COMMUNITY DETECTION AND MINING IN SOCIAL MEDIA

Aishatu Ibrahim Birma and Vandi Musa Valentine

Department of Computer Science and Software Engineering, XJTLU University, Suzhou, China.

ABSTRACT

Participating networks and social media have emerged by mobilizing people in many creative ways over the last decade. Millions of users are activating, identifying, working and socializing on the Internet, expressing new types of cooperation, communication and intelligence that have not been imagined much earlier. Social Media do impact ideas and emotions, as well as lot of fake news and rumours. With some of the characteristics mentioned above are my reasons for choosing this topic in other to have a clear understanding on how SNA are used to analyse massive amount of data.

KEYWORDS

Community detection, Social media, large scale network, vertices, clusters.

1. INTRODUCTION

Community detection is an important tool to analyse complex networks, facilitating the study of mesoscopic groups often associated with organizational structure and network performance. In today's situation, social media is a new field for many researchers. In social media, the data generated on the user side is huge. In order to maintain user generated data, there are many mining tasks in social media mining.

There are many social networking sites where users build their own communities according to their interests. As we all know, social media is a huge virtual world, many users have their personal data, and they connect to different types of groups. To understand the user's behavior, we need to understand the user's background. It is not easy to identify single use behaviors in social networks, so community detection is needed in social networks. Many researchers have done a lot of work in the field of social network.

The problem that community detection attempt to solve is the identification of cluster and vertex that are more densely connected to each other than to the rest of the network. Detecting and analyzing the nature of social networks has led to important discoveries ina wide range of domain. Finally the extensive use of social media such as Flicks, Twitter, YouTube, Facebook, WeChat etc. are responsible for creating multiple networks.

2. PROBLEM DEFINITION

Social networks are huge and complex, because most researchers do not consider the community knowledge stratum, that is, the ground truth in their methods is well known to everyone, and everyone has its important role in the formation of community and social groups. This is what

David C. Wyld et al. (Eds): ICAITA, CDKP, CBW, SOFT, CMC, Signal, MLT-2023
pp. 73-81, 2023. CS & IT - CSCP 2023 DOI: 10.5121/csit.2023.130606

researchers can do better to work through community testing with two parameters; influence and user attributes. This is where a social network that can maximizes its impact. The main focus in this work is detection. On the effect of user's influence on the community. It is well known that the most influential users with this increase in community mobility affect more community detection problems i.e. the scalability of large networks.

3. LITERATURE REVIEW

Word "Community" has been widely used in various contexts in literature and with different connotations. Social studies is probably the oldest institution to use the concept of community to refer to a group of people who share an interest or activity (e.g., a community of practice). Once networks are widely accepted and accepted as a way of studying social interactions and processes (Wasserman and Faust [109], Scott [97]), the concept of community is related to networks of human actors, making it possible to exhibit structural propertiesof the specified combining.

3.1. Elements of a Social Media network

Ecosystems of social media use include different types of objects that are connected through various types of interactions and relationships.Social media networks provide elegant representation of social media data, including online objects as their vertices and their relationships / interactions as edges. The social network vertices can represent different types of participants, such as users, content items (e.g. posts, images, videos) and metadata items (such as topic categories, tags). Furthermore, the characteristics of social networks can vary, such as flexibility, weight, paths and multi-directionality (i.e., more than two entities), depending on the processes of network creation.

3.2. Social Media Network Creation

In practice, the creation of social media network begin with the process of transactions carried out and recorded in social media applications.Each of these transactions usually involves different entities; for example, the label allocation in the Facebook involves a user, photos, and labels, while comments on blog articles involve commentators, the article and review texts. In this way, a link (edge) is created between the elements of the same transaction on the underlying network, and the resulting social media network creates a direct subset of online transactions.

3.3. In this Research, the following Hypothesis will be Conceded in Order to

Demonstrate form of collaboration. Sway of opinion and emotions.

What are people using it for?

Studying of human interaction and collaborative behaviour in an unparalleled scale.

4. DATA SETS

4.1. Data Set Descriptions

Dataset is a set of data in the form of a table. The columns represent attributes while the rows represent an entry of a person. The types of data sets been used such as the name attribute is from Facebook, Twitter and instagram. The combination of data set is between the collaboration of

members created, but it is divided into subgroups because of some disputes. This data set is published online by the Hawaii kaggle. Data set descriptions include:
Contains 250 nodes. Have edges.

