A Novel Technique in Data Mining Via Investing Social Analysis Tools

Assist. Prof. Dr. AbdulkareemMerhej Radhi[1],Gaith A. Majeed[2]
Al-Nahrain University[1,2], College of Information Engineering, Iraq
abdulkareemradhi@gmail.com[1], ghaithit90@gmail.com[2]

***Abstract*-With the rapid development of Internet technologies in the last two decades and the tremendous progress in information and communication engineering, and the necessity need for most people and different communities to exchanging information,, ideas and opinions on a particular subject via the Internet using social media like Facebook, Twitter …etc. This paper presents a new approach in mining data from this medium, which contains a large and various data. This approach relies on investing social analysis tools to determine the complex relationships between users of social networks as well as using the Java platform and adapting them to deal with the Arabic language for its users. The main contributions of the proposed system are to create an efficient environment to improve the productivity of different communitiesand analyzing the behavior and social relationships of the individuals and recipients which improves the level of performance to achieve positive and successful outputs.**

Keywords - Mining,Social Network,Crawler,SNA.

I. INTRODUCTION

A Social Network is the innovative version of social relations on the internet with millions and millions of users everyday which is consist of individuals (nodes) connected by Edges(relations) [1].SNA is being a necessary tool for media, inquiries and students which needs to display social relationships regarding to network theory consisting of individuals and relations in terms of Graph-based structures which is very complex, it's operate on multiple levels from individuals up to organizations and take an important role in solving problems and accomplishing goals.Visual representations are very useful to understand network data and extract the result of the analysis.SNA play a dominant role as a key technique in latest sociology, communication, economics, information science and different studies.The importance of social network analysis came from its difference from traditional social studies, that assume it is the attributes of individuals whether they are friendly or not, etc. In addition, to test organizations interact with each other as well as connections between employees [1].Social networks are growing on the web day after day by their size and number therefore social networking becoming the biggest inclinations on the web with millions of people on it, and hundreds of web-based social networks like Facebook, twitter, Instagram, etc. [2].

II. RELATED WORK

This section presents review of related effort to the Social Network Analysis:

-N.Ghali, M. Panda, A. E.Hassanien, A. Abraham And V.Snasel in 2012 [2]Shows an online visualization that is focused on the topics of joint publications.-P. Singer in 2011 [3]applied a context to detect bi-directional links between social and its content. -D.Cai, Z. Shao, X. He, X. Yan, J.Han in 2005 [4]proposea new method for learning an optimal linear combinationfor different communities.-A. Bohn, I.Feinerer, K.Hornik and P.Mair in 2011 [5] applied "content-based SNA" to describepeople's interests.-B. Hoppe, C.Reinelt in 2010 [6]offered a framework for hypothesizing different types of leadership networks and uses caseexamples to identify outcomes typically associated with each type of network. -A.Mislove, M.Marcon, K. P. Gummadi, P.Druschel, B. BhattacharjeeIn 2007 [7]presents measurement study andanalysis of the structure of multiple online social networks.-A. Chin in 2007 [8]definedapproach for identifying communities in blogs.-X. Shi in 2007 [9] build query networks from web search engine with nodes (individuals) and edges represent and deduct segmented query sessions.

III. AIMS OF THIS RESEARCH

We can summarize the main aims of this research as follows:

1- Designing and implementing an integrated web application for SNA and its dataset.
2- Propose and implement a new model for crawling data from Facebook.
3- Transform different and heterogeneous data to suitable form for the proposed software.
4- Representation Web Communities, which is important to understand the network data.
5- Filter community data to Nodes and its attributes and Defining and Clustering different communities.
6- Implementation of an efficient SNA metrics to analyze data and express the results of the analysis as SNA report, which is useful to the user, or the instructor of specific community.

IV. CRAWLER

A web crawler or a spider is a system for the majority downloading of web pages. They are one of the main components of web search engines where systems that assemble a mass of web pages and index them. Figure[1] shows the basic web crawler block diagram.

