, in which these probabilities are then used to re-estimate the parameters. These parameter-estimates are then used to figure out the distribution of hidden variables in E step. This algorithm is divisive algorithms and begins with just only one cluster and then the single cluster splits into two or more clusters that have higher dissimilarity.

V. ALGORITHM

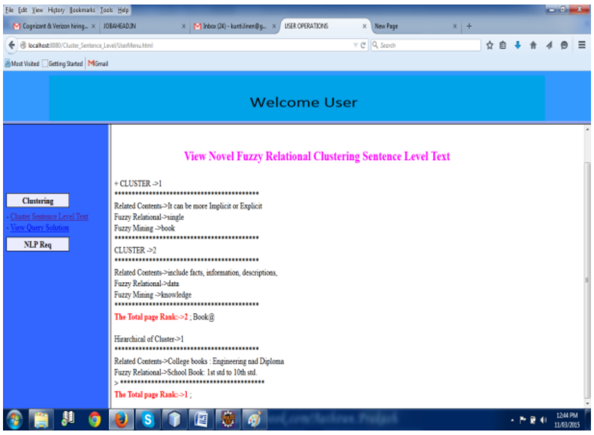
The algorithm has the following steps.
1. Initialization: random initialization and normalization of cluster membership. Also Mixing coefficient is initialized.
2. Expectation: Page Rank values get calculated
3. Maximization: updating the mixing coefficients found in E step

```
Algorithm 1. The FRECCA algorithm.
Input: Pairwise similarity values $S = \{s_{ij} | i = 1, \ldots, N,$
       $j = 1, \ldots, N\}$ where $s_{ij}$ is the similarity between
       sentences $i$ and $j$. Number of clusters, $C$.
Output: Cluster membership values $\{p_i^m | i = 1, \ldots,$
       $N, m = 1, \ldots, C\}$
// INITIALIZATION
// initialize and normalize membership values
for $i = 1$ to $N$
    for $m = 1$ to $C$
        $p_i^m = \mathrm{rnd}$            // random number on [0, 1]
    end for
    for $m = 1$ to $C$
        $p_i^m = p_i^m / \sum_{j=1}^{C} p_i^j$    // normalize
    end for
end for
for $m = 1$ to $C$
    $\pi_m = 1/C$                 // equal priors
end for
repeat until convergence
  // EXPECTATION STEP
  for $m = 1$ to $C$
    // create weighted affinity matrix for cluster m
    for $i = 1$ to $N$
      for $j = 1$ to $N$
        $w_{ij}^m = s_{ij} \times p_i^m \times p_j^m$
      end for
    end for
    // calculate PageRank scores for cluster m
    repeat until convergence
      $PR_i^m = (1 - d) + d \times \sum_{j=1}^{N} w_{ji}^m (PR_j^m / \sum_{k=1}^{N} w_{jk}^m)$
    end repeat
    // assign PageRank scores to likelihoods
    $l_i^m = PR_i^m$
  end for
  // calculate new cluster membership values
  for $i = 1$ to $N$
    for $m = 1$ to $C$
      $p_i^m = (\pi_m \times l_i^m) / \sum_{j=1}^{C} (\pi_j \times l_i^j)$
    end for
  end for
  // MAXIMIZATION STEP
  // Update mixing coefficients
  for $m = 1$ to $C$
    $\pi_m = \frac{1}{N} \sum_{i=1}^{N} p_i^m$
  end for
end repeat
```

Fig 2 FRECCA Algorithm

VI. RESULT ANALYSIS

Hierarchical fuzzy clustering is technique of cluster gathering. Each cluster is assigned with a membership value. The main drawback of FRECCA is time complexity. The work can be extended to cluster probabilistic data. The performance evaluation of the proposed HFRECCA clustering algorithm is based on certain performance metrics. The performance metrics used to measure results are Partition Entropy Coefficient (PE), Purity and Entropy, V-Measure, Rand Index and F-Measure.


File Edit View History Bookmarks Tools Help
Cognizant & Verizon hiring...
JOBAHEAD.IN
Inbox (24) - kurti.linen@g...
USER OPERATIONS
New Page
localhost:8080/Cluster_Sentence_Level/UserMenu.html
Search
Most Visited
Getting Started
Gmail
Clustering
- Cluster Sentence Level Text
- View Query Solution
NLP Req
12:44 PM
11/03/2015

Welcome User

View Novel Fuzzy Relational Clustering Sentence Level Text

+ CLUSTER ->1
****
Related Contents->It can be more Implicit or Explicit
Fuzzy Relational->single
Fuzzy Mining ->book
****
CLUSTER ->2
****
Related Contents->include facts, information, descriptions,
Fuzzy Relational->data
Fuzzy Mining ->knowledge
****
The Total page Rank:->2 ; Book@

Hirarchical of Cluster->1
****
Related Contents->College books : Engineering nad Diploma
Fuzzy Relational->School Book: 1st std to 10th std.
> ***
The Total page Rank:->1 ;

Fig 3 Hierarchical clusters created

This is how the hierarchy of clusters is being created after application of the proposed system.Also users can add Query if any data searched is unavailable.The admin can update the database so as to be able to make searchable the keyword.

