APPLICATION OF CLUSTERING TO ANALYZE ACADEMIC SOCIAL NETWORKS

K. Sobha Rani[1], KVSVN Raju[2] and V.Valli Kumari[3]
[1]Dept. of Information Technology, MVGR College of Engineering, Vizianagaram.
sobharani.t@gmail.com
[2]Anil Neerukonda Inst. of Technology & Sciences, Visakhapatnam.
kvsvnraju@gmail.com
[3]Department of CSSE, College of Engineering, Andhra University, Visakhapatnam.
vallikumari@gmail.com

ABSTRACT

Social network is a group of individuals with diverse social interactions amongst them. The network is of large scale and distributed due to involvement of more people from different parts of the globe. Quantitative analysis of networks is need of the hour due to its' rippling influence on the network dynamics and in turn the society. Clustering helps us to group people with similar characteristics to analyze the dense social networks. We have considered similarity measures for statistical analysis of social network. When a social network is represented as a graph with members as nodes and their relation as edges, graph mining would be suitable for statistical analysis. We have chosen academic social networks and clustered nodes to simplify network analysis. The ontology of research interests is considered to measure similarity between unstructured data elements extracted from profile pages of members of an academic social network.

KEYWORDS:

Social Network Analysis, Clustering, Graph Mining, RDF

41. INTRODUCTION

Different kinds of network exist viz. social, technological, business etc., all of which share similar attributes like being distributed, continuously growing and of large scale. There are some interesting quantifiable measures that help analyze these networks, like number of nodes, connectivity, centrality, clustering coefficient and degree distribution [14]. These networks can be modeled as random graphs, scale-free networks and hierarchical networks. Due to the property of being scale free, the social networks continuously expand with addition of new nodes and the relation amongst the nodes vary with time and frequency of interaction between them. Also the probability of a node being influential in the network is not uniform due to their difference in attributes and interactions. The social network data is not even suitable to be stored in relational database; hence different approaches are followed to store and analyze unstructured social network data.

1.1 ACADEMIC SOCIAL NETWORKS

With the growth of the internet and the World Wide Web, social networks have become influential. One of the most effective channels for obtaining information is the informal network of collaboration of colleagues and friends etc. The use of social networks is widespread in employers' recruiting and in workers' job-seeking, pursuit of hobbies and building collaboration within or between organizations. A person can have different types of information: personal profile with fields like homepage, field of interest and hobbies, contact information including

address, email, telephone and fax number. However, the information is usually stored in heterogeneous and distributed web pages. The web pages follow a standard framework called as RDF [26] to represent properties of members. With the development of social network and the wide range of services provided by them, members have transformed from content consumers to content providers, which has lead to vast amounts of data to be stored and processed.

Visual representation of social networks is important to understand the network structure and identification of its members and their attributes. Social network analysis maps relationships between individuals in social networks. Such individuals are often persons, but may be groups, organizations, web sites or citations between scholarly publications. Information about the relative importance of nodes and edges in a graph can be obtained through centrality measures, widely used in disciplines like sociology. For example, eigenvector centrality [13] uses the eigenvectors of the adjacency matrix to determine nodes that tend to be frequently visited.

Academic social networks provide a platform for scholars / researchers to publish their work and share knowledge with their peers. In addition to this they can create and update their own profiles. These networks also rank the researcher's achievement based on their efforts as a contributor to the network. Some sample academic social networks include; Arnetminer(www.arnetminer.org) [23], that provides comprehensive search and mining services for researcher in social networks, Microsoft Academic Research (academic.research.microsoft.com) that is a free academic search engine developed by Microsoft Research Asia [23]. They provide many innovative ways to explore scientific papers, conferences, journals and authors, connecting millions of scholars, students.