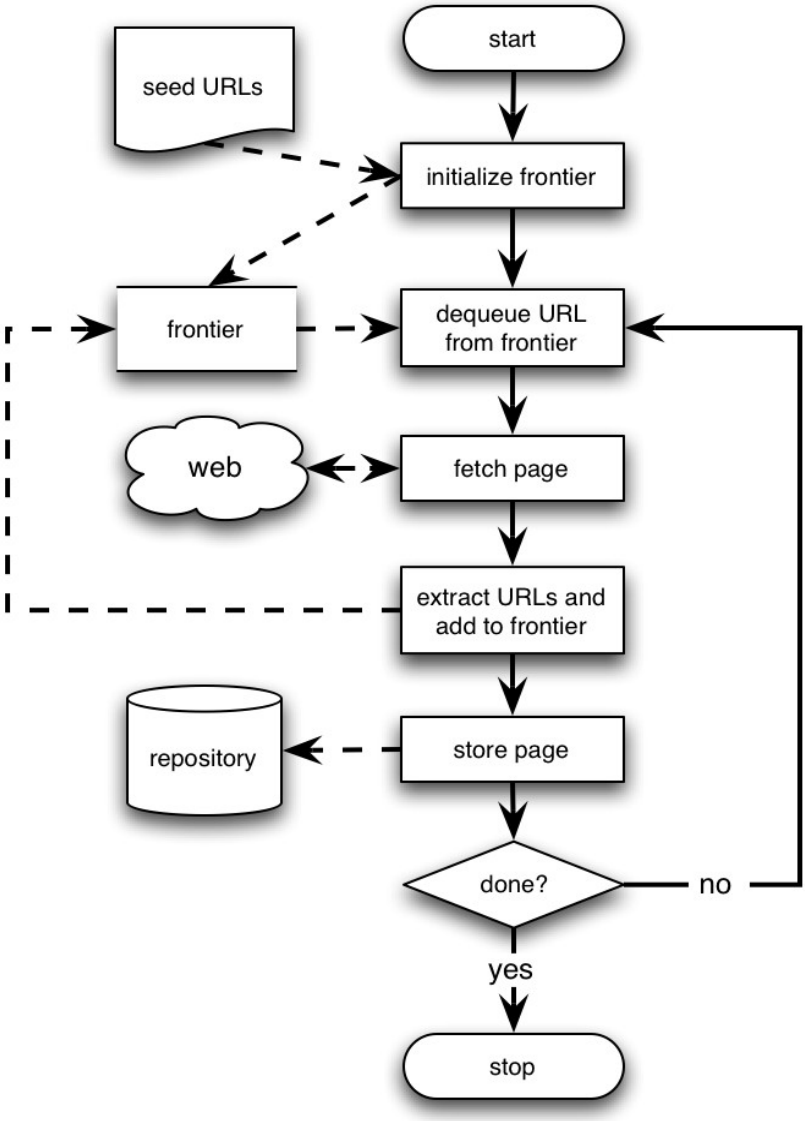


Figure [1] Basic Web Crawler

V. PROPOSED SYSTEM

The proposed system have multi step implementation from paper design to final output of the system, therefore the following shows this system steps in more details and its block diagram in figure 2:

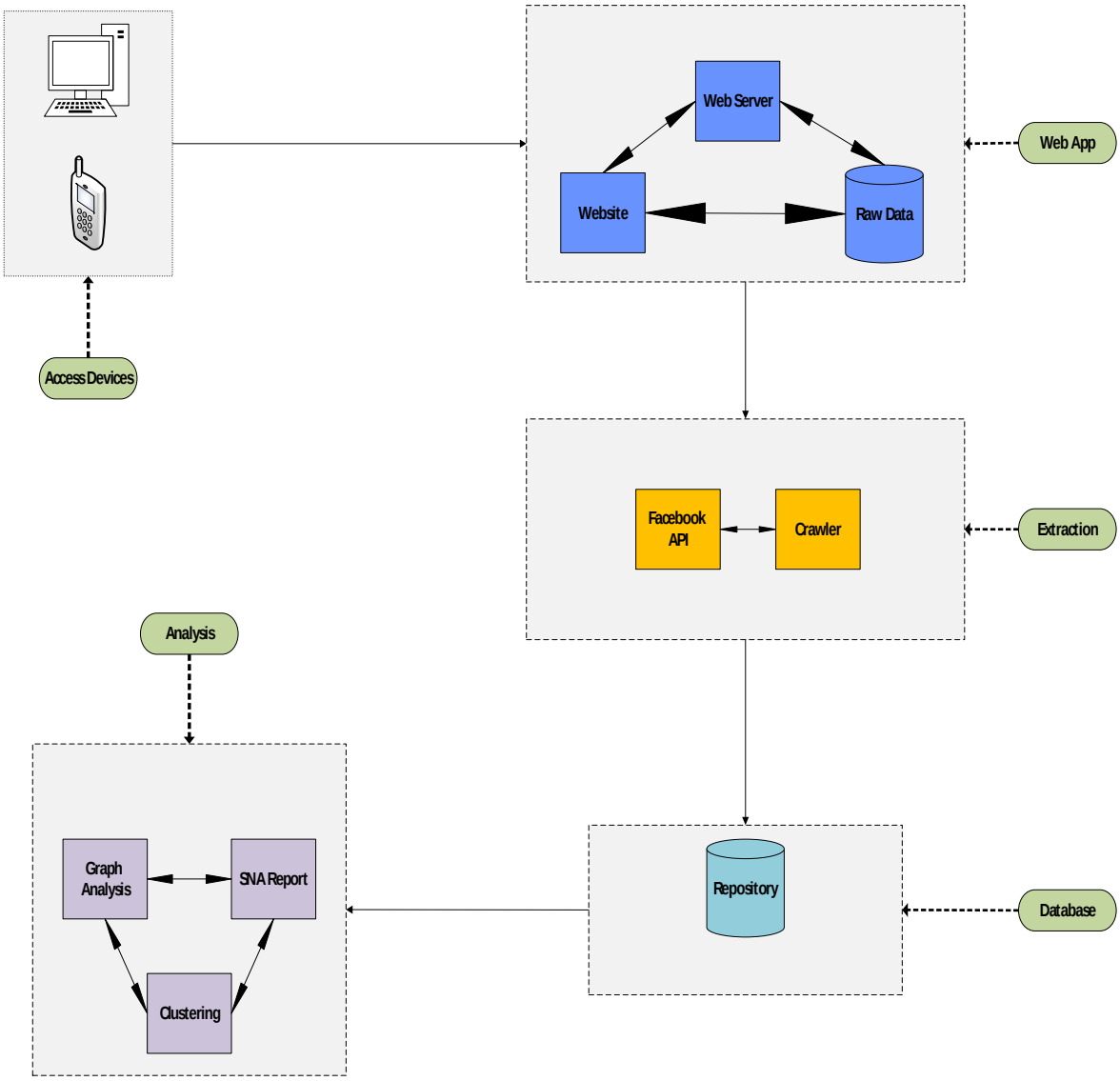


Figure [2]Proposed SNA System Block Diagram

VI. CRAWLER DESIGN AND IMPLEMENTATION

The most crucial issue facing this research is "What is the data required to analyze and How to get it", thereforeto give this research a point of contribution against many SNA works the crawler was built to extract and analyze Facebook pages real data. Due to Facebook limitations to access its real data, must register as a Facebook developer in order to use Facebook Graph API, therefore many works done to design and implement a crawler program to extract the desired data. JavaScript language used as a main language for crawler. JavaScript has many limitations but it is compatible with web pages and has high performance, so the JavaScript Object Notation(JSON) used to overcome the read file issue to make JavaScript read the SNA database properly. The crawler extracted data and insert it directly to an SNA database to analyze it later. The proposed crawler designed and implemented to diagnose extracted non-Latin languages so that it became more suitable with Arabic language to analyze Arabic Facebook data,figure [3] shows a block diagram of the proposed crawler.

