ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 7, July 2015

Clustering Sentence-Level Text via a Novel Nebulous Relational Clustering Algorithm

G.Hemalatha [1], H.Lookman Sithic [2]

Research Scholar, Dept. of CS, Muthayammal College of Arts & Science, Rasipuram, Namakkal, India[1]

Associate Professor, Department of BCA, Muthayammal College of Arts & Science, Rasipuram, Namakkal, India[2]

ABSTRACT: In comparison with hard clustering methods, in which a pattern belongs to a single cluster, fuzzy clustering algorithms allow patterns to belong to all clusters with differing degrees of membership. This work presents a novel fuzzy clustering algorithm that operates on relational input data; i.e., data in the form of a square matrix of pairwise similarities between data objects. The algorithm uses a graph representation of the data, and operates in an Expectation-Maximization framework in which the graph centrality of an object in the graph is interpreted as a likelihood. Results of applying the algorithm to sentence clustering tasks demonstrate that the algorithm is capable of identifying overlapping clusters of semantically related sentences, and that it is therefore of potential use in a variety of text mining tasks. We also include results of applying the algorithm to benchmark data sets in several other domains.

KEYWORDS: Fuzzy relational clustering, natural language processing, graph centrality.

I. INTRODUCTION

Sentence clustering plays an important role in many text processing activities. For example, various authors have argued that incorporating sentence clustering into extractive multi document summarization helps avoid problems of content overlap, leading to better coverage However, sentence clustering can also be used within more general text mining tasks. For example, consider web mining , where the specific objective might be to discover some novel information from a set of documents initially retrieved in response to some query. By clustering the sentences of those documents we would intuitively expect at least one of the clusters to be closely related to the concepts described by the query terms; however, other clusters may contain information pertaining to the query in some way hitherto unknown to us, and in such a case we would have successfully mined new information. Irrespective of the specific task (e.g., summarization, text mining, etc.), most documents will contain interrelated topics or themes, and many sentences will be related to some degree to a number of these. The work described in this paper is motivated by the belief that successfully being able to capture such fuzzy relationships will lead to an increase in the breadth and scope of problems to which sentence clustering can be applied. However, clustering text at the sentence level poses specific challenges not present when clustering larger segments of text, such as documents. We now highlight some important differences between clustering at these two levels, and examine some existing approaches to fuzzy clustering. Clustering text at the document level is well established in the Information Retrieval (IR) literature, where documents are typically represented as data points in a high dimensional vector space in which each dimension corresponds to a unique keyword , leading to a rectangular representation in which rows represent documents and columns represent attributes of those documents (e.g., tf-idf values of the keywords). This type of data, which we refer to as "attribute data," is amenable to clustering by a large range of algorithms. Since data points lie in a metric space, we can readily apply prototype-based algorithms such as k-Means , Isodata , Fuzzy c-Means (FCM) ,and the closely related mixture model approach , all of which represent clusters in terms of parameters such as means and co-variances, and therefore assume a common metric input space. Since pairwise similarities or dissimilarities between data points can readily be calculated from the attribute data using similarity measures such as cosine similarity, we can also apply relational clustering algorithms such as Spectral Clustering and Affinity Propagation, which take input data in the form of a square matrix (often referred to as the "affinity matrix"), where wij is the (pairwise) relationship between the ith and jth data object.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 7, July 2015

II. LITERATURE SURVEY

A novel method for simultaneous key phrase extraction and generic text summarization is proposed by modeling text documents as weighted undirected and weighted bipartite graphs. Spectral graph clustering algorithms are used for partitioning sentences of the documents into topical groups with sentence link priors being exploited to enhance clustering quality. Within each topical group, saliency scores for key phrases and sentences are generated based on a mutual reinforcement principle. The key phrases and sentences are then ranked according to their saliency scores and selected for inclusion in the top key phrase list and summaries of the document. The idea of building a hierarchy of summaries for documents capturing different levels of granularity is also briefly discussed. Our method is illustrated using several examples from news articles, news broadcast transcripts and web documents. Partitioning a large set of objects into homogeneous clusters is a fundamental operation in data mining. The *k*-means algorithm is best suited for implementing this operation because of its efficiency in clustering large data sets. However, working only on numeric values limits its use in data mining because data sets in data mining often contain categorical values. In this paper we present an algorithm, called *k*-modes, to extend the *k*-means paradigm to categorical domains. We introduce new dissimilarity measures to deal with categorical objects, replace means of clusters with modes, and use a frequency based method to update modes in the clustering process to minimise the clustering cost function. Tested with the well known soybean disease data set the algorithm has demonstrated a very good classification performance. Experiments on a very large health insurance data set consisting of half a million records and 34 categorical attributes show that the algorithm is scalable in terms of both the number of clusters and the number of records We present a statistical similarity measuring and clustering tool, SIMFINDER, that organizes small pieces of text from one or multiple documents into tight clusters. By placing highly related text units in the same cluster, SIMFINDER enables a subsequent content selection/generation component to reduce each cluster to a single sentence, either by extraction or by reformulation. We report on improvements in the similarity and clustering components of SIMFINDER, including a quantitative evaluation, and establish the generality of the approach by interfacing SIMFINDER to two very different summarization systems.