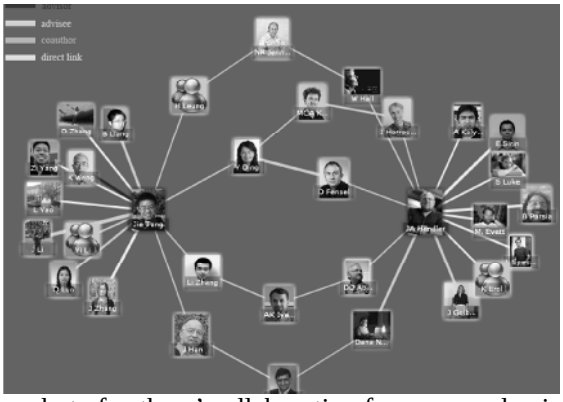

Figure 1. Snapshot of authors' collaboration from an academic social network

1.2 GRAPH REPRESENTATION OF SOCIAL NETWORKS

As stated by Han[18], information of a social network is heterogeneous and the multi-relational data can be represented as a graph or network. Nodes/Vertices (V) are objects i.e. members of the network and Edges(E) are links which are either uni-directional or bi-directional that represent the relationship or interaction amongst the nodes. The complexity of mining real world datasets w.r.t. traditional data mining is that the data is multi-relational, heterogeneous and semi-structured. Our approach of clustering simply depends on the principle that adjacent nodes of a selected node in a graph, also tend to be adjacent to each other, i.e. if B and C are adjacent nodes of A, there is a probability of B and C to be adjacent. Sub-Graph pattern mining is an important method which helps in clustering analysis to simplify the dense social network. Network parameters are identified by Network Analyzer [20] as number of connected components, network diameter, path length, number of neighbors, density etc. The complex parameters that are derived from the above are degree distribution, neighborhood connectivity, clustering coefficient, betweenness centrality, closeness centrality etc.

1.3 CLUSTERING

Clustering, also referred as unsupervised classification is the process of identifying groups of nodes in a network based on the similarity of attributes of the nodes. This process also redefines the network topology which makes the network analysis simple, fast and efficient. The clustering of nodes in the network results in identification of sub-graphs which will give us scope to have an intense analysis at individual nodes and the network properties using graph metrics.

The important metric needed for our analysis, the *clustering coefficient* is defined by [19], [21] as a measure of degree to which nodes in a graph tend to cluster together. The local clustering coefficient of node quantifies how its neighbors tend to form a complete graph, where as the global clustering coefficient is based on triplets of nodes. The average clustering coefficient of the network is the average of local clustering coefficients of all nodes in the network.
The local clustering coefficient c_i is defined as

$$c_i = \frac{|\{e_{jk}\}|}{k_i(k_i-1)} \quad v_j v_k \quad N_i e_{jk} \quad E$$

Where k_i is the out-degree of vertex i, and
$N_i = \{v_j \; e_{jk} \; E\}$, is the set of neighbors of vertex i.

However, as the academic social network is treated as an undirected graph due to the collaboration amongst them, c_i is normalized as $c_i' = 2c_i$

Different nodes in the network can be clustered based on frequently used similarity measures like Euclidean Distance, Cosine Similarity and Jaccard Coefficient. [22]

Euclidean distance is the distance between any two points (x,y) in the sample space of multiple dimensions.

$$d(x,y) = \sqrt{\sum_{k=1}^{n}(x_k - y_k)^2}$$

Euclidean distance will be small for almost similar documents and very high for non-similar web pages. The reason behind this is that all terms in the document may not be similar. Hence, we use pair wise similarity measures for matching tag values in the RDF format of web document.

Cosine similarity is used to compare two documents that are represented as vectors. Jaccard coefficient computes the probability that two nodes *x* and *y* will have a common neighbor *k*, given *k* is a neighbor of either *x* or *y*. This metric is used to compute document similarity in information retrieval and it does not consider term frequency and placement order.

$$\cos(x,y) = \frac{x.y}{\|x\|.\|y\|} \qquad \text{Jaccard}(x,y) = \frac{x \cap y}{x \; y}$$

Hence we are comparing the values for the selected attribute tags in RDF document.

1.4 FOAF SPECIFICATION

Resource Description Framework(RDF) is a family of specifications maintained by World Wide Web Consortium to represent meta data in XML format. FOAF (Friend of a Friend) is an RDF based schema to describe persons and their social network in a semantic way[15]. The design objective of FOAF is to allow integration of data across different applications. The project is based on machine readable web home pages for people, companies and other data. As social networks deal with events that occur between nodes of the network, the object oriented treatment of a node contains different attributes and methods. All this data is standardized based on RDF framework and the FOAF vocabulary. An RDF dataset contains a triple of *subject*, *predicate* and *object.*

For example, the sample data element is of the form <*foaf:Person*> *John* </*foaf:Person*> represents the *subject* of the RDF data set as foaf(Acronym for Friend Of A Friend), *predicate* or the property as "Person" and it's *object* as the value of the property with "John". Some of the FOAF vocabulary definitions selected for our analysis of academic social network are presented in Table 1. For example *foaf:knows* attribute specifies that our selected node know the person specified in those tag definitions, *foaf:interest* specifies the research interests of the respective node.

