Fast and improved clustering technique with user profile information for correlated probabilistic graphs

G.Priyadharshini [1], M.Usha [2]

[1] *Student, Master of Philosophy,* [2] *Head of the Dept, Dept of Computer Application, KG College of Arts and Science, Coimbatore, Tamil Nadu.*

[1]priyadharshinimphil123@gmail.com, [2]usha1231231@gmail.com

Abstract

Objectives: The main objective of this work is to achieve the better efficiency and accuracy of the clustering along with the user profile information.

Methods: Partially Expected Edit Distance Reduction (PEEDR) technique is used for adding or eliminating vertices from the clusters. Correlated Probabilistic Graph Spectral (CPGS) is used to progress the quality of cluster. Improved attractiveness-based community clustering is a weighted clustering approach which enhances the clustering performance in superior.

Findings: The proposed method achieves high performance in terms of precision, recall and accuracy.

Application/Improvements: The proposed system is done by using improved attractiveness-based community clustering (IACC). It performs the clustering process based on the weight value node and edge in the network. The weight of node implies the core degree of the person in the network, and the weight of edge means the attractiveness between the two nodes. Additionally, this method performs the efficient graph clustering technique which combines the user profile of users.

Keyword: Clustering, correlated, probabilistic graph, improved attractiveness-based community clustering.

1. Introduction

In modern days, the data mining plays an important role such as road networks, protein-protein interaction network and social networks. The information from variety of applications usually shows an inherent property of uncertainty and it is modelled as probabilistic graphs. There are two edges in the probabilistic graph named as conditionally independent and conditionally dependant. In preceding research [1], the scenario suggested for improving the efficient similarity search in graph database. This paper examines the sub node similarity search in the high probabilistic graph records. Based on the occurrences of adjacent edges the method can compute the correlated of graph nodes. In several applications, co occurrences and common segregation between adjacent edges are ensured by efficient graph techniques [2].

The clustering technique is extensively utilized in diversity graph analysis applications. The most prominent applications are such as community discovery and index structure. This scenario is focused on the clustering of correlated probabilistic graphs which targets to split the vertices into various disconnected clusters. In [3] the scenario recommended an effective machine learning method named as heuristic clustering approach. This approach is used for dealing with the discovery of group of users in community networks. Also this is used to determine the complexities of probabilistic protein-protein interaction network.

In [4], the scenario is discussed about the traffic uncertainty and probabilistic path queries concept in road networks. Path queries are used to discover the shortest path in the road network structure efficiently. It confines the uncertainty like travelling time among two vertices, the edge weight and sample set in traffic. The main objective of this scenario is providing solutions for path queries in huge scale probabilistic road networks. The best first search technique is sued to improve the shortest path problems effectively. However it has issue with computational complexities. In [5] the scenario recommended the model which is used to develop the correlations in network scheduling. The simulation based model in this scenario is estimating the schedule networks while activity durations are correlated. The uncertainty issues are handled by using proposed scenario's model which improves the system performance efficiently. However it has issue with cost complexities.

In [6] the authors are suggested the concept of clustering of data. The clustering concept is based on the unsupervised machine learning approach. Typically, the clustering algorithm is focused on the separation of similarity data into one group and dissimilar data into another group. In recent research, the statistical based cluster algorithm access the wide range of community group more effectively. In this scenario, the techniques are introduced named as incremental clustering which targets on higher similarity. However it has issue with pattern recognition and searching complexities.

In [7] George Kollios, Michalis Potamias and Evimaria Terzi discussed high probabilistic graph concepts by using clustering techniques. The probabilistic graph clustering is as similar as clustering paradigm graph which has several applications. This technique is used to determine the difficulties in probabilistic protein-protein communication network and also determine the users group in community groups. Another method named as edit distance based algorithm which is the form of graph clustering into probabilistic graphs. This scenario found the relationship among objective function and correlation clustering. By using protein-protein communication system and ground truth information the techniques are determined the accurate number of clusters as well as recognize the protein connections. Also this clustering method is able to deal with huge size of social network users more efficiently. However the scenario is problem with computational time.

In [8] the technique is introduced named as spectral clustering. This algorithm is used for producing high quality of clustering results. The main motivation of this scenario is developing the algorithms to solve the computational issues as well as large dimensional datasets. Extend the spectral clustering series through developing the standard structure for fast estimate spectral clustering. In particular this structure is depends on the theoretical analysis which produces a numerical classification of the effect of confined distortion on the mis-clustering rate. It is also used to enhance k-means clustering based on random projection trees. Still it has problem in terms of inaccuracy results.