- Is a directed type of graph.
- Is dynamic in nature, nodes have the ability to learn. Properties change over time, nodes can adapt.
- Is unwieghted i.e. breadth-first search of the graph ensures that when we first reach node v, we can be sure that we have found the shortest path to it.

Attributes	Description
Name	It is one of the attributes of dataset it shows the uniqueness of the node
Age	It indicates the range at which people are clustered based on purpose .i.e. there choice to purpose of usage.
Gender	It also help in indicating often used and in this dataset it is indicated as male=1 and female=0.
Collaboration	It show how the community where been detected and it is indicated such as Facebook=1, Twitter=2, and Instagram=3.
Often used	It showsspecific use duration such as per day, week and month. It is indicated as per day=1, week=2 and month=3.
Purpose of use	It indicate what the individuals are using it for i.e. news=1, friendship=2 and lastly Businesses =3.

5. METHODOLOGY

5.1. Evaluation Measures

Evaluation measures are the parameter through which communities detected by the algorithm is as per the ground truth or not. In this analysis I will be using the normalized mutual information. Mutual information is used to determine the similarity between two clustering's from different datasets. Computer-generated benchmarks are often used to initiate networks of well-defined communities. In other to be able to detect how nodes form communities and connect to each other;

Link analysis: is a method used to evaluate relationships between networks. Contacts can be identified between different nodes (objects), including organizations, people, and deals.

Community detection: network nodes can easily be grouped into a set of nodes (potentially overlapping) if any nodes are tightly bound nodes.

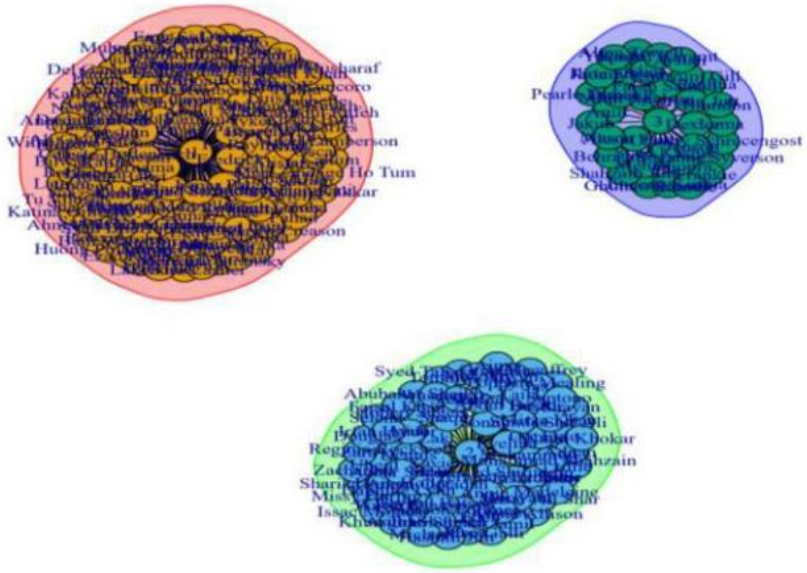
Proximity measures describe the similarity or dissimilarity between nodes.

Clustering of network using clustering functions. The coefficient of a vertex (node) in the graph indicates how close neighbours are to being a clique (complete graph) which are composed from a set of vertices and edges between them.

6. IMPLEMENTATION

6.1. Community Detection

It is formed by individuals such that those within a group interact with each other more frequently than with those outside the group. It detects groups in the network for which group membership of individuals is not explicitly provided.In this analysis We are using the collaborative attribute to detect how the community is formed based, implicit group that are formed by their social interaction. The algorithm use for the community detection is “clustering edge betweenness”. Below is a picture representation of the output clusters.

