VOL. 11, NO. 11, JUNE 2016 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

ESTIMATION OF THE EFFECTIVENESS OF CLUSTER ON ATTRIBUTE AND STRUCTURAL SIMILARITIES

Saravanan Venkataraman Tirumalai
Department of Computer Science, College of Computer and Information Sciences, Majmaah University, Majmaah, Kingdom of Saudi Arabia
E-Mail: s.tirumalai@mu.edu.sa

ABSTRACT

The goal of this study is to evaluate the qualities of various clusters and their improvements using some techniques. A comparative analysis is carried to identify the quality cluster. Distance measures are presented for measuring the distance functions. The selection of centroids, influence function, density function, pseudo inverse and other related concepts are mathematically explained for utilizing in this work. Density and entropy expressions are provided to exhibit the quality clusters. For the purpose of discussing the quality of various clusters the densities, entropies and weights relating to S, SA and W clusters are computed. The improvement of cluster quality based on above concepts have been analysed. A comparison is made among S, SA and W clusters and pointed out the quality clusters

Keywords: graph clustering, clusters, distance measures, density, entropy, cluster quality.

1. INTRODUCTION

Clustering is a vital process to distinguish the objects according to either their shapes or characters. Clustering is used as unsupervised learning process and its goal is to discover a new set of categories. Farley and Raftery (1998) have suggested two broad groups of clustering methods as Hierarchical and partitioning methods. Han and Kamber (2001) have proposed additional three methods such as density based methods, model based methods and grid based methods. In addition to that soft computing method including fuzzy clustering and Evolutionary process for clustering have been presented in Estivill-Castro and Yang (2000).

The qualitative and quantitative data such as economic data, managerial data, chronological data, clinical data, industrial data and so on are the major applications of clustering. Clustering process separates the data into two or multiple number of clusters.

Graph clustering is mainly utilized in network analysis. The purpose of clustering is to partition the graph into several connected components. Graph clustering measuresvertex closeness based on connectivity as well as structural similarities. Traditional data clustering measures the distance between two data points. When separating the clusters, outliers may occur. If it is so, Dendrogram is used to detect the outliers. Vertices belong to one group are having identical attribute values but it is not so for the vertices belong to different groups. This gains homogeneous attribute values within clusters.

The existing information networks are social, sensor and biological networks. In social networks, vertex properties describe the activities of a person or a valuable component but the topological structures represent the relationship among a group of persons or a set of components. When clusters are formed from a large graph, select the vertices which are closely connected and having similar characteristics belong to a particular cluster. This implies that a graph clustering generates clusters which have cohesive intracluster structure with homogeneous vertex properties. The clustering outcomes contain densely connected components within clusters.

Clustering method balances the attribute and structural similarities. Vertex distances and similarities have been measured by using any one the existing methods, in particular, random walk principle. In the literature, there are various types of random walks such as Neighborhood random walk, Unified random walk, Monte Carlo random walk and so on.

Consider the viewers of research papers on computer science in PLOS ONE Journal. 700 research papers which are published from Jan 2007 to Dec 2011 are listed out. Research papers collaborate with each other may have different characters such as topics, subtopics and prolific values. Research papers give relationship between/among viewers. During a short period of one month, it is observed that there are 17,356 viewers who viewed the above said papers. In the graph context, viewers represent the nodes (vertices) and research papers represent the edges.

The structure and attribute similarities along with viewers and research papers are stated as follows:

Structure: Any cluster is having only structures which give the outcomes based on vertex connectivity, that is, research papers relationship. Viewers within clusters are closely connected. However, in one of the clusters, viewers have different topics.

Attribute: Some clustering results are arrived based on attribute similarities (topics). Viewers within clusters belong to the same topic. Here, the research papers relationship may lose due to the partitioning so that viewers are isolated in one of the clusters.

Structure and Attribute

The clustering results are obtained based on both structures and attributes similarities and the results balance structural and attribute similarities. Viewers within one cluster are closed connected and homogenous on research topics.

The techniques adopted in this paper are listed below:

- Apply a unified random walk distance measure to combine attribute and structural similarities.
- Theoretical methods are given to boosten the presentations of attribute similarity to the unified random walk distances for studying the closeness of the vertices.
- Apply a weight self-adjustment method to analyze the degree of contributions of attributes in random walk distances.
- Apply inverse matrix to reduce large number of matrix multiplications.
- Perform suitable experiments using designed clustering algorithm.

