Volume 5, Issue 8, August 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

A Fuzzy Relational Clustering Algorithm

Kirti M. Patil*
Department of Computer Engineering,
ARMIET, Shahapur, Thane (E).India

Dr. Jagdish. W. Bakal
Department of Computer Engineering,
SSJCOE, Dombivali, Thane (E), India

Abstract- Cluster analysis groups objects into a set. The groups are form on the basis of similarities and dissimilarities of objects. Sentence clustering is used in multi-document summarization or text mining. Sentence clustering helps to avoid content overlapping problem. In sentence clustering domains, a sentence is related to more than one theme in document or set of documents. Due to this, the proposed system will be captures fuzzy relationships to increase the scope of problems. In this system, the PageRank and EM algorithm used for sentence level text clustering.

Keywords: Clustering, Partitional, Hierarchical clustering, PageRank, EM algorithm.

I. INTRODUCTION

Clustering is the process which groups a set of objects in such a way that objects in the same group are more similar to each other than the objects in the other group. Clustering is the unsupervised pattern into clusters. Clustering is a set of clusters which contains all objects in the data set. It specifies the relationship of the clusters to other clusters. Clustering can be classified as hard clustering and soft clustering.

II. CLUSTERING ALGORITHMS

Clustering can be divided into various types of clustering

Partitional Clustering algorithm is a set of data objects groups such that the data object is in exactly one cluster. Each cluster may be represented by centroid.

Hierarchical Clustering is the type of cluster analysis. This method seeksto form a hierarchy of clusters.Hierarchical clustering algorithms were developed for to overcome the disadvantages of partitional clustering algorithms. Hierarchical clustering is divided as:

Agglomerative: The working of this method is based upon the grouping of the data which is depend upon the nearest distance measure of all pairwise distance between the data point. Agglomerative clustering is the bottom up Approach.

Divisive:- It is top-down approach. This method starts with one object and then split groups into smaller cluster until every object all in one cluster. At every step divisive method divides the data objects in disjoint cluster and follows the pattern until the data objects fall into separate cluster.

III. PROPOSED WORK

A. PageRank Algorithm

PageRank algorithm is a graph centrality based algorithm. A graph-centrality algorithm is used to find the importance of node in a graph. This is determined by the relation between the nodes. PageRank is an algorithm which measures the importance of website pages.

PageRank assigns a numerical weight to each element or sentence of a set of documents which are hyperlinked and measures its relative importance in the sentences. The PageRank score is used to measure the centrality of that cluster.

In this algorithm, the rank value or rank score shows an importance of a sentence in document. The PageRank score assigns a numerical value between 0 and 1, and defined as:

$$PR(V_i) = (1-d) + d \times \sum_{j \in In(V_i)} \frac{1}{|Out(V_j)|} PR(V_j)$$

Where, In ($V_{i)}$ is the set of vertices of graph

$V_{j\ is}$ the set of vertices pointed by V_j

d is the damping factor which set around 0.8 to 0.9

The PageRank algorithm can modified for weighted undirected edges:

$$PR(V_i) = (1-d) + d \times \sum_{j=1}^{N} \left(w_{ji} \frac{PR(V_j)}{\sum_{k=1}^{N} w_{jk}} \right)$$

Where, w_{ji} is the similarity between the Vi and V_j.

Here, PageRank algorithm assigns the rank value or score to the fetched all url's from Google search page after the query processing.

© 2015, IJARCSSE All Rights Reserved

pageid	urlsequence
0	http://archive.financialexpress.com/news/the-new-n...
1	http://scroll.in/article/747708/the-daily-fix-sett...
2	https://www.facebook.com/narendramodi&sa=U&amp...
3	http://timesofindia.indiatimes.com/topic/Narendra-...

Fig.1 URL's fetched by search page with Rank Value

B. Expectation-Maximization Algorithm

Expectation-Maximization (EM) is a distance based algorithm. EM algorithm is same as k-means algorithm; the difference is the membership degree. EM algorithm computes the probabilities of cluster membership which is based on one or more probability distribution.

This algorithm is used to find the parameters of mixture of Gaussian. Each iteration consists of two steps that are E-step and M-step

E-Step: - PageRank value is calculated by E-step for every object of each cluster.

M-Step: -This step updates the mixing coefficient based on the membership values which are calculated by E-Step.

The EM algorithm shows clusters patterns which are generated when the urls are extracted from the urls of Google search page for the query.