III. EXISTING SYSTEM

The vector space model has been successful in IR because it is able to adequately capture much of the semantic content of document-level text. This is because documents that are semantically related are likely to contain many words in common, and thus are found to be similar according to popular vector space measures such as cosine similarity, which are based on word co-occurrence. However, while the assumption that (semantic) similarity can be measured in terms of word co-occurrence may be valid at level, the assumption does not hold for small-sized text fragments such as sentences, since two sentences may be semantically related despite having few, if any, words in common. A limitation of this approach is the high dimensionality introduced by representing objects in terms of their similarity with all other objects.

Disadvantages of existing system:

The major disadvantage of the algorithm is its time complexity. Despite its success, the Euclidean requirement in RFCM was considered restrictive, and various alternatives have been proposed.

IV. PROPOSED SYSTEM

We first describe the use of Page Rank as a general graph centrality measure, and review the Gaussian mixture model approach. We then describe how Page Rank can be used within an Expectation-Maximization framework to construct a complete relational fuzzy clustering algorithm. The final section discusses issues relating to convergence, duplicate clusters, and various other implementation issues. Since Page Rank centrality can be viewed as a special case of eigenvector centrality, we name the algorithm Fuzzy Relational Eigenvector Centrality-based Clustering Algorithm (FRECCA).

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 7, July 2015

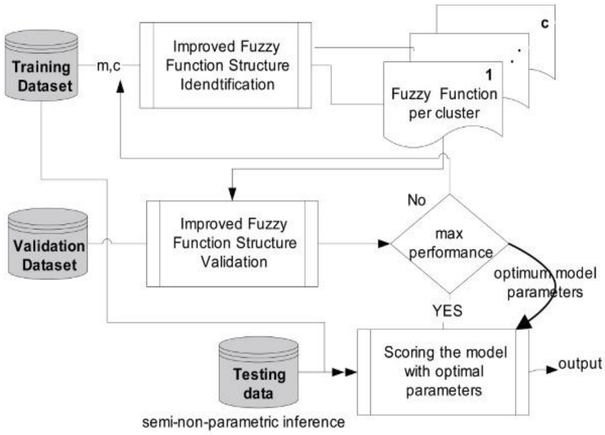

Fig 1.fuzzy logic system architecture

ADVANTAGES OF PROPOSED SYSTEM:

Able to achieve superior performance to benchmark Spectral Clustering and k-Medoids algorithms when externally evaluated in hard clustering mode on a challenging data set of famous quotations.

V. FUZZY RELATIONAL CLUSTERING

Fuzzy Relational Clustering Unlike Gaussian mixture models, which use a likelihood function parameterized by the means and co-variances of the mixture components, the proposed algorithm uses the PageRank score of an object within a cluster as a measure of its centrality to that cluster. These PageRank values are then treated as likelihoods. Since there is no parameterized likelihood function as such, the only parameters that need to be determined are the cluster membership values and mixing coefficients. The algorithm uses Expectation Maximization to optimize these parameters. We assume in the following that the similarities between objects are stored in a similarity matrix S ¼ fsijg, where sij is the similarity between objects i and j. Initialization, we assume here that cluster membership values are initialized randomly, and normalized such that cluster membership for an object sums to unity over all clusters. Mixing coefficients are initialized such that priors for all clusters are equal.