This paper is divided into 7 sections. The section 1 describes the concepts of clustering, clusters and the techniques. Section 2 explains the problem and findings. Section 3 is devoted to review of literature. Distance measure and matrix application are presented in section 4. Section 5 describes clustering process, density function and weight self-adjustment. Evaluation process of cluster quality is explained in section 6. The section 7 gives the conclusion.

2. FORMULATION OF GRAPH

Consider a triplet$G = \{V, E, \Lambda\}$is an attributed graph. Here, $V, E \ \ and \ \Lambda = \{\alpha_1, \alpha_2, \dots, \alpha_m\}$are the set of vertices, the set of edges and the set of m attributes associated with vertices in V respectively. The size of the vertex set is $|V| = N$. Each vertex $\gamma_i \epsilon V$is associated with an attribute vector $[\alpha_1(\gamma_i), \alpha_2(\gamma_i), \dots, \alpha_m(\gamma_i)]$. An attributed graph G is partitioned into k disjoint sub graphs $G_i = \{V_i, E_i, \Lambda\}, \ \ i = 1, 2, 3, \dots, k, [\cup_{i=1}^{k} V_i = V, V_i \cap V_j = \varphi]$ by attributed graph clustering.

Clustering of G gains the balance between two key properties mentioned below:

- Vertices within cluster are close to each other but, that between clusters are distinct from each other.
- Vertices within cluster have same attribute values but, that between clusters have different attribute values.

In this paper, cluster quality comparison is carried out. Effectiveness of clusters are found and efficient cluster is identified.

Graph clustering techniques have been analyzed in many directions and primarily concentrated on topological structures. Since literature survey related to this topic is very huge, the important studies are listed here.

Ng and Han (1994) have developed clustering large applications based on Random search and this method identified candidate cluster centroids by using repeated random samples of the original data. Farley and Raftery (1998) have suggested the division of clustering methods, namely hierarchical and partitioning methods. Jain *et al* (1999) have proposed density-based clustering which employ non-parametric methods such as searching for bins with large counts in a multidimensional histogram of the input instance space. Shi and Malik (2000) have discussed graph clustering problems on normalized cuts and derived effective results.

Han and Kambar (2001) have recommended another three methods, viz, density-based methods, model-based clustering and grid-based methods. Jeh and Widom (2002) have designed a technique called 'SimRank measure' which helps to measure the similarity between two vertices. The concept of random walk has been used to measure vertex distances and similarities. Strehl and Ghosh (2002) have studied Ensemble analysis which improves classification accuracy and the general quality of cluster solution. They have also discussed the availability of multiple segmentation solutions within an ensemble and the method is Meta clustering algorithm and is based on the notion of "clustering on clusters"

Tong *et al* (2006) have formulated an algorithm for fast random walk computation. Pons and Latapy (2006) have proposed short random walks of length 'l' to measure the similarity between two vertices in a graph for community detection.Sun *et al* (2007) have proposed graph scope which is used to discover communities in large graphs and to detect the changing time of communities.Tsai and Chui (2008) have developed a feature weight self-adjustment mechanism for k-means clustering on relational datasets. Here, an optimization model is designed to find feature weights in which the partitions within clusters are minimized and that between clusters are maximized. Tian *et al* (2008) have discussed graph summarization analogue to partition a graph according to attribute values.

Orme and Johnson (2008) have discussed ensemble analysis for improving k-means cluster analysis and the methods have been described with the help of numerical illustrations. Zhou *et al* (2009) have proposed graph clustering algorithm based on both structural and attribute similarities and estimated the effectiveness of SA cluster as compared with other three clusters, through experimental analysis. Raj and Singh (2010) have summarized and described the types of clusters and different clustering methods. Zanghi *et al* (2010) have adopted generative process and proposed a probabilistic model to cluster attributed graphs. Cheng *et al* (2011) have studied graph clustering using unified random walk distance measures. A comparative analysis of clusters and their efficiencies have been carried out.

3. DISTANCE MEASURES

The distance measure is defined as the distance between two objects $O_1$and $O_2$from the universeof objects denoted as $d\ (O_1, O_2)$, which is always non- negative real number. Distance measures are used to obtain the similarity or dissimilarity between any pair of objects. In

general, distance measures are used for Numeric attributes [Minkowski metric (Han and Kamber (2001))], Binary attributes, Nominal attributes, Ordinal attributes and Mixed type attributes.