C. A Fuzzy Relational Clustering Algorithm (FRECCA)

A fuzzy relational clustering approach is used to produce clusters with sentences, where each of them corresponds to some content. The output of clustering indicates the strength of the association among the data elements. This algorithm that is a novel fuzzy relational clustering algorithm (FRECCA) is proposed by Andrew Skabar and Khaled Abdalgar [1]. This algorithm is divided into three steps: Initialization, Expectation and Maximization.

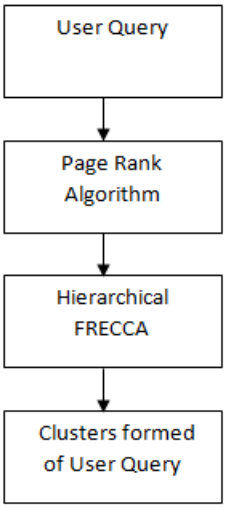

Fig. 2. FRECCA Clustering Process

IV. SOFTWARE & HARDWARE REQUIREMENT

Software Requirements:

- Operating System: Windows 7.
- Language : ASP.NET/C#
- Front End : Microsoft Visual Studio 2010, .Net Framework
- Databases Used: My SQL

Hardware requirements:

- Processor: Intel Core i3-370M Processor 2.40GHz. (& onwards).
- Memory (RAM) : 1GB RAM(32 bit)
- Hard disk : 40GB
- Internet access

V. CONCLUSION

Hierarchical fuzzy clustering algorithm plays an important role in the sentence level text clustering. This algorithm works on the relational data which is in the form of relational matrix. This proposed algorithm uses the PageRank and EM algorithm for the sentence text clustering.

The PageRank and EM algorithms are used for ranking of the urls fetched by the Google. EM algorithm which works in two steps i.e. E-step and M-step are creates the clusters patterns with their probabilities and re-estimates the parameters.

© 2015, IJARCSSE All Rights Reserved

ACKNOWLEDGEMENT

We thank to all the authors for the information provided.

REFERENCES

[1] Andrew Skabar and Khaled Abdalgader, "Clustering Sentence Level Text Using A Novel Fuzzy Clustering Algorithm". January 2013, IEEE Transactions on Knowledge and Data Engineering, Vol. 25, no. 1, pp 62-75.

[2] Yuhua Li,David Mclean, ZuhairBandar,James D. Oshea & Keeley Crokett, "Sentence Similarity Based on Semantic Nets and Corpus Statistics",Aug. 2006 IEEE Trans. Knowledge and Data Eng vol. 8, no. 8 pp. 1138-1150.

[3] P. Corsini ,B. Lazzerini,F.Marcelloni.,"A New Fuzzy Relational Clustering Algorithm Based On The Fuzzy C-Means Algorithm", 24 ,April 2004,Soft Computing, pp-439-447.

[4] G. Erkan and D.R. Radev, "LexRank: Graph-Based Lexical Centrality As Salience In Text Summarization", 2004., J. Artificial Intelligence Research, vol. 22, pp. 457-479.

[5] Jianbo Shi and Jitendra Malik, "Normalized Cuts And Image Segmentation". August 2000, IEEE Transaction on Pattern Analysis And Machine Intelligence, Vol 22, No.8, pp-888-905.

[6] M.S.Yang, "A Survey Of Fuzzy Clustering", 1993, Math. Computer Modelling, vol. 18, no. 11, pp 1-16.

[7] J.C. Bezdek, "Cluster Validity with Fuzzy Sets," J. Cybernetics, vol. 3, no. 3, pp. 58-72, 1974.

[8] Raymond Kosala and Hendrik Blockeel, "Web Mining Research: A Survey". 2000, ACM SIGKDD Explorations Newsletter, vol. 2, no. 1, pp. 1-15.

[9] J.C. Bezdek, "Mathematical Models for Systematics and Taxonomy,"Proc. Eighth Int'l Conf. Numerical Taxonomy, pp. 143-166, 1975.

[10] A. Budanitsky and G. Hirst, "Evaluating WordNet-Based Measuresof Lexical Semantic Relatedness," Computational Linguistics, vol. 32, no. 1, pp. 13-47, 2006.

[11] Brendan J. Frey* and Delbert Dueck, "Clustering By Passing Messages Between Data Points", 2007, Science, vol. 315, pp. 972-976.

[12] A.K. Jain, M.N. Murty, and P.J. Flynn, ªData Clustering: A Review,º ACM Computing Surveys, vol. 31, no. 3, pp. 264-323,1999.

[13] M. Jordan and R. Jacobs. Hierarchical mixtures of experts and the em algorithm. *Neural Computation*, 6:181–214, 1994.

[14] C.F.J. Wu. "On the convergence properties of the em algorithm". *The Annals of Statistics*, 11(1):95–103, 1983.

© 2015, IJARCSSE All Rights Reserved