Text Processing Using Fuzzy Relational Clustering

[1]J.Sakunthala Devi, [2]G. Umamaheswara Rao, [3]B. Kameswara Rao

[1,2,3]Dept. of CSE, Vizag Institute of Technlogy, Dakamarri, Visakhapatnam, AP, India

Abstract

Fuzzy clustering algorithms let configurations to belong to all clusters with different degrees of membership. A novel fuzzy clustering algorithm that works on relational input data; i.e., data in the arrangement of a square matrix of pairwise resemblances among data objects. The algorithm uses a graph representation of the data and functions in an Expectation-Maximization framework in which the graph centrality of an entity in the graph is interpreted as a possibility. Results of relating the algorithm to sentence clustering errands determine that the algorithm is capable of categorizing coinciding clusters of semantically related sentences and that it is consequently of latent use in a variety of text mining tasks.

Keywords

Hierarchical Fuzzy Relational Clustering, Fuzzification Degree, Hard Clustering, Soft Clustering

I. Introduction

The objective of text summarization is to extant the most significant information in a shorter form of the original text while keeping its foremost content and assistances the user to rapidly understand huge volumes of information. Text summarization statements together the problem of choosing the most significant units of text and the problem of generating coherent summaries. This procedure is suggestively different from that of human based text summarization since human can seizure and relate deep connotations and themes of text documents while automation of such an ability is very problematic to implement.

II. Related Work

Documents that are semantically connected are probable to comprise many words in common, and consequently are found to be alike according to popular vector space measures such as cosine resemblance, which are based on word co-occurrence. Though the assumption that semantic resemblance can be measured in terms of word co-occurrence may be valid at the document level, the statement does not embrace for small-sized text fragments such as sentences, since two sentences may be semantically associated despite having words in common. Somewhat than representing sentences in a common vector space, these measures describe sentence similarity as some function of word-to-word similarities, where these resemblances are in turn usually derived either from distributional information from some corpus-based measures or semantic information represented in external sources.

III. Existing Method

Clustering text at the document level is well recognized in the Information Retrieval (IR) literature, where documents are characteristically signified as data points in a high dimensional vector space in which each dimension resembles to a unique keyword, leading to a rectangular representation in which rows represent documents and columns represent attributes of those documents. The vector space model has been successful in IR as it is able to sufficiently capture much of the semantic content of document-level text.

IV. Disadvantages

The results frequently writhed from uncertainty in the optimization algorithms. A constraint of existing method is the high dimensionality hosted by representing objects in terms of their similarity with all other objects.

V. Proposed Method

The objective purpose of Fuzzy is to categorize a data point, cluster centroid has to be closest to the data point of membership for assessing the centroids, and typicality is used for easing the objectionable effect of outliers. The function is composed of two expressions, the first is the fuzzy function and uses a distance exponent and the second is prospect function and uses a typical fuzziness weighting exponent; but the two coefficients in the objective purpose are only used as exhibitor of membership and typicality. The aim is to determine nonlinear relationships among data, kernel methods use embedding mappings that map features of the data to new feature spaces.

VI. Advantages

Clever to attain superior performance when externally assessed in hard clustering mode on a challenging data set of famous quotations and applying the algorithm has demonstrated that it is capable of identifying overlapping clusters of semantically related sentences. Comparisons with the ARCA algorithm on each of these data sets suggest that FRECCA is proficient of identifying softer clusters than ARCA without sacrificing concert as assessed by external measures.

VII. System Architecture

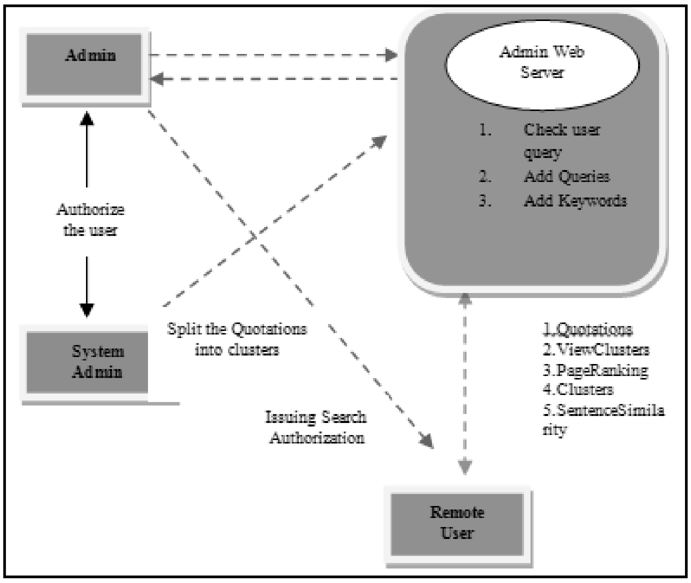


Fig. 1:

VIII. Modules

A. User Module

The user login and register for the definite query search, NLP Request and to cluster sentence level text using FRECCA algorithm.

