Review Social Network Analysis and Mining: Link Prediction

Kalum P. Udagepola and Fatima Chiroma

***Abstract*—The rapid generation and uncontrollable accumulation of the social network data has raised a real issue now, because the data are vast, noisy, distributed, unstructured and dynamic. Since this data can be mined by using web mining techniques, social network analysis and link prediction algorithms, in this article we try to understand the social structure and issues surrounding mining social network data. We will also be looking at the link prediction problems in dynamic social networks and the important techniques that can be applied as an attempt for a resolution.**

***Index Terms*—Social network, social network analysis, Link prediction, web mining.**

I. Introduction

The social network, which is a social structure comprising of people who are linked by different types of interdependency, has changed the nature of information in terms of volume, availability and importance[1]. According to Gupta *et al*. [2], a social network is denoted by a graph fig. 1, where the nodes represent a user and edges represent relationships among nodes which are the users [2].

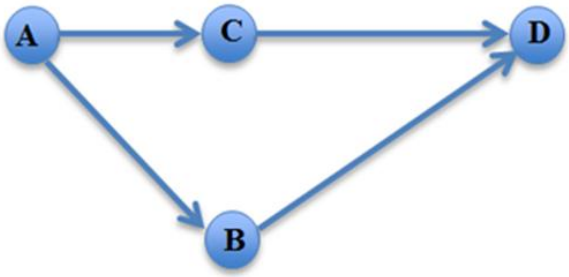

Fig. 1. Social network.

It is normally formed by continuous interactions between people. Gupta *et al*. [2] claimed that social networks are dynamic in nature, as they grow over time through the addition of new users, creation of new relationships and ending of some old relationships [2]. Nandi *et al*. [1] strongly agrees that social networks are dynamic and even went further to state that the data generated from online social network is vast, noisy and distributed as well therefore to analyze such complex and dynamic social network data appropriate data mining techniques are required [1].

Through the Social Network, participants publish or publicly reveal a lot of personal information (communications, relationships and behaviors) and this information have real values that can be mined, utilized and monetized. With this information interested parties can examine, determine and even predict things about a user or group of users for many purposes such as to improve decision making, control costs and minimize risks. This data also has research implications that will enable researchers to improve the design and robustness of several systems such as recommender engines. Users on social networks publish personal information that has real economic values. With this information, interested parties can examine and predict things about a user or group of users for many purposes. Such as to improve decision making, control costs and minimize risks. This scenario is showing Fig. 2 as a Link Prediction Causal Loop.

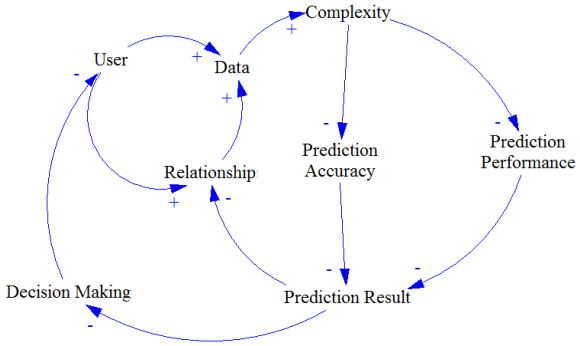

Fig. 2. Link prediction causal loop.

Social networks are highly dynamic objects as they grow and change quickly over time through the addition of new edges, signifying the appearance of new interactions in the underlying social structure [3]. As social networks continue to grow, excess data needs to be mined and Chakrabarti [4] claimed that web mining is the most suitable [4]. Web mining is an application of data mining (which is the technique of discovering and extracting useful information from large data sets or databases) used to discover and extract useful information from the web" [5]. There are three web mining techniques namely the web content mining, web usage mining and web structure mining. These techniques have been used to mine contents, discover interaction patterns between users and identify structures and links between web pages.