The E-step calculates the PageRank value for each object in each cluster. PageRank values for each cluster are calculated as described in , with the affinity matrix weights wij obtained by scaling the similarities by their cluster membership values; i.e., wm ij ¼ sij * pmi * pmj ; where wm ij is the weight between objects i and j in cluster m, sij is the similarity between objects i and j, and pmi and pmj 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 object depends not only on its similarity to that other object, but also on its degree of membership to the cluster. Likewise,an object's entitlement to receive a contribution depends on its membership to the cluster. Once PageRank scores have been determined, these are treated as likelihoods and used to calculate cluster membership values. Since there is no parameterized likelihood function as such, the only parameters that need to be determined are the cluster membership values and mixing coefficients. The algorithm uses Expectation Maximization to optimize these parameters.

Maximization step.

Since there is no parameterized likelihood function, the maximization step involves only the single step of updating the mixing coefficients based on membership values calculated in the Expectation Step.

The pseudocode is presented in Algorithm 1, where wm ij , sij, pmi , and pmj are defined as above, m is the mixing coefficient for cluster m, PRmi is the PageRank score of object i in cluster m, and lmi is the likelihood of object i in cluster m.

VI. ALGORITHM IN FRECCA SYSTEM

Graph-Based Centrality and PageRank:

The basic idea behind the PageRank algorithm is that the importance of a node within a graph can be determined by taking into account global information recursively computed from the entire graph, with connections to high-scoring nodes contributing more to the score of a node than connections to low-scoring nodes. It is this importance that can then be used as a measure of centrality. In both TextRank and LexRank, each sentence in a document or documents is represented by a node on a graph. However, unlike a web graph, in which edges are Unweighted, edges on a document graph are weighted with a value representing the similarity between sentences. The PageRank algorithm can easily be modified to deal with weighted undirected edges.

Mixture Models and the EM Algorithm

The algorithm we present is motivated by the mixture model approach, in which a density is modeled as a linear combination of C component densities in the form , where are called mixing coefficients, and represent the prior probability of data point x having been generated from component m of the mixture. Assuming that the parameters of each component are represented by a parameter vector _m, the problem is to determine the values of the components of this vector, and this can be achieved using the Expectation-Maximization algorithm. Following random initialization of the parameter vectors _m, m { 1; . . . ; C, an Expectation step (E-step), followed by a Maximization step (M-step), are iterated until convergence. The E-step computes the cluster membership probabilities.

Fuzzy Relational Clustering

Unlike Gaussian mixture models, which use a likelihood function parameterized by the means and co-variances of the mixture components, the proposed algorithm uses the PageRank score of an object within a cluster as a measure of its centrality to that cluster. These PageRank values are then treated as likelihoods. Since there is no parameterized likelihood function as such, the only parameters that need to be determined are the cluster membership values and mixing coefficients. The algorithm uses Expectation Maximization to optimize these parameters.

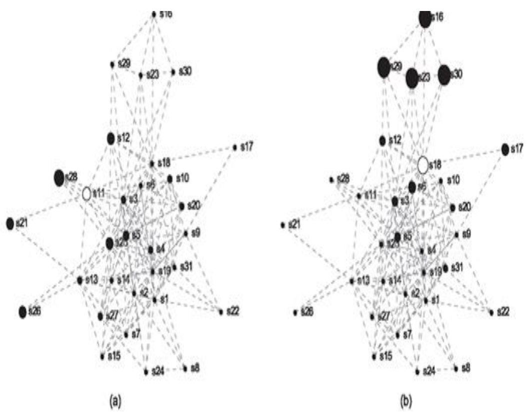

Fig 3.Clustering Sentence using Fuzzy Relational.

VII. COMPUTATION

Algorithm

// INITIALIZATION
// initialize and normalize membership values
 for i = 1 to N
 for m = 1 to C

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 7, July 2015

```
        pmi=rnd           // random number on [0, 1]
      end for
      for m = 1 to C
pmi= pmi/€ pji          // normalize
      end for
   end for
 for m = 1 to C
π 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
                 wmij =sij * pmi* pmj
             end for
         end for
 // calculate PageRank scores for cluster m
repeat until convergence
      PRmi=(1 _ d) +d * PNj¼1 wmji              (PRmj=PNk/1 wjk)
       end repeat
 // assign PageRank scores to likelihoods
          Lmi  = PRmi
      end for
 // calculate new cluster membership values
 for i = 1 to N
         for m = 1 to C
                 pmi= (πm * lmi)/€j=1(πj *Pji)
          end for
end for
 // MAXIMIZATION STEP
// Update mixing coefficients
         for m = 1 to C
           πm = 1/NPNi=1 pmi
         end for
 end repeat
```

