Investigation and Analysis of Research Gate User's Activities using Neural Networks

Omar Alheyasat
Department of Computer Engineering, Al-Balqa' Applied University, Jordan

Abstract: *Online Social Networks (OSNs) have been proliferating in the past decade as general-purpose public networks. Billions of user's are currently subscribing by uploading, downloading, sharing opinions and blogging. Private OSNs emerged to tackle this issue. Research Gate (RG) is considered as one of the most popular private academic social networks for developers and researches in the internet. The current study consists of two parts. The first part is a measurement study of user's activities in RG and second part deals with the relationship between user's profile data and their links. To this end, a sample of one million RG user's records was. To facilitate this analysis, three layers back-propagation neural network models were generated. The purpose of this network is to show the correlation between user profiles data and the number of their followers. The results show that there is a high positive relationship between user's followers and research activities 'publication, impact factor, total number of publication views and citation'. In addition, the results indicated that the number of questions and answers (activity) of a user have low correlation with the corresponding followers. The present results demonstrate that the question/answer contributions of researchers are limited, which therefore, needs more collaboration from the RG researchers.*

Keywords: *RG, neural networks, OSN, regression, measurement, crawling, follower.*

Received December 20, 2013; accepted December 23, 2014; published online April 1, 2015

1. Introduction

With millions or billions of subscribers, Online Social Networks (OSNs) evolved to be one of the most success sites on the Internet [29]. The first start of OSNs was in 1995 with Classmates.com [28]. Since, that day, the idea of OSN has been growing with the evolution of the Internet. Nowadays, the Internet embraces many massive OSNs, such as, Twitter and Facebook. The users of OSNs are involved in two main activities; downloading and uploading. In uploading activities, subscribers register, update their profile, add content and write blogs. In download activities, user's read their friends profiles and blogs. In addition, they can download contents, such as, photos, videos and music. In addition to, these two activities, users may become friends and follow each others. This features of OSN gains its publicity.

With billions of OSNs user's, it is very hard to find whom to follow. In addition, it is very hard to discover fake and anonymous users. These issues necessitate the construction of private and special purpose OSNs.

Research Gate (RG) has emerged as one of the first private social networks for researchers and developers. The registration process of RG requires a private domain Email address. This property increases the credibility of RG. Besides, it eliminates fake and anonymous accounts. RG was constructed to increase collaboration between researchers all over the world. It was launched in 2008. Nowadays, RG has more than three millions registered researchers and developers [24].

RG is a special type of OSNs. It adds many new features and activities to the normal OSNs. First, it is private and constructed for scientist and researchers. Second, it has a build-in forum, which allows researchers to upload question and answers. Third, it allows developers and researchers to search for jobs. Fourth, it maximizes the collaboration between researchers by sharing publications, opinions and expertise. Finally, Blogs on user's pages have vanished and replaced with the question/answer forum.

The present work mainly consists of two parts. The first part deals with the contributions and collaborations of the researchers (i.e., how many activities are in RG and how social is RG user's). The second part deals with the relation between user's profile and the number of followers they have (i.e., what makes a user more popular to be followed). The study started by crawling a sample of one million user's of RG out of more than three millions RG users. The harvested information are public information that can be found through any search engine, with no interest to harvest private data of RG user's such as, user's networks, publications PDFs or personal information. To this end, the study implemented a back-propagation neural network to examine the relationship between user's features and their followers. The neural model was fed with various types of inputs and studied the patterns between these inputs and the number of followers as an output. This method allowed us to find the influence of user's attributes on their popularity in RG.

The rest of this paper is organized as follows. The next section provides a brief review of the literature that has been conducted in the field of OSNs. It has been followed by presentation of our experiment environment and its configuration, then data analysis. Subsequently, discussion of results, contribution and future work is presented.

2. Related Works

With the tremendous growth of the World Wide Web, OSNs have become important tool for communication in specialized communities according to the user's specific interests. Web-applications have been studied heavily in the past decade [9, 24]. Social networks and graph analysis have been previously studied by many authors [1-28]. Some studies focused on the analysis of OSNs [1, 5, 6, 7, 8, 10, 11, 14, 15, 16, 17, 22, 23, 30].