3.1 Measure of unified distance

It is assumed that each vertex is associated with a set of attributes $\Lambda = \{\alpha_1, \alpha_2, \dots, \alpha_m\}$ is an attributed graph. The distance function from the vertex γ_i to $\gamma_j$is defined as

$$d(\gamma_i, \gamma_j) = \alpha d_S(\gamma_i, \gamma_j) + \beta d_A(\gamma_i, \gamma_j) \quad (1)$$

Where d_S and $d_A$denote structure and attribute distances respectively. α and β are weighted factors. It is cumbersome to set the parameters of the equation (1). So we have to use another measure, namely, unified distance measure to combine the structure and attribute similarities. Now, two important definitions with relevant explanations are given.

Definition 1: Attribute augmented graph

Consider an attributed graph $G = \{V, E, \Lambda\}$. Let the domain of an attribute $a_i$is $Dom(a_i) = \{a_{i1}, a_{i2}, \dots, a_{in_i}\}$ and $|Dom(a_i)| = n_i$ is the size of the domain fora_i. An attribute augmented graph is defined as $G_a = (V \cup V_a, E \cup E_a)$,where $V_a = \{\gamma_{ij}\}_{i=1\,j=1}^{m\quad n_i}$ isa set of attribute vertices.

An attribute vertex $\gamma_{ij} \in V_a$explains that the attribute $i$take the j^{th} value. The two edges $(\gamma_i, \gamma_j) \in E$ and $(\gamma_i, \gamma_{jk}) \in E_a$are respectively called as structure and attribute edges. These two edges are significantly different. The attributes $\{\alpha_1, \alpha_2, \dots, \alpha_m\}$ have different importance and different degree of contributions in random walk distance. It is assumed that ω_0 be the weight of structure edge and $\{\omega_i\}_{i=1}^{m}$ be the weights of attribute edges $\{a_i\}_{i=1}^{m}$respectively.

Define the transition probabilities between the vertices.

- Let $P\gamma_i, \gamma_j$be the transition probability from vertex $\gamma_i$to vertex $\gamma_j$through a structure edge.
- $P\gamma_i, \gamma_{jk}$be the transition probability from vertex $\gamma_i$to vertex γ_{jk}through an attribute edge.
- $P\gamma_{ik}, \gamma_j$be the transition probability from vertex$\gamma_{ik}$to vertex $\gamma_j$through an attribute edge.
- $P\gamma_{ip}, \gamma_{jq}$be the transition probability from vertex $\gamma_{ip}$to vertex $\gamma_{jq}$and its value is zero since there is no edge between two attribute vertices.

Let $P_A$be the transition probability matrix of G_a and it is formed by using the above said transition probabilities interms of weights of structure and attribute edges.

Definition 2: Unified neighborhood random walk distance

Given the length of the random walk as 'l' with the probability of restart $c \in (0,1)$. The unified neighborhood random walk distance $d(\gamma_i, \gamma_j)$ from $\gamma_i$to $\gamma_j$in $G_A$is defined as

$$\mathrm{d}(\gamma_i, \gamma_j) = \sum_{\substack{\tau: \gamma_i \to \gamma_j \\ \delta \le l}} P_A(\tau) c\,(1-c)^{\delta} \quad (2)$$

where τ is the path from γ_i to $\gamma_j$whose length is denoted as δ with transition probability $P_A(\tau)$. The equation (2) can be written in matrix form as

$$\mathrm{R_A^l} = \sum_{\mathrm{r}=0}^{\mathrm{l}} \mathrm{c}\,(1-\mathrm{c})^{\gamma}\,\mathrm{P_A}^{\gamma} \quad (3)$$

Here, $R_A$is the neighborhood random walk distance matrix and $P_A$is the transition probability matrix for graph G_a.

The right hand side of equation (3) requires large number of matrix multiplications to compute random walk distance matrix R_A^l. Since the matrix multiplications in a finite series of higher power it is very difficult, so consider the square matrix $B = (1-c)P_{\mathrm{A}}$and use the following property of square matrix [Manning *et al* (2008)].

Property: IfB is a square matrix and $I - B$ is invertible, then the sum of the finite series of a square matrix is given by

$$\sum_{i=0}^{k} B^i = (I-B)^{-1}(I - B^{k+1}) \quad (4)$$

where, I is the identity matrix.

When the entries of B are very small in magnitude, $(I-B)^{-1}$ is approximately equal to $(I+B)$. Now, apply the property (equation (4)) in the equation (3) which becomes

$$R_A^l = c\,\{I - (1-c)P_A\}^{-1}[I - \{(1-c)\,P_A\}^{l+1}] \quad (5)$$

The equation (5) can be easily computed since the number of matrix multiplications is less as compared with that of equation (3).

