Survey on social networking analysis and clustering

[1]Rejna A N, [2]Kalaiarasi R
[1] School of computer science, Tamilnadu open university, chennai ,[2] School of computer science, Tamilnadu open university, chennai.

Abstract**—*****This paper analyses the corpus of e-learning research publications, using tools of Social Network Analysis (SNA), to develop a network structure of e-learning research. We applied clustering and centrality measurement to the e-learning network to identify the influential research articles in the field. This study identifies significant core clusters of e-Learning research that centre around (i) e-learning models (ii) adoption and acceptance of e-learning methodologies and tools, and (iii) e-learning using social media and highlights areas for future research in the field.***

Index Terms—***social networking ,clustering, K-means algorithm*****.**

I. INTRODUCTION

Social network research is concerned with studying patterns of social structure. Social networks can be obtained from archival data such as journal articles, newspapers, court records, minutes of executive meetings, biographical records, patterns of citation among scholars. Social network analysis (SNA) is the process of investigating social structures through the use of network and graph theories. It characterizes networked structures in terms of nodes (individual actors, people, or things within the network) and the ties, edges, or links (relationships or interactions) that connect them. There are many tools for interactively visualizing and analyzing small networks such as Sci2 tool, Pajek, Gephi and so on. We can make use of any of these tools for constructing the network and for further analysis.

Relations of real-world entities are often represented as networks, such as social networks connected with friendships or co-authorships. Social network is a one-mode network in which the relation of friendship is measured on a single set of people. In many cases, real social networks contain denser parts and sparser parts. Denser sub networks correspond to groups of people that are closely connected with each other. Detecting communities from given social networks are practically important. Communities will help us understand the structures of given social networks. Communities are regarded as components of given social networks, and they will clarify the functions and properties of the networks. Communities are regarded as components of given social networks, and they will clarify the functions and properties of the networks.

Definitions of community can be classified into the following three categories.

- Local definitions
- Global definitions
- Definitions based on vertex similarity.

Local definitions: Local definitions of community can be further divided into self-referring ones and comparative ones. The former considers the sub network alone, and the latter compares mutual connections of the vertices of the sub network with the connections with external neighbors.

Global definitions: Global definitions of community characterize a sub network with respect to the network as a whole. These definitions usually starts from a null model, in another words, a network which matches the original network in some of its topological features, but which does not display community structure.

Definitions Based on Vertex Similarity : A definition of the last category is based on an assumption that communities are groups of vertices which are similar to each other.

II. LITERATURE SURVEY

M. Girvan and M. E. J. Newman(2002) analyzed the social networks to find the community structure by using hierarchical clustering. They measured the edge weight between nodes and also performed centrality measures thereby finalize how closer the nodes are ,which edge acts as a bridge between the clusters. They have applied community finding algorithm to collaboration network of scientists e Santa Fe Institute, an interdisciplinary research center in Santa Fe, New Mexico[2].

Aaron Clauset, M. E. J. Newman, and Cristopher Moore(2004) present a hierarchical agglomeration algorithm for detecting community structure which is faster than many competing algorithms. They represent item listed in Amazon.com sites as a network and measured the rate of purchasing in each year [3].

Jain, A.K(2010)explains about data clustering using K-means algorithm. According to him extensions can be made to K-means algorithm by involving minimum cluster size and splitting and merging clusters. They conclude that, we need to develop clustering methods that lead to stable solutions and should achieve a tighter integration between clustering algorithms and the application needs.[7]

III. CLUSTERING

Clustering is a process of partitioning a set of data (or objects) in a set of meaningful sub-classes, called clusters. It helps users understand the natural grouping or structure in a data set. Clustering has a wide range of applications, from spatial data analysis to market research. In document clustering, the distance measure is often also Euclidean distance. Different distance measures give rise to different clusters. Thus, the distance measure is an important means by which we can influence the outcome of clustering .Fig 1. Shows cluster can vary in size, shape and pattern.