Benevenuto *et al*. [1] analyzed the user workloads in OSNs, where data was collected from a social network aggregator website in Brazil, which enables user's to connect to multiple social networks with a single authentication.

Buccafurri *et al*. [3] performed an investigation of bridges in Social Inter-networking Scenarios (SISs), authors showed that bridges can allow users to state a number of their role in a SIS. In a different study [28] the authors studied the influence of common attributes and the relative importance of different characteristics at different institutions of social structure of Facebook "friendship" networks at one hundred American colleges and universities at a single point in time. They illustrated how microscopic and macroscopic perspectives give complementary insights on the social organization at universities. Recommendation of similar user's, resources, attributes and the identification of the challenges and evaluation of social influence were studied and investigated in depth in large scale social networks in many studies [14, 15, 16, 17].

Other studies [2, 19, 21, 31, 32] where investigated in order to analyze, predict, learn, and evaluate different user personalities, attributes, common behavioral patterns, ranking methodology in different OSNs. The common character between all of the above mentioned studies is that they investigated large scale social networks where everybody can sign in (public social networks) and use the services delivered by the specified social network.

In this work, RG was investigated and analyzed. RG considered as a private academic online information society social-network. RG have more than 3 millions scientists specialized in different fields. User's eligible to sign in to RG must be a member of an organization such as universities, institutes, etc. To the best of our knowledge, this study is the first to analyze the RG characteristics and user's behaviors and activities. The motivation behind this analytical study is to investigate level of contribution that RG adds to OSNs. In addition, this study seeks to find the main interest that user's of RG are looking for (i.e., how they find their colleagues and how they create links to other user's). The result of this work can be used to generate recommendations for user's to find their peers.

3. The Experiment

To harvest the data, two web-crawlers have been developed specifically to meet the goals of the study. The first crawler generates the seed list that the second crawler utilizes to harvest the data. The first crawler harvested random links of researchers from RG through the alphabetical index. The output of this crawler is a seed list of one million links. This seed list was divided into two lists, because we have two locations that we used to conduct our experiment. These lists have been fed into the second crawler. Figure 1.

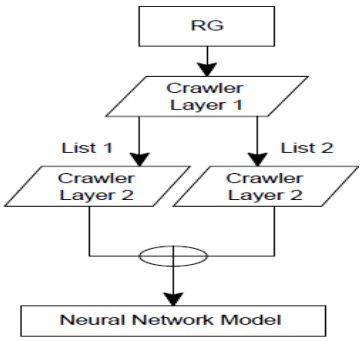


Figure 1. Crawling steps.

To speed-up the crawling time of the second crawler, a distributed-threaded crawler was implemented. This crawler consisted of ten computers (nodes) distributed over two locations. The mission of these nodes is to collect the results from the other nine nodes. This node in turn called the coordinator. The coordinator node has been implemented as a Remote Procedure Call server (RPC). Each location has been fed with one of the seed list that has been generated from the first crawler. Figure 2 shows the experiment setup.

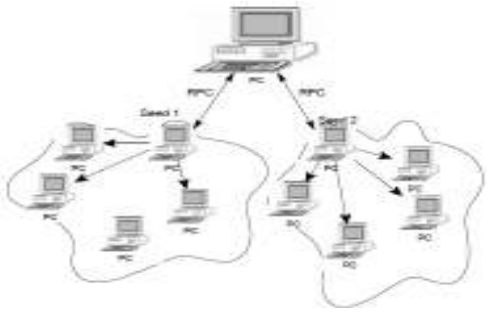


Figure 2. Experiment environment.