Additionally, Ting stated that social network analysis is a vital technique that will aid the understanding of social behaviors, social relationships and social structure [6]. "Social network analysis is the mapping and measuring of relationships and flows between people, groups, computers and other connected information/knowledge entities which provides both a visual and a mathematical analysis of human relationships" [1]. It is similar to web mining but it is about

Manuscript received March 25, 2016; revised July 10, 2016.

K. P. Udagepola is with the American University of Nigeria and School of Information Technology and Computing, Yola, Nigeria (e-mail: kalum.udagepola@ aun.edu.ng).

F. Chiroma is with the American University of Nigeria and Department of Information Systems, Yola, Nigeria (e-mail: Fatima.Chiroma@aun.edu.ng).

data extraction from different resources [6]. Other important and popular techniques that are used for social network mining are the data mining techniques such as sequential pattern analysis, visualization, association rule generation, clustering and classification which have been used by several researchers.

Additionally, Borgatti [5] claimed that in the past the real problem with social network analysis (and mining) was its inability to statistically test hypotheses however this is less of a problem now with the advent of permutation tests [5]. Although, that is not to say that the social network analysis and mining problem has been eradicated in fact there are still several problems surrounding it and even Nandi *et al.* [7] stated that the problems related to mining social networks are still in the infancy state as such there is subsequent need to develop techniques for further improvements [7].

Although, to specifically mine data to make predictions, link prediction algorithm is additionally required because it is a "link mining task that tries to find new edges within a given graph" [8] and attempt to imitate them. According to Yu *et al.* [9], most social media predictions can be done better by expert human agents however there is need to automate predictions [9]. For example, automated predictions are unbiased, cost effective, more accurate and performs better than human agents.

The problem statement here is given any social network graph at a time t (Fig. 3), accurately predict the new relationships that will be created on the network after an interval of time t1.

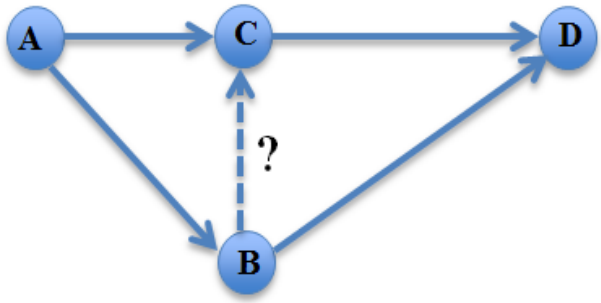

Fig. 3. Link Prediction at time t1.

II. Related Work

Research interest in the area of Social Network Analysis specifically link prediction (scalability) is increasing due to its importance. Previous research works have revealed several problems with link prediction such as scalability, accuracy and efficiency/performance. Gupta *el al.* [2] and Nandi *et al.* [7], in particular have recently worked on related research and they have emphasized on the importance of finding a solution to the scalability problem of link prediction. Fire *et al.* has also done some research where he claimed that the massive growth of social networks has brought about several research directions in which the structural and behavioral properties of large-scale social networks are examined [10]. Additionally, Gupta *et al.* [2] did some work which agrees with Fire *et al.* claim but they were more specific as they believe that the dynamic change forms the base of link prediction algorithms which involves trying to understand the process of the dynamic changes and trying to replicate them [2]. Additionally, Nandi *et al.* [7] have identified seven other key representative research issues in there research namely influence propagation, community or group detection, expert finding, recommender systems, behavior and mood analysis, predicting trust and distrust among individuals, and opinion mining [1].