In case, if the matrix $C = \{I - (1-c)\mathrm{P_A}\}$ is not invertible, the above method cannot be directly applied to solve the problem. One of the best techniques to compute matrix inverse of a noninvertible matrix is Pseudo inverse [Penrose (1956)].

Consider the matrix $C = U\sum \mathrm{V^T}$and its Pseudo inverse is$C^{-1} = V\sum^{-1}\mathrm{U^T}$. The Pseudo inverse$\sum^{-1}$of the diagonal matrix $\sum$ is computed by taking the reciprocal of each non zero elements on the diagonal and keeping the zero diagonal elements.

4. CLUSTERING PROCESS

Clustering process has the duty of separating the data into different clusters with same or different

characters. Group clustering has been studied by many experts in different directions based on either attribute edges or structure edges. Xu *et al* (2007) have analyzed graph partitioning on topological structures.

Tian *et al* (2008) have discussed graph clustering algorithm on vertex attributes. Most of the researchers on clustering grouping have analyzed homogeneous graphs. The appropriate formulation and techniques are not enough available to distinguish the different developments of attribute and structure edges since the attribute augmented graph is heterogeneous.

The unified neighborhood random walk model covers all paths through structure as well as attribute edges in clustering attributed graph. By the principle of this model, if two vertices belong to the same cluster, then the random walk distance is too lengthy, but if two vertices are placed in different clusters, then the random walk distance is small or tending to zero. This shows that there is no neighborhood random walk path between two vertices.

4.1 Selection of centroids

The selection of good initial centroid is more powerful than that of randomly selected initial centroids. Good initial centroids are easily partitioning clustering algorithms. If the l-step neighborhood of a vertex $\gamma_i$is dense, then many vertices are reachable from $\gamma_i$within l steps and its probability is very high. Otherwise, vertex $\gamma_i$is not agood one. In order to select the centroids, define the density function of vertex.

The density function of a vertex $\gamma_i$is the sum of the influence functions of $\gamma_i$on all vertices in V.
The influence function is stated as

$$f_B^{\gamma_j}(\gamma_i) = 1 - e^{-\frac{1}{2\sigma^2}\{d(\gamma_i,\gamma_j)\}^2} \quad (6)$$

Hence, the density function is written as

$$f_B^D(\gamma_i) = \sum_{\gamma_j \epsilon V}\left[1 - e^{-\frac{1}{2\sigma^2}\{d(\gamma_i,\gamma_j)\}^2}\right] \quad (7)$$

It is noted that the influence of γ_i on γ_j is proportional to the random walk distance from $\gamma_i$toγ_j. We know that larger random walk distance gives more influence. If γ_i has a large density value, then γ_i connects to many vertices.

By using the density functions given in equation (7), the vertices are arranged in descending order of their densities and select the top$k$ vertices whose initial centroids are stated as $\{c_1^0, c_2^0, \dots, c_k^0\}$. After a large number of iterations are performed, the k centroids in the t^{th}iteration are$\{c_1^t, c_2^t, \dots, c_k^t\}$.

Let c*$\in \{c_1^t, c_2^t, \dots, c_k^t\}$ be a closest centroid with largest random walk distance fromγ_i.

$$c^* = arg\ \max_{c_j{}^t}\ d(\gamma_i, c_j{}^t) \quad (8)$$

The centroid must be updated with the most centrally situated vertex in each cluster. For evaluating that vertex, the average point $\overline{\gamma_l}$of a cluster $V_i$is obtained by using the relation

$$R_A^l(\overline{\gamma_l}, \gamma_j) = \frac{1}{|V_i|}\sum_{\gamma_k \epsilon V_i} R_A^l(\gamma_k, \gamma_j), \quad \gamma_j \epsilon V \quad (9)$$

Here,$R_A^l(\overline{\gamma_l})$is the average random walk distance vector for clusterV_i. The centroid (c_i^{t+1})in cluster $V_i$in the $(t+1)^{th}$iteration is formed as

$$c_i^{t+1} = \arg\min_{\gamma_j \in V_i}\left\|R_A^l(\gamma_j) - R_A^l(\overline{\gamma_l})\right\| \quad (10)$$

The equation (10) shows that the random walk distance vector of the centroid c_i^{t+1} is the closest to the cluster average. It is noted that the clustering process iterates until the clustering objective function converges.