The output of this crawler is a file that consists of tuples. Each tuple represents one RG user attributes. These tuples consist of 10 separated values. These values can be shown in Table 1. The information values in the tuple represents; the cumulative citation number for the author (cit.), the number of RG user's that viewed his publications (view.), the number of

RG user publications (pub.), the number of RG user's that viewed his profile (pro.), the cumulative impact factor for all publications of the user (im.), the field of specialty of the RG user (maj.), number of RG user's that are following him (fing.), the number of followers for this RG user (fer.), the link for the RG user, the number of the questions and answers that the RG user have contributed (q.) and the number of questions and answers that the RG user's have been viewed for him (qisview.). The values of these tuples were harvest from each user. Each user has three different pages; the main page, the contribution page and the state page. To speed-up the process, the second and the third pages were visited if the first page is not empty. Subsequently, the output file was fed to the statistical and neural network analyzers. (Data was crawled in June 2013).

Table 1. The tuple values.

Cit	Vw	Pub	Pro	Im	Maj	Fng	Fer	Q	Qview

4. Results

4.1. Neural Network Model

In this work, one million RG user's records were crawled. Each record consists of 10 parameters. To investigate and study the collected data, a back-propagation neural network model was implemented to study the relationship between the input and the output of the network. This network consists of three layers. Figure 3 shows the implemented network. The setup configuration of the parameters of the neural network model was adopted from [7]. That configuration was appropriate for such collected data.

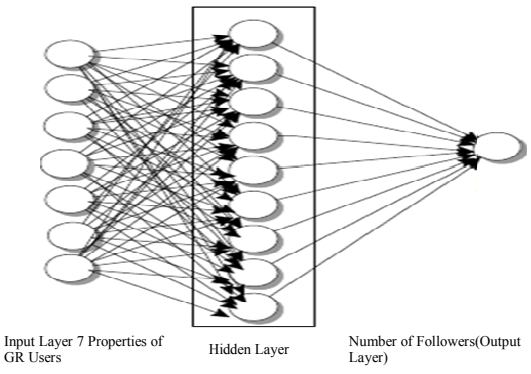


Figure 3. Back-propagation neural network model.

Correlation matrix among the 10 collected variables was developed to investigate the relationship between the number of follower's variable and the other nine attributes, as well as among the attributes. These properties include (cit, fing, im, maj, pro, qis, qisview, view and fer). Many models were developed, each model with different combinations of attributes have been fed separately as an input to the network. Subsequently, the network has been trained, validated, and tested. Finally, the quality of each model was evaluated based on regression equation between output and targeted output showing the correlation coefficient. Table 2 shows parameter configuration of the network.

Table 2. Configuration of neural network model parameters.

Parameter	Value
Number of Inputs	7
Epochs	500
Hidden Layers	1
Number of Nodes	10
Function	Triangular

Studying the relation between the inputs and the outputs for different variables is known as generic empirical retrieval problem. This means a simple mapping process between a set of input variables and an output and can be performed by utilizing different mathematical tools, conventional methods such as regression analysis, or artificial intelligent techniques such as neural networks. Linear regression is an appropriate tool for developing many mapping algorithms and functions. Regression both linear and non-linear can estimate the relationships and correlation among inputs and outputs [23]. Linear regression is well-known and an easy tool to implement model for mapping functions. However, the most important limitation of linear regression is that, it works over a broad range of variability if the function which that represents is a linear relationship. If the function is nonlinear, linear regression can be inaccurate [16]. To implement sophisticated and accurate mapping functions between inputs and outputs, non-linear regression is adopted [7]. One of the easiest methods to generate non-linear regression functions is neural networks. Neural networks are commonly known to be used for data classification showing the relationship with logistic regression, neural networks typically used a logistic activation function and output values from 0 to 1 same as logistic regression. However, due to the importance of neural networks to analyze complex models, non-linear hypothesis is desirable for many real world problems, such as, regression [5, 20]. In this work, neural network was adopted to investigate the interrelationships among variables. The results of this experiment are consisting mainly from two parts. In the first part, users activities have been studied. To facilitate this experience, three main parameters have been used. In the second part, regression values of the neural network model have investigated. These two parts are discussed in the following sections.

