JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

A New Approach for Data Scrubbing in Social Networks

[1] Annamraju Priyanka, [2] Dr. Sreedhar Bhukya and [3] Dr. Ch. Niranjan Kumar
[1] M.Tech student, [2,3] Professors
Department of Computer Science & Engineering
Sreenidhi Institute of Science and Technology, Hyderabad, India

Abstract : Facebook is the largest social media platform globally with billions of monthly active users world wide.The enormous amount of information generated and gathered on the Facebook platform is referred by the Facebook .The process of identifying the inconsistencies, incomplete data and updating them is called as Data Cleaning or Data Scrubbing. This process is applied for Facebook data .Network analysis technology applied for the Cleaned data. This technology allows for identification of Key nodes and their relationships and the measurement of Network properties. By using the network, our proposed approach determine who the key players are, comprehend how the community is organised, find out about hidden trends, and learn more about how the network is structured overall. A branch of network analysis known as "social network analysis" (SNA) focuses on understanding social interactions and relationships between people or organisations by identifying unique users.

IndexTerms - Facebook data, Data Cleaning, Network Analysis, Social Network Analysis(SNA).

I. INTRODUCTION

The large amount of data generated and gathered on the Facebook platform is referred to as Facebook data[1][2][3]. This information includes multiple aspects of user involvement, activities, profiles, and content posted on the site. Facebook gathers and keeps this information for a variety of uses, such as enhancing user experience, focusing adverts, and performing data analysis. The different types of data include are user profile, social connections, posts and status updates, Interactions, messages and chats, Groups, events, advertisements, usage and behavioural data. Facebook users are constantly changing every day as the users are increasing. The problems faced in Traditional analytical methods are lack of understanding of structure, Incomplete Information. It was challenging to identify underlying patterns and dynamics in the data. Limitations in relational dependencies-based on forecasting or prediction of outcomes. It was difficult to visualize and describe intricate relationships and interactions in a way that made sense. Identifying the relationships between only the individuals is possible and difficult for larger ones. Fig[1] refers to connecting to the people through an application and forms a network .That connections might refer to family, friends, organisations or groups. It also observed that they might be friends mutually also. In general if someone checks your profile immediately it shows a notification that will get a suggestion to send a request. There is a possibility of even rejecting the request or just remove the request based on users interest. User can send request any time to anyone. If your account is public then chances of requests will be more than in private.

Figure1: FacebookData set

II. RELATED WORK

Data Cleaning[4] is the initial step in Data preprocessing[5]. Data Cleaning[4][12][13] process is first step applied for Facebook data [2][3]to clean by removing duplicate values ,error values, inconsistent records and replace them by modifying and deleting the dirty data. Data purification is done by using data wrangling tools. we can observe that data loading from different sources and collecting it in fig[2]. It represents the different data is then cleaned[5] and shown results in the fig[2] by reducing data storage and provision of quality data. The next step is data preprocessing [5][11] that is the data is converted into an adjacency matrix by using python programing language. That adjacency matrix is used to obtain a network graph. Adjacency matrix is generated because it refers of edge weights whether they are connected or not. Here comes our technology Network Analysis[6].Connecting the people through a network of image data set is done and referred in fig[3].Social Network Analysis [7][8][9]form a network which gives the person having highest number of connections and also provides the minimum connections that a single person can have.

Figure2: Data Cleaning Image

Figure3: Facebook data using Network analysis image

III. METHODOLOGY

3.1. Data Sources

we have downloaded the data set from Kaggle. The data set is a collection of ids generated at the time when one sends the request and the other accepts. It also includes information about the date and month of the id generated. It refers to repeated ids in the Facebook data[1][2]set. The data set that we are using Facebook data[3] 163383 rows and 3 columns. The data set is of year 2007.we are considering one month data that is January with 8152 rows and 3 columns. In these first and second column contains the ids of the people or friends. The third column refers to data, month and year at the time of id generated. we have downloaded the data set from Kaggle.