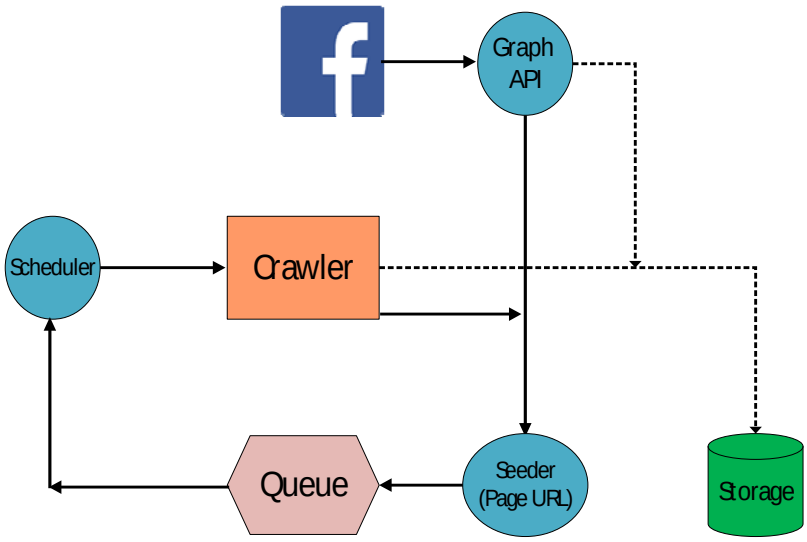


Figure [3]Proposed Crawler block diagram

VII. SYSTEM LAYOUT

The main objective of SNA is to analyze the Facebookdata through the use of the latest web technologies and analysis metrics. In order for the system to work properly and efficiently there must be a well-designed specifications for each function of its subsystems. SNA application uses three-stages Web Client / Server architecture as shown in figure [4]the three stages explained below:

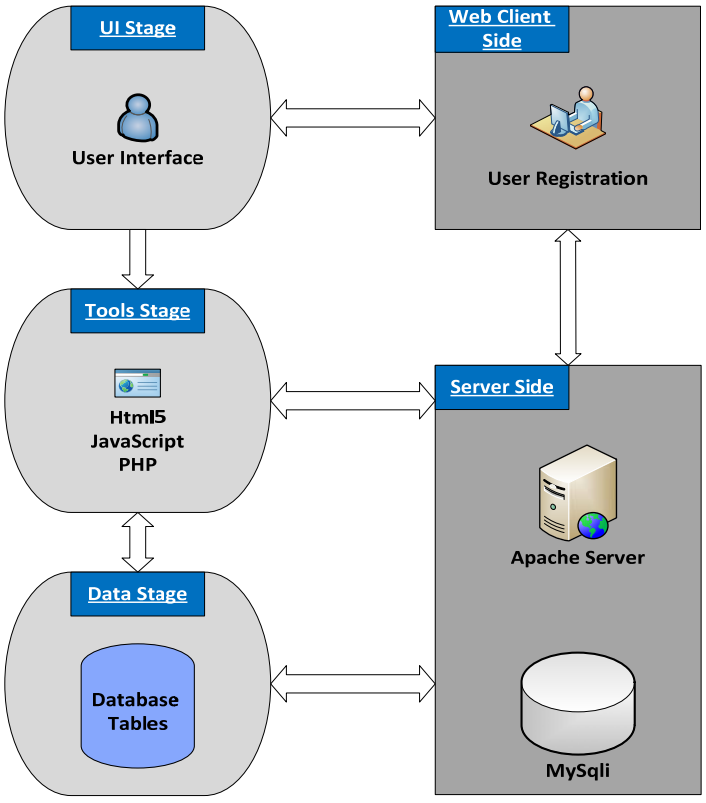


Figure [4] SNA System Layout

VIII. SYSTEM ARCHITECTURE

Figure [5] shows the complete system architecture for the SNA application. It shows all its webpages that are involved in the application. SNA application has two types of users; each type authorized to access specific pages and allowed to do particular jobs.Analyzers access the pages that allow them to analyze the data online. On the other hand, System admins access the whole database, edit its data and access the main pages that control the application that forming the core module of the system.