4.2 Weight self-adjustment

As defined earlier, $\omega_0$isthe weight of structure edge and $\omega_1, \omega_2, \dots, \omega_m$are the weights of attribute edges which are relative to ω_0. Initially fix the values of $\omega_0$and $\omega_i(i = 1, 2, \dots, m)$. Assuming $\omega_0^0 = 1.0$ and $\omega_1^0 = \omega_2^0 = \cdots = \omega_m^0 = 1.5$.

Let $W^t = \{\omega_1^t, \omega_2^t, \dots, \omega_m^t\}$be the weights of attribute edges in the t^{th}iteration. An increment $\Delta\omega^t$is weight update of attribute $a_i$between the $t^{th}$and$(t+1)^{th}$iterations. The weight of $a_i$in the $(t+1)^{th}$ iteration is defined as the average of weight in the t^{th}iteration and its increment. That is,

$$\omega_i^{t+1} = \frac{1}{2}(\omega_i^t + \Delta\omega_i^t) \quad (11)$$

For determining the weight increment $\Delta\omega_i$, design a vote mechanism under the following condition. If a_i has a good clustering tendency, then the weight ω_i of $a_i$is increased. Otherwise, the weight ω_i should be decreased.
Define vote measure as

$$vote_i(\gamma_p, \gamma_q) = \begin{cases} 1, & \text{if vertices share the same value on } a_i \\ 0, & \text{otherwise} \end{cases} \quad (12)$$

The numbers of vertices within clusters that share attribute values with the centroids on $a_i$are counted andthen the increment weight $\Delta\omega_i^t$ is computed by using the relation:

$$\Delta\omega_i^t = \frac{\sum_{j=1}^k \sum_{\gamma \epsilon V_j} vote_i(c_j, \gamma)}{\frac{1}{m}\sum_{p=1}^m \sum_{j=1}^k \sum_{\gamma \epsilon V_j} vote_p\,(c_j, \gamma)} \quad (13)$$

Substitute the equation (13) in the equation (11) and get the expression for the adjusted weight $(\omega_i^{t+1})$of the i^{th} attribute $a_i$in the$(t+1)^{th}$ iteration as

$$\omega_i^{t+1} = \frac{1}{2}\left[\omega_i^t + \frac{m\sum_{j=1}^{k}\sum_{\gamma\epsilon V_j} vote_i(c_j,\gamma)}{\sum_{p=1}^{m}\sum_{j=1}^{k}\sum_{\gamma\epsilon V_j} vote_p(c_j,\gamma)}\right] \quad (14)$$

The adjusted weight ω_i^{t+1} is changed as compared with the weight in the previous iteration based on the value of $\Delta\omega_i^t$.

If $\Delta\omega_i^t >= < \omega_i^t$, then $\omega_i^{t+1} >= < \omega_i^t$ respectively. This implies that the attribute $a_i$makes an increasing, equal or decreasing contributions to the random walk distance. It is concluded that for increasing iterations, the weight of an attribute increases and the random distance between vertices which have same attribute value on $a_i$increases.

5. EVALUATION PROCESS

In the literature, there are many clusters according to the nature of data. In graph clustering, there are only structure and attribute clusters. They are embodied in the experiment to compare and identify the efficient cluster by using the methodologies given in the sections 4 and 5 for the dataset given under.

5.1 Viewers' dataset: 17, 356 viewers

We use the data drawn from the net in connection with the number of viewers of research papers and the number of research papers published in PLOS ONE. 17,356 viewers who viewed the papers during one month period are selected and 700 papers are selected which are published from 2007 January to 2011 December. We build a graph in which the nodes represent the viewers and edges represent the research papers relationship.

5.2 Clusters

Here, we briefly define the clusters and important measures.

- **S-cluster:** This baseline clustering algorithm considers topological structure only. Random walk distance is used to measure vertex closeness.
- **SA-cluster:** This algorithm considers both structural and attribute similarities.
- **W-cluster:** This algorithm combines both structural and attribute similarities through the distance function given in equation (1) whose weighted factors are considered as $\alpha = 0.4$ and $\beta = 0.6$

For analysing the quality of clusters $\{V_i\}_{i=1}^{k}$, we use the density and entropy measures whose formulae are respectively given as

$$D[\{V_i\}_{i=1}^{k}] = \frac{1}{|E|}\sum_{i=1}^{k}\left|\{(\gamma_p,\gamma_q)|\gamma_p,\gamma_q\epsilon V_i j(\gamma_p,\gamma_q)\epsilon E\}\right| \quad (15)$$

And

$$E[\{V_i\}_{i=1}^{k}] = \sum_{r=1}^{m}\frac{\omega_r}{\omega}\sum_{j=1}^{k}\frac{|V_j|}{|V|}\left\{-\sum_{n=1}^{n_r}\varepsilon_{rjn}\log_2\varepsilon_{rjn}\right\} \quad (16)$$

where$\varepsilon_{rjn}$is the percentage of vertices in cluster j on attribute $a_r$with value a_{rn}.