3.2. Data Cleaning Approach For Facebook

Data cleaning[4] is an iterative process that calls for thorough analysis and data manipulation. Prior to analysis or modelling , its seeks to increase data quality, reduce mistakes and increase the dataset’s general readability. Data cleaning[12][13] also known as data cleansing or data scrubbing[4].Python programming language is used .Its design philosophy prioritises code readability through the use of extensive indentation in accordance with the off-side rule. It is garbage-collected and dynamically typed. It supports a wide range of programming paradigms, including structured (especially procedural), object-oriented, and functional programming. Because of its extensive standard library, it is frequently referred to as a "batteries included" language. It may be accessed by anyone via www.python.org, which is its official website. The creation of new Python modules and functions is the passionate pursuit of a huge community[5] of people worldwide. The definition of "open-source" simply means that anyone can download the source code for nothing. An open-source web tool called Jupyter Notebook enables you to create and share documents with real-time code, equations, visuals, and text. For interactive computing and data exploration, it is a well-liked tool among data scientists, academics, and developers. Web-based interface for Jupyter Notebook allows for the creation and management of notebooks. The computing engine that runs the code in a notebook is known as a kernel. By selecting the cell and clicking the "Run" button, or by utilising keyboard shortcuts, you can run the code inside a code cell. The output is shown below the cell after the code has been delivered to the kernel for execution.

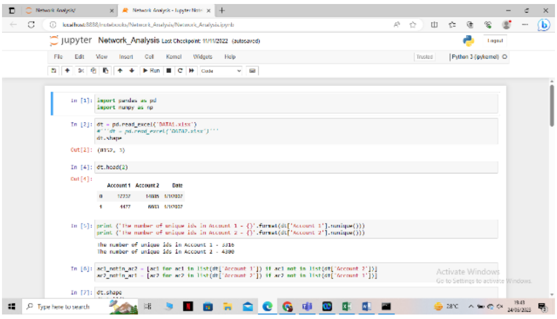
3.2.1. Remove Duplicate Data

Duplicate data occurs when the data is store multiple times in the data set .human error is the cause for multiple entry of data rather than storing it in a single cell. It refers to removal of repeated and just storing the individual values and ensure that value appears only once in the data after cleaning. Below refers to a process to remove duplicate values or generating unique values. These step by step process achieve unique values out of a list.

```
Step1:start
Step2:Create an empty list to store the unique values.
Step3:Iterate every element in original list.
Step4:Check if present value in the unique list.
Step5:If not matched then add it to unique list.
Step6:Return the unique list values.
Step7:End
```

3.2.2. Unique Values generation

By using the above process in terms of python code obtains unique values .This unique values gives you how many connected to the particular person and reduces the data set size. Fig[4] Represents code to generate unique values using python programming.

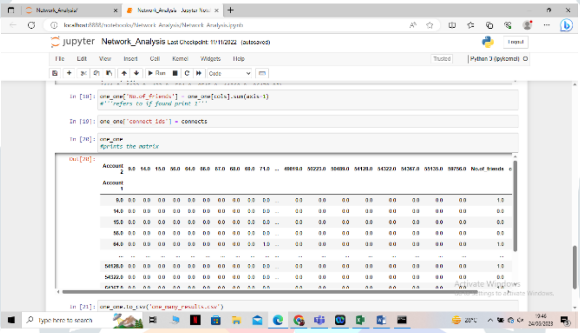

jupyter Network_Analysis Last Checkpoint: 11/11/2022 (autosaved)
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code
Logout

```
In [1]: import pandas as pd
        import numpy as np

In [2]: dt = pd.read_excel('DATA1.xlsx')
        #'''dt = pd.read_excel('DATA2.xlsx')'''
        dt.shape

Out[2]: (8152, 3)

In [4]: dt.head(2)

Out[4]:
     Account 1  Account 2      Date
  0      17237      14085  1/1/2022
  1       4477       6603  1/1/2022

In [5]: print ('The number of unique ids in Account 1 - {}'.format(dt['Account 1'].nunique()))
        print ('The number of unique ids in Account 2 - {}'.format(dt['Account 2'].nunique()))

        The number of unique ids in Account 1 - 3316
        The number of unique ids in Account 2 - 4380

In [6]: ac1_notin_ac2 = [ac1 for ac1 in list(dt['Account 1']) if ac1 not in list(dt['Account 2'])]
        ac2_notin_ac1 = [ac2 for ac2 in list(dt['Account 2']) if ac2 not in list(dt['Account 1'])]

In [7]: dt.shape
```

