ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 20, April 2016

An Efficient Evolutionary Clustering Using Strengthened Density based Algorithm

L. Visalatchi[1] and Dr.M.Balamurugan[2]

[1]*Associate Professor,Department of Information Technology,*
Dr.Umayal Ramanathan College for Women, Karaikudi, India
Email: Visaramki@gmail.com

[2]*Associate Professor,Department of Computer Science, Engineering and Applications,*
Bharathidasan University, Tiruchy - 23, India.
Email: mmbalmurugan@gmail.com

Abstract: **Clustering is one of the most common data mining tasks. It is a descriptive technique providing summary of available data by grouping them according to certain similarities. Moreover, due to the continuous flow and frequent changes in the input data, makes the traditional clustering producing static information, inefficient. In this paper, we study the problem of finding clusters over time period. As evolutionary clustering addresses the clustering problem along with optimization of time period, we choose evolutionary clustering for our work. Evolutionary clustering provides the evolution behavior of the clusters over time period which enables us to derive some important and interesting information. For example, the evolutionary collaborating groups of research area from DBLP dataset could be identified and how they evolve as time goes on is also studied.**

Key words: **Evolutionary clustering, density based algorithm, heterogeneous network, dynamic network.**

I. INTRODUCTION

A heterogeneous information network is a dynamic network consist of multiple type of objects, it needs optimization of parameters to discover novel, significant and meaningful information. Examples of dynamic networks include network traffic data, bibliographic data, dynamic social network data, and time-series microarray data. Clustering in dynamic networks is employed with various techniques. One such is density based algorithm which finds the neighbor nodes based on the density parameters. Density based algorithms are suitable for any type of objects and it is able to identify clusters of arbitrary shapes, handling noise and is also fast. In this paper, we propose an efficient strengthened density based algorithm that uses structural similarity along with weight which is labeled on the links based on the strength of the link.

II. RELATED WORK

Graph partitioning mean partitioning the graph according to certain criteria, into a subset of graphs. Given a graph G={V,E} where V is a set of nodes or vertices and E is a set of edges drawn between nodes, the aim of graph partitioning is to create a subset of G into k disjoint sub-graphs $G_t = \{V_t, E_t\}$. The number of sub-graphs k may be known a priori or not known. In the problem of finding clusters using graphs many methods are prevalent. One such method is probabilistic generative method which uses maximum likelihood technique to evaluate the posterior probability of a

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 20, April 2016

node in a cluster and it uses expectation maximization approach to evaluate the priors. The drawback of this method is, it is suitable and effective only for the situation where we know the number of clusters to be generated in advance.

In contrast to the above situation we use density base clustering, adopted from the SCAN algorithm proposed by XU et.al which uses the neighborhood notions as criteria to find clusters. The number of clusters that may be generated is not fixed and it is not known a priori. Inspired by the density algorithm SCAN, in addition to the already existing density notions, we introduced certain features of computing weight based on the strength of the links between the neighbor nodes to find the quality clusters of papers from the four area subset of real dataset DBLP.

III. PROPOSED WORK

A. *Strengthened Density based Algorithm*

Input: G(V,E,T), Similarity threshold, minimum no. of objects

Output: CS={C_i}

1. Find the local clusters for each timestamp with the structural similarity which uses matchTerm() to identify a dense subgraph.
2. T-Partite graph is constructed by connecting two local clusters, with help of the weight labeled on the links which is computed using simWeight().
3. Now the similarity between clusters of two adjacent timestamp is computed and the weight is labeled on the inter cluster edges of two timestamps.
4. Therefore, now it is easy to apply the density based notions of finding common neighbors(direct reachable and indirect reachable)from the core node.

Initially our algorithm finds the local clusters using the structural similarity instead of Euclidean distance. With the structural similarity based on density notion, matchTerm() takes two parameters, ε and *minobjs*. ε is the similarity threshold and *minobjs* is the minimum number of objects that should be matched. The node with the neighbors greater than minobjs is treated as core node. With the dense subgraph obtained, a t-partite graph is constructed. Two local clusters of a adjacent time stamp network may be connected if they have a non-zero similarity. Again the weight of the links is computed using simWeight() which returns the weight. The similarity σ(v,*w*) becomes non-zero only if *v* is directly connected to *w* with an edge. The value of σ(v,*w*) ranges from 0.0 to 1.0 and especially becomes 1.0 when the number of terms match is >= 5. In the local clusters of the same partite the edges that have weight of 0.3 and above plays the important role. In calculating the degree of the vertex let us take into account only the edges having the weight of 0.3 and above. This helps in finding the degree of the graph with dominant vertex. Thus the weight and structural similarity is integrated to find the common neighbors which includes both direct and indirect reachable.

IV. EXPERIMENTAL SETUP

We use a subset of real dynamic network data set, the DBLP data. For ease of computing and to reduce the complexity in interpreting the results obtained, we extracted a four papers of area DB,DM,IR,ML

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 20, April 2016

from the year 2012 to 2015. This is called as four area dataset. We regard papers as nodes, terms as edges and years as timestamps. We give weight affinity to those papers which have similar terms accordingly. Length of the title of the paper is split and the stop words are filtered.

A link between two paper nodes occur only if atleast there is three matching term between them accordingly the weight is given for the links. If there is a match of three words, the weight is given 0.3, and accordingly the weight affinity is increased and a maximum of ten terms is given a weight affinity 1.0. With this set up a t-Partite graph is constructed initially and the t-partite graph is used as an input for connecting the two local clusters of adjacent time stamp, with which clusters are generated using density based notions.