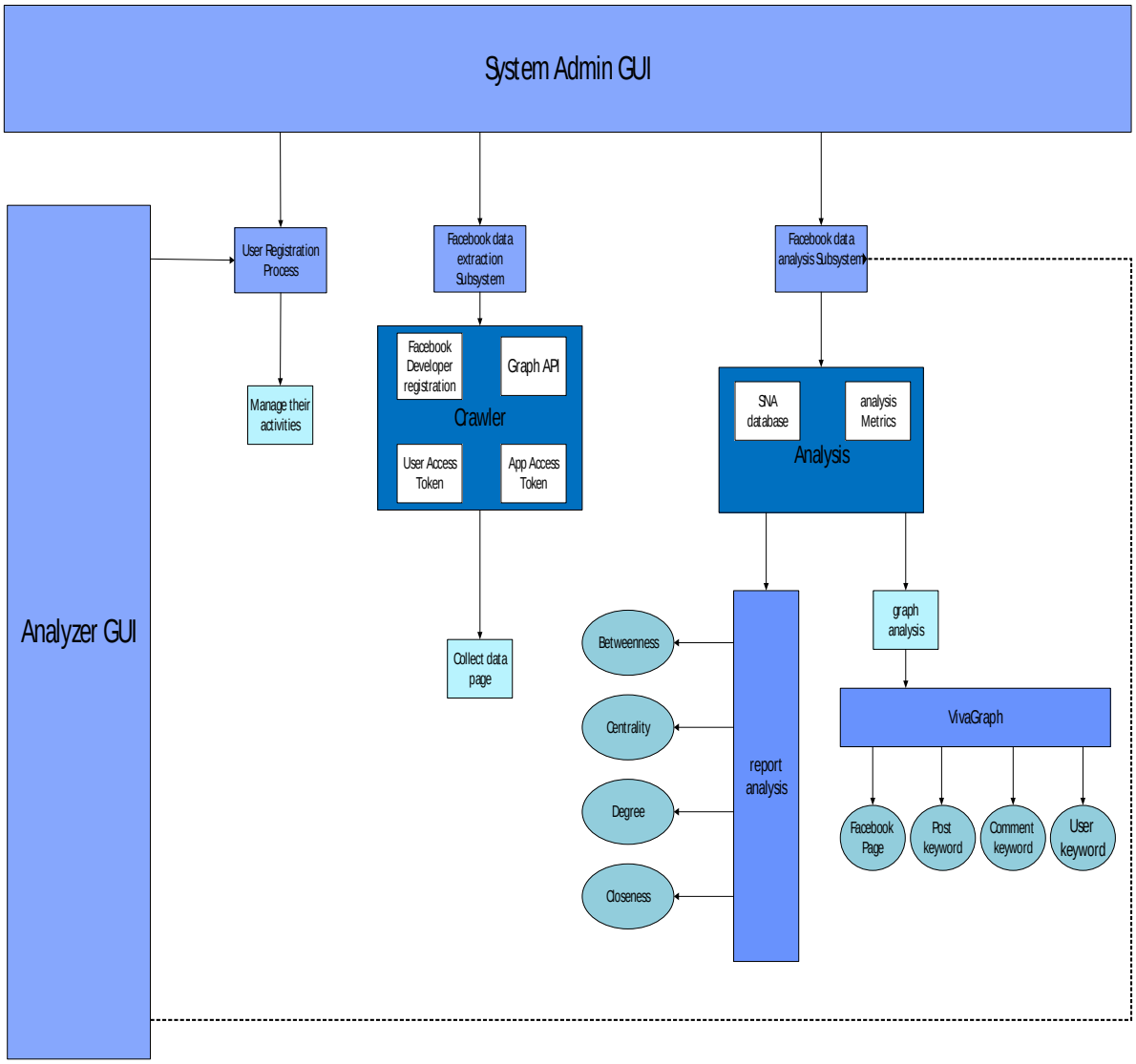


Figure [5]System Architecture

IX. DATA EXTRACTION

This subsystem involve proposedwell-designedFacebookcrawler to extract all desired information from Facebook pages.This subsystem divided into two stages:

A. *Developer registration*

It is very important to get permission to extract data officially from Facebook, so after registration the SNA developer team can access and deal with Facebook API which is called Graph API.So they can build any software on Facebook or deal with Facebook data; however, it's necessary to get App Access Token and User Access Token to access these data because without Access Token couldn't get anything from Facebook.In order to generate user and app access tokens after getting User ID there is an Access Token Tool in Facebook Graph API.This kind of access token is needed when the app calls an API to read, modify or write a specific Facebook data.Facebook user Access Token expire like any other user Access Token and should get long lived User Access Token (2 Months expire in each request).Now can use this User Access Token to link with Graph API and use it in the crawler software.

X. PROPOSED CRAWLER SOFTWARE

This software is responsible forextracting data from Facebook, so in this SNA system the proposed crawler is designed and programmed in JavaScript and PHP languages to be more compatible with Web application. This crawler get data directly from online Facebook Database via Graph API 2.7 then restructure it to be compatible with SNA system database and capable to deal with Arabic data. Finally,these data became a seed for MySQL database of SNA system.

XI. System Environment Implementation

The system implemented as browser / server application using PHP 5.5.12 as a server side programing languages. Apache 2.4.9 is used as the Web server and MySQL 5.6.17 as a database server. The system consists of a main PC (Admin) and several Devices (PCs or Mobiles) for the systemclients, which are all connected by a network (Intranet or Internet).