Table 1. Elements of FOAF document considered to measure similarities.

foaf:Document	foaf:Person	foaf:knows
foaf:Image	foaf:Homepage	foaf:mbox
foaf:Organization	foaf:interest	foaf:name
foaf:PersonalProfileDocument	foaf:mbox	foaf:publications

Sample RDF Document from FOAF project

```
<rdf:RDF
        xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
        xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
        xmlns:foaf="http://xmlns.com/foaf/0.1/"
        xmlns:admin="http://webns.net/mvcb/">
<foaf:PersonalProfileDocumentrdf:about=]"">
        <foaf:makerrdf:resource="#me"/> <foaf:primaryTopicrdf:resource="#me"/>
        <admin:generatorAgent rdf:resource="http://keg.cs.tsinghua.edu.cn/tj/cs/foaf_creator"/>
        <admin:errorReportsTordf:resource="mailto:jery.tang@gmail.com"/>
</foaf:PersonalProfileDocument>
<foaf:Personrdf:ID="me">
<foaf:name></foaf:name>
<foaf:title>Professor</foaf:title>
<foaf:homepagerdf:resource="http://www.cs.uiuc.edu/~hanj/"/>
<foaf:phonerdf:resource="tel:(217) 333-6903 "/>
<foaf:interest>Natural Language Processing</foaf:interest>
<foaf:interest>Machine Learning</foaf:interest>
<foaf:knows>
        <foaf:Person>
                <foaf:name>Wenmin Li</foaf:name>
                <foaf:homepage rdf:resource="http://arnetminer.org/person/wenmin-li644700.html"/>
        </foaf:Person>
</foaf:knows>
----------/*  list of all persons */
<foaf:publications rdf:resource="YizhouSun,HongboDeng,Jiawei Han: Probabilistic Models for Text
Mining.: Mining Text Data: 259-295"/>
----------  /* list of all publications */
</rdf:RDF>
```

The above FOAF document extracted from the academic social network[24] presents the co-authors and publications of a researcher along with his personal attributes like email id and contact number. The co-authors' names and homepage resources are specified in <foaf:name> and <foaf:homepage> which are presented as child nodes of <foaf:person> which in turn is the child node of <foaf:knows>.

Rest of the document is organized as follows. Section 2 gives the related work, and problem description is given in Section 3. Architecture of our proposed method is discussed in Section 4.

Experimental data set and result analysis are discussed in Section 5. We conclude our work in Section 6 with possible future scope of research.

2. RELATED WORK

Subgraph discovery in relational graphs to retrieve important subgraphs known as SkyGraph by [1] is done by considering number of vertices and edge connectivity to identify frequent patterns. Relation extraction in social networks using similarity developed by [2], the process clusters similar entity pairs according to the collective context in the web documents and assigns a new relation label. Peter Mika [3] demonstrated the application of semantic web technology using FOAF documents and RDF framework to analyze social networks by considering graph metrics also for analysis. Social Action Prediction by [4] have used SNA measures to aggregate networks using link structure and find patterns across different links.

Model based clustering proposed by [5] states that the probability of a tie between two actors depends up on distance between them in an unobserved Euclidean social space. I-HsienTing[6] has described the different types of web mining i.e. web content mining, web structure mining and web usage mining. Information propagation in a social network based on strong and weak ties is demonstrated by [7] based on clustering coefficient. Frequent pattern mining proposed by [8] searches for typical patterns of structural change in dynamic social networks. In the work by Peter Mika [9], Semantic annotation of Wikipedia is implemented using Natural Language Processing techniques. Alan Mislove[11] addresses the issue of large scale measurement study and analysis of the structure of multiple online social networks. Co-author relation is represented in matrix form by [12] to cluster set of coauthors. User profile representation and personalized content retrieval proposed by [16] uses semantic clustering to identify user clusters.

Of all the works mentioned above, multiple similarity measures were not considered and the clustering techniques will not be compatible to different categories of social networks. But, our proposed model covers different similarity measures applied to the basic attributes of FOAF document and the retrieval process involve graph mining techniques. This model suits wide categories of social networks, as the RDF frame work is standardized representation of service offered by the social networks.