V. RESULTS AND DISCUSSION

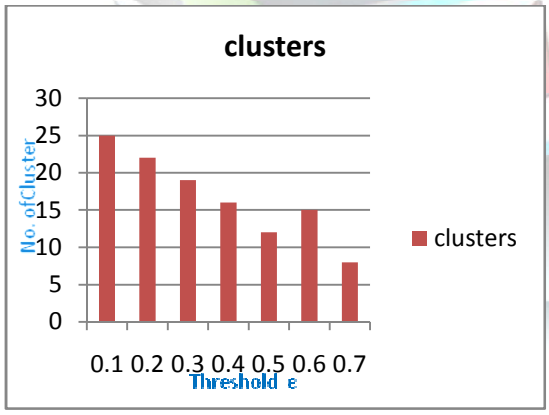

Fig. 1 No. of Clusters derived varies with threshold value ε

Our study discovers that a wide range of clusters arise instead of stable number of clusters when we vary the density threshold ε. To get approximated top k clusters we can use moderate threshold value ε=0.5~0.7. For temporal smoothness α is introduced which controls the direct and indirect weight affinity and therefore restricts the number of cluster created.

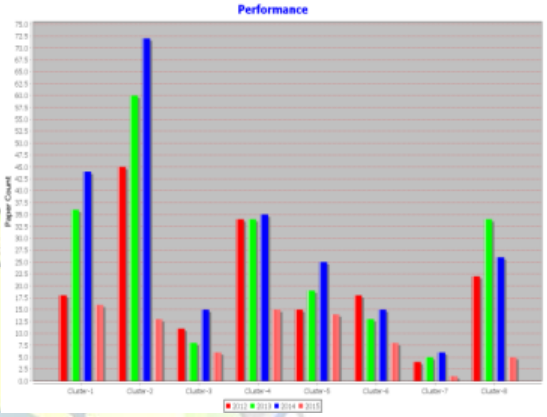

Fig. 2 Density of papers in the clusters derived for the four area dataset

With ε=0.7 , minobjs=7 , alpha=0.7, we got eight quality clusters which have the dominant paper as given in the table 1.

TABLE I
MAJOR PAPER DOMAIN IN EACH CLUSTER

Cluster 1	Data Mining
Cluster 2	Information Retrieval
Cluster 3	Information Retrieval
Cluster 4	Machine Learning
Cluster 5	Machine Learning
Cluster 6	Machine Learning
Cluster 7	Machine Learning
Cluster 8	Image processing

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 20, April 2016

VI. CONCLUSION

The proposed algorithm discovers clusters of collaborating research area over time period successfully. To reduce the complexity, only the adjacent time stamp partite are considered in finding the similarity (i.e T_i and T_{i+1}). In future, the algorithm could be extended to find the research hierarchy of prominent research areas from DBLP and the large dataset could be scaled by deploying with Hadoop cluster architecture.

REFERENCES

1. Chakrabarti. D, Kumar.R, and Tomkins.A 2006, Evolutionary clustering KDD pages 554-560
2. Y, Hino.K, Song.X, B.Tseng and Zhou.D 2007, Evolutionary spectral clustering KDD SIGKDD
3. pages 153 - 162
4. Lin.Y.R,Sundaram.H,TsengB.L and Zhu.S,2008, FacetNet: A framework for analyzing communities and their evolutions in dynamic networks. In Proc. pages 685-694.
5. Donn Morrison and Ian McLoughlin and Alice Hogan and Conor Hayes 2012. Evolutionary clustering and analysis of user behavior in online forums AAAI
6. Ester.M, Kriegel H.P, Sander.J and Xu.X. . 1996, A density – based algorithm for discovering clusters in large spatial database with noiseIn proc KDD, pages 226-231.
7. Han.J and Kamber.M. 2006. Data Mining Concepts and Techniques. Second edition, MorgKaumann,
8. Kim.M and Han.J 2009. A Particle and Density based Evolutionary clustering method for dynamic networks VLDB pages 622-633.
9. Liu.H, Z.Nazeri, J.Zhang, and Tang.L. 2008, Community evolution in dynamic multimode networks.
10. In proc. KDD,pages 677-685.
11. Sun.Y,Han.J, Zhao.P,Yin.Z,Cheng.H and Wu.T 2009 Rankclus: Integrating clustering with Ranking for heterogeneous information network analysis EDBT pages 565-576
12. Sun.Y, Han.J, Yintao Yu.2009. Ranking-Based Clustering of Heterogeneous Information Networks with star network schema KDD EDBT pages 149-160
13. Sun. Y., Jiawei Han ,Manish Gupta and CharuC.Agarwaal. 2010.Evolutionary Clustering and Analysis of bibliographic networks KDD ASONAM pages 63-70
14. Sun, Y.,JieTang,J.Han, Bo Zhao and Manish Gupta. 2010.Community evolution detection in dynamic heterogeneous networks. ACM MLG-10 pages 137-146
15. Sun, Y.,JieTang,J.Han, Cheng chen and Manish Gupta. 2013.Co-Evolution of multi-typed objects in dynamic star networks IEEE
16. VikramPudi GVR,Ravi Shankar, Kiran.2010. Evolutionary clustering using frequent itemsets. ACM
17. Vijaykumar S[1], Dr. M. Balamurugan[2], Ranjani K[3], Big Data: Hadoop Cluster Deployment on ARM Architecture, International Journal of Advanced Research in Computer and Communication Engineering (IJARCCE), Vol. 4, Special Issue 1, June 2015, ISSN 2278-1021 & 2319-5940.
18. S. Vijaykumar[1] ,M. Balamurugan[2], S.G. Saravanakumar[3], Unique Sense: Smart Computing Prototype, Procedia Computer Science, Volume 50, 2015, Pages 223-228, ISSN 1877-0509, http://dx.doi.org/10.1016/j.procs.2015.04.056.