ISSN XXXX XXXX © 2017 IJESC

Research Article Volume 7 Issue No.12

Semi Supervised and Unsupervised constraints Using Novel Fuzzy Relational Clustering Algorithm

Kirthika .K. M[1], Meenachi V.R[2], Hebziba Jeba Rani[3], Barani .G[4], Suganya .V[5]
Assistant Professor[1, 2, 3, 4, 5]
Department of Computer Science & Engineering
Sri Ramakrishna Institute of Technology, Coimbatore,Tamil Nadu, India

Abstract:
The proposed novel constrained co clustering method to achieve two goals. First, the combination of both information theoretic co-clustering and constrained clustering to improve clustering performance. Thus saving the effort and cost of using manually labeled constraints.Second ,the development of a two-sided hidden Markov random field (HMRF) model to represent both document and word constraints, that use an alternating Expectation-Maximization (EM) algorithm to optimize the model.A novel Fuzzy Clustering Algorithm that operates on relational input data, data in the form of a square matrix of pairwise similarities between data objects. A graph representation of the data,operates in an Expectation-Maximization framework in which the graph centrality of an object in the graph is interpreted. Result of applying the algorithm to sentence clustering tasks demonstrate that the algorithm is capable of identifying overlapping clusters semantically , and that it is therefore of potential use in a verify of text mining tasks.

Keywords: HMRF-hidden Markov random field, EM -Expectation Maximization,Fuzzy Clustering Algorithm.

1. INTRODUCTION

Data mining, the extraction of the hidden predictive information from the large database is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools to predict the trends of the business for the knowledge- driven decisions. It is mainly used to discover new correlations, patterns and trends by shifting large amounts of data stored in repositories. Its techniques are designed to interpret data to provide additional information to assist in understanding the data which, in turn, can facilitate handling and estimation.clustering is a popular technique for automaticallyorganizing or summarizing a large collection of text; there have been many approaches to clustering

a. Classification

Classification is a collection consists of dividing the items that make up the collection in to the categories or classes. The goal of the predictive classification isto accurately predict the target class for each record in new data that is not in the historical data.

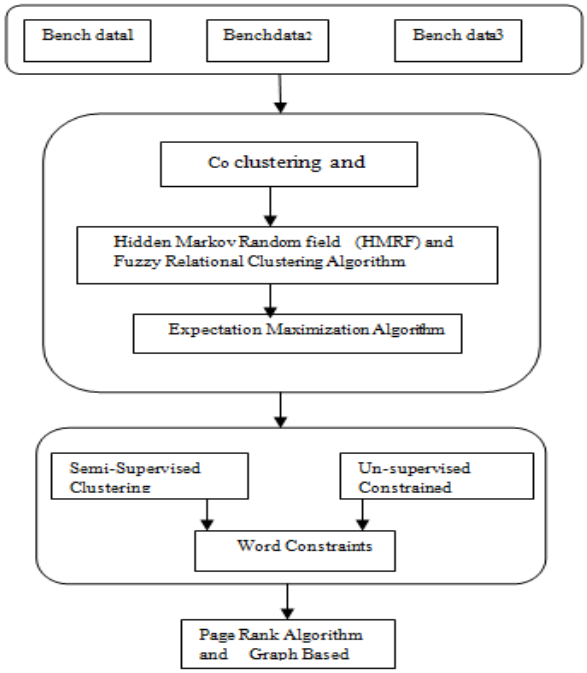

Figure.1. Architecture diagram for Semi-supervised Clustering and Un-supervised Constrained

b.Regression

Regression is similar to the classification models. The major difference between them is regression is that the numerical or the continuous target attributes and for classification is that the discrete or categorical target attributes. If the target attribute contains continuous values or integers which is of inherent order, a regression technique can be used. If the target attribute contains categorical value, that is string or integer values where order has no significance. Continuous target can be turned in to a discrete target by binning, in which the regression problem is being solved using the classification algorithm.

c.Clustering

Clustering aims to partition a set of records into several groups such that similar records are in the same group according to some similarity function, identifying similar sub populations in the data.