3. PROBLEM DESCRIPTION

Academic social networks an offshoot of social networks provide a channel through which researchers and academicians know about the current trends of research and find experts in the concerned domain. In some applications like Arnetminer[24], Microsoft Academic Research[23] etc. the representation of connectivity highlights the relation of co-authorship by identifying researchers as nodes and relations as edges in the graph. Due to the huge number of people involved in the highly dense network, it is difficult to identify the group of researchers working in a single domain. The present offered services are limited to profile extraction and graphical representation, which need lot of time to scan the entire database and produce the required result. However, if the graph is clustered based on research interests and further on the similarity of their profiles, it would make the process of information retrieval faster and efficient.

The attributes of a researcher are shown in Fig.2, where the different attributes like affiliated university, research interests would be considered for measuring similarity to form cluster of users from the existing dense network.

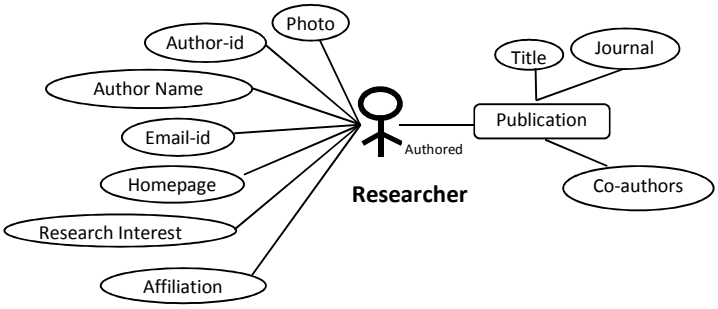

Figure 2. Researcher Profile

The member of the academic social network, i.e. the author/researcher is represented with the multi dimensional dataset with possible values as specified here.

A={‘name’, ’email-id’, ’affiliation’, ’research interests’(R), ’publications’(P),’coauthors’(C)} where
R is a set representing identified research interests like
= {‘data mining’, ‘natural language processing’, ‘ontology, ’machine learning’…}
P =Set of publications with attributes of “title”, “journal” and “co-authors”.
And
C=Set of coauthors collaborated for publications

In the representation, the first two attributes of name and email are considered as unique attributes to identify a researcher in the network and an ID is assigned based on this key pair. The research interests are again a set of values as a researcher may work in multiple disciplines. In order to measure similarity between these set of attributes we have considered the ontology graph by Peter Mika[3] Fig.3 represents the part of ontology graph of research interests, which shows the relation between different research topics in the domain of computer science. The last two attributes are interrelated because publications have list of co-authors as attributes and author’s profile page has their corresponding list of publications. Each profile page in RDF format has elements of <foaf:knows> to represent co-authors and these names can also be extracted from value specified in <foaf:publications>

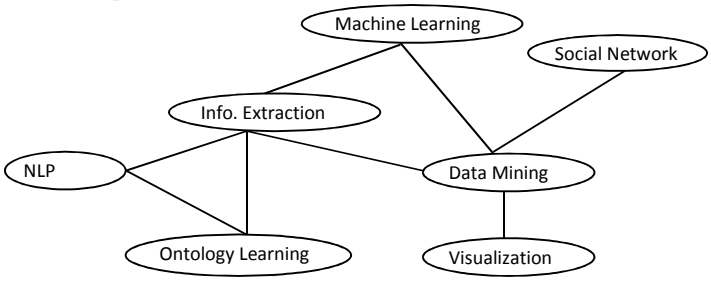

Figure 3. Ontology of Some Research Interests

After completing the preprocessing phase of data, the next phase of problem to be addressed is to represent the attributes in vector format and then compute the similarity among them. The attributes are categorical, hence we have to transform them to binary for faster access and to reduce the storage space. The affiliation and research interest attributes are to be enumerated to transform them to binary format. The publications and co-authors have to be uniquely identified to minimize duplication of data and the IDs can be further used for information retrieval and attribute matching.

The similarity measures we have used for our clustering principle are i) Euclidean distance(d) ii) Jaccard Coefficient(j) and iii) Cosine Similarity(cos(x,y)). To have a good initiative and wide scope for analysis, we have to first identify target nodes of the entire network. The graph metric called Clustering Coefficient(c_i) can be used to identify most influential nodes in the entire network. A node tends to cluster more nodes in the neighborhood, if it has more frequency of interaction with adjacent nodes in the network. Nodes with high clustering coefficient should be considered as there is scope for getting more data of co-authors and publications. Also, a least frequently accessed node in the network may be termed as an outlier in the analysis as it does not tend to be in a cluster with any prominent node.