Gupta *et al.* [11] further went to briefly discuss the importance of link prediction in other fields as well as sufficient related works that have been carried out over the years to solve the link prediction problems. They proposed a method for link prediction i.e. to predict the likelihood of a link between two nodes, based on the information about their neighbors and the connections between them. Additionally, they used mathematical formulations to comprehensively analyze/test the proposed method. They implemented a prediction algorithm. They went further to state how the research can identify the efficiency of the proposed method based on the results by stating that the degree to which the value exceeds 0.5 indicates how better the algorithm performs. The result of the analysis provides an informative rather than conclusive result, which serves as an evidence of the validity of the proposed method. The result can also be used as a benchmark or means of comparison to better comprehend the method. On the other hand, the authors did however fail to state the size of the dataset they used for the analysis of the proposed method. Since that proposed method is applied to a social network which is known to be dynamic, knowing the size of the data used is very important because the proposed method may not be scalable or applied to certain sized data. Additionally, stating the logic behind the choice of social network is also important as this will aid in knowing which type of social network that proposed method can be applied to. The method will contribute to the ongoing researches in social network analysis and mining, as well as other fields such as bioinformatics, security and electronic commerce as stated by the authors. Additionally, it also has the potential to be extended to other domains or categories of social network mining e.g. community detection.

Vaghela *et al.* [12] seek to provide an effective and efficient link prediction method for directed networks which is very relevant to research that are related to the development of link prediction models for social networks. If the method they were proposing meets the goal of their research then it will aid in the analysis, design and evaluation of link predictions to improve its performance and accuracy in other fields as well. Based on the critical review of the article, it has been observed that they introduced the link prediction problem and the problem statement as stated by Liben-Nowell and Kleinberg [3]. The problem statement as defined by Liben-Nowell and Kleinberg is that given a snapshot of a social network at time t and a future time t′, the problem is to predict the new friendship links that are likely to appear in the network within the time interval [t, t′]. However, they did discuss the general problem focus of the article which is the link prediction problem in the context of evolving co-authorship networks, this problem focus is however too broad. This is because the link prediction problem that they authors are referring to as defined by Liben-Nowell and Kleinberg is about the extent to which the evolution of a social

networks can be modelled using features intrinsic to the network itself and based on the use of the term "features" there seem to be more than one feature (as such making the focus of the research broad). They have discussed the research focus and the link prediction problem as part of the introduction but failed to discuss the categories of the problem there but instead it was discussed under the related work section, which may be acceptable by some readers but generally disorganized. These measures include Newmans common neighbors, Jaccard's Index and Adamic/Adar Metric to name a few. Additionally, review of the related work also aided in finding a benchmark and narrowing down the measures they required for the research. They were able to propose a link prediction method that seems straightforward and easy to grasp. This steps include the location of a similar node of the target, the identification of the candidate linked nodes and they ranking of those candidate. Additionally, the proposed method was claimed to have been used to conduct an experiment to evaluate the accuracy of the method using real-world micro blog data. However, the experiments or its overview have not been presented in the journal as such raising doubts about its validity. Although, they proposed a method they claimed is "effective and efficient" because of the aggregation of three categories similar nodes with different weighs which contains more useful information to recommend interesting followers and the consideration of similar users which can improve accuracy performance; they did not however show proof that the analysis or test has been carried out neither have they provided the result of the analysis. They did however give a conclusion based on their claimed results but again the proposed method cannot be taken seriously if they authors themselves cannot show or present prove of its effectiveness and efficiency to the readers. The research does have a potential to provide solution to the claimed link prediction problem. Additionally, the result presented is vague and misdirecting therefore it can neither be used as informative or conclusive. Including a brief overview of the claimed or scientific analysis/testing of the proposed method would have validated the claim of the result.

Furthermore, in Fire *et al*. [10] research they also described the link prediction (problem) as a "problem of predicting the existence of hidden links or creating new ones in social networks" [10]. They further stated the relevance of this problem to different circumstances as in recent years several algorithms have been proposed so as to solve the problem but majority of which were merely tested on bibliographic or co-authorship data sets [10]. Although not disputing Fire's claim for the description of the link prediction problem, Nandi *et al.* [7] stated that research on link prediction has evolved over the years however the main concern is the scalability of solutions that have were proposed for link predictions. Fire *et al*. did however state that in addition to the "link prediction problem", "existing link prediction techniques lack the scalability required for full application on a continuously growing social network" [10]. Gupta *et al.* [2] similarly agrees that there is a scalability problem with dynamic networks as they earlier stated that link prediction algorithms involves trying to understand the process of the networks dynamic changes and try to replicate them [2].