VIII. RESULT ANALYSIS

The results show that the best average precision, recall and f-measure to summaries produced by the fuzzy method. Certainly, the experimental result is based on fuzzy logic could improve the quality of summary results that based on the general statistic method. In conclusion, we will extend the proposed method using combination of fuzzy logic and other learning methods and extract the other features could provide the sentences more important.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 7, July 2015

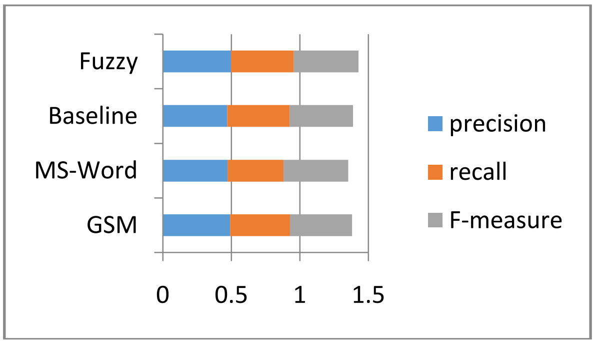

Fig 4.Performance graph

IX. CONCLUSION AND FUTURE WORK

The FRECCA algorithm was motivated by our interest in fuzzy clustering of sentence-level text, and the need for an algorithm which can accomplish this task based on relational input data. The results we have presented show that the algorithm is able to achieve superior performance to benchmark Spectral Clustering and k-Medoids algorithms when externally evaluated in hard clustering mode on a challenging data set of famous quotations, and applying the algorithm to a recent news article has demonstrated that the algorithm 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 capable of identifying softer clusters than ARCA, without sacrificing performance as evaluated external measures. Our main future objective is to extend these ideas to the development of a hierarchical fuzzy relational clustering algorithm.

REFERENCES

1. V. Hatzivassiloglou, J.L. Klavans, M.L. Holcombe, R. Barzilay, M. Kan, and 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, and 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 and 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 and 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.ybernetics, 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, and 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 and D. Dueck, "Clustering by Passing Messages between Data Points," Science, vol. 315, pp. 972-976, 2007.
14. S. Theodoridis and K. Koutroumbas, Pattern Recognition, fourth ed. Academic Press, 2008.
15. C.D. Manning, P. Raghavan, and H. Schu¨ tze, Introduction to Information Retrieval. Cambridge Univ. Press, 2008.
16. 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.
17. R. Mihalcea, C. Corley, and C. Strapparava, "Corpus-Based and Knowledge-Based Measures of Text Semantic Similarity," Proc. 21st Nat'l Conf. Artificial Intelligence, pp. 775-780, 2006.
18. D. Wang, T. Li, S. Zhu, and C. Ding, "Multi-Document Summarization via Sentence-Level Semantic Analysis and Symmetric Matrix Factorization," Proc. 31st Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 307-314, 2008.
19. C. Fellbaum, WordNet: An Electronic Lexical Database. MIT Press, 1998.
20. E.H. Ruspini, "A New Approach to Clustering," Information Control, vol. 15, pp. 22-32, 1969.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 7, July 2015

BIOGRAPHY

G.HEMALATHA was born on 02.05.1990 in Tamilnadu, India. She received BCA 2010 degree from vivekanandha college of arts science for women, elayampalayam, Affiliated to periyar university,salem, Tamilnadu, India. she received Master of computer science 2012 degree from Trinity college for women, Namakkal, Affiliated to periyar university-Salem,Tamilnadu, India. She is pursuing M.phil (full time) degree from Muthayammal college of Arts & Science, in Periyar University, Salem, Tamilnadu, India. She interested area is DATAMINING.

Mr.H.LOOKMAN SITHIC M.S(IT).,M.Phil.,[Ph.D]., He received him MS(IT) degree from Jamal Mohamed College, Bharathidasan university and M.Phil(c.s)degree from Periyar University, Salem. He is having 14 years of Experience in Collegiate Teaching and He is the Associate professor in depart of BCA in Muthayammal College of Arts and Science, Rasipuram, Affiliated by Periyar University, Salem, Tamilnadu, India. His main research interested include Data Mining.