2. Materials and Methods

2.1. Distance-Probability-Threshold Clique (DPTC) construction

In this module, we have to construct Distance-Probability-Threshold Clique (DPTC) which is centred in the singleton cluster. It is described as probability distribution which generates the relationships between community groups. It has two notations and for two vertices *vi* and *vj* in a correlated probabilistic graph *G* = {*V, E, P, F*}, if there is a route *rk* from *vi* to *vj*, then we describe the distance from *vi* to *vj* as the number of hops in this route, indicated as *drk (vi, vj)*. Then describe the likeness among *vi* and *vj* as the continuation possibility of this transmit, represented as *simrk (vi, vj)*. The specified threshold dt and probability threshold α, then describe a subgraph C of G as a Distance-Probability-Threshold Clique (DPTC). If some couple of vertices vi, vj ∈ VC, there exists a direction rk from vi to vj fulfilling the needs drk (vi, vj) $\leq$ dt and simrk (vi, vj) $\geq$ α.

2.2. PEEDR clustering algorithm

In this module, the algorithm named as Partially Expected Edit Distance Reduction (*PEEDR*) which is used for finding correlations based on probability conditions. The PEEDR algorithm starts a cluster along with one vertex. Then for every vertex that is adjacent to the cluster, it is eliminated to the cluster if it decreases the expected edit detachment from G to the present cluster graph. The above mentioned process is iteratively used till could not enlarge the cluster. Then select a vertex from the unclustered vertices and repeat the above process to create another cluster. This process is repeated till each vertices of G are gathered into clusters. Then we can obtain the final cluster

Algorithm

Input: correlated probabilistic graph G = {V, E, P, F} a distance threshold d_t, probability threshold α

Output: cluster graph Q

1. Sort the vertices of G in descending order of their degress
2. Initialize i←0, b←true;
3. Initialize a virtual cluster C', where $V_{C'}$←V;
4. While ($V_{C'} \neq \emptyset$) do
5. Select the vertex $v' \in V_{C'}$ with the highest degree
6. Establish a DPTC cluster C_i centered with v' according to d_t and α, and set $V_{C'}$=$V_{C'}\backslash V_{C'}$
7. While (b=true) do
8. b←false
9. for ($v_j \in V_{C'} \cap adj(C_i)$) do
10. if(isReduce Edit(v_j, C_i)) do
11. $V_{C'}$←$V_{C'}\backslash\{v_j\}$, $V_{C'}$←$V_{C'} \cup\{v_j\}$;
12. b←true
13. end
14. end
15. end
16. i←i+1
17. end
18. return A cluster graph composed of clusters C_j(j=1,2……i-1)

2.3. CPGS clustering algorithm

This algorithm is useful for multi dimensional data which means we can cluster the nodes for large social network. In this module, we consider objects as DPTC which has to be cluster based on the highest degree of the vertices. For example the user one placed as source node then user 2 and user 3 might be placed as target node for 1. We can compute by using this algorithm based on the maximum similarity and minimum distance for still more number of users. Hence communication has been improved with accuracy. It considers spaces and by k-means algorithm it will cluster the nearest neighbor nodes in the network. Hence it achieves number correlated graph nodes with greater accuracy results. This algorithm determines based on the correlated graphs in the social networks. In existing scenario this algorithm yields maximum accuracy.

Algorithm

Input: a correlated graph G = {V, E, P, F} a distance threshold d_t, probability threshold α, cluster number K

Output: K disconnected clusters

1. initialize a virtual cluster C' $where$ $V_{C'} \leftarrow \{v | v \in G\}$;
2. m←0
3. while($V_{C'} \neq \emptyset$) do
4. m=m+1
5. select the vertex v' with the highest degree in $V_{C'}$
6. employ algorithm 1 to build a DPTC on v_i according to α and d_t denoted as C_m
7. end
8. discover the k nearest DPTC of each DPTC and compute the weight among them
9. establish a Laplacian matrix according to the k nearest neighbor DPTCs of each DPTC, and compute the k eigen vectors corresponding to the biggest eigen values of matrix
10. represent each of the m DPTC through a point P_i in a k-dimensional space, where the coordinate of P_i is P_i (U_{1i}, U_{2} , … … … … U_{Ki}) and cluster the m points into k clusters along with k-means algorithm
11. cluster the m DPTCs into K clusters according to the cluster result of their corresponding points. Each vertex belongs to the cluster that its DPTC belongs to;
12. return K disconnected clusters