II. METHODOLOGIES

d.Co clustering

Most co clustering algorithms deal with dyadic data, for example the document and word co-occurrence frequencies.The dyadic data can be modeled as a bipartite graph, and then spectral graph theory is adopted to solve the partition problem.The co-occurrence can also be encoded in co-occurrence matrices and the matrix, factorizations are utilized to solve the clustering problem.The document and word co-occurrence can also be formulated as a two-sided generative model using a Bayesian interpretation.

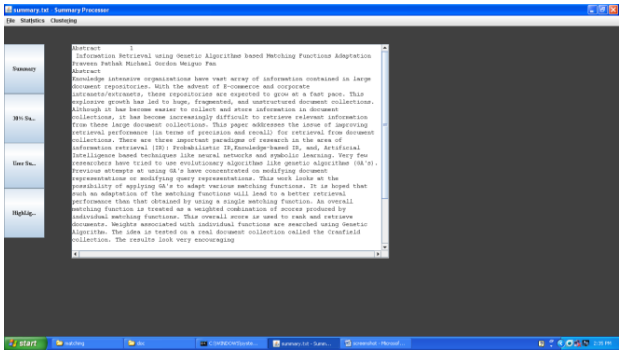

Figure.2. Summary is Choosen for Clustering

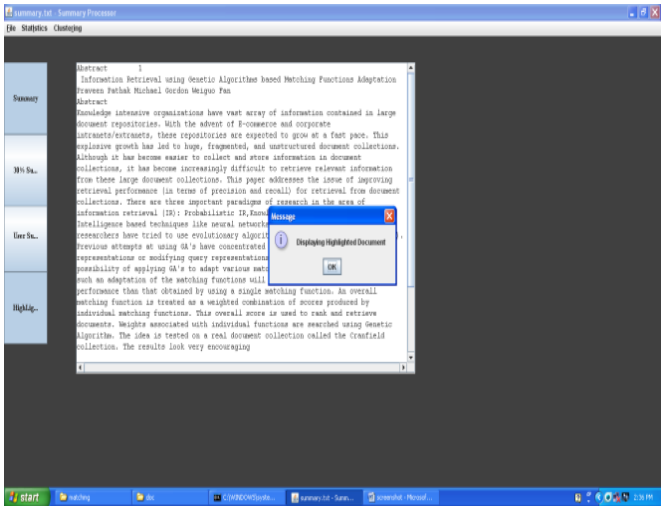

Figure.3. Diplaying The Highlighted Document

e.Semi-Supervised Clustering

Semi-supervised clustering with labeled seeking points and semi-supervised clustering with labeled constraints. Constraints –based clustering methods often use pair wise constraints.In co clustering ,for text data, co clustering can not only show the relationship between document and word cluster, but also leverage the knowledge transferred between the two sides. There are some initial efforts on extending the existing co clustering methods to semi-supervised co clustering and constrained co clustering. Most of the these methods are based on matrix factorizations that optimize a sum squared residues-based objective function.

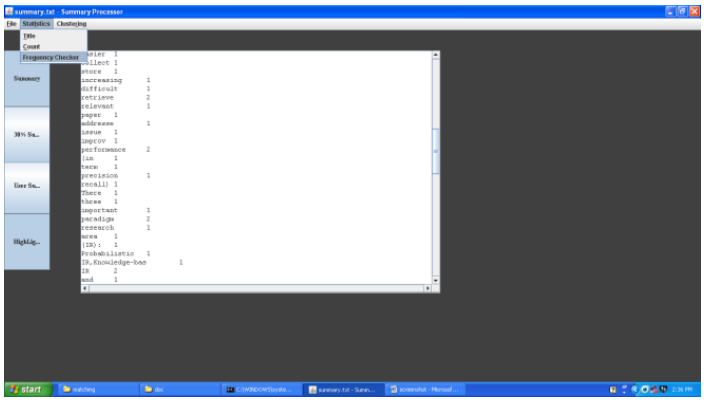

Figure.3. Frequency Check

f.Unsupervised Constrained Clustering

Some research has been conducted to handle constraints automatically derived based on either human provided meta data or existing knowledge sources .Demonstrated that the ACM keyword taxonomy can help cluster scientific papers using a non-negative matrix factorization approach. it constructs the word constraints based on the word categories learned from the auxiliary corpus. It is added with the must-link for document when two documents have many overlapped NEs. While for word constraints we add must-links if the two words are close to each other semantically, which is measured by a Word Net-based semantic similarity.