4.2. Profiles Statistics

Table 3 shows statistics of the considered variables. The statistics clearly indicate the skewness of data. The average number of followers and the average number of links that a user generates are approximately the same. This means that the original user follows most of his followers. The variance between the maximum value and the average value of the total number of profiles views shows that some users are more popular than others in RG. Finally, the contribution in question/answer environment is limited in RG as observed from Table 3. The average value is less than one and the maximum number is 744 which

is considered small compared with the number of blogs that other OSNs user's write.

Table 3. Profile's statistics.

Data	MAX	Avg	STD
Number of Publications	3127	12.262	36.365
Number of Profile Views	2124314	212.3	3625.4
Number of Publication Views	1000000	272	1945.13
Cumulative Impact Factor of Researchers	11069	27.58	111.72
Number of Following	1148	7.5	15.7999
Number of Followers	808	7.57	15.3483
Total Number of Citation	39601	44.16	283.36
Number of Question/Answer	744	0.282	4.1634
# of Question Views	74958	9.28	216.82

4.3. User Activities

To study the activity of the users, three different parameters were studied. First, the RG score. RG gains a score for each user according to its profile. This score is calculated according to its publications, questions/answer, profile views, and the data profile. Second, the number of question/answer that user's update. As mentioned, the blogs in RG has been eliminated and a forum has been added instead. This number can be used to indicate the activity and the contribution of the user's in the RG network. Third, the participation network. In OSNs, each user can participate in a network of user's that links to each other. Users add others when they find interested data in their profiles. In addition, the number of users that follow a user or the number of user is this parameter links can be used as a measurement of the user activities.

Figure 4 shows the CDF of the RG score of the users. RG starts to calculate the RG score for users after the completion of their profile. It is easy to observe from this figure that about 60% of the users have a zero RG score. This figure also demonstrates that most of the users are not active even to complete their profiles.

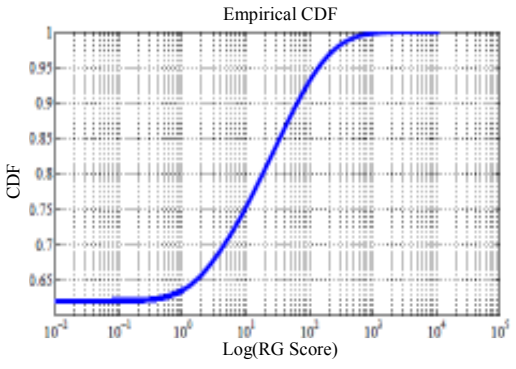


Figure 4. CDF of Users' versus RG score.

Figure 5 shows the CDF of the question/answer environment. This figure shows that over 95% of the users have no contribution in the question/answer environment. In addition, we can also observe that, the highest contribution of any user in this field did not even exceed 900 blogs or comments. Figure 3 also demonstrates that the question/answer environment requires more focusing and efforts on RG.

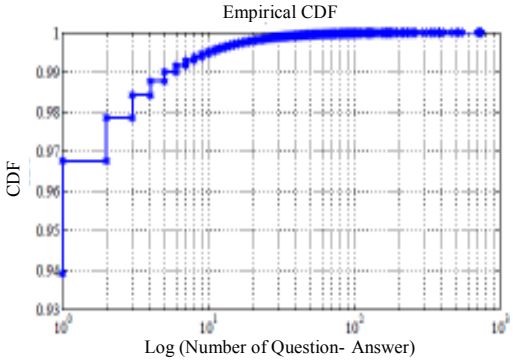


Figure 5. CDF of Question-answer environment.

Finally, Figure 6 shows the CDF of the number of followers, and the number of users that a user follows. 80% of the users have less than 10 followers in their networks. However, 75% of the users are following less than 10 users. This figure illustrated that the links in RG are not in symmetry as in Facebook. You can follow someone whom is not following you. In addition, 98% of all users are following less than 100 users and less than 100 users are following them. This figure demonstrates that users participate in small networks and they have less interest in social connections in RG.