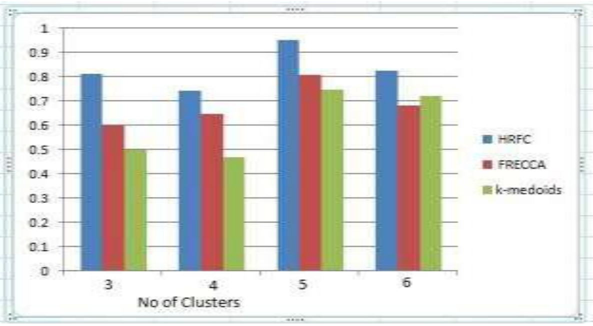


Fig 4 Comparison with ARCA

After comparing with ARCA, it is found that the performance of HFRECCA is superior. The dataset used can be dynamic ally added and used as per user convenience.

VII. CONCLUSION

Sentence clustering is one of best clustering technique for relational data inputs. Effectiveness of the algorithm is based on feature selection and accuracy of input. The algorithm can apply to asymmetric matrices and is not sensitive to cluster membership value initialization. The proposed HFRECC algorithm can be applicable to any

DOI 10.17148/IJARCCE.2016.56138

type of input relational data. It is used to analyse hierarchical relation between sentences. It can identify the overlapping clusters. Future objective is to extend the idea of probabilistic based fuzzy relational clustering algorithm.

ACKNOWLEDGEMENT

I would like to thank all people who help me in different way. Especially, I am thankful to my guide and P.G. coordinator **Prof. Amrit Priyadarshi** for his continuous support and guidance in my work. Also, I would like to thank H.O.D. of Computer Engineering Department, **Prof. Sayyad G. G.** for motivating me. Lastly, I thank to IJARCCE who have given opportunity to publish my paper.

REFERENCES

[1] A. Skabar and K. Abdalgader, “Clustering sentence-level text using a novel fuzzy relational clustering algorithm," Knowledge and Data Engineering, IEEE Transactions on, vol. 25, no. 1, pp. 62-75, 2013

[2] V. Hatzivassiloglou, J. L. Klavans, M. L. Holcombe, R. Barzilay, M.Y. Kan, and K. McKeown, “Simfinder: A flexible clustering tool for summarization." Proceedings of the NAACL Workshop on Automatic Summarization, 2001

[3] C.D. Manning, P. Raghavan, and H. Schutze, Introduction to Information Retrieval,Cambridge Univ. Press, 2008.

[4] Y. Li, D. McLean, Z.A. Bandar, J.D. O’Shea, and K. Crockett, “Sentence Similarity Based on Semantic Nets and Corpus Statistics,” IEEE Trans. Knowledge and Data Eng., vol. 8, no. 8, pp. 1138-1150, Aug. 2006.

[5] Huang, G. and Sheng, J. 2012. Measuring Similarity between Sentence Fragments, 4th International Conference on Intelligent Human-Machine Systems and Cybernetics, 978-0-7695-4721-3/12 $26.00 ©2012 IEEE.

[6] T. Zhang, Y. Y. Tang, B. Fang, and Y. Xiang, “Document clustering in correlation similarity measure space," Knowledge and Data Engineering, IEEE Transactions on, vol. 24, no. 6, pp. 1002-1013, 2012

[7] Buckley, P. J. 1995.”A Hierarchical Clustering Strategy for Very Large Fuzzy Databases”, 0-7803-2559-1/95 $4.00 0 IEEE.

[8] Shehata, S., Karray, F. and Mohamed, S. Kamel, “An Efficient Concept-Based Mining Model for Enhancing Text Clustering” IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 22, NO. 10, 2010.

[9] A.P. Dempster, N.M. Laird, and D.B. Rubin, “Maximum Likelihood from Incomplete Data via the EM Algorithm,” J. the Royal Statistical Soc. Series B (Methodological), vol. 39, no. 1, pp. 1-38, 1977.

DOI 10.17148/IJARCCE.2016.56138