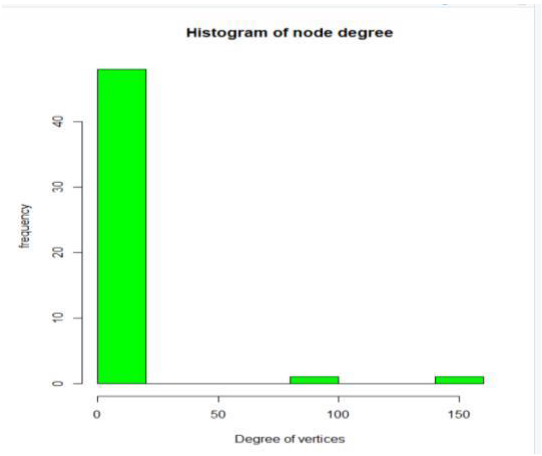


According to this picture the red cluster indicates the community form by Facebook while the green indicates the Instagram group and lastly the purple is based on Twitter group formed by the attribute “collaboration “in the dataset.

6.2. Link Analysis

It is assigning class labels to entities based on their link properties, e.g. iterative classification, relaxation labelling link-based object ranking (LOR). Correlate the quantitative assessments related to each factor with link-based metrics, e.g. Page Rank, HITS etc. Using link to establish higher –order relationships among entities. Such as

Degree of a node which is the number of edges incident on it. In-degree is the number of outgoing edges while Out-degree is the number of outgoing ones and the total degree is equal to in-degree + out-degree. So in this analysis the highest degree is 1 with the total number of connection as 193 nodes, followed by 2=160 then 3=107 and lastly 4=28 with the least.



There are 3 nodes with the degree less than 50 with less connection and there is some node which has a high degree between 150 and 200 that much connection.

Density the graph with high density is more connected and can resist link failures in this graph the density is

Edge betweenness below is a generated output from the age betweenness algorithm.

```
> edge.betweenness(network)
  [1] 0.1250000 0.1250000 0.1250000 0.2000000 0.3333333 0.2000000 0.5000000 0.3333333 0.1666667
 [10] 1.0000000 0.3333333 1.0000000 0.5000000 1.0000000 0.1250000 0.3333333 0.2000000 1.0000000
 [19] 1.0000000 0.5000000 0.2500000 0.5000000 1.0000000 0.2000000 0.5000000 0.3333333 1.0000000
 [28] 0.2000000 0.2000000 0.2500000 0.2000000 0.1250000 0.3333333 0.2500000 1.0000000 0.3333333
 [37] 0.2000000 0.3333333 0.3333333 0.2500000 0.2000000 0.1250000 0.3333333 0.2500000 0.3333333
 [46] 0.2500000 0.2500000 0.1428571 0.3333333 0.2000000 0.2000000 0.2000000 0.1666667 0.3333333
 [55] 0.5000000 0.2000000 0.1250000 0.5000000 0.1250000 0.5000000 0.3333333 0.3333333 0.2500000
 [64] 0.2500000 0.3333333 0.2000000 0.2000000 1.0000000 0.1250000 0.1250000 0.1250000 1.0000000
 [73] 1.0000000 1.0000000 0.2500000 0.2000000 0.2000000 0.3333333 0.2000000 0.3333333 0.5000000
 [82] 0.1250000 0.3333333 0.2000000 0.2500000 0.2000000 0.5000000 0.3333333 1.0000000 0.3333333
 [91] 0.1250000 0.1250000 0.2000000 0.3333333 1.0000000 0.1250000 0.3333333 0.2000000 0.1250000
[100] 0.3333333 0.3333333 0.1250000 0.2000000 0.1428571 0.2000000 0.1428571 0.5000000 0.1111111
[109] 0.3333333 0.1250000 0.2500000 0.2000000 0.1250000 0.3333333 0.1111111 0.2500000 0.2000000
[118] 0.5000000 0.1428571 0.2000000 0.5000000 0.5000000 0.1428571 0.3333333 0.5000000 0.5000000
[127] 1.0000000 0.2500000 0.1250000 0.2000000 1.0000000 0.5000000 0.5000000 0.1428571 0.3333333
[136] 0.1428571 0.3333333 0.5000000 0.1111111 0.1250000 0.5000000 0.1250000 0.1666667 0.2000000
[145] 0.2000000 0.1250000 1.0000000 0.3333333 0.2000000 0.1428571 0.5000000 0.1250000 0.1250000
[154] 0.1250000 0.1428571 0.2000000 0.2000000 1.0000000 0.2500000 0.1250000 0.2500000 0.3333333
[163] 0.1111111 0.1250000 0.2000000 0.2500000 0.5000000 0.2500000 0.2000000 0.2000000 0.2500000
[172] 0.1666667 0.5000000 0.1111111 0.5000000 0.2000000 0.2500000 0.2000000 0.2000000 0.2000000
[181] 0.5000000 0.2000000 0.2500000 1.0000000 0.1666667 0.5000000 0.3333333 0.2000000 1.0000000
[190] 0.1250000 0.2000000 0.1250000 0.3333333 0.2000000 0.1250000 0.2500000 0.2000000 0.3333333
[199] 0.2000000 0.1428571 0.2000000 0.1428571 0.5000000 0.1111111 0.3333333 0.1250000 0.5000000
[208] 0.2000000 0.1250000 0.3333333 0.1111111 0.2500000 0.2000000 0.5000000 1.0000000 0.2000000
[217] 0.2000000 0.2500000 0.1428571 0.3333333 0.5000000 0.1250000 0.5000000 0.2500000 0.1250000
```