In their work, Nandi *et al.* [7] concluded that most of the existing work on link prediction focuses on building models that are based solely on the structure of the network, while others were built based on categorical attributes of the nodes but they strongly believe that algorithms and techniques can still be developed that will provide more accuracy and speed that is based on implicit social networks which is formed due to the daily interaction between users i.e. dynamic networks [7]. They further suggested that the key factors to consider while designing a new propagation model is its reliability on very large number of parameters but most importantly it should have the ability to handle the problem of scalability [7], [13]-[17].

III. Result/ Analysis

The result of the analysis will provide an informative rather than conclusive result, which will serve as an evidence of the validity of the proposed method. The presentation of the results will be well structured and readable, while the result itself will be vital to readers that plan on using the proposed method as it will give an idea on what to expect. The result can also be used as a benchmark or means of comparison to better comprehend the method. Since this proposed method will be applied to a social network which is known to be dynamic, knowing the size of the data used is very important because the proposed method may not be scalable or applied to certain sized data. Additionally, the logic behind the choice of social network will be presented as it will aid in knowing which type of social network this proposed method can be applied to.

IV. Conclusion

The best approach to study the question is to use the quantitative approach/method to carry out a comparison analysis of the different link prediction algorithms in a social network using performance metrics as a base and similarity as a measure of calculation. Furthermore, the main reason for using the quantitative approach is because a conclusive and more descriptive result is required to arrive at an accurate conclusion as such a standard measurement is used to avoid unfairness. Additionally, the data that will be used from Facebook to analyze the algorithm is structured in the form of numbers so as to get a result that can be used to measure the accuracy of the link prediction algorithms.

Research in this area is in its infancy stage and the interest is increasing due to its importance. Therefore, it is essential for researchers to fine tune the approaches and processes used for Social Network Analysis and Mining especially link prediction, in order to obtain efficient and accurate results. That is the purpose of this article as it will contribute to overcoming the challenges in social network data analysis and mining by attempting to resolve the link prediction problem thru proposing a scalable framework/model based on the analysis carried out. Furthermore, if link prediction is effective and accurate for dynamic social networks the results of the analysis will be beneficial to other network domains. That is, this research will as well benefit other fields in obtaining maximum accuracy, such as Bioinformatics,

E-commerce and Security. It has become critical that link prediction algorithms are accurate on small and large datasets or dynamic networks.

ACKNOWLEDGMENT

The authors sincerely thank Dr. Jean-Paul Cleron, Dr. Ferdinand Che and Dr. Mathias Mbu Fonkam from American University of Nigeria.