IX. Input Dataset

The input dataset is taken from the previously extracted information that is presented in the paper itself. The dataset is the assortment of data. Most usually a dataset resembles to the contents of a single database table or a single statistical data matrix where every column of the table indicates a particular variable and each row resembles to a given member of the dataset in query.

X. Fuzzy Clustering

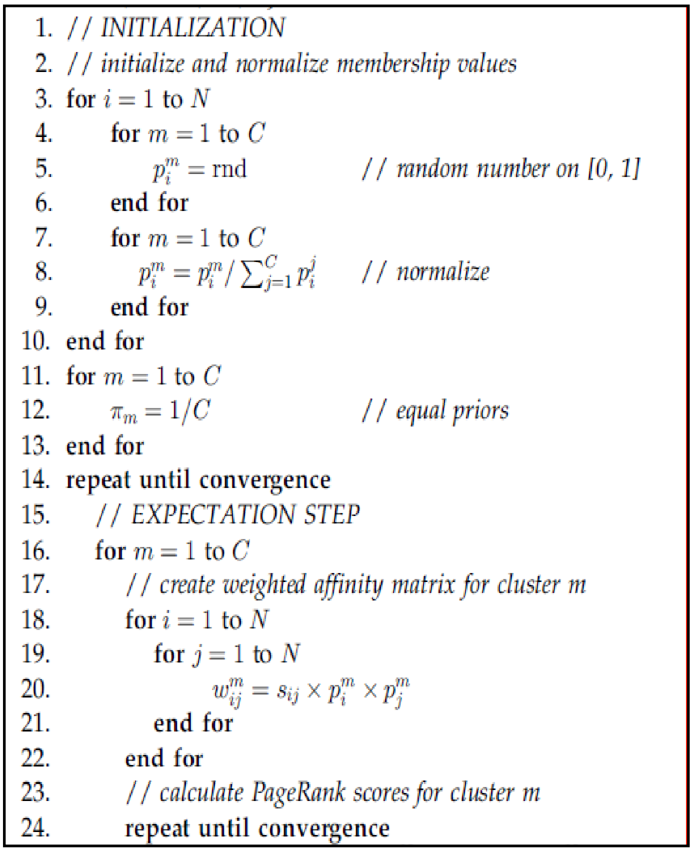
Clustering text at the document level is well established in the Information Retrieval (IR) literature. Documents are characteristically represented as data points in a high-dimensional vector space in which each dimension resembles to a unique keyword, leading to a rectangular representation in which rows represent documents and columns represent attributes of those documents.

XI. Page Rank

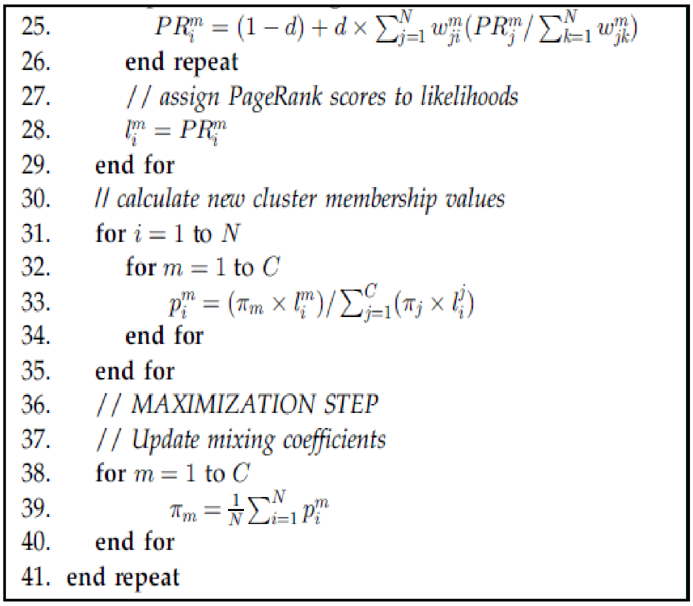
Applying Page Rank algorithm to each cluster and interpreting the Page-Rank score of an object within cluster as a possibility, we can then use the Expectation-Maximization (EM) framework to regulate the model parameters such as cluster membership values and mixing coefficients. The consequence is a fuzzy relational clustering algorithm which is generic in nature and can be functional to any domain in which the relationship between objects is expressed in terms of pair wise similarities.