5.3 Interpretation of results

Based on the methodologies given in sections4, 5, 6.1 and 6.2, the experiments are performed on Matlab using java program

- From the experiments, the following results are gathered. Figure-1 [Table-1] reveals that the density of clusters decreases as the number of clusters (k) increases irrespective of three clusters. The density of SA cluster is less than that of S cluster. The density of W cluster is very low as compared with that of other two clusters in each k.
- Similarly, Figure-2) [Table-2)] shows that the entropy of clusters decreases as the number of clusters increases. As in the case of density, the entropy of SA cluster is much lower than the entropy of S cluster but greater than that of Wcluster.In this juncture, it is difficult to apply weighted distance function in W cluster to achieve a good balance between attribute and structural similarities.
- Since the cluster qualities based on both density and entropy of SA cluster lies between that of S and W clusters, the cluster qualities of SA cluster are computed iteratively, in terms of density and entropy on viewer's data set. Figures 3(a) and 3(b) show that the cluster qualities improve in terms of density and entropy iteratively. It shows that weight self-adjustment method is the effective method for analyzing the improvement of cluster quality. In each iteration, cluster quality improves based on both density and entropy when the number of clusters (k) increases. For increasing iteration, cluster quality improves on entropy for fixed values of k. It is concluded that the cluster quality on entropy achieves more improvement but that on density does not achieve upto the level of entropy.
- Figure-4) [Table-4)] shows the trend of the weight on viewers data set for different k values. The prolific weights decrease as the values of k increase at each iteration point. On the other hand, weight for each k increases as iteration increases while for higher k with higher iteration, the weights may decrease. It is observed that, when k is small $(k = 50)$, a cluster with many viewers mixed up has a diverse distribution of research papers. It implies that nodes

are not served as good clustering attributes when k is very small.

- Figure-5) [Table-5)] shows the effectiveness of three clusters. It is observed that W cluster is 2 times slower than S cluster and SA cluster is 2 to 6 times slower than S cluster. This reveals that S cluster is having faster running times and concluded that S cluster is the most effectiveness as compared with SA and W clusters.

Table-1. Cluster quality comparison using density on viewers dataset.

Cluster / *k*	***S***	***SA***	***W***
50	0.56	0.47	0.38
100	0.50	0.45	0.30
150	0.48	0.42	0.22
200	0.45	0.40	0.17
250	0.43	0.40	0.15

Table-2. Cluster quality comparison using entropy on viewers dataset.

Cluster / ***k***	***S***	***SA***	***W***
50	3.25	2.65	1.20
100	3.10	2.30	0.50
150	2.90	2.10	0.50
200	2.86	2.00	0.15
250	2.85	2.00	0.10

Table-3. Cluster quality on viewers dataset using density and entropy.

k / **Iteration**	**50**		**100**		**150**		**200**		**250**	
	D	**E**	**D**	**E**	**D**	**E**	**D**	**E**	**D**	**E**
1	0.62	3.53	0.54	3.14	0.47	2.75	0.41	2.52	0.38	2.25
2	0.58	2.97	0.51	2.95	0.42	2.31	0.37	2.00	0.35	1.81
3	0.56	2.92	0.46	2.81	0.36	2.28	0.34	1.82	0.31	1.71
4	0.53	2.85	0.47	2.76	0.38	2.24	0.32	1.74	0.29	1.53
5	0.55	2.82	0.47	2.76	0.40	2.21	0.30	1.61	0.28	1.42
6	0.55	2.82			0.40	2.21	0.33	1.55	0.26	1.37
7							0.33	1.55	0.28	1.32
8									0.28	1.32

Table-4. Weights of attribute edges on viewers dataset.

k / **Iteration**	**50**	**100**	**150**	**200**	**250**
1	1.53	1.47	1.44	1.39	1.31
2	1.68	1.54	1.50	1.47	1.36
3	1.74	1.57	1.53	1.51	1.45
4	1.81	1.70	1.64	1.58	1.40
5	1.88	1.76	1.68	1.53	1.34
6	1.92	1.80	1.74	1.42	1.31

Table-5. Effectiveness of clusters.