The final and important issue to be considered is the matching criteria of the attributes based on the RDF specifications. The properties and the corresponding RDF tags are specified in Table 2 and comparison process of attribute values should not consider the position of occurrence of the matching pattern as the frequency and relative position will not have any influence on the similarity. The resultant vectors of attributes can be sequentially processed to compute the similarities among the profile pages.

Table 2. Criteria for similarity measures

Criteria	**Property**	**RDF specification**
C1	belong to same affiliation	foaf:Organization
C2	share same set of research interests	foaf:interest
C3	co-author a publication	foaf:Publications
C4	same adjacent node in the network i.e. has same co-author	foaf:knows

4. ARCHITECTURE

Our framework started with selection of profile documents of researchers from an academic social network and extracting useful attributes for analysis as shown below.

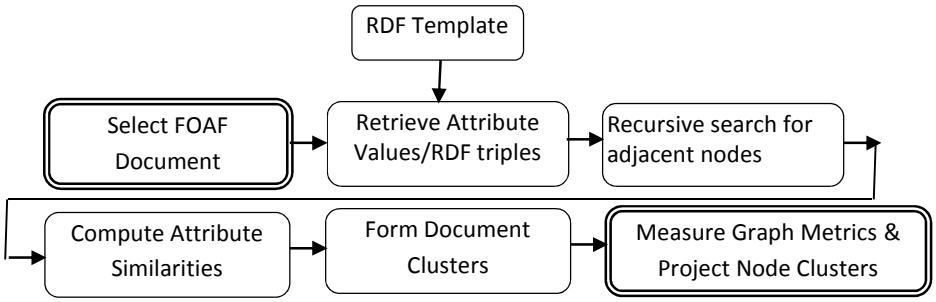

Figure 4. Data Pre-Processing and Clustering

4.1 ATTRIBUTE RETRIEVAL

The profile documents in FOAF format are collected and stored by giving a unique ID to the document as well as the node in the network. Then it is parsed using RDF query language to retrieve selected attributes and stored as vectors representing tokens of RDF format. The affiliation is a single attribute, but the other attributes are having multiple values due to the activities of the researcher. The vectors of publications, research interests and coauthors are stored as binary format shown in Table 3a. and 3b.

Table 3a. Vector of Author Vs Research Interests

Author ID	Affiliation	Research Interests(0,1)							
		R1	R2	R3	R4	R5	R6	R7	R8
49618	MSRA	1	0	1	1	0	0	0	1
49695	UI	1	0	1	0	1	1	0	1
41529	USC	0	1	1	0	0	0	1	0

Table 3b. Vector of Author Vs Publications

Author ID	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
37780	1	1	1	1	1	0	1	1	1	1
42219	1	1	0	1	1	0	0	1	1	0
37736	0	1	0	0	0	0	0	0	0	0
37670	1	0	0	1	0	1	0	0	0	0

4.2 RECURSIVE NODE SEARCH

Once the co-authors are identified, based on their frequency of interaction i.e. number of publications their documents are selected and parsed. To minimize the data to be processed and neglect the passive nodes, we have taken the primary attribute of publication count as the filtering criteria. Publications retrieved are again stored as a vector containing their own ID and the authors' ID. A 2D vector of publications vs. authors containing binary attributes is formed and stored for further analysis.

4.3 SIMILARITY COMPUTATION

In this phase of similarity computation, the binary attributes are compared and three similarity measures of i) Euclidean distance (Sim_{ed}) ii) Jaccard Coefficient (Sim_{jc}) and iii) Cosine Similarity(Sim_{cs}) are computed for each document with reference to all other documents. The traditional DBSCAN method is used for clustering as this technique does not need prior information about the number of clusters to be formed. Also, this technique is insensitive to the ordering of the poins

4.4 GRAPH PROJECTION

The original data set and the newly formed clustered data are now given as input to NodeXL[25] and graphs are projected to compute the graph metrics for the network. It is observed that graph density and clustering coefficient decreases due to the reduction in number of nodes in an individual cluster. Resultant graphs are shown in Figure 6.

5. DATASETS AND RESULT ANALYSIS

The experimentation was performed on different data sets that were taken from Arnetminer and Co-astro physics datasets and we acquired an efficient mechanism for clustering the dense graph. Our analysis with data set downloaded from[24] comprise author names as a pair that indicate a collaboration i.e. a research publication, the snapshot of few instances is shown in Table 4 and Table 5. We have formed a vector by extracting attributes from researcher profile that were shown in Fig. 2

The data set sample shown in Table 4 contain 898 records of association instances of research publications resulted due to collaboration of 350 authors in the area of "Data Mining". Different authors' publication details in research interests of "machine learning", "information retrieval" etc. are shown in Table 5.