2.4. ACC Algorithm

For weighted graph clustering, we propose an attractiveness-based community clustering algorithm. It is an amalgamation algorithm, the merger between clusters could be considered while the attractiveness of clusters (as the edge weight) is bigger than the densities of clusters (as the node weight). ACC algorithm is designed to make some breakthrough on the time complexity of community detection for large social networks

ACC algorithm can be divided into two main steps, iterating between the two steps to get clusters:

1. Merge the pair of clusters which has the largest attractiveness.

2. Calculate or update the cluster density and cluster attractiveness matrix;

Executing the update of cluster density and attractiveness matrix, and the cluster merger process iteratively, until the structure of clusters does not change, or there is only one cluster left.

1) Initialize community attractiveness matrix S , S is a n-order matrix, the matrix elements S i j denotes the attractiveness between the node i and j;

2) According to the newest attractiveness matrix S , for each cluster i, find cluster j which would get the highest attractiveness with i, and record their attractiveness with S i j , and calculate the density of the cluster i and cluster j, respectively denoted by Wi, Wj;

3) Merge the communities satisfy the expression (1);

4) If any of the following happens, skip to step 7): Case one: the structure of clusters does not change Case two: only one cluster left

5) Update the cluster attractiveness matrix S ;

6) Repeat from step 2) to 5);

7) Stop the iterative process, save the result of clusters as communities and return.

2.5. User profile based clustering

In our proposed system, we are proposing the efficient graph clustering technique which combines the user profile of users. In other words, we are analyzing user profile information for clustering process. This consideration of user profile information is improving the clustering accuracy and enhances the performance. In the case of large social network, the clustering process is taken the long time for clustering in the existing system. In order to overcome this drawback in the existing system, we are including of user profiles is used for well reduce the time complexity and improve the performance of the system.

The user profile information like friends list of users, contains information about education, interests of a user, etc. according to the information of this user profile we can perform the graph clustering effectively. From this consideration of user profile, we can obtain the number of common neighbours for the users. This is important factor for clustering process in order to reduce the long time in large social network data. Thus in our proposed system, we are performing the similarity between the user profiles. Based on this similarity value we can cluster in our proposed system.

3. Results and Discussion

In this section, we compared the performance metrics for existing system and proposed system. The performance metrics are such time factor, precision and recall values. The metrics are performed by using the methods such as PEEDR and ACC for existing and proposed scenario respectively. Both of the algorithms are implemented for improving the accuracy of the scenario. However the proposed scenario yields highest accuracy rather than existing scenario.

3.1. Precision

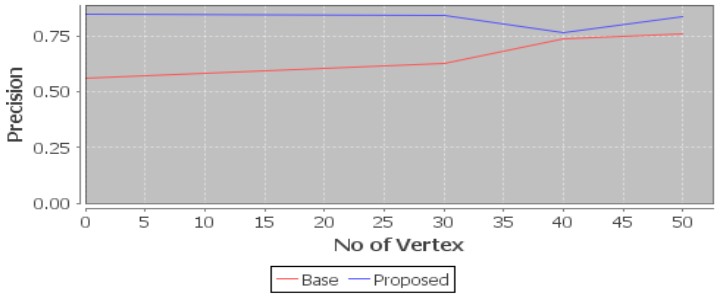
Precision is defined as the Percentage of correct predicted results from the set of input terms. The precision value should be more in the proposed methodology than the existing approach for the better system performance.

Precision is calculated by using following equation

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive+False Positive}}$$

In this graph, x axis is taken for two methods of and y axis is taken for precision. From the Figure.1 the proposed scenario shows the highest precision rather than existing method. The proposed ACC method is used to improve the clustering performance in superior.