XII. Algorithm Used

PageRank can be used within an Expectation-Maximization framework to build a complete relational fuzzy clustering algorithm. The final section deliberates issues relating to convergence, duplicate clusters, and various other implementation issues. Since PageRank centrality can be observed as a special case of eigenvector centrality, we name the algorithm Fuzzy Relational Eigenvector Centrality-based Clustering Algorithm (FRECCA).



```
// INITIALIZATION
// initialize and normalize membership values
for $i = 1$ to $N$
    for $m = 1$ to $C$
        $p_i^m = \text{rnd}$                    // 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
```

XIII. Conclusion

The method takes FRECCA as a reference to group the sentences, where the page rank value of a particular word or a synonym is taken into concern so as to find out the cluster name. The improvement has logins for Admin as well as particular users, where the users can request details about any kind of information. The users may send queries to the admin. The admin has privileges to enhance extra keywords and make the method dependable to solve for various sentences.

References

[1] V. Hatzivassiloglou, J.L. Klavans, M.L. Holcombe, R. Barzilay, M. Kan, K.R. McKeown,"SIMFINDER: A Flexible Clustering Tool for Summarization", Proc. NAACL Workshop Automatic Summarization, pp. 41-49, 2001.

[2] H. Zha,"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, 2002.

[3] D.R. Radev, H. Jing, M. Stys, D. Tam, "Centroid-Based Summarization of Multiple Documents", Information Processing and Management: An Int'l J., Vol. 40, pp. 919-938, 2004.

[4] R.M. Aliguyev,"A New Sentence Similarity Measure and Sentence Based Extractive Technique for Automatic Text Summarization", Expert Systems with Applications, Vol. 36, pp. 7764- 7772, 2009.

[5] R. Kosala, H. Blockeel,"Web Mining Research: A Survey", ACM SIGKDD Explorations Newsletter, Vol. 2, No. 1, pp. 1-15, 2000.

[6] G. Salton,"Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer", Addison-Wesley, 1989.

[7] J.B MacQueen,"Some Methods for Classification and Analysis of Multivariate Observations", Proc. Fifth Berkeley Symp. Math. Statistics and Probability, pp. 281-297, 1967.

[8] G. Ball, D. Hall,"A Clustering Technique for Summarizing Multivariate Data", Behavioural Science, Vol. 12, pp. 153-155, 1967.

[9] J.C. Dunn,"A Fuzzy Relative of the ISODATA Process and its Use in Detecting Compact Well-Separated Clusters", J.

Cybernetics, Vol. 3, No. 3, pp. 32-57, 1973.

[10] J.C. Bezdek,"Pattern Recognition with Fuzzy Objective Function Algorithms", Plenum Press, 1981.

[11] R.O. Duda, P.E. Hart, D.G. Stork,"Pattern Classification", second ed. John Wiley & Sons, 2001.

[12] U.V. Luxburg,"A Tutorial on Spectral Clustering", Statistics and Computing, Vol. 17, No. 4, pp. 395-416, 2007.

[13] B.J. Frey, D. Dueck,"Clustering by Passing Messages between Data Points", Science, Vol. 315, pp. 972-976, 2007.

[14] S. Theodoridis, K. Koutroumbas, Pattern Recognition", fourth ed. Academic Press, 2008.

[15] C.D. Manning, P. Raghavan, H. Schu¨ tze,"Introduction to Information Retrieval", Cambridge Univ. Press, 2008.

Mrs. J.Sakunthala Devi is a student of Vizag institute of Technology. Presently she is pursuing her M.Tech from this college and she received her B.Tech from JNTUK, Kakinada. Her area of interest includes Data Mining and Object oriented Programming languages, all current trends and techniques in Computer Science.

Mr. G. Umamaheswara Rao, M.Tech (CSE) from JNTU is working as Assistant Professor, Vizag Institute of technology. He has 4 years of teaching experience. His area of Interest includes Data Warehouse and Data Mining, information security, flavors of Unix Operating systems and other advances in computer.

Mr. B. kameshwara rao, M.Tech, (Ph.D) working as HOD (CSE), Vizag Institute of technology. He has 9 years of teaching experience.. His area of Interest includes Data Warehouse and Data Mining, information security, flavors of Unix Operating systems and other advances in computer.