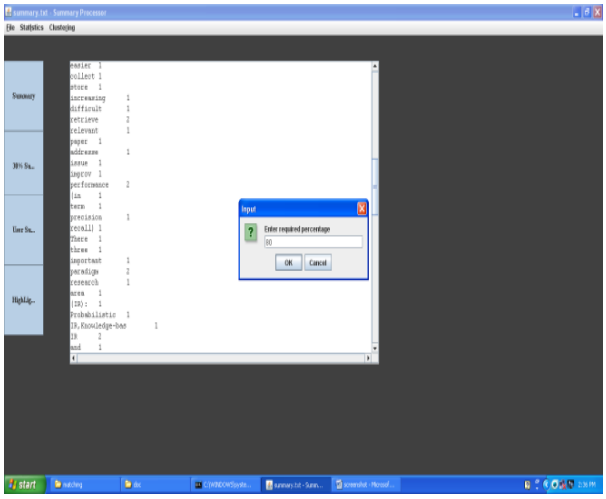

Figure.4. Choosing Required Percentage of Summary

g.Fuzzy Relational Clustering

The proposed algorithm uses the Page Rank score of an object within a cluster as a measure of its centrality to the cluster. The

Page Rank values are then treated as likelihoods, Since there is no parameters that need to be determined are the cluster membership values and mixing coefficients.

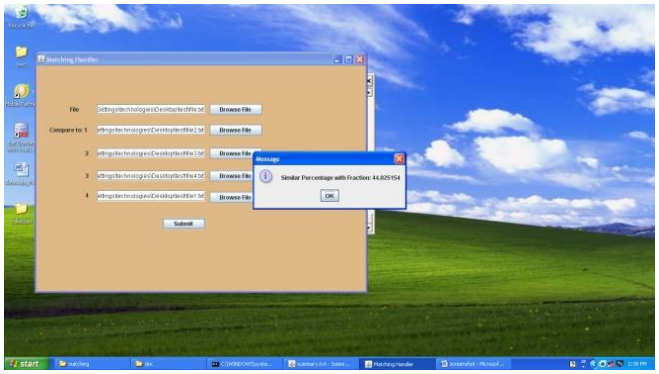

Figure.5. Clustering Analysis

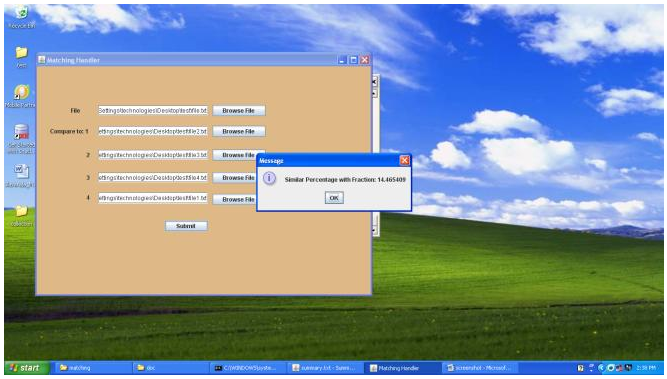

Figure.6. Analysis Of Clustering File Fraction

h. Graph-Based Centrality and Page Rank

The page Rank algorithm is important 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.

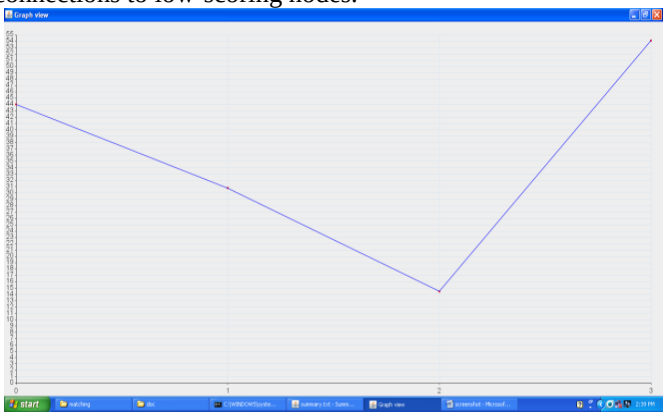

Figure.7. Gragh Based Centrality And Page Rank

III CONCLUSION AND FUTURE WORK:

The demonstration to construct various document and word constraints and apply them to the constrained coclustering approach that automatically incorporates various word and document constraints into information the or it ccoculstering.The evaluations on two benchmark data sets demonstrated the effectiveness of the proposed method for clustering textual documents. The algorithm consistently outperformed all the tested constrained clustering and coclustering methods under different conditions. The investigation of unsupervised constraints is still preliminary. Further investigate whether better text feature that can be automatically derived by using natural language processing or information extraction tools.Also interested in appling CITCC to other text analysis application such as visual text summarizations.

IV. REFERENCES

[1]. A. Jain, M. Murty, and P. Flynn, "Data Clustering: A Review," ACM Computing Surveys, vol. 31, no. 3, pp. 264-323, 1999.

[2]. Y. Cheng and G.M. Church, "Biclustering of Expression Data," Proc. Int'l System for Molecular Biology Conf. (ISMB), pp. 93-103,2000.

[3]. I.S. Dhillon, "Co-Clustering Documents and Words Using Bipartite Spectral Graph Partitioning," Proc. Seventh ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD),pp. 269-274, 2001.

[4]. I.S. Dhillon, S. Mallela, and D.S. Modha, "Information-TheoreticCo-Clustering," Proc. Ninth ACM SIGKDD Int'l Conf. KnowledgeDiscovery and Data Mining (KDD), pp. 89-98, 2003.

[5]. H. Cho, I.S. Dhillon, Y. Guan, and S. Sra, "Minimum Sum-SquaredResidue Co-Clustering of Gene Expression Data," Proc. FourthSIAM Int'l Conf. Data.Mining (SDM), 2004.

[6]. Semi-Supervised Learning, O. Chapelle, B. Scho¨lkopf, and A. Zien,eds. MIT Press, http://www.kyb.tuebingen.mpg.de/ssl-book,2006.

[7]. S. Basu, I. Davidson, and K. Wagstaff, Constrained Clustering:Advances in Algorithms, Theory, and Applications. Chapman & Hall/CRC, 2008.

[8]. R.G. Pensa and J.-F.Boulicaut, "Constrained Co-Clustering of Gene Expression Data," Proc. SIAM Int'l Conf. Data Mining (SDM),pp. 25-36, 2008.

[9]. F. Wang, T. Li, and C. Zhang, "Semi-Supervised Clustering via Matrix Factorization," Proc. SIAM Int'l Conf. Data. Mining (SDM),pp. 1-12, 2008.

[10]. Y. Chen, L. Wang, and M. Dong, "Non-Negative Matrix Factorization for Semi-Supervised Heterogeneous Data Co-Clustering," IEEE Trans. Knowledge and Data Eng., vol. 22, no. 10,pp. 1459-1474, Oct. 2010.

[11]. A. Banerjee, I. Dhillon, J. Ghosh, S. Merugu, and D.S. Modha, "AGeneralized Maximum Entropy Approach to Bregman Co-Clustering and Matrix Approximation," J. Machine Learning Research, vol. 8, pp. 1919-1986, 2007.

[12]. Y. Song, S. Pan, S. Liu, F. Wei, M.X. Zhou, and W. Qian, "Constrained Co-Clustering for Textual Documents," Proc. Conf. Artificial Intelligence (AAAI), 2010.

[13]. C. Ding, T. Li, W. Peng, and H. Park, "Orthogonal NonnegativeMatrix T-Factorizations for Clustering," Proc. 12th ACM SIGKDDInt'l Conf. Knowledge Discovery and Data Mining, pp. 126-135, 2006.

[14]. H. Shan and A. Banerjee, "Bayesian Co-Clustering," Proc. IEEEEight Int'l Conf. Data Mining (ICDM), pp. 530-539, 2008.

[15]. P. Wang, C. Domeniconi, and K.B. Laskey, "Latent Dirichlet Bayesian Co-Clustering," Proc. European Conf. Machine Learning andKnowledge Discovery in Databases (ECML/PKDD), pp. 522-537, 2009.