Figure4: Unique values generation in notebook

3.2.3. Creating Adjacency Matrix

The links or associations between nodes or vertices in a graph are represented by an adjacency matrix, which is a square matrix. A gragh in graph theory is made up of collection of vertices, also known as nodes, and set of edges, which join up pair of vertices. In unweighted graph values given as 0 or 1. Zero represent no edge or no connection and One represent there is an edge connecting to relevant vertices. Here in the Fig[5] adjacency matrix is loaded into a csv file that is further used for network formation. The downloaded matrix can be used in to identify the number of connections.


jupyter Network_Analysis Last Checkpoint: 11/11/2022 (autosaved)
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code
Logout

```
In [18]: one_one['No.of_friends'] = one_one[cols].sum(axis=1)
         #'''refers to if found print 1'''

In [19]: one_one['connect ids'] = connects

In [20]: one_one
         #prints the matrix

Out[20]:
 Account  9.0 14.0 15.0 56.0 64.0 86.0 67.0 68.0 69.0 71.0 ... 49819.0 50223.0 50489.0 54128.0 54322.0 54367.0 55135.0 59756.0 No.of_friends
       2
 Account
       1
     9.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0 ...     0.0     0.0     0.0     0.0     0.0     0.0     0.0     0.0           1.0
    14.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0 ...     0.0     0.0     0.0     0.0     0.0     0.0     0.0     0.0           0.0
    15.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0 ...     0.0     0.0     0.0     0.0     0.0     0.0     0.0     0.0           0.0
    56.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0 ...     0.0     0.0     0.0     0.0     0.0     0.0     0.0     0.0           0.0
    64.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  1.0 ...     0.0     0.0     0.0     0.0     0.0     0.0     0.0     0.0           1.0
     ...  ...  ...  ...  ...  ...  ...  ...  ...  ...  ... ...     ...     ...     ...     ...     ...     ...     ...     ...           ...
 54128.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0 ...     0.0     0.0     0.0     0.0     0.0     0.0     0.0     0.0           1.0
 54322.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0 ...     0.0     0.0     0.0     0.0     0.0     0.0     0.0     0.0           0.0

In [21]: one_one.to_csv('one_many_results.csv')
```

Figure5: Creation of Adjacency matrix in notebook

3.2.4. Minimum Number of connections

From the data set we got 5278 rows and 5278 columns as it is an adjacency matrix. the minimum is calculated by a formula '=min(Range)'.Gives you the minimum number of connections that is zero for the data set. Initially as it counts the zeroes also as individual element. after removal of zeroes as it doesn't have any connections with other they are removed hence our minimum number of connections is one.

3.2.5. Maximum Number of connections

Similarly, formula for maximum number of connections is '=max(Range)'.That is 29 is the maximum connections that two person is having 29 connections individually.

3.2.6. Average Number of Connections

Similarly, Formula for Average number of connections is '=AVERAGE(Range)'.that is1.543 is the average connections that one single person can have.

3.3. NETWORK ANALYSIS

A discipline of study called network analysis[6], commonly referred as Network science or Graph theory. Focuses on comprehending and examining the connections between different things or nodes. To gain important insights and knowledge, it entails investigating the patterns, structures, and dynamics of networks[14][15]. In Network analysis[6] objects are referred to as nodes, and the connections between them as edges. The relationships represented by these linkages might take many different forms, including friendships, partnerships, information exchange, or dependency. Network analysis[6] provides a framework to investigate the behaviour, influence and communication within complex systems by examining the pattern and characteristics of these relationships.

3.3.1. SOCIAL NETWORK ANALYSIS

A subfield of network analysis[6] called "Social Network Analysis(SNA)"[10] focuses on understanding the social links, structures, and interactions among people, groups, organisations, and other social entities, In order to analyse and visualise social networks[15] It offers a number of methodologies and approaches that expose the dynamics, patterns, and characteristics of social systems. Within a network node(Vertices) represent people, things, or social agents. They can be individual, groups, communities, or any other type of analytical unit. The interconnections or connections between nodes are represented by edges(Links).These ties can take many different forms, including friendships, teamwork, channels of communication, co-authorships, and other social ties. Measures of centrality quantify a node's significance or influence inside a social network[8].Degree centrality, betweenness centrality, and eigenvector centrality are common measurements of centrality. Visualisations are frequently used in social network analysis[9][10] to display and examine social networks[14]. Graph-based visualisations show edges as lines or curves and nodes as points or forms, making it possible to spot patterns, clusters, and important nodes.

