Advances in Computer Science and Information Technology (ACSIT)
p-ISSN: 2393-9907; e-ISSN: 2393-9915; Volume 3, Issue 4; April-June, 2016, pp. 330-334
© Krishi Sanskriti Publications
http://www.krishisanskriti.org/Publication.html

Critical Analysis and Behavior Exploration of Social Network Data Analysis Tools

Divya[1], Sanjeev Dhawan[2] and Kulvinder Singh[3]

[1]M.Tech.(SoftwareEngineering), University Institute of Engineering and Technology, Kurukshetra University, Kurukshetra, Haryana, India
[2]Faculty of Computer Science and Engineering,University Institute of Engineering and Technology, Kurukshetra University, Kurukshetra, Haryana, India
[3]Faculty of Computer Science and Engineering, University Institute of Engineering and Technology, Kurukshetra University, Kurukshetra, Haryana, India
E-mail: [1]simplydeevya@gmail.com, [2]rsdhawan@rediffmail.com, [3]kshanda@rediffmail.com

Abstract—*A generous variety of software are developed within moment of time to assist and facilitate qualitative and quantitative study in a social network. This paper furnishes crisp overview on some well-known software of social network mainly Weka, UCINET, NodeXL, Gephi, R Package, RapidMiner. Weka is powerful for performing data mining job, whereas UCINET includes numerous analytical competent tools in exploring and computing structures of social network, Gephi in social network provides enhanced visualization and speed up exploration. NodeXL presents fine visualization and simplifies data analysis in network, R environment facilitates statistical computing of the Social network. RapidMiner gives 500plus operator for assisting data mining jobs and also carry out predictive analytics.*

Keywords*: Social Network analysis, tool, visualization.*

1. INTRODUCTION

Social network is a structure of enormous nodes, that are tie up with one or more than one types of interdependency mainly companionship, same interest, monetary trade, or connections of trust, knowledge and each type gives rise to a corresponding network. Social networks is focusing on various stages, from families up to the stage of nations, and carry a significant task in determining the way problems being solved, businesses are run, and also shows the degree to which individuals accomplished their goals, this all reflected from study of numerous fields of academics. As a result Social network become popular among user and Due to this extreme popularity, handling a data is become a challenging task. In performing this challenging job social network tools are used to represent the nodes and links in the network, and to analyze the network data. Social network analysis has now stimulated from being a evocative metaphor to an analytic approach to a paradigm, with having its own theoretical statements, methods, social network analysis tools, and researchers. Social network software allows researchers to investigate large networks. In this paper an attempt has been made to describe some popular tools such as Weka, Node xl, UCINET, Gephi, R, RapidMiner for handling huge dataset. These software provide mathematical functions and algorithm that can be applied to network model and used to identify, represent, analyze, visualize nodes and links from various formats of input data. The output data can be saved in external files. Whereas a visual representations assist in understanding social network data and providing the result of an analysis [1][2][11].

2. SOCIAL NETWORK ANALYSIS SOFTWARE

2.1 Weka

A Weka is an influential Machine learning/data mining tool. It is known for open source tool written in Java by the Waikato University from New Zealand. In 1996 its first public version 2.1 was out. Weka software is useful for the educational system, research and applications. Weka version 3.6 come out with 49 preprocessing instruments, 79 organization and regression algorithms, 8 clustering algorithms, 3 algorithms use for finding connection rules and 3 graphic interface. It provides array of data mining jobs such as data preprocessing, clustering, classification, regression, visualization, and attribute selection. Weka well-matched for almost every modern computing platform, as it is implemented in java and moreover well-suited for developing new machine learning schemes. Weka technique is work on supposition of data is available as one flat file where data is depicted by a preset number of attributes. The data files are associated with Arff [4][12].

2.2 Node XL

NodeXL is known for providing basic network analysis jobs. It incorporate in network data workflow steps that are follow as collect, store, analyze, visualize and publish its datasets, by using broadly framework of spreadsheet. It permits user to manually perform steps or allow to automatically calling up in a solo operation. Users are able to manually explore networks and find out insights into their structures and related texts content with an ease use of interface. Automated reports can be effortlessly formed and scheduled for repeated regular update and generation. These reports enclose with an automated textual abstract of messages coupled with a network with the description and visualization of a structure without the need of human involvement and refinement. Users also able to publish and share NodeXL data sets through NodeXL Graph Gallery. Reports can be effortlessly shared with people who are not NodeXL users via email or the web. An automation settings can helpful to exchanged to facilitate beginner to use proficient level configuration [5][9].