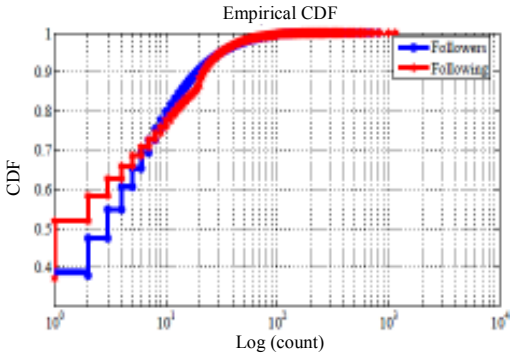


Figure 6. CDF of number of followers and the number of users that a user follows.

4.4. Neural Network Model Results

In this work, one million RG records were crawled. The objective of this study is to investigate the relationship between the collected properties of RG users and the number of followers of those users. Prior to running the experiments using neural networks, and based on statistical analysis, RG crawled records with zero followers were eliminated farther analysis. And randomly divided into three groups; 60% for training the neural network; 20% are used for testing, and the remaining 20% was used for validation.

MatLab neural Network tool has been utilized in this section. This tool has been used in massive amount of research paper. This fact made it an accurate tool for neural network studies.

Figure 7 shows the correlation matrix among variables showing the correlation coefficients and the significant level. The correlation matrix used to guide the development of the various neural network models

by understanding the strength of the linear relationship between each two variables.

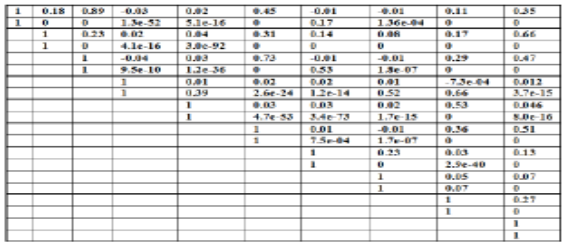
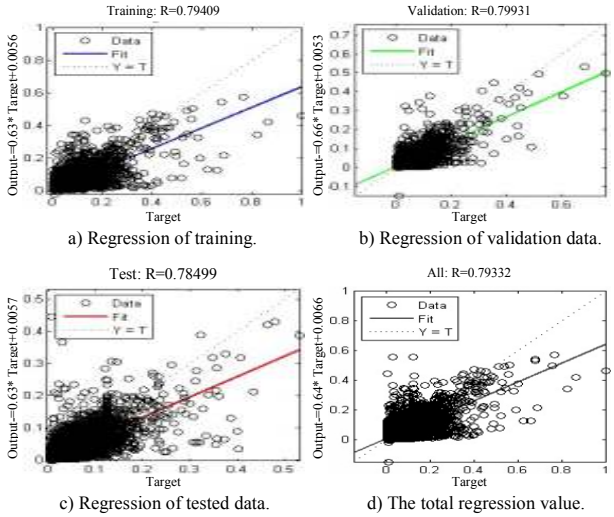


Figure 7. The Correlation Matrix.

Figure 8 shows the quality of the developed neural network model using regression equations, where the R value represents the correlation coefficient indicating the strength of the linear relationship between the target and the output. The plots also show the regression equation between the target and the output.



a) Regression of training. b) Regression of validation data.

c) Regression of tested data. d) The total regression value.

Figure 8. Regression results.

The plots are indicating that about 80% of the variation among the variables is explained by the regression line, which is an acceptable result. The values R for the trained, validate and test are very close indicating the stability of the developed model. Also, the inclination of the line shows that the developed model under estimate the output, especially at the higher values of outputs.

The above developed model was the ultimate model reached after testing many other models. Table 4 shows the statistics of other initial models.

Table 4. Statistics of All Tested Models.