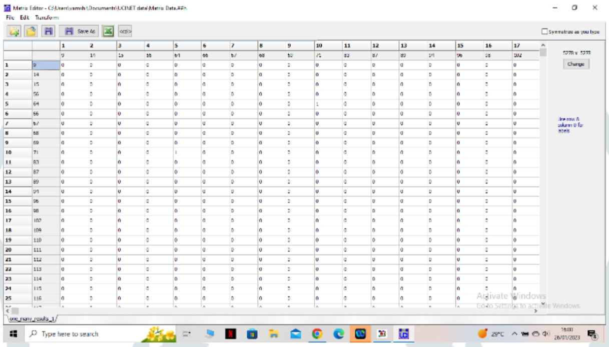
3.3.2. UCINET Application

Software for visualising and analysing social networks[15] is available under the name UCINET(University of California, Irvine Network) .For analysing and researching social networks[14], graph data, and relational data it offers a variety of tools and features. Researchers, social scientists, and analysts across fields frequently use UCINET.It offers a number of visualisation methods for graphical representations of social networks[15].Users can alter the properties , colours, and layouts of nodes and edges. Researcher exploration and understanding of network relationships and architecture is aided by visualisations.

3.3.3. Network formation

A potent tool for studying social networks [14][15] is UCINET, sometimes referred to as the UCINET software programme. First we have to upload the csv file of adjacency matrix to ucinet data. For that we need to use the extension. programme "##.h"or "##.d" only extensions with these kind will have the access in ucinet application.

Choose an appropriate data file with extension that you have store your adjacency matrix data. once the data is selected from main menu choose 'Net draw' and then from that select network .That will open the Net draw paper in which we can visualize the network and can adjust the coloring and nodes etc. In this process our adjacency matrix converted a network by net draw. We can zoom in/out, pan, select nodes/edges, and perform other actions to explore the network.



		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
		9	14	15	56	64	66	67	68	69	71	83	87	89	91	96	98	102
1	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	56	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	64	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
6	66	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	67	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	68	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	69	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	71	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
11	83	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	87	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	89	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	91	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	96	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	98	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	102	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	109	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	110	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	111	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	112	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
22	113	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	114	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
24	115	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
25	116	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure7 : Ids Connected

IV. RESULTS

In this section, we obtained a reduced data set that is initially taken 8152 rows and found the unique matrix is of 5278 rows and 5278 columns. Out of which obtained atleast one connection with unique ids are 3316 and unique ids with no connection is 1962.In fig[7] shows the ids connected to the other id. Table[1] represents ids and number of friends connected individually is shown as a sample .similarly we can identify the rest of the data connections. Therefore around 4836 rows been reduced. Based on that Adjacency matrix our network is formed in fig[7] and fig[8] shows how the persons connected without color labelling and with color labelling. hence our Network formed .

S.no	ID(Vertex)	No of friends or No of Edges	Connected ids
1	9	1	640
2	71	14	62,320,1643,1986,2195,2321 2330,2575,2769,3925,7043,8431 11998,15013
3	83	4	2982,6126,6272,26617
4	243	3	253,257,262
5	8644	9	3621,3630,3714,5030,8624,11784, 14097,28326,29089
6	11538	4	1594,5077,11844,36882
7	12581	1	3485
8	36122	3	9907,19488,37770
9	47229	2	47288,49819
10	59756	1	13606