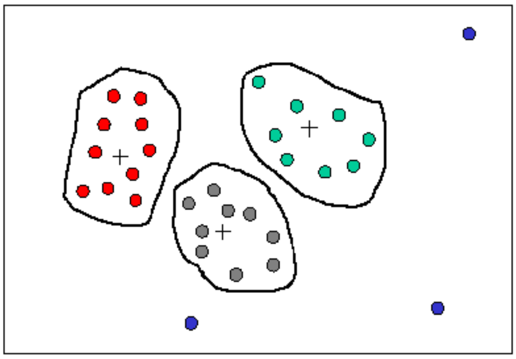

Fig 1. Clustered Data in groups

Clustering can be broadly classified into two groups: hard clustering and soft clustering. If each data is assigned to a single cluster uniquely, it becomes hard clustering. If there is a possibility of finding the data in more than one clusters, then it is soft clustering.

Clustering algorithm can be classified into two groups: hierarchical and partitional. In hierarchical clustering, similar data will be grouped together to form clusters. Each cluster is distinct from each other. Partitional clustering divides the set of data into groups such that each group contains at least one data and each data belongs to exactly one group.

The most widely used and the simplest partitional algorithm is K-means. One main issue is that we should decide the number of clusters in advance before running the algorithm. In social network data we have to find the criteria in which clusters are to be made. The number of clusters is given as a parameter before clustering.

k-means clustering algorithm

The k-means clustering algorithm is known to be efficient in clustering large data sets. This algorithm is one of the simplest and best known unsupervised learning algorithm. The k-means algorithm aims to partition a set of objects, based on their attributes/features, into k clusters, where k is a predefined constant. The algorithm defines k centroids, one for each cluster. The centroid of a cluster is formed in such a way that it is closely related, in terms of similarity (where similarity can be measured by using different methods such as Euclidean distance or Extended Jacquard) to all objects in that cluster. Technically, what k-means is interested in, is the variance. It minimizes the overall variance, by assigning each object to the cluster such that the variance is minimized. The main steps of K means algorithm are as follows (Jain and Dubes, 1988):

1.Select an initial partition with K clusters; repeat steps 2 and 3 until cluster membership stabilizes.

2. Generate a new partition by assigning each pattern to its closest cluster center.

3. Compute new cluster centers. iv. Repeat from (ii) until a stopping criterion is reached.

IV. CENTRALITY MEASURES

Centralization is property of the overall network and it is defined as the variation in the centrality scores of the vertices in the network .It shows the extent to which there is a center and a periphery. Centrality will be measured based on the degree of the nodes in the network. Centrality measurements will be applied on a network represented as a Graph, G=(V,E) with |V| vertices and |E| edges. Different types of centrality measures are degree centrality, closeness centrality, between centrality and eigenvector centrality.

The degree centrality of a vertex x, for a given graph is defined[11] refer to equation (1)

$$C_D(x)=deg(x) \qquad (1)$$

Deg(x),degree of vertex(x),which is the number of links incident upon vertex, v.

The closeness centrality of a vertex x is the average length of the shortest path between vertex, x and all other vertices in the graph. Closeness was defined by Alex Bavelas (1950) as the reciprocal of the farness as in [11], refer to equation (2)

$$C(x) = \frac{1}{\sum_y d(y,x)} \qquad (2)$$

where d(y,x) is the distance between vertices x and y.

Betweenness centrality of a vertex x represents number of times the vertex acts as a bridge along the shortest path between two other nodes or between two connected components in a graph. The betweenness of a vertex v in a graph G=(V,E) with v vertices is computed as follows as in [11]

1. Node For each pair of vertices(s, t),compute the shortest paths between them.
2. For each pair of vertices(s, t),determine the fraction of shortest paths that pass through the vertex in question
3. Sum this fraction over all pairs of vertices(s, t).