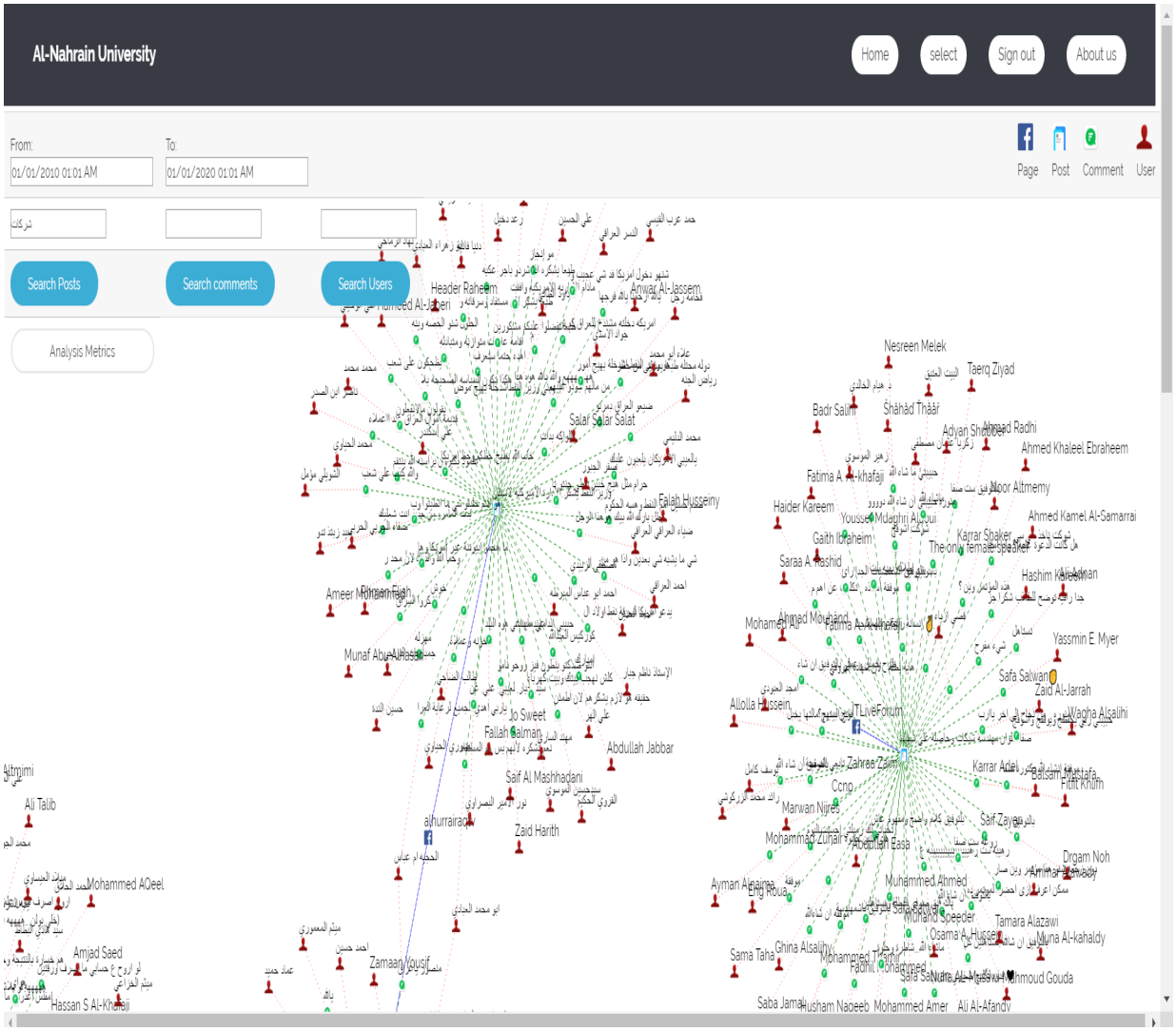
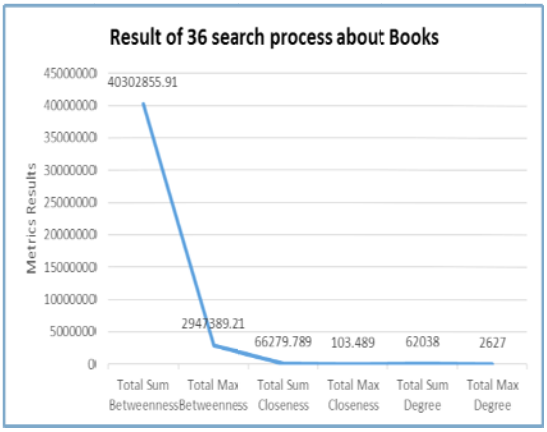


Figure [6] Searching for Posts keyword

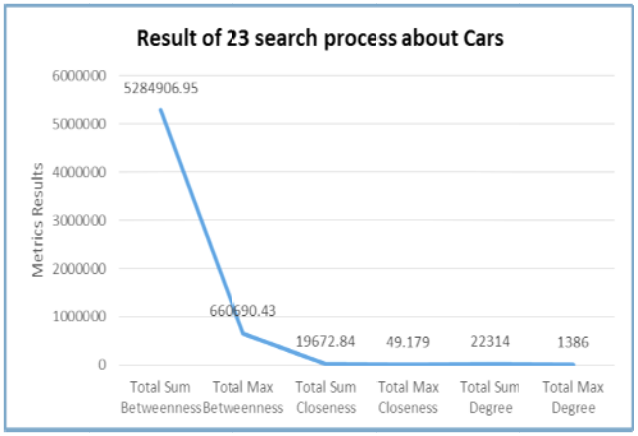
The challenge facing any social network analysis is the analysis of Arabian communities & Arabic language, therefore this give more contribution to the research& build a complete report to the whole Iraq community rather than specific small community.The challenges here is how to analyze broad Iraqi people so multiple steps would do this:

1. Choose the top ranking Iraqi pages, the verified pages and the more popular pages to be crawled through different times**[Posts from 2012-11-28 22:25:43 to 2017-03-20 16:44:06]**.
2. Accumulation of Big Data by crawling the mentioned pages through many periods and gathering 118,011 Comments and 4,489 Posts from 157 Iraqi Pages to build a sample of Iraqi community.
3. Iraqi community classified to 17 different category about human needs (Company, Tourism and places, Music, Sport, Books & Education, Cars, Technology, TV News & Cinema, Health, Food, Religion, Political, Photographic, Government, Shopping/Retail & Game, Financial and Others).
4. Searching for 471 keywords for all categories that should build a huge variety database of human needs.
5. Get analysis report and graph from the proposed system for each search state then store the results in excel sheets to build the final big report. The results stored for each metric of Betweenness, Closeness and Degree then take the summation and maximum value of each metric from the graph of each keyword.

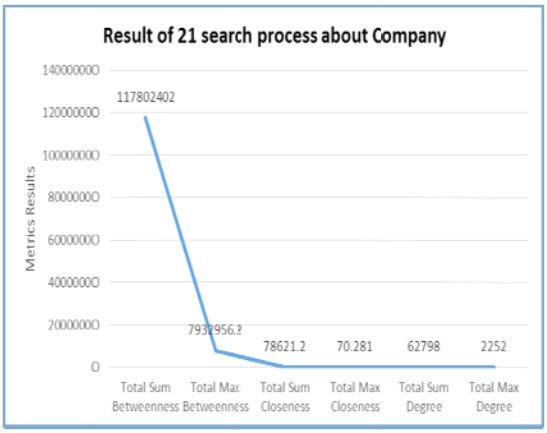
The following charts showing the total calculations of each category:



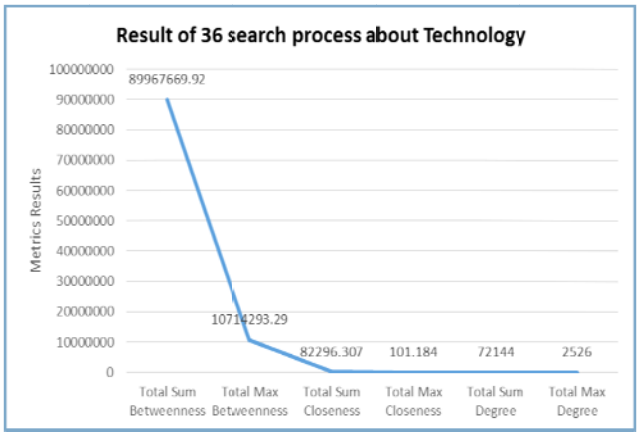
a



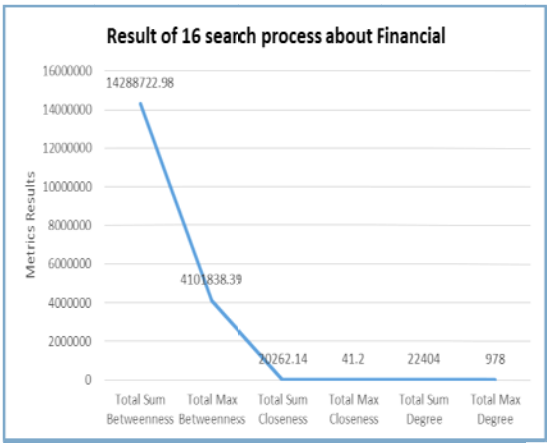
b



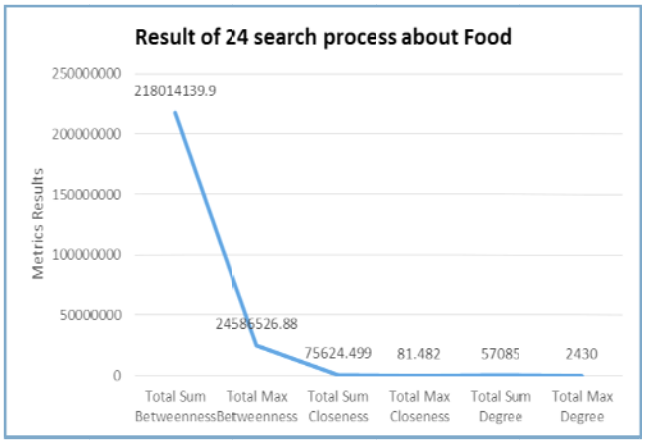
C



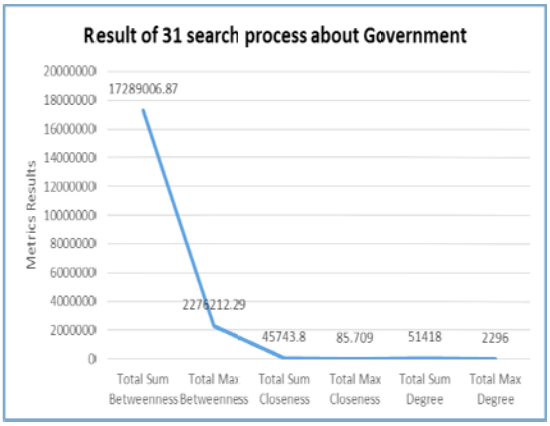
d



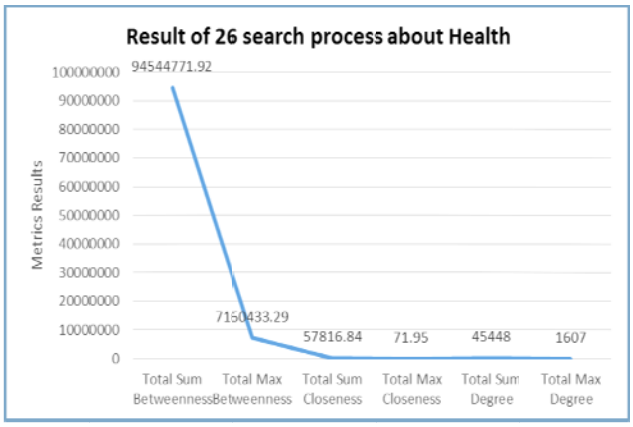
E



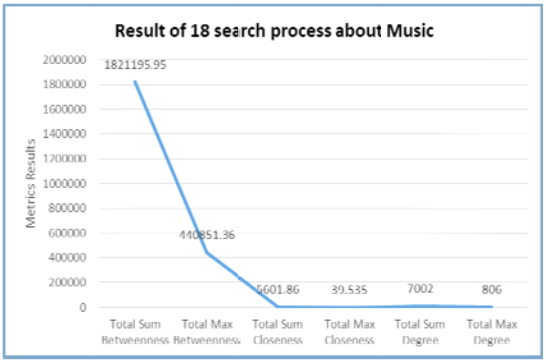
f


Result of 31 search process about Government
Metrics Results
17289006.87
2276212.29
45743.8
85.709
51418
2296
Total Sum Betweenness
Total Max Betweenness
Total Sum Closeness
Total Max Closeness
Total Sum Degree
Total Max Degree

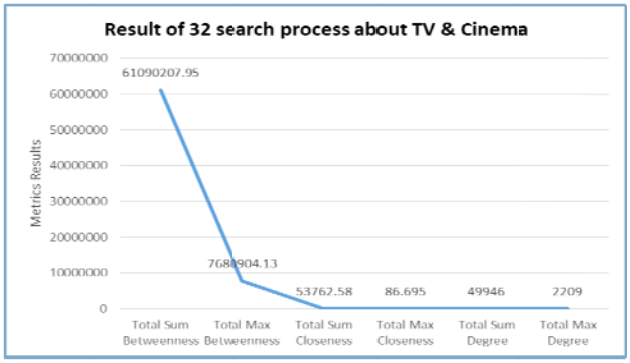
G


Result of 26 search process about Health
Metrics Results
94544771.92
7150433.29
57816.84
71.95
45448
1607
Total Sum Betweenness
Total Max Betweenness
Total Sum Closeness
Total Max Closeness
Total Sum Degree
Total Max Degree