2.3 Gephi

Gephi is known for open source interactive network exploration and visualisation tool for all variety of networks, dynamic and hierarchical graphs. A flexible and multi-task architecture carry new possibilities to work with complex data sets and create valuable visual outcomes. It provides effortless and broad access to network data and allocates them for spatializing, filtering, navigating, manipulating and clustering. It left a CPU free for other computing via computer graphic card. Immense consideration has taken for the extensibility of the software. An algorithm, filter or tool can be easily added to the program, with little programming experience. It provides two kinds of algorithms named force-based algorithms and multi-level algorithms [3][8].

2.4 R package

R is a programming language comprise with some common features with functional and object oriented programming language. It is a foremost software environment for statistical computing and graphics technique such as linear and non-linear modelling, classical statistical tests, time-series analysis, classification, clustering, and many other techniques. Additional functionality is added because R allows users to describe new functions which allow expertise statistical techniques, graphical devices, as well as import/export capability to many external data formats. An international team of statisticians and computer scientists is maintained and distributed the R package [7][13].

2.5 UCINET

UCINET tool is a Windows software package specialized for the analysis of data in social network. The package provides the tools to analyze 1-mode or 2-mode data, and it can handle a maximum of two millions nodes, but practically most of the procedures become too slow to run when networks work with 5000 plus nodes. It has numerous network analytical tools, some tool named as centrality measure, permutation-based statistical analysis and many others. This tool also fine in carrying out an ample range of data transformations named as sub-graphs and sub-matrices, merging datasets, permutation and sorts, transposing and reshaping, recodes, linear transformations and some more. In addition, matrix analysis routines also offered by such as matrix algebra and multivariate statistics. [8][14]

2.6 Rapid miner

RapidMiner was incepted in 2001, it is an interactive open source data mining tool and provides 500plus operators for different jobs of data mining. WEKA is fully integrated as matter of fact through a WEKA Extension but RapidMiner is not integrated in WEKA. An extension for R is available. RapidMiner is able to work with different sources such as files and databases. The graphical user interface of RapidMiner is easier and more efficient to use in comparison with the WEKA Explorer when working with reusable blocks and trying to establish a connection to a database. All RapidMiner sources were written in Java and are using the GNU Affero General Public License (AGPL). Commercial developers have to obtain a proprietary license (OEM license) if AGPL does not fit their requirements [6][7].

3. CENTRALITY MEASURE IN SOCIAL NETWORK ANALYSIS

It is eminent for giving a coarse indication of social power of network node. Centrality measure is work on the basis of, how well node is "connect" in a network. Measures of centrality are as follow "Degree", "Betweenness", and "Closeness”.

3.1 Degree Centrality

Degree centrality is defined as a node having number of links. It accentuates to that nodes who having high degrees. If directed network is to be considered then two separate measures of degree centrality are defined one is indegree and other one is outdegree. A sum of number of links directed to the node is termed defined to indegree measure, and sum of number of link that the node directs to other nodes is termed as outdegree measure.

The degree centrality for node “v” is calculated as:

$$C_D(v) = \frac{\deg(v)}{n-1}$$

Characteristics of node with high degree centrality:

- It usually an active member in network.
- It generally a connector in a network.
- It may be in a benefited position in network.
- It may have alternative options to satisfy organizational needs, and therefore, it may be less dependent on other nodes which eventually reduce interdependency.
- It may identify as a third party node [10].

3.2 Betweenness Centrality

Betweenness centrality is defined as node position within a network in terms of its ability to create bonds with other nodes or groups in a network. Vertices that take place on many shortest paths between other vertices hold higher betweenness than those that do not have it.

Betweenness centrality is calculated for node "v" as:

$$C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

Characteristics of node with high betweenness centrality:

- It holds the powerful position in a network.
- It fallout to a single point of failure
- It has huge amount of influence over what happens in the network [10].

3.3 Closeness Centrality

Closeness centrality is defined in connected graph as an inverse of average distance to all other nodes. It measures a node that how fast it can contact other nodes in a network.

Closeness Centrality is Calculated for node "v" as:

$$C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

Characteristics of a node with a high closeness centrality:

- It has fast access to other nodes in the network.
- It has short path to other nodes.
- It is close to other nodes.
- It has high visibility to a network whatever happens in it. [10]

4. RESULT AND ANALYSIS

An analyzed result is summarized in Table- 1. Tools are detailed in following evalution criteria: developer of tool, implementation language, source of access, website on which they are available, operating system, main functionality of these tools, input/output format, Interfaces, network handled by them, their goals and drawbacks.