Table 4.Sample Associations for a research interest of "Data Mining"
(Author names are shortened for ease of access)

S.No.	Author-1	ID-1	Author-2	ID-2
1	JY	37671	DW	47586
2	JP	37729	DS	46574
3	TF	37802	ST	43265
4	SC	37921	RR	46647

Table 5. Publications, Coauthors and Affiliation details of some nodes in the network

S.No.	Author ID	Research Interests	Publications ID	Coauthors' ID			
1	37669	Machine Learning, NLP	P23,P56,P45,P27	37802	47586	89618	89695
2	37729	Statistical Learning	P45,P34,P76	37921	86574	89491	521667
3	37802	NLP, Information Retrieval	P23,P74,P34	43265	37669	47948	49887
4	37921	Statistical Learning, Info.Retrieval	P32,P45,P58, P97	89541	37729	95060	37842

Fig. 5 represents a sample representation of egocentric network of researchers visualized using NodeXL[25]. The colors of the node represent the grouping based only on research interests and edge weight represents the #papers coauthored by them. Fig. 6.a. represents the dense graph representing all nodes of the network, where as Fig.6.b. represent the simple subgraph identified by selecting nodes of a single cluster. The other geometric shapes in the figure in different colors represent the other clusters that are collapsed before visualization.

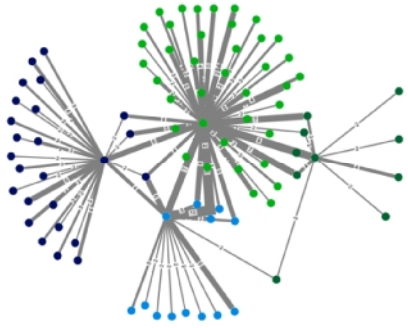

Figure 5. Sample Egocentric network representing collaboration of researchers

For computation of distance measures, the author Vs. research interests and author Vs publications are converted as adjacency matrices having binary attributes of 0 or 1.

Terms compared for similarity between RDF formats of different FOAF documents:

a) <foaf:organization> = { "Microsoft Research Asia", "University of California", "University of Wisconsin",……}
b) <foaf:interests>={"Data Mining", "Natural Language Processing", "Statistical Analysis",……}
c) <foaf:knows> → <foaf:name> = { "JW", "AS",…….}
d) <foaf:publications>{"Yizhou Sun,Hongbo Deng,Jiawei Han: Probabilistic Models for Text Mining: Mining Text Data", "Zhenhui Li,Jingjing Wang,Jiawei Han: Mining event periodicity from incomplete observations",……..}

For example, let us take for the set of research interests, the vector of authors 37669 and 37729 are {1,0,1,1,0,0,1} and {1,0,0,0,1,1,0,1}, the Cosine Similarity measure is computed as given as

$$Sim_{cs} = \frac{(1*1 + 0*0 + 1*0 + 1*0 + 0*1 + 0*1 + 0*0 + 1*1)}{\sqrt{(1^2 + 0^2 + 1^2 + 1^2 + 0^2 + 0^2 + 0^2 + 1^2)} + (1^2 + 0^2 + 0^2 + 0^2 + 1^2 + 1^2 + 0^2 + 1^2)} = 0.71$$

Here the set of research interests is represented by binary values of 0 and 1, where the presence of 1 represent the researcher having the specific research interest. Using this similarity measure the clustering algorithm is implemented and the document clusters are formed. Similarly the algorithm also computed the other two measures of Euclidean distance and Jaccard Coefficient. The results are compared and integrated for further analysis. Sample calculation of Euclidean distance for the above example is given here.

$$Sim_{ed} = \sqrt{((1-1)^2 + (0-0)^2 + (1-0)^2 + (1-0)^2 + (0-1)^2 + (0-1)^2 + (0-0)^2 + (1-1)^2} = 2$$

The clustering process is repeated by considering the other attribute matrices of publications and co-authors and the end result is the integration of all similarity measures for efficient clustering. All the different dimensions are mapped to their equivalent binary attributes for simplifying our analysis. The results are compared and crosschecked with "Entropy" for error minimization and to measure the quality of clustering.