Cluster / *k*	*S*	*SA*	*W*
50	75	182	150
100	90	273	175
150	105	550	200
200	110	645	260
250	120	732	294

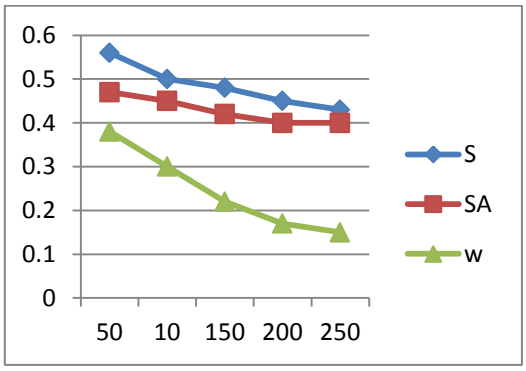

Figure-1. Cluster quality comparison using.

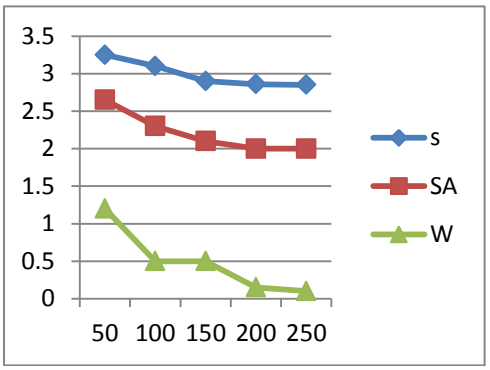

Figure-2. Cluster quality comparison using.

Density on viewers dataset entropy on viewers dataset

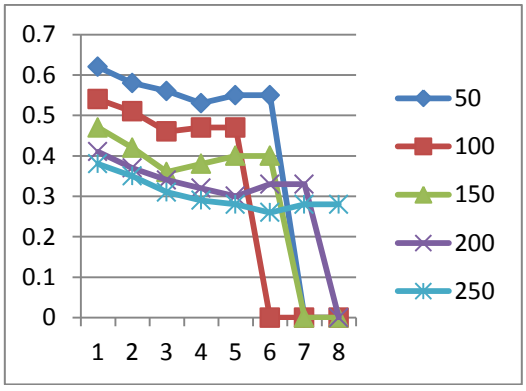

Figure-3(a). Cluster quality using density on viewers dataset.

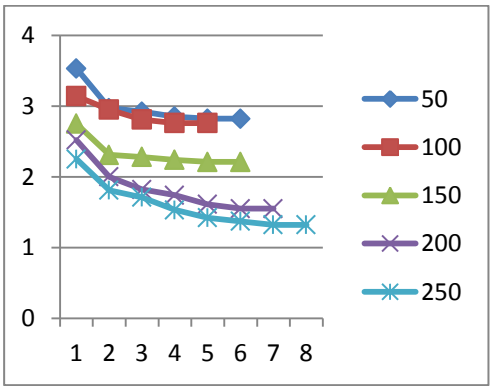

Figure-3(b). Cluster quality using entropy on viewers dataset.

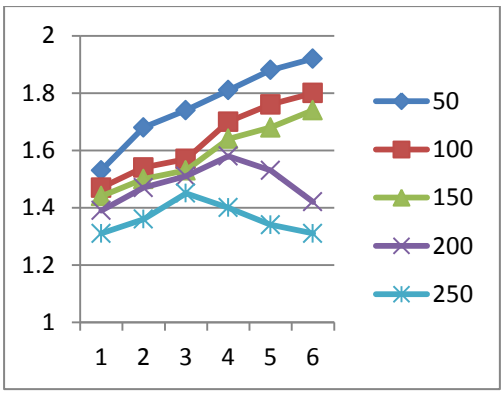

Figure-4. Weight on viewers dataset.

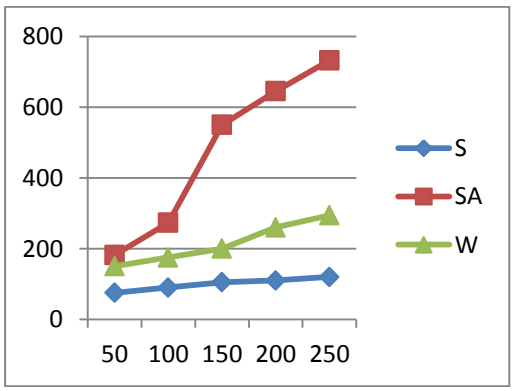

Figure-5. Effectiveness of clusters.