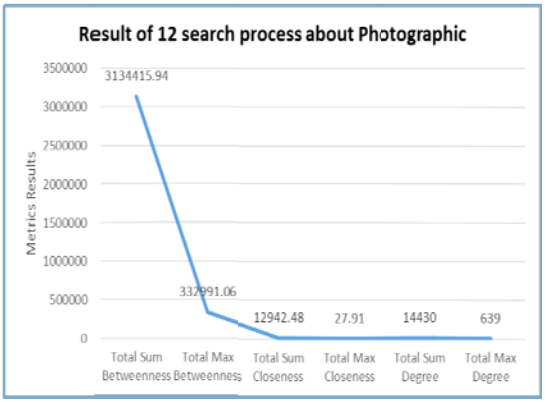
h


Result of 18 search process about Music
Metrics Results
1821195.95
440851.36
6601.86
39.535
7002
806
Total Sum Betweenness
Total Max Betweenness
Total Sum Closeness
Total Max Closeness
Total Sum Degree
Total Max Degree

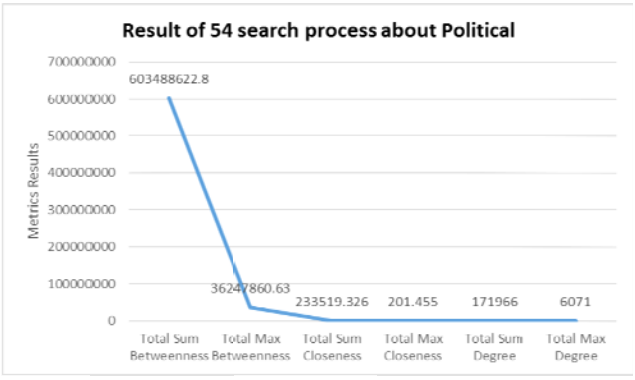
I


Result of 32 search process about TV & Cinema
Metrics Results
61090207.95
7684904.13
53762.58
86.695
49946
2209
Total Sum Betweenness
Total Max Betweenness
Total Sum Closeness
Total Max Closeness
Total Sum Degree
Total Max Degree

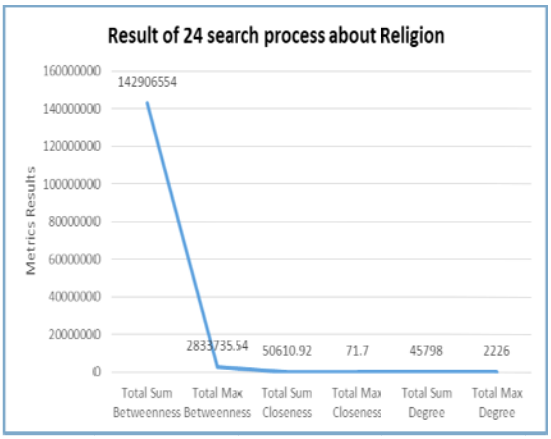
j


Result of 12 search process about Photographic
Metrics Results
3134415.94
332991.06
12942.48
27.91
14430
639
Total Sum Betweenness
Total Max Betweenness
Total Sum Closeness
Total Max Closeness
Total Sum Degree
Total Max Degree

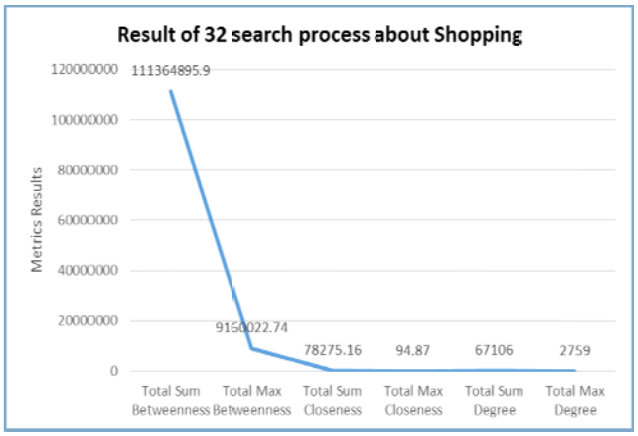
K


Result of 54 search process about Political
Metrics Results
603488622.8
36247860.63
233519.326
201.455
171966
6071
Total Sum Betweenness
Total Max Betweenness
Total Sum Closeness
Total Max Closeness
Total Sum Degree
Total Max Degree

