User behavior Based Link Prediction in Online Social Networks

Srilatha P and Manjula R
School of Computer Science & Engineering (SCOPE),
VIT University, Vellore
Tamil Nadu, India
Email: sreelatha.pulipati@gmail.com

Abstract— **In recent years, online social networks such as Facebook, Twitter, LinkedIn are most popular visited sites over the internet. Presently, there is a great interest in understanding and studying the relationships among the users in social networks. Existing link prediction methods predicts the links based on the topological structure features and the node attribute features but overlook the benefit other features such as clicks, shares, likes, forwards and comments could bring to any social network. To address this gap we propose a link prediction method based on user actions with the post which includes clicks, shares, likes, forwards and comments. In this paper, we propose link prediction model combining user action metrics and topological structure metrics. The proposed metric can bridge the gap between the existing methods and propose a new metric for defining link prediction. This is a work in progress paper, further as future direction, implementation of the proposed metrics on standard datasets by suitably training the classifiers is the topic of investigation.**

Keywords—link prediction; attribute based methods; online social networks

I. INTRODUCTION

Online social networks such as Google plus, LinkedIn, Facebook and Twitter have huge number of users and the number of users are growing every year[1]. Such a potential and diverse network of users is attracting researchers from both academia and industry to study and understand the utilization of these networks for economic and social benefits[2]. Link prediction aims at prediction of future links between the users[3]. Examples of link recommendation includes "people you may know" on Facebook, LinkedIn etc. Members of these social networks site use these social network platforms to communicate with others by posting, commenting, liking, forwarding etc. and in turn the owners of these social networks generate huge revenue by displaying advertisements on the candidate nodes (profiles) based on their preference or choice (also termed as recommender systems in literature, a particular case of link prediction problem) which drives ads to reach more users[4]. In that way one can think that the link recommendation or link prediction provides double folded benefits for the users as well as for the social networks.

Existing methods in link prediction mainly focus on the link likelihood by considering user attributes and topological structures, and predicts the links with the highest linkage likelihood[5]. But they overlook the benefit of analyzing the other existing attribute features these social networks provide such as clicks, shares, likes and comments etc., to predict links. This motivates the fact to consider these already existing features and predict links more aptly and accurately to recommend or predict links and there is a need for defining new metrics apart from the topological features of the online social networks

The paper is organized as follows: Section 2 deals with the preliminaries and related work in link prediction problem. Section 3 defines the new metrics and using it along with the existing topological structures. Section 4 ends the paper with conclusion.

II. PRELIMINARIES AND RELATED WORK

A social network is modelled as a graph for easy analysis[3]. A graph $G = (V, E)$ represents a social network with V vertices and E edges where vertices V refers to the users of the network and edges E refers to the relationship among users. The possible forming of links in the graph based on topological and attribute features constitute the link prediction problem. Various measures have been defined in literature for topological features based link prediction and attribute features based link prediction. Topological based link prediction algorithms are defined on the structural property of the graph[6] such as local structure, global structure and quasi local structure whereas the attribute based link prediction are defined on the attribute values[7] of the nodes such location, age, college, interest etc.,

We briefly review the commonly used structural link prediction indices. For a detailed note on Similarity based link prediction methods in social networks, readers are referred to [8]. Lorrain and White proposed the common neighbor based link prediction techniques. One of the most commonly used technique in link prediction. Two nodes are likely to have a link if they many common neighbors. This method of link prediction is also called as Friend-of-a-Friend[9] link prediction method.

Common neighbor Index[10] for two nodes a and b is given by

$$S_{CN}(a,b) = |\Gamma(a) \cap \Gamma(b)|$$

Where, a and b are nodes and $\Gamma(a)$ and $\Gamma(b)|$ represent set of neighbours for nodes a and b respectively.

Salton proposed an index[11] to measure the cosine similarity between rows of adjacency matrix having nodes a and b. It is also called as Salton cosine index

$$S_{Salton}(a,b) = \frac{|\Gamma(a) \cap \Gamma(b)|}{\sqrt{k_a \times k_b}}$$

Where, a and b are nodes, $\Gamma(a)$ and $\Gamma(b)|$ represent set of neighbours for nodes a and b respectively and k_a and k_b represent degrees of nodes a and b respectively.

Jaccard introduced an index[12] to compare similarity and diversity in sample sets. It is the ratio of common neighbours of nodes a and b to all the neighbours of nodes a and b. Jaccard index is given as follows

$$S_{Jaccard}(a,b) = \frac{|\Gamma(a) \cap \Gamma(b)|}{|\Gamma(a) \cup \Gamma(b)|}$$

Where, a and b are nodes, $\Gamma(a)$ and $\Gamma(b)|$ represent set of neighbours for nodes a and b respectively.

Sørenson index[13] is calculated as ratio of twice the common neighbors of nodes a and b to total degree of nodes a and b. Sørenson index is given as follows

$$S_{Sørenson}(a,b) = \frac{2|\Gamma(a) \cap \Gamma(b)|}{k_a + k_b}$$

Where, $\Gamma(a)$ and $\Gamma(b)|$ represent set of neighbours for nodes a and b respectively. k_a and k_b represent degrees of nodes a and b respectively.

Hub promoted Index[14] is calculated as a ratio of common neighbors of nodes a and b to the minimum of degrees of nodes a and b. Hub Promoted Index (HPI) is calculated as follows

$$S_{HDI}(a,b) = \frac{|\Gamma(a) \cap \Gamma(b)|}{min\{k_a, k_b\}}$$

Where, a and b are nodes, $\Gamma(a)$ and $\Gamma(b)|$ represent set of neighbours for nodes a and b respectively. k_a and k_b represent degrees of nodes a and b respectively.

Hub depressed Index[15] is calculated as a ratio of common neighbors of nodes a and b to the minimum of degrees of nodes a and b. Hub Depressed Index (HDI) is calculated as follows

$$S_{HDI}(a,b) = \frac{|\Gamma(a) \cap \Gamma(b)|}{max\{k_a, k_b\}}$$

Where, a and b are nodes, $\Gamma(a)$ and $\Gamma(b)|$ represent set of neighbours for nodes a and b respectively. k_a and k_b represent degrees of nodes a and b respectively.

These indices only provide the topological structure based similarity and doesn't take the attribute based similarity measures. In the next section we add the additional attribute based similarity measures to be used along with topological features.

III. PROPOSED NEW LINK PREDICTION METHOD

Existing social network platforms provide features such as sharing, commenting, liking the content shared by other users. These attributes can be used along with the existing attributes to improve the link prediction accuracy[4]. The below proposed metrics are extensively used in online advertising or digital marking by companies like Google through AdSense program and Facebook through Facebook Ads[16]. They use these metric to maximize the advertisement reach and only allow legitimate spending of customer budget. But using these metrics in link prediction problem in addition to node attributes is a novel approach.

A. User Action Metrics

Three types of User action metrics are defined, namely Engagement rate, Reach Rate and Impression rate.

1) Engagement rate:

It measures users' interaction with the post and promotion of the post to others circle of friends. It is a key metric for discovering how people engage with the post by sharing, commenting, likening or by clicking the post. The amount of engagement the post receives makes it feasible to understand the node's interest in the subject and similar nodes can be recommended for forming links. Relatively these are considered as potential users for recommendation.

Engagement Rate (*ER*) is calculated as the sum of likes, comments and shares made by a node,

$$ER = |\,likes| + |comments| + |shares|$$

where, |likes| is the number of pages liked by the node (i.e., user), |comments| is the number of comments made by the user and |shares| is the number of posts or content shared by the user.

2) Reach rate:

Reach is also another key metric that social media marketers use to in any product or brand awareness. It is more accurate measure than page likes. Since all the people who like the page may not see the posts and many users who do see the post may not like the page. Total reach rate is calculated from organic reach, paid reach and viral reach. Organic reach is defined as the number of unique people who saw the post in news feed. Paid reach is defined as the number of unique people who saw the post from Ads or the suggested posts. Viral reach is defined as the number of unique people who saw the post published by a friend. For example, if a fan likes, comments or share the post, their friends see the post even if they are not fans of the page. If total reach rate increases numbers of Ad click also increases which in turn generates revenue to the face book.

Total Reach (*TR*) is calculated as

$$TR = |organic\ reach| + |\ paid\ reach| + |viral\ reach|$$

Impression Rate:
Impression rate is a key metric to understand how frequently users are exposed to the post content. Impressions are calculated by counting number of times the content associated with the page is displayed. Total impression rate is calculated from organic impression, paid impression and viral impression. Organic Impressions is defined as number of times the content was displayed in a user's ticker, news feed. Paid impression is defined as number of times the content was displayed through Ads. Viral Impression is defined as number of times content associated with the page was displayed directed by a friend by liking, commenting and sharing.

Total Impression rate (*TI*) is calculated as

$$TI = |organic\ impression| + |paid\ impression| + |viral\ impression|$$

IV. Conclusion

In this paper we have defined new attribute based similarity based namely Engagement rate, Reach rate and Impression rate. This is a work in progress paper. Using these attribute based similarity measures in addition to the existing topological similarity measures increases the link prediction rate. As a future direction, we are working in direction of implementing a classifier and training the classifier for predicting the link based on the defined metrics in this paper and testing it on data sets publicly available.

References

[1] A. Perrin, "Social Media Usage: 2005-2015," *Pew Research Center*, no. October. 2015.

[2] R. N. Lichtenwalter, J. T. Lussier, and N. V Chawla, "New perspectives and methods in link prediction," *Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discov. data Min. - KDD '10*, p. 243, 2010.

[3] D. Liben-Nowell and J. Kleinberg, "The link prediction problem for social networks," *Proc. twelfth Int. Conf. Inf. Knowl. Manag. - CIKM '03*, pp. 556–559, 2003.

[4] Z. Li, "Utility-based link recommendation in social networks.," *Dissertation Abstracts International Section A: Humanities and Social Sciences*, vol. 75, no. 4-A(E). ProQuest Information & Learning, p. No–Specified, 2014.

[5] B. Chen and L. Chen, "A link prediction algorithm based on ant colony optimization," *Appl. Intell.*, vol. 41, no. 3, pp. 694–708, 2014.

[6] L. Lu and T. Zhou, "Link prediction in complex networks: a survey," *Phys. A Stat. Mech. its Appl.*, vol. 390, no. 6, pp. 1150–1170, 2010.

[7] W. H. Hsu, J. Lancaster, M. S. R. Paradesi, T. Weninger, and N. Hall, "Structural Link Analysis from User Profiles and Friends Networks : A Feature Construction Approach," *Ratio*, pp. 75–80, 2007.

[8] M. R. Srilatha P, "Similarity Index based Link Prediction Algorithms in Social Networks: A Survey," *J. Telecommun. Inf. Technol.*, vol. 2, pp. 87–94, 2016.

[9] Z. Zhang, Y. Liu, W. Ding, W. Huang, Q. Su, and P. Chen, "Proposing a new friend recommendation method, FRUTAI, to enhance social media providers' performance," *Decis. Support Syst.*, vol. 79, pp. 46–54, 2015.

[10] F. Lorrain and H. White, "Structural Equivalence of Individuals in Social Networks," *J. Math. Sociol.*, vol. 1, no. 1, pp. 49–80, 1971.

[11] G. Salton and M. J. McGill, "Introduction to modern information retrieval.," *Introduction to modern information retrieval*. 1983.

[12] P. Jaccard, "Étude comparative de la distribution florale dans une portion des Alpes et des Jura," *Bull. del la Société Vaudoise des Sci. Nat.*, vol. 37, no. JANUARY 1901, pp. 547–579, 1901.

[13] T. Sorensen, "A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons," *Biol. Skr.*, vol. 5, pp. 1–34, 1948.

[14] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A. L. Barabasi, "Hierarchical Organization of Modularity in Metabolic Networks," *Science (80-.).*, vol. 297, no. 30 August 2002, pp. 1551–1555, 2002.

[15] Y. L. He, J. N. K. Liu, Y. X. Hu, and X. Z. Wang, "OWA operator based link prediction ensemble for social network," *Expert Syst. Appl.*, vol. 42, no. 1, pp. 21–50, 2015.

[16] Simply Measured, "The Simply Measured Blog." [Online]. Available: http