Figure 1. Precision comparison



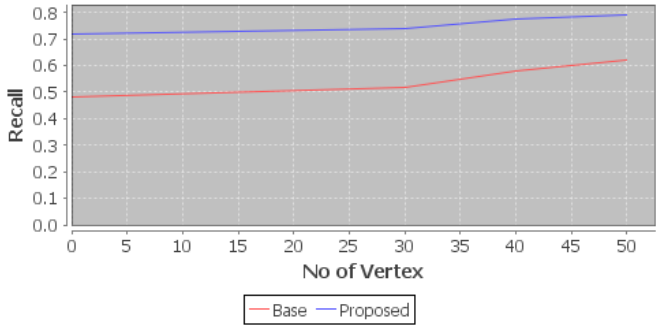
3.2. Recall

The recall or true positive rate (TP) is the proportion of positive cases that were correctly identified, as calculated using the equation:

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive+True Negative}}$$

In this graph, x axis is taken for two methods of and y axis is taken for recall. From the Figure.2 the proposed scenario shows the highest recall rather than existing method. The proposed ACC method is used to improve the clustering performance in superior.

Figure 2. Recall comparison

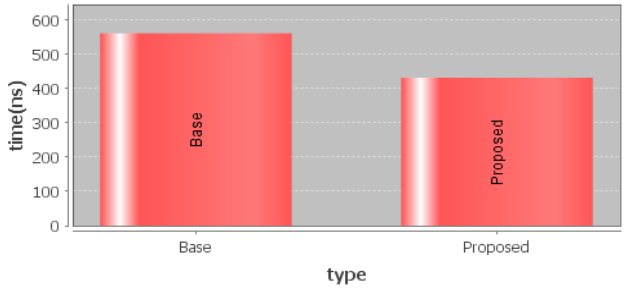


3.3. Time factor

The time factor is calculate by using the execution time of existing and proposed methodologies.

In this graph, x axis is taken for two methods of and y axis is taken for time factor. From the Figure.3 the proposed scenario shows lower time complexity rather than existing method. The proposed ACC method is used to improve the clustering performance in superior.

Figure 3. Time comparison



4. Conclusion

In data mining, the efficient and effective algorithms are introduced as well as evaluated the clustering concepts based on the unsupervised learning approaches. In existing scenario, the PEEDR algorithm is used to improve the clustering performances. It used the concept of CPGS to reduce the complexity of system. However it has issue with accuracy of the clustering. To enhance this concept in proposed scenario, the method involved named as ACC which is used for improving the precision and recall values in higher. It is also used to reduce the time complexity significantly in this scenario. The proposed scenario yields higher performance rather than existing scenario. The performance is superior in terms of time complexity, precision and recall values. Hence the proposed method is higher efficiency by using ACC algorithm rather than existing scenario.

5. Acknowledgement

Authors declare that the work is original and there is no conflict of interest.

6. References

1. Yuan Ye, Guoren Wang, Lei Chen, Haixun Wang. Efficient subgraph similarity search on large probabilistic graph databases, *Proceedings of the VLDB Endowment.*2012; 5(9), 800-811.
2. Yu Gu, Chunpeng Gao, Gao Cong, Ge Yu. Effective and efficient clustering methods for correlated probabilistic graphs, Knowledge and Data Engineering. *IEEE Transactions on*. 2014; 26(5), 1117-1130.
3. Balaji, Vani shree, Naveena. Probabilistic graphs using clustering algorithm with efficient performance, on I*nternational Journal of Innovative Research in Computer and Communication Engineering*. 2015; 3(2), 784-790.
4. Hua Ming, Jian Pei. Probabilistic path queries in road networks: Traffic uncertainty aware path selection, *Proceedings of the 13th International Conference on Extending Database Technology*. ACM, 2010, 347-358.
5. Wang, Wei-Chih, A. Laura Demsetz. Model for evaluating networks under correlated uncertainty-NETCOR, *Journal of Construction Engineering and Management*. 2000; 126(6), 458-466.
6. Jain, K. Anil, M. Narasimha Murty, J. Patrick Flynn. Data clustering: A Review, ACM computing surveys (CSUR).1999; 31(3), 264-323.
7. Kollios George, Michalis Potamias, Evimaria Terzi. Clustering large probabilistic graphs, Knowledge and Data Engineering, *IEEE Transactions* on. 2013; 25(2), 325-326.
8. Yan Donghui, Ling Huang, I.Michael Jordan. Fast approximate spectral clustering, *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2009; 907-916.

The Publication fee is defrayed by Indian Society for Education and Environment (iSee). www.iseeadyar.org

Citation:

G.Priyadharshini, M.Usha. Fast and improved clustering technique with user profile information for correlated probabilistic graphs. *Indian Journal of Innovations and Developments.* 2015; 4 (5), September.