Table 1: Comparison between different social network tool

Software	Weka	NodeXL	Gephi	R package	Ucinet	RapidMiner
Developer	University of Waikato from new Zealand	Social media Research foundation	University of Technology of Compiègne in France	R-core team, university of Auckland	Lin freeman, Martin Everett and Steve Borgatti	Rapid Miner
Written in	Java	#C, .Net	Java	C++, Fortran ,R	BASIC,turbo pascal	Java
Source of access	Open	Open	Open	Open	Paid	Paid
Website	www.cs.waikato.ac.nz/~ml/weka	Nodexl.codeplex.com	Gephi.org	r-project.org	Analytic Technologies	Rapidminer.com
Operating system	Windows ,OS X, Linux	Windows	Linux , windows, Mac OS X	Windows, Linux, Mac	Windows	Cross platform
Main functionality	Machine learning and data mining	Data analysis and visualization	Visualization	Social network analysis within a versatile and popular R environment	Social network analysis and visualization	Statistical analysis, data mining, predictive analytics
Input format	.arff, .libsvm,.csv	email, .csv (text), .txt, .xls (Excel), .xslt (Excel 2007), .net (Pajek), .dl (UCINet), GraphML.	GraphViz(.dot), Graphlet(.gml), GUESS(.gdf), LEDA(.gml), NetworkX(.graphml, .net), NodeXL(.graphml, .net), Pajek(.net, .gml), Sonivis(.graphml), Tulip(.tlp, .dot), UCINET(.dl), yEd(.gml), Gephi (.gexf), Edge list(.csv), databases	R will read in almost any format data file	Excel, DL, text, Pajek .net, Krackplot, Negopy, proprietary (##.d & ##.h)	.arff, .xrff, .csv, .xls, SQL databases

Output format	.csv,html, GNUPlot, LaTeX, Plain Text(default)	.csv (text), .txt, .xls (Excel), .xslt (Excel 2007), .dl (UCINet), GraphML	GUESS(.gdf), Gephi(.gexf), .svg, .png	R has write capability for most data formats	Excel, DL, text, Pajek .net, Krackplot, Mage, Metis, proprietary (##.d & ##.h)	.csv, .xls.
GUI/ Command line	Both	Both	GUI	Both	GUI	GUI
Network handle	Several thousand	1000 – 10.000 nodes	50.000 nodes	Several thousand	32,767 nodes (with some exceptions)	Several million
Betweenness centrality	No	Yes	Yes	No	Yes	No
Degree centrality	No	Yes	Yes	No	Yes	No
Closeness centrality	No	Yes	Yes	No	Yes	No
Goal	Facilitate developing techniques of machine learning.	To support students who are learning social network analysis and professionals interested in applying network analysis to business problems	Help data analysts to make hypothesis, discover patterns, isolate structure singularities or faults during data sourcing.	R is very much a vehicle for newly developing methods of interactive data analysis	Built for speed.	Speed delivery and reduce errors by nearly eliminating the need to write code
Drawback	• Traditional algorithms need to have all data in main memory • Big datasets are an issue	• Big data sets are an issue. • Limited functionality.	• Need Improvements for integrated the data structure to support grouping and navigation within a network hierarchy • User interface to users' need is an issue • Development of new features, especially filters, statistics and tools is required	• R's main problem is its language, which, although highly extendable, is also a difficult one to learn thoroughly enough to become productive in DM	• Most of procedures are too slow to run more networks larger than 5000 nodes	• Suppor t for deep learning methods and some of the more Advanced specific machine learning algorithms is currently limited

The first point is developer of these tools, Weka is developed in university of Waikato, NewZealand. NodeXl is developed by social media Research foundation, gephi is developed in University of Technology of Compiègne in France. R package is developed by R core team, university of Auckland. Ucinet is developed by Lin freeman, Martin Everett and Steve Borgatti. RapidMiner is developed by Rapid Miner Company. Weka, gephi and RapidMiner tool is written in Java, whereas nodeXl written in #c, .net and R package implemented in C++, fortran, R. UCINET written in BASIC and Turbo pascal. Next evaluation criteria is licensing. It appears that Weka, Gephi, R package tool are Licensed under GNU General Public License and NodeXL is licensed Under Microsoft Public License. These four tools are freely available. Ucinet and rapidMiner tool are paid. UCINET License is governed by law of Massachusetts and Rapid Miner is using Affero General Public License (AGPL).