REFERENCES

[1] G. Nandi and A. Das, "A Survey on using data mining techniques for online social network analysis," *International Journal of Computer Science*, vol. 10, no. 6, pp. 52-57, November 2013.
[2] S. Gupta, S. Pandey, and K. K Shukla, "Comparison analysis of link prediction algorithms in social network," *International Journal of Computer Applications*, vol. 111, no. 16, pp. 27-29, February 2015.
[3] D. Liben-Nowell and J. Klienberg, "The link prediction problem for social networks," *Journal of the American Society for Information Science and Technology*, vol. 58, no. 7, pp. 1019–1031, May 2007.
[4] S. Chakrabarti, *Mining the Web: Discovering Knowledge from Hypertext Data*, California: Morgan Kaufmann Publishers, 2003.
[5] D. Hand, H. Mannila, and P. Smyth, *Principles of Data Mining*, Massachusetts: MIT Press, 2001.
[6] I. H. Ting, "Web mining techniques for on-line social network analysis," in *Proc. International Conference on Service Systems and Service Management*, 2008, pp. 1–5.
[7] G. Nandi and A. Das, "Online social network mining: Current trends and research issues," *International Journal of Research in Engineering and Technology*, vol. 3, no. 4, pp. 346-350, April 2014.
[8] D. Garcia-Gasulla and U. Cortes, "Link prediction in very large directed graphs: Exploiting hierarchical properties in parallel," in *Proc. 3rd Workshop on Knowledge Discovery and Data Mining Meets Linked Open Data-11th Extended Semantic Web Conference*, 2014, pp. 1-13.
[9] S. Yu and S. Kak, "A survey of prediction using social media," *CoRR, abs/1203.1647*, 2012.
[10] M. Fire, L. Tenenboim, O. Lesser, R. Puzis, L. Rokach, and Y. Elovici, "Link prediction in social networks using computationally efficient topological features," in *Proc. 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third International Conference on Social Computing (SocialCom)*, 2011, pp. 73-80.
[11] N. Gupta and A. Singh, "A novel strategy for link prediction in social networks," in *Proc. ACM: CoNEXT Student Workshop'14*, 2014, pp. 12-14.
[12] Y. Vaghela and M. B. Chaudhari, "Link prediction in social mining," *International Journal for Technological Research in Engineering*, vol. 1, no. 9, pp. 885-887, May 2014.
[13] P. Wang, B. Xu, Y. Wu, and X. Zhou, "Link prediction in social networks: The state-of-the-art," *Science China Information Sciences*, vol. 58, no. 1, pp. 1-38, January 2015.
[14] E. Raju and K. Sravanthi, "Analysis of social networks using the techniques of web mining," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 2, no. 10, pp. 443–450, October 2012.
[15] F. M. Facca and P. L. Lanzi, "Recent developments in web usage mining research," *LNCS 2737, Ch Data Warehousing and Knowledge Discovery*, pp. 140-150, 2003.
[16] E. Ferrara, "Mining and analysis of social networks," Ph.D. dissertation, Department of Mathematics, University of Messina, Italy, 2012.
[17] X. Y. Li, "A deep learning approach to link prediction in dynamic networks," in *Proc. the 2014 SIAM International Conference on Data Mining*, 2014, pp. 289-297.

Kalum Priyanath Udagepola is a professor as well as the head of the Department Computer Science & Software Engineering and Principal Research Investigator at the American University of Nigeria (AUN). In last two decades, his experience in teaching, research and management roles has supported across eight countries (Sri Lanka, China, Japan, Australia, USA, South Korea, Saudi Arabia and Nigeria). He was the chief scientist at Scientific Research Development Institute of Technology Australia at Australia and professor at King Abdulaziz University (KAU), Saudi Arabia. It is a critical leadership role in the senior ranks of the Nigerian academy, which has prioritized the field of ICT as a key enabler of progress. He is also the founder of "Why & How" series in MENA Region, which is giving to faculty members to enable to reach their wisdom. He received his B.S. and M.Sc. in computer science degrees from University of Colombo and then his Ph.D. degree in computer science from the Harbin Institute of Technology. He authored four books and more research papers in the areas of databases, GIS and mobile computing. The papers reflected his broad and deep knowledge in applied technologies, focusing on the interface between cutting-edge systems and real-world user needs. He is currently on the editorial board of the many Journals and Conferences. He has reviewed more than 300 research papers, convened international conferences and contributed to the development of global information system standards. He has also held senior leadership roles in the subcontinent, the Asia-Pacific, the Middle East and African, ensuring that his name is widely recognized in both the leading research nations and the key emerging market. Prof. Udagepola is recipients of Australian, China & Sri Lanka Awards. He received charted scientist status in 2013. He is a fellow in the British Computer Society & Australian Computer Society as well as a senior member in IEEE. He is a certified assessor in many professional certification bodies.

Fatima Chiroma holds a B.Sc. degree in computer science from the American University of Nigeria and a M.Sc. degree in software development from Coventry University, United Kingdom. She is reading her PhD degree in computer science at the American University of Nigeria. Her research interests are in data and web mining, social network analysis and artificial intelligence.

Bioinformatics