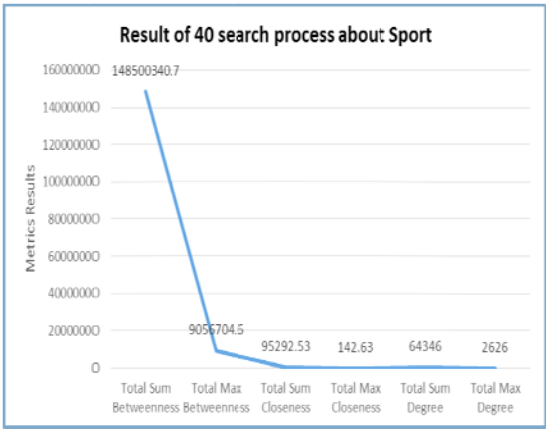
l



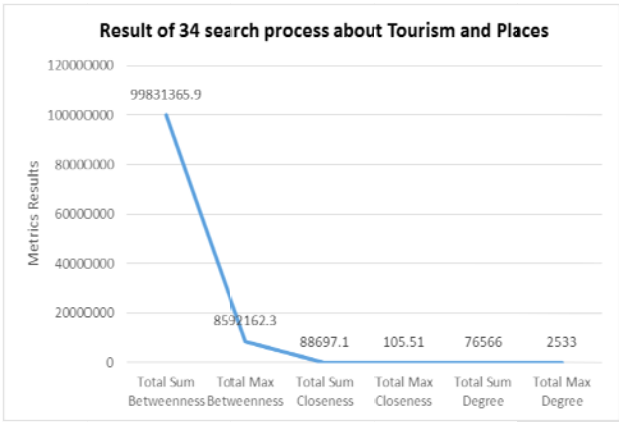
M



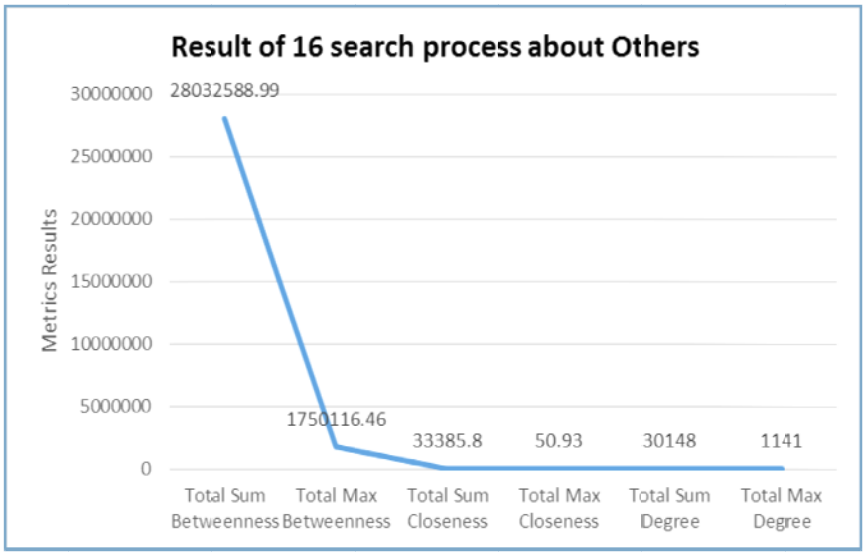
n



O



p



q

Figure [7] [a,q] Categories Metrics for Iraq Community

Drawing a chart for each metric to represent the whole report of all categories, which shows the most interested trends for Iraqi community as shown in figures [14].Sounds good from the following analysis charts that the Iraqi community cares about political news at the first grade and superior to all other interesting categories cause the current situation of our country. The other categories varies from Food, Sport to Photographic and Music at the lowest level. This shows the difference between SNA and traditional Social Statistic Analysis because"Kadim Al Sahir, Barbie" Pages represent the highest ranking pages in Iraq as pages like count equal to "12973401,8349944" respectively but SNA shows the highest trends in Iraq is the Political news interactivity.

XII. CONCLUSIONS

This research proposed a Facebook network analysis system to analyze datasets crawled from Facebook online database by proposed crawler subsystem. The proposed system is designed and implemented as a web application to be practicable on different machines like PCs and smart phones. It analyzes data using Social Network Analysis (SNA) metrics.The proposed system is an integrated system for crawling data from online Facebook database and store it offline in its well-designed and suitable database then analyze these data using SNA metrics rather than using separated different systems to do this job.The proposed system uses Graph theory to analyze crawled big data because it cannot be analyzed using traditional statistical social analysis methods.The proposed crawlerextract and insert data directly to a suitable database without intervention from end user, and it fully supporting non-Latin languages ex. Arabic.Using SNA metrics to analyze big data by its nodes and relations rather than statistics.Analyze these crawled data using SNA metrics to identify Iraqi people trends in social media at present-days.Adding a Value Added Service (VAS) for analyzing call log of mobile phones and identify the relations between indices (Numbers).

REFERENCES

[1] Creativecommons.Org," Social network analysis: Theory and applications",2011.
[2] Ghali, A. Abraham. ," Computational social networks: Mining and visualization, Compute",Soc. Networks Min. Vis., vol. 9781447140, pp. 1–385, 2012.
[3] S. Philipp," Time Series Analysis of Online Social Network Data and Content", Master's Research Graz University of Technology,26 September,2011.
[4] D. Cai," Mining Hidden Community in Heterogeneous Social Networks, University of Illinois at Urbana Champaign, University of Chicago, pp. 58–65, 21,2005.
[5] Bohn, I. Feinerer, and K. Hornik.(2011). Content-Based Social Network Analysis of Mailing Lists, R J., vol. 3, no. 1, pp. 11–18.
[6] B. Hoppe and C. Reinelt, "Social network analysis and the evaluation of leadership networks", Leadersh. Q., Elsevier Inc., vol. 21, no. 4, pp. 600–619, 2010.
[7] A. Mislove, M. Marcon, K. P. Gummadi, and B. Bhattacharjee,"Measurement and analysis of online social networks", Proc. 7th ACM SIGCOMM Conf. IMC '07 California USA, pp. 29–42,2007.
[8] Chin and M. Chignell, "Identifying communities in blogs: roles for social network analysis and survey instruments", Int. J. Web Based Communities, vol. 3, no. 3, p. 345-363, 2007.
[9] Shi. Xiaodong, " Social Network Analysis of Web Search Engine Query Logs", Ann Arbor, vol. 1001 PP. 48109, University of Michigan, 2007.

Authors Biography

[1] Dr. Abdulkareem Merhej Radhi is Assist. Prof. and Doctorial Philosophy in Artificial Intelligence. Supervisor of many M.Sc. students in Information Engineering Colleges rather than Science Colleges. Lecturer in Al-Nahrain University. Director of Computer Center and AvinCina for E-Learning. Interested in Data Security, Soft Computing, Distributed Database, Engineering Analysis, Wireless Networks, Data Mining and Social Network Analysis.

[2] Mr. Majeed is M.Sc. in Information Engineering College and he is interested in Social Network Analysis and Data Mining.