Table 6. Similarity Measures for some documents

S.No.	**Document Pair**	**Euclidean Distance** Sim_{ed}	**Jaccard Coefficient** Sim_{ic}	**Cosine Similarity** Sim_{cs}
1	37669 & 37729	2	0.32	0.71
2	37802 & 37729	2.76	0.54	0.35
3	37669 & 37802	2.45	0.48	0.38

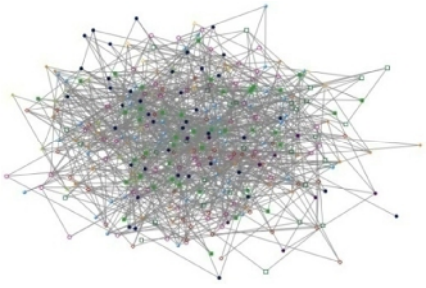

Fig.6a. Nodes of Complete Graph

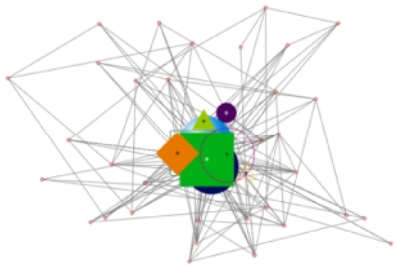

Fig.6b. Nodes of a single clustered graph

The resultant graph measures are given in Table 6, which project the variation in those measures caused due to clustering. Graph metrics of complete graph and clustered graph are computed and tabulated in Table 7 and 8. It is observed that with small size of the cluster, the measures and in turn the graph properties do not vary. But as the size of cluster increases, the properties like clustering coefficient will vary from that of the original graph.

Table 7.Graph Metrics of the clustered network

Graph Metric	Value
Graph Type	Undirected
Vertices	274
Unique Edges	500
Max. Vertices/Connected Component	272
Max. Edges in a Connected Component	499
Maximum Geodesic Distance (Diameter)	11
Average Geodesic Distance	4.3
Graph Density	**0.014**
Modularity	0.545
Average Degree	3.6
Average Betweenness Centrality	449.6
Average Closeness Centrality	0.008
Average Clustering Coefficient	0.02

Table 8.Graph Metrics of each cluster

Group id	#Vert	#Edg	Graph Density
G1	34	43	**0.077**
G2	29	36	**0.089**
G3	37	47	**0.071**
G4	18	18	**0.118**
G5	14	14	**0.154**
G6	27	35	**0.100**
G7	25	28	**0.093**
G8	27	33	**0.094**
G9	22	26	**0.113**
G10	22	22	**0.095**
G11	17	19	**0.140**
G12	2	1	**1.000**

6. CONCLUSION & FUTURE SCOPE

We have proposed a mechanism of clustering that can be applied academic social networks. The similarity measures we have considered for clustering are suitable for the attributes of such a dataset. However, our process has to be improvised in order to be compatible to any social network due to the standard format of different unstructured data and its versatile attributes represented by RDF framework. We hope that with a little effort from academia, the traditional data mining algorithms will be applicable to suit the current requirements of infinitely large scale social networks and satisfy the search criteria of millions of users of social networks.

7. BIBLIOGRAPHY

[1] Apostolos N. Papadopoulos, ApostolosLyritsis and YannisManolopoulos, *"SkyGraph: an algorithm for important subgraph discovery in relational graphs"*, Data Min Knowl Disc (2008), DOI 10.1007/s10618-008-0109-y

[2] Junichiro Mori, Takumi Tsujishita, Yutaka Matsuo and Mitsuru Ishizuka, *"Extracting Relations in Social Networks from the Web Using Similarity Between Collective Contexts"*,ISWC 2006, LNCS 4273, pp. 487–500, 2006. Springer-Verlag Berlin Heidelberg 2006

[3] P. Mika, *"Flink: Semantic Web Technology for the Extraction and Analysis of Social Networks,"*Journal of Web Semantics, vol. 3, no. 2, 2005, pp. 211–223;

[4] Adam Perer, Ben Shneiderman, *"Balancing Systematic and FlexibleExploration of Social Networks"*, IEEE Trans. On Visualization and Computer Graphics, Vol. 12, No. 5, Sep/Oct'06

[5] Mark S. Handcock, Adrian E. Raftery and Jeremy M. Tantrum, "*Model-based clustering for social networks*", J. R. Statistical Society, A (2007) 170, Part 2, pp. 301–354

[6] Ting, I. H., "*Web Mining Techniques for On-line Social Networks Analysis*", In Proc. of the 5th Intnl. Conf.on Service Systems and Service Management, Melbourne, Australia, 30, 2008

[7] Zhao, Jichang, Junjie Wu, XuFeng, HuiXiong, and KeXu. "*Information propagation in online social networks: a tie-strength perspective*." Knowledge and Information Systems: 1-20.