Test	Input									Output	Results			
Try	Cit	Fng	Im	Maj	Pro	Pub	Q	Qw	View	Fer	Train	Validate	Test	All
1	N	Y	N	N	N	Y	N	N	N	Y	0.76	0.74	0.76	0.75
2	N	Y	Y	N	N	Y	N	N	N	Y	0.75	0.76	0.76	0.76
3	Y	Y	Y	N	N	Y	N	N	N	Y	0.74	0.74	0.74	0.74
Select	Y	Y	Y	Y	Y	Y	Y	N	N	Y	0.79	0.80	0.79	0.79
4	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	0.78	0.78	0.78	0.78

5. Conclusions

With more than three million users, RG is considered as one of the most popular academic OSNs in the world. In this work, user activities in RG were studied. Back-propagation neural network models were generated to analyze RGs users' behavior. Regression of neural network has been adopted in order to investigate the relationship between users followers and their profiles. To facilitate the study, a distributed web crawler that consists of two layers and ten nodes has been implemented. The nodes in the crawler communicate with each other by using RPC. A sample of one million records of RG users has been harvested. The observation of current results proves a high correlation between user's followers and the research activities 'publication, impact factor, total number of publication views and citation'. This observation revealed that most of the users are interested in users with higher research skills. In addition the results demonstrated that RG requires more contribution of the users in the question/answer field.

The future work will consider studying the RG question/answer environment in more details. User links in the question/answer environment will be utilized for this environment to create a network that connects users with each other. This network will reveal the characteristics of question/answer in RG.

References

[1] Benevenuto F., Rodrigues T., Cha M., and Almeida V., "Characterizing User Navigation and Interactions in Online Social Networks," *Information Sciences*, vol. 195, no. 15, pp. 1-24, 2012.

[2] Bringmann B., Berlingerio M., Bonchi F., and Gionis A., "Learning and Predicting The Evolution of Social Networks," *IEEE Intelligent Systems*, vol. 25, no. 4, pp. 26-35, 2010.

[3] Buccafurri F., Foti D., Lax G., Nocera N., and Ursino D., "Bridge Analysis in a Social Internetworking Scenario," *Information Sciences*, vol. 224, pp. 1-18, 2013.

[4] Chen L., Zeng W., and Yuan Q., "A Unified Framework for Recommending Items, Groups and Friends in Social Media Environment Via Mutual Resource Fusion," *Expert Systems with Applications*, vol. 40, no. 8, pp. 2889-2903, 2013.

[5] Han J., Kamber M., and Pei J., *Data Mining: Concepts and Techniques*, USA: Morgan Kaufmann Publishers Inc., San Francisco, 2011.

[6] Hao F., Pei Z., Zhu C., Wang G., and Yang L., "User Attractor: An Operator for the Evaluation of Social Influence," *Future Generation Computer Systems*, vol. 5, pp. 458-465, 2012.

[7] Heaton J., *Introduction to Neural Networks for Java*, 2008.

[8] Hui-Yi H. and Hung-Yuan P., "Use Behaviors and Website Experiences of Facebook Community," *in Proceedings of International Conference on Electronics and Information Engineering*, Kyoto, pp. 379-383, 2010.

[9] Jeon H., "A Reference Comments Crawler for Assisting Research Paper Writing," *the International Arab Journal of Information Technology*, vol. 11, no. 5, pp. 493-499, 2014.

[10] Kim W., Jeong O., and Lee S-W., "On Social Web Sites," *Information Systems*, vol. 35, pp. 215-236, 2010.

[11] Kurt M., Berberler M., and Ugurlu O., "A New Algorithm for Finding Vertex-Disjoint Paths," *the International Arab Journal of Information Technology*, vol. 12, no. 6, pp. 550-555, 2014.

[12] Leskovec J., Huttenlocher D., and Kleinberg J., "Predicting Positive and Negative Links in Online Social Networks," *in Proceedings of the 19th International Conference on World Wide Web*, North Carolina, USA, pp. 641-650, 2010.

[13] Li T. and Xiao N., "Solving QBF with Heuristic Small-world Optimization Search Algorithm," *the International Arab Journal of Information Technology*, vol. 12, no. 4, pp. 370-378, 2015.

[14] Li Y., Hsiao H., and Lee Y., "Recommending Social Network Applications Via Social Filtering Mechanisms," *Information Sciences*, vol. 239, pp. 18-30, 2013.