Table1: Number of friends connected to each other

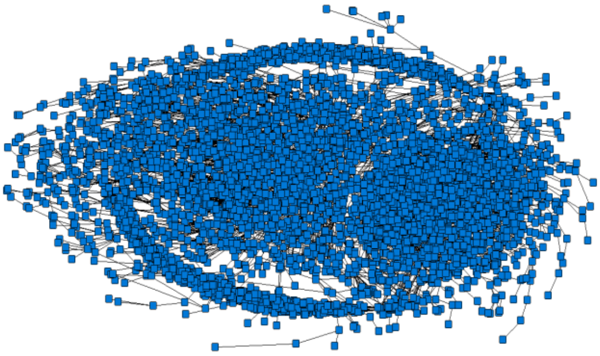


Figure8: Network Graph

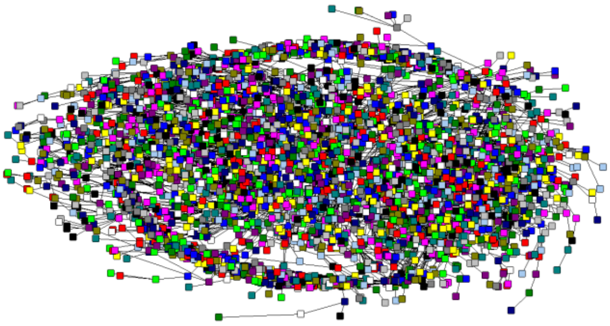


Figure 9: Network Graph with color labeling

V. CONCLUSION

In our work we have obtained a graph that provides the community graph[5] in a visualization manner that is social graph[10] that refers to the data connected with unweighted graph. It gives you about how the person having different relations with on single person and what are the criteria to become the maximum or highest connected person in the graph. From the initial data set there is a reduction of 4836 rows. It is applicable for larger data sets as well it can eliminate the data who does not have connection with other.

REFERENCES

[1] L. C. Boldt et al, “Forecasting Nike’s sales using Facebook data,”2016 IEEE International Conference on Big Data,Washington ,DC,USA,2016.
[**2**] R.R. Mukkamala, J.I. Sorensen,A.Hussain and R.Vatrapu, “Detecting Corporate Social Media Crises on Facebook Using Social Set Analysis,”2015 IEEE International Congress on Big Data,New York,NY,USA,2015.
[**3**].N.H.Egebjerg, N.Hedegaard, G.Kuum, R.R. Mukkamala and R.Vatrapu, “Big Social Data Analytics in Football:Predicting Spectators and TV Ratings from Facebook Data”,2017 IEEE International Congress on Big Data, Honolulu,Hl,USA.2017.
[4] P. Deshpande, A. Rasin, R. Tchoua, J. Furst, D. Raicu and S. Antani, "Enhancing Recall Using Data Cleaning for Biomedical Big Data,"2020 IEEE 33rd International Symposium on Computer-Based Medical Systems(CBMS),Rochester,MN,USA,2020.
[5] S. Sharma and A. Bhagat, "Data preprocessing algorithm for Web Structure Mining," 2016 Fifth International Conference on Eco-friendly Computing and Communication Systems(ICECCS),Bhopal,India,2016q.
[6] Guo, H. Liang, S. Ai, C. Lu, H. Hua and J. Cao, "Improved approximate minimum degree ordering method and its application for electrical power network analysis and computation," Tsinghua Science and Technology,2021.
[7] T. Crnovrsanin, C. D. Correa and K. -L. Ma, "Social Network Discovery Based on Sensitivity Analysis,"2009 International Conference on Advances in Social Network Analysis and Mining, Athens,Greece,2009.
[8] J. Hu, M. Liu and J. Zhang, "A semantic model for academic social network analysis,"2014 IEEE/ACM International Conference on Advances in Social Network Analysis and Mining,Beijing,china,2014.
[9] P. Lubarski and M. Morzy, "Measuring the Importance of Users in a Social Network Based on Email Communication Patterns, 2012 IEEE/ACM International Conference on Advances in Social Network Analysis and Mining,Istanbul,Turkey,2012.
[10] T. -T. Kuo, J. -J. Yeh, C. -J. Lin and S. -D. Lin, "Designing, Analyzing and Exploiting Stake-Based Social Networks," 2010 International Conference on Advances in Social Network Analysis and Mining,Odense,Denmark,2010.
[11] H. Huang, B. Wei, J. Dai and W. Ke, "Data Preprocessing Method For The Analysis Of Incomplete Data On Students In Poverty," 2020 16th International Conference on Computational Intelligence and Security(CIS),Guangxi,China,2020.
[12] L. Meng and F. Yu, "RFID data cleaning based on adaptive window,"2010 2nd International Conference on Future Computer and Communication,Wuhan,China,2010.
[13] bV. Kumar and C. Khosla, "Data Cleaning-A Thorough Analysis and Survey on Unstructured Data", 8th International Conference on Cloud Computing, Data Science & Engineering(confluence),Noida,India,2018.
[14] D. K. Singh, R. A. Haraty, N. C. Debnath and P. Choudhury, "An Analysis of the Dynamic Detection Algorithms in Complex Networks,"2020 IEEE International Conference on Industrial Technology(ICIT),Buenos Aires,Argentina,2020.
[15] S. Ahajjam, M. El Haddad and H. Badir, "Influentials identification for community detection in complex networks,"2016 4th IEEE International Colloquium on Information Science and Technology(CiSt),Tangier,Morocco,2016.