[8] Bringmann, Bjoern, Michele Berlingerio, Francesco Bonchi, and ArisitdesGionis. "*Learning and predicting the evolution of social networks*", Intelligent Systems, IEEE 25, no. 4 (2010): 26-35.

[9] Mika, Peter, MassimilianoCiaramita, Hugo Zaragoza, and JordiAtserias. "*Learning to tag and tagging to learn: A case study on wikipedia*." IEEE Intelligent Systems 23, no. 5 (2008): 26-33.

[10] Luis R. Izquierdo, Robert A. Hanneman, "*Introduction to the Formal Analysis of Social Networks using Mathematica*"

[11] Mislove, Alan, MassimilianoMarcon, Krishna P. Gummadi, Peter Druschel, and Bobby Bhattacharjee. "*Measurement and analysis of online social networks*", InProc. of the 7th ACM SIGCOMM conf. on Internet measurement, pp. 29-42. ACM, 2007.
[12] Said, Yasmin H., Edward J. Wegman, Walid K. Sharabati, and John T. Rigsby. *"RETRACTED: Social networks of author–coauthor relationships*", Computational Statistics & Data Analysis 52, no. 4 (2008): 2177-2184.
[13] On-line textbook by Robert A. Hanneman and Mark Riddle, "*Introduction to Social Network Methods*", Department of Sociology at the University of California, Riverside.
[14] Charu C Aggarwal, "*Social Network Data Analytics*", e-ISBN 978-1-4419-8462-3DOI 10.1007/978-1-4419-8462-3 Springer, New York.
[15] http://xmlns.com/foaf/spec/20100809.html
[16] Cantador, Iván, and Pablo Castells. "*Building emergent social networks and group profiles by semantic user preference clustering*", In Proc. of the 2nd International Workshop on Semantic Network Analysis, pp. 40-53. 2006.
[17] http://lrs.ed.uiuc.edu/tse-portal/analysis/social-network-analysis/
[18] Han Jiawei Han and Micheline Kamber, , "*Data Mining: Concepts and Techniques*", 2nd Ed. ISBN 1-55860-901-6
[19] http://projects.skewed.de/graph-tool/doc/clustering.html
[20] http://med.bioinf.mpi-inf.mpg.de/netanalyzer/help/2.6.1/index.html
[21] http://en.wikipedia.org/wiki/Clustering_coefficient
[22] http://inside.mines.edu/~ckarlsso/mining_portfolio/similarity.html
[23] www.academic.microsoft.research.com
[24] http://aminer.org
[25] http://nodexl.codeplex.com/
[26] http://www.foaf-project.org
[27] Batagelj, Vladimir, and Andrej Mrvar. "*Pajek—analysis and visualization of large networks*." In Graph Drawing, pp. 8-11. Springer Berlin/Heidelberg, 2002.

Authors

K.Sobha Rani is currently working as an Associate Professor in the Department of Information Technology, MVGR College of Engineering, Vizianagaram, Andhra Pradesh. She has a teaching experience of 8 years and has practical exposure to work on different platforms since 1996. She is a research scholar working in the areas of data mining, privacy preservation and social networks.

Dr.KVSVN Raju is one of the first generation educationists of CSE in India and is a Professor & Director (R&D) at Anil Neerukonda Institute of Technology and Sciences(ANITS), Bheemunipatnam, Andhra Pradesh. Earlier he worked for 32 years at Dept of CSSE, College of Engineering, Andhra University, as Assistant Professor to Professor. His research areas include Data Engineering, Security Engineering, Software Engineering and Web Engineering and he supervised so many Ph.D scholars in these areas. His professional body member ships include IEEE, IETE, CSI, ISTE and Inst. of Engineers.

Dr. Valli Kumari is a Professor at Dept. of CSSE, College of Engineering, Andhra University and is also Head of the Department, Dept. of Computer Engineering, College of Engineering for Women, AU. She has a total of 23 years teaching experience. She won "Gold Medal" for *Best Research* in 2008. She is a fellow of IETE and life member of CSI, ISTE, CRSI etc. Her research areas include Network Security, Privacy Preservation, Image Processing and Web Mining. She is certified by Microsoft in VC++ and IBM in DB2 and is well versed with so many other technologies.