Weka is specialized in machine learning/data miming task work on window, OS X, Linux Platform. It has Highly Interactive Interface and Handle several thousand of nodes but big datasets is an issue for Weka tool. NodeXL is known for Visualization and Data analysis and also have both interfaces

GUI as well as Command Line. It handles several thousand nodes and can measure centrality of nodes. But big data set cannot be handled with NodeXL. NodeXl goal is to support students who are learning social network analysis and professionals interested in applying network analysis to business problems. Gephi is used for visualization purpose and can handle 50,000 nodes of network. It is well suited for Windows, Linux, Mac OS X Platforms. Its goal is to help data analysts to make hypothesis, discover patterns, isolate structure singularities or faults during data sourcing and also to calculate centrality measures. It also consists of numbers of drawback as follows: firstly, it needs improvement for integrated data structure to support grouping and navigation within a network hierarchy. Secondly user interface to users' need is an issue. Third one is development of new features, especially filters, statistics and tools is required. Next tool is R package, it is popular for analyzing social network with versatile R environment. It also handles several thousand nodes and providing GUI interface to user. R is very much a vehicle for newly developing methods of interactive data analysis but its language is difficult to learn. UCINET is specialized in Social network analysis and visualization and can handle 32,767 nodes. Ucinet is built for speed and work on windows platform. It has drawback for network have larger than 5000 nodes, because procedures become too slow to run. RapidMiner is used for Statistical analysis, data mining, predictive analytics having GUI interface. It can handle millions of nodes. It delivers speed and reduces errors by nearly eliminating the need to write code but machine algorithm is very limited.

5. CONCLUSION

This paper conversed over social network analysis tools named as Weka, NodeXL, Gephi, UCINET, R package and RapidMiner. These tools have their own specialization including visualization, analysis and data mining. All tools are efficient in providing good data visualization. For the present research work, Weka tool is taken into consideration for data mining analysis. Although RapidMiner and R package tools all are also known for good data mining jobs. But Weka have some good points over R package and RapidMiner. First, Weka provides three graphical user interfaces i.e. the Explorer for exploratory data analysis to support preprocessing, attribute selection, learning, visualization, the Experimenter that provides experimental environment for testing and evaluating machine learning algorithms, and the Knowledge Flow for new process model inspired interface for visual design of KDD process. A simple Command-line explorer which is a simple interface for typing commands is also provided by Weka . Second, Weka is best suited for mining association rules. Third, it is also suitable for developing new machine learning schemes. Fourth, Weka have database connection using JDBC with any RDBMS Package. Weka is best suited for data mining analysis and easy to learn for naive users.

REFERENCES

[1] Wellman, Barry and S.D. Berkowitz, eds., "*Social Structures:A Network Approach",*Cambridge: Cambridge University Press, 1988

[2] Linton Freeman," *The Development of Social Network Analysis"*, Vancouver: Empirical Press, 2006.

[3] Mathieu Bastian and Sebastien Heymann, Mathieu Jacomy," Gephi : An Open Source Software for Exploring and Manipulating Networks", Association for the Advancement of ArtificialIntelligence, 2009

[4] R. Robu* and V. Stoicu-Tivadar*," Arff Convertor Tool for WEKA Data Mining Software",IEEE, 2010, pp.247-251

[5] Hansen, D., B. Shneiderman, and M.A. Smith. 2010. Analyzing Social Media Networks With NodeXL: Insights From a Connected World.Elsevier Science.

[6] Rapid-I GmbH. RapidMiner 5.0. Manual. Dortmund, Germany, 2010.

[7] Hilda Kosorus, J¨urgen H¨onigl, Josef K¨ung, " Using R, WEKA and RapidMiner in Time Series Analysis of Sensor Data for Structural Health Monitoring", 22nd International Workshop on Database and Expert Systems Applications, 2011, pp.306-310.

[8] Ioana-Alexandra APOSTOLATO˙,"An overview of Software Applications for Social Network Analysis", International Review of Social Research, vol. 3,2013, pp. 71-77.

[9] Marc A. Smith, "NodeXL: Simple Network Analysis for Social Media", IEEE, 2013, pp. 89-93.

[10] Jyoti Sunil More and chelpa lingam, "Reality Mining based on Social Network Analysis", International Conference on Communication, Information & Computing Technology (ICCICT), 2015,

[11] Divya, Dr. Kulvinder Singh, Dr. Sanjeev Dhawan, " Threshold Based Mechanism to Detect Malicious URL's in Social Networks", IOSR Journal of Computer Engineering (IOSR-JCE), 2016, pp. 18-21.

[12] Remco R. Bouckaert, Eibe Frank, Mark Hall, Richard Kirkby, Peter Reutemann,Alex Seewald, David Scuse, WEKA Manual for Version 3-9-0, University of Waikato, Hamilton, New Zealand, 2016.

[13] W. N. Venables, D. M. Smith and the R Core Team, An Introduction to R, manual version 3.3.0, 2016

[14] UCINET: Social Network Analysis Software. http://analytictech.com/