The betweenness can be defined as in [11] refer to equation(3):

$$C_{B(v)} = \sum_{s\neq v\neq t\in V} \frac{\sigma_{st}(v)}{\sigma_{st}} \qquad (3)$$

Eigen vector centrality is a measure of the influence of a node in a network. It assigns relative scores to all nodes in the network based on whether the nodes are connected to more influential nodes or not

For a graph G , Eigen vector equation is defined as

$$Ax = \lambda x \qquad (4)$$

Where A is the adjacency matrix , λ is the Eigen value. The greatest Eigen value gives the centrality measure.

V. PROPOSED METHOD

The first stage was to compile a dataset of e-learning publications, followed by network creation and clustering the network using tools of the social network analysis. Web of Science database, one of the largest collection of academic publications and contains research papers published in peer-reviewed journals were used as the bibliographic database for interrogation. The keywords “e-learning OR “electronic learning OR “eLearning”, which appeared in the publication title, abstract or keyword sections, were used to identify a collection of representative research articles from the bibliographic database. Selected publications were categorized in the article, written in English and a “journal” publication to suit our requirements. Manual checking was also implemented to ensure the relevancy, which produced a list of articles related to e-learning research. The second stage created a citation network of e-learning, constructed based on the concept that an edge exists between any two publications when one publication was cited on another publication.

Clustering will be implemented on e-learning citation network using K-means algorithm by making use of social networking tools. Centrality Measures shows the influence of individual research articles on each cluster. The most influential articles contributing to the e-learning research domain were able to be identified through it. Core articles in each year has been identified based on their citations by using degree centrality

CONCLUSION

E-learning research papers under study were clustered using K-means Algorithm. Papers were identified based on published year from 2003 to 2019 and made different clusters accordingly. Centrality measures can apply to find the core papers in each cluster. Database is limited to Web of science. Future enhancement can be made by extending the search to other databases.

REFERENCES

[1] Jain, Anil K., Dubes, Richard C., 1988. Algorithms for Clustering Data. Prentice Hall

[2] Girvan, M., Newman, M.E.J.: Community structure in social and biological networks.Proc. Natl. Acad. Sci. 99(12), 7821–7826 (2002)

[3] Aaron Clauset, A., Newman, M.E.J., Moore, C.: Finding community structure in very large networks. Phys. Rev. E 70, 066111 (2004)

[4] Merriam-Webster Online Dictionary, 2008. Cluster analysis. .

[5] Zhang, S., Wang, R.-S., Zhang, X.-S.: Identification of overlapping community structure in complex networks using fuzzy-means clustering. Phys. A 374(1), 483–490 (2007)

[6] Jain, A.K., Topchy, A., Law, M.H.C., Buhmann, J.M., 2004. Landscape of clustering algorithms. In: Proc. Internat. Conf. on Pattern Recognition, vol. 1, pp. 260–263. JSTOR, 2009.

[7] Jain, A.K.: Data clustering: 50 years beyond k-means. Pattern Recogn. Lett. 31(8),651–666 (2010). Award winning papers from the 19th International Conference on Pattern Recognition (ICPR)

[8] Steinhaeuser, K., Chawla, N.V.: Identifying and evaluating community structure in complex networks. Pattern Recogn. Lett. 31(5), 413–421 (2010)

[9] Mr. Anand Khandare , Dr. A.S. Alvi Efficient Clustering Algorithm with Improved Clusters Quality IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 18, Issue 6, Ver. V (Nov.-Dec. 2016), PP 15-19

[10] Jun Wu, Member, IEEE, Mianxiong Dong , Member, IEEE, Kaoru Ota, Member, IEEE, Jianhua Li, and Zhitao Guan, Member, IEEE Big Data Analysis-Based Secure Cluster Management for Optimized Control Plane in Software-Defined Networks IEEE transactions on network and service management, vol. 15, no. 1, march 2018

[11] https://en.m.wikipedia.org>wiki>Centrality