6. CONCLUSIONS

In this paper, the problem of large graph clustering is analyzed and the cluster quality is to be identified. The unified distance measure, mathematical concept, clustering process, clusters, density, entropy, weighted self-adjustment and dataset are explained. Cluster quality of clusters are compared and inferred that cluster quality of SA cluster lies between that of other two clusters. The test revealed that cluster quality based on entropy achieves more improvement. The trend of the weight on dataset is discussed. The quality cluster is justified based on density and entropy measures.

REFERENCES

Cheng, H, Zhou Y, and Yu, J. X. 2011. 'Clustering Large Attributed Graphs: A balance between Structural and Attribute Similarities', TKDD. Vol. 5, No: 2, Article12 (1-33).

Estivill-Castro, V and Yang, J. 2000. A Fast and Robust General Purpose Clustering Algorithm', Pacific Rim Intl. conf. on Artificial Intelligence, Vol. 1886, pp. 208-218.

Farley, C and Raftery, A. E. 1998. How many clusters? Which clustering method? Answers via Model based cluster analysis', Technical report No: 329, Dept. of Statistics University of Washington. 41(8): 578-588.

Han, J and Kamber, M. 2001. Data Mining: Concepts and Techniques', Morgan Kaufmann Publishers.

Jain, A. K, Murty, M. N and Flynn, P. J. 1999. Data Clustering: A Survey', ACM computing Surveys. 31(3): 264-323.

Jeh, G and Widom, J. 2002. SimRank: a measure of structural-context similarity' Proceedings of the ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining (KDD'02). pp. 538-543.

Manning, C.D, Raghavan, P and Schutze, H. 2008. Introduction to Information Retrieval' Cambridge University Press.

Ng, R and Han, J. 1994. Very large databases', Proceedings of the 20th Intl. Conf. on very large databases, VLDB endowment Berkeley CA, 144155.

Orme, B and Johnson, R. 2008. Improving $K-$Means cluster analysis: Ensemble analysis instead of Highest Reproducibility Replicates. Proceedings of the 2008 Saw tooth software conference, Sequim WA.

Pons, P and Latapy, M. 2006. Computing communities in large networks using random walks, Journal of Graph Algorithms and Appns. 10(2): 191-218.

Penrose, R. 1956. On best approximate solution of linear matrix equations' Math. Proc. Cambridge Phil. Soc. 52: 17-19.

Raj, Pand Singh, S. 2010. A survey of clustering techniques', Intl. J. Comp. Appns. 7(12): 1-5.

Strehl, A and Ghosh, J. 2002. Clustering Ensembles - knowledge reuse framework for combining multiple partitions', J. Machine Learning Research. 3(2): 583-617.

Shi, J and Malik, J. 2000. Normalized cuts and image segmentation' IEEE Trans. Pattern Analysis and Machine Intelligence. 22(8): 888-905.

Sun, J, Faloutsos, C, Papadimitriou, S and Yu, P.S. 2007. Graphscope: parameter-free mining of large time-evolving graphs' Proc. Of the ACM SIGKDD Intl. Conf. on Knowledge Discovery in Databases (KDD'07). pp. 687-696.

Tian, Y, Hankins, R.A and Patel, J.M. 2008. Efficient aggregation for graph summarization' Proceedings of the ACM-SIGMOD Intl. Conf. on Management of Data (SIGMOD'08), pp. 567-580.

Tong, H, Faloutsos, C and Pan, J.Y. 2006. Fast random walk with restart and its applications' Proceedings of the Intl. Conf. on Data Mining (ICDM'06), pp. 613-622.

Tsai, C.Y and Chui, C.C. 2008. Developing a feature weight self-adjustment mechanism for a k-means clustering algorithm' Computational Statistics and Data Analysis. 52: 4658–4672.

Xu, X, Yuruk, N, Feng, Z, andSchweiger, T. A. J. 2007. Scan: A Structural Clustering Algorithm for Networks' Proceedings of the ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining (KDD'07). pp. 824-833.

Zanghi, H, Volant, S and Ambroise, C. 2010. 'Clustering based on random graph model embedding vertex features', Pattern Recognition letters. 31(9): 830-836.

Zhou, Y, Cheng H and Yu, J. X. 2006. Graph clustering based on structural/Attribute similarities', PVLDB. 2(1): 718-729.