[15] Malik H. and Malik A., "Towards Identifying the Challenges Associated with Emerging Large Scale Social Networks," *Procedia Computer Science*, vol. 5, pp. 458-465, 2011.

[16] Menke W., *Geophysical Data Analysis: Discrete Inverse Theory*, Academic Press, 1989.

[17] Meo P., Nocera A., Terracina G., and Ursino D., "recommendation of Similar Users, Resources and Social Networks in a Social Internetworking Scenario," *Information Sciences*, vol. 181, no. 7, pp. 1285-1305, 2011.

[18] Mislove A., Marcon M., Gummadi K., Druschel P., and Bhattacharjee B., "Measurement and Analysis of Online Social Networks," *in Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement*, New York, USA, pp. 29-42, 2007.

[19] Ortigosa A., Carro R., and Quiroga J., "Predicting User Personality by Mining Social Interactions in Facebook," *the Journal of Computer and System Sciences*, vol. 80, no. 1, pp. 57-71, 2014.

[20] Pao H., "A Comparison of Neural Network And Multiple Regression Analysis In Modeling Capital Structure," *Expert Systems with Applications*, vol. 35, no. 3, pp. 720-727, 2008.

[21] Parthasarathi J., Sundararaman K., and Rao G., "Perisikan: An Intelligent Framework for Social Network Data Analysis," *in Proceedings of International Conference on Communications and Information Technology*, Hammamet, pp. 13-16, 2012.

[22] Rahnama, A., and Madni A., "Relational Attribute Integrated Matching Analysis (Raima): A Framework for the Design of Self adaptive Egocentric Social Networks," *IEEE Systems Journal*, vol. 5, no. 1, pp. 80-90, 2011.

[23] Regression analysis, available at: http://en.wikipedia.org/wiki/Regression-analysis, last visited 2015.

[24] Researchgate, available at: http://www.researchgate.net/, last visited 2015.

[25] Salamanos N., Voudigari E., Papageorgiou T., and Vazirgiannis M., "Discovering Correlation between Communities and Likes in Facebook," *in Proceedings of International Conference on Green Computing and Communications*, Besancon, pp. 368-371, 2012.

[26] Sakata S. and Yamamori T., "Topological Relationships between Brain and Social Networks," *Neural Networks*, vol. 20, no. 1, pp. 12-21, 2007.

[27] Saeid M., Abd Ghani A., and Selamat H, "Rank-Order Weighting of Web Attributes for Website Evaluation," *the International Arab Journal of Information Technology*, vol. 8, no. 1, pp. 30-38, 2011.

[28] Traud A., Mucha P., Porter M., and Porter M., "Social Structure of Facebook Networks," available at: http://arxiv.org/pdf/1102.2166.pdf, last visited 2011.

[29] Trusov M. and Bodapati A., "Determining Influincial Users in Internet Social Networks," *the Journal of Markiring Research*, vol. 47, no. 4, pp. 643-658, 2010.

[30] Wongyai W. and Charoenwatana L., "Examining the Network Traffic of Facebook Homepage Retrieval: An end User Perspective," *in Proceedings of the International Joint Conference on Computer Science and Software Engineering*, Bangkok, pp. 77- 81, 2012.

[31] Zhang Y. and Cao F., "Analysis of Convergence Performance of Neural Networks Ranking Algorithm," *Neural Network*, vol. 34, pp. 65-71, 2012.

[32] Zhao X., Yuan J., Li G., Chen X., and Li Z., "Relationship Strength Estimation for Online Social Networks with the Study on Facebook," *Neurocomputing*, vol. 95, pp. 89-97, 2012.

Omar AlHeyasat is an Associate Professor at the Computer Engineering Department at the Faculty of Engineering, Al-Balq'a Applied University. He received his PhD in Computer Engineering from Vinnitsia National Technical University, Vinnitsia, Ukraine, His research interests include social networks analysis, artificial intelligence, image and signal processing, CPU and GPU scheduling.