Using the maximalCliqueEnumerator function it was detected that the no of maximalClique is =80. Meaning within a clique there are also sub-cliques.

Vertex Betweenness clustering and edge betweenness clustering we used this clustering algorithm to find the connection between nodes, similarity and connectedness.

6.3. Proximity Measures

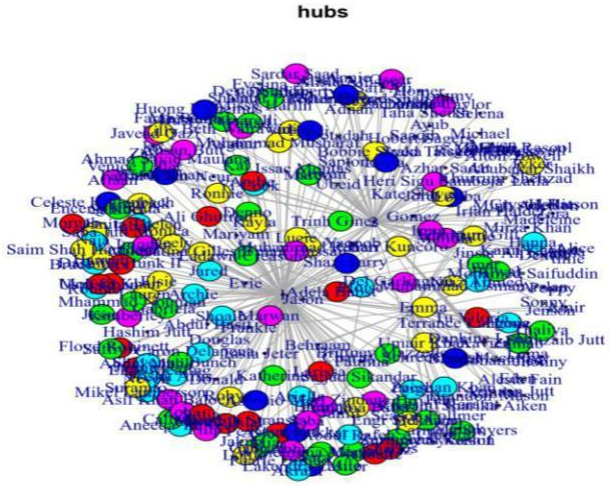
Page rank

Below is a picture of the page rank which depicts the vertex that has the highest page rank in this simulation. According to the result vertex 1 has the highest ranking in the network.

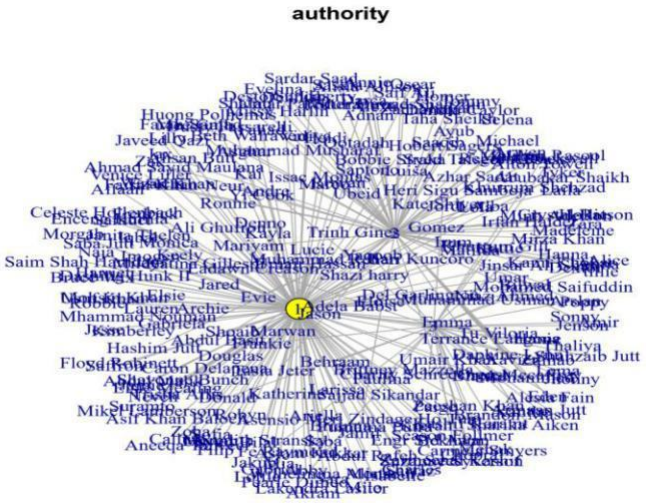
Strength as Page Rank compute ranking at crawling time so it responds to user queries faster. One of the weaknesses is that Rank sinks problem occur when network pages get in infinite connection loop.

HIT

Here the input is given by the adjacency matrix representing the set of items.The value defines the number of iterations to be performed and then an output of hub and authority is done. The hub and vector allow application to decide which vector is most interesting and highly efficient. Base on the picture below the vertex point to many other vertices (High out degree) in this case that is what makes it a hub.



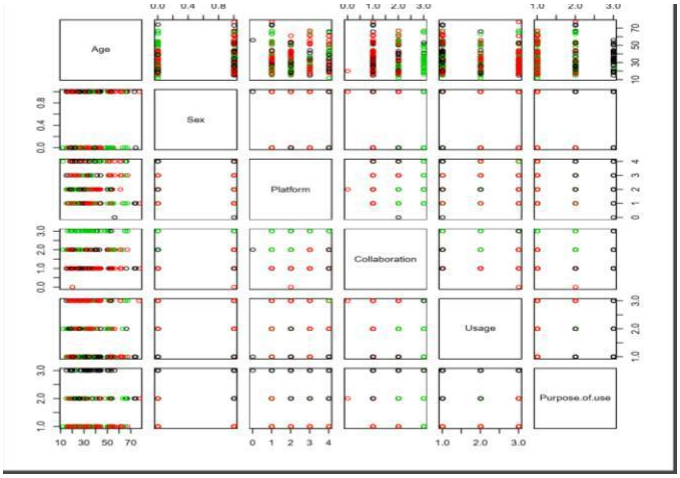
A vertex is considered an authority if it has many connected vertices at a (High indegree). While nodes marked in yellow in this diagram have high authority based on the high indegree pointing to it.



Its strengths lie in generating its ability to rank pages according to the query string, resulting in relevant authority and hub pages. Weakness of More Query Time: The query time evaluation is expensive. As HITS calculate rank of pages at query time,so queries take longer to respond.

6.4. Clustering of Networks

Clustering is based on similarity using the K-means clustering algorithm. Each cluster is associated with a centroid(center point), each node is assigned to the cluster with the closest centroid. On this section will be analysing attributes using K-means clustering. Below is a picture representation of the attributes clusters.



Note that the equation is clustered using k-means, on the result above each attribute has been iterated and clustered base on their connection to each other. Some of its strength are it detect outliers within the system. K-means clustering weakness is quality of classification measured by Rand index.

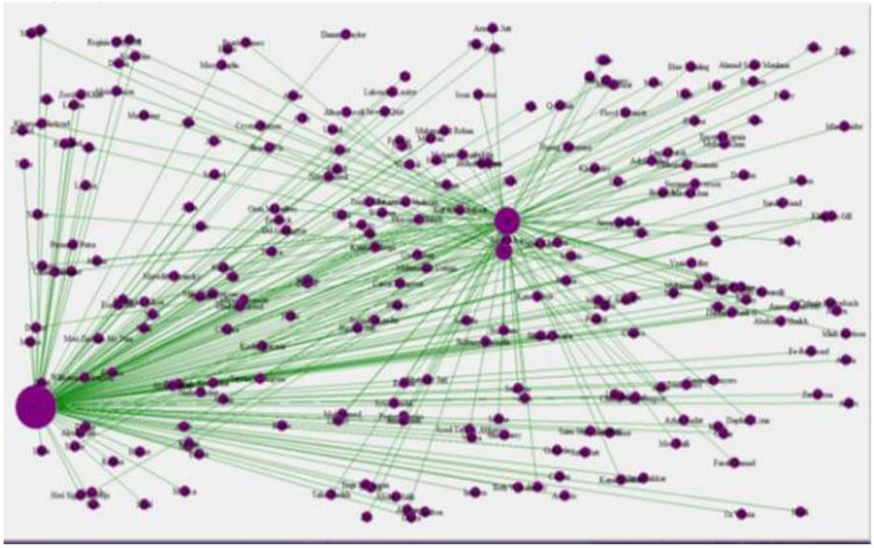
We also use SNN clustering, one of the main advantages of SNN is that, it considers not only the direct relationship between vertices, but also the indirect relationship.

7. RESULT

Base on the dataset been used and the analysis result, some of the hypothesis that should be conceded to demonstrate forms of collaboration are been detected, the method in which it has been identified based on the dataset is that they use different platform with the help of trending news, businesses advert and through friend of friend which help forming a new community. It is also identified that lots of people use the social media platform for mostly this reasons such as news. Social media is one of the fastest means of distributing information.

Another hypothesis that the research answer is the studying of human interactions and collaborative behaviour in an unparalleled scale. Due to coming together and interactions among members in a platform they get to easily study the behaviour of people around them by the frequent mode of communication between parties. This leads to sway of opinion and emotions among trusted members of a certain group or community even with the lack of face-to-face coming in contact with each other.

8. CONCLUSION



Above was the full graphical representation of the dataset using the name attribute before the analysis began. According to it the nodes are sparsely distributed and not much connection among vertices. In this study I have a clear understanding on how social network analysis(SNA) is done and how dataset is analysed using different algorithms and some certain function in other to correlate them, or even have a better understanding on how they are done, in real world using data mining. It also helps in understanding why people behave and react to some certain information's. Community detection also plays an important role in SNA world, which is one of the main reasons why data are been mined before storing to have a clear understanding about the world in action.

With the help of clusters it makes it easier and fast to do analysis base on the recommendation and classifications of attributes by the data set. At last R programming is one of the essential tools for data mining in social network even through it both has it strength and weakness on every algorithm been used, but they are mainly based on the type of dataset set you are using. Among all the clustering algorithm we found out that k-mean is the most essential and accurate function because it tends to find the centroid of each cluster and it keep iterating until a good spot is detected.

ACKNOWLEDGMENT

We would like to thank Dr. Kok Hoe WONG who assists in coaching; whose idea was helpful in building the project and Associate professor Paul Craig whose writing skills where really of great help throughout the process.

REFERENCES

[1] Shubha Chaturvedi, Anurag Jain , "Community Detection on Social Media: A Review Sweta Rai*, "*International Journal of Scientific Research Engineering Technology* , pp. Volume 3, Issue 2,2017.(Community detection in Social Media, 2018)

[2] Community detection in Social Media, 28 May 2018. https://www.researchgate.net/publication/233790771_Community_detec tion_in_Social_Media .pdf . (Tang, 2010)

[3] C. Giatsidis, " Graph Mining Tools for Community Detection & Evaluation in Social Networks & the Web,"*international world web confrence*, May 2013. (Oehlert)
[4] L. Tang, "Community Detection and Mining in Social Media Morgan & Claypool Publishers, Yahoo! Labs Huan Liu, Arizona State University," 2010. (Regina J.J.M Van Eijinden et. al, 2016)
[5] D. T. Wang, "A Beginner's Guide to Social Computing in Python Shanghai interenational study university.".
[6] G. W. Oehlert, Statistics 5401 34. Multidimensional Scaling School of Statistics 313B Ford Hall 612-625-1557.
[7] Regina J.J.M Van Eijinden et. al, "Computing in human behavior the socail mdia disorder scale," 2016. (Carley, 2003)
[8] A Novel Density based improved k-means Clustering Algorithm – Dbkmeans K. Mumtaz1 and Dr. K. Duraiswamy 2, 1 Vivekanandha Institute of Information and Management Studies, Tiruchengode, India 2 KS Rangasamy College of Technology, Tiruchengode, India. (Dudas)
[9] K. Wilson, Graph-based Proximity Measures Kevin A. Wilson, Nathan D. Green, Laxmikant Agrawal, Xibin Gao, Dinesh Madhusoodanan, Brian Riley, and James P. Sigmon, North Carolina State University.
[10] Pooja Devi1 , Ashlesha Gupta2 , Ashutosh Dixit3, "Comparative Study of HITS and PageRank Link based Ranking Algorithms," 2014.
[11] N. Grover, " Comparative analysis of page rank and hit algorithms ,instituted of technology and management," 2012.
[12] b. K. M. Carley, "Dynamic Network Analysis," 2003.
[13] Faraz Zaid et. al, "Analysis and Visualization of Dynamic Networks," 2014.
[14] b. P. M. Dudas, " Cooperative, Dynamic Twitter Parsing and Visualization for Dark Network Analysis".

AUTHORS

Aishatu Ibrahim Birma received MSc. Applied Informatics from Xi'an Jiaotong Liverpool University and B.Engr. Computer Science and Information Technology from Liaoning University of technology, in 2019 and 2017 respectively from PR, China. She is currently a Researcher, also a Lecturer with Borno State University, and her research interest includes AI and Data analysis.

Vandi Musa Valentine received MSc. Applied Informatics from Xi'an Jiaotong Liverpool University and B.Engr. Electronics and Information from Liaoning University of technology, in 2019 and 2017 respectively from PR, China. He is currently a Research Engineer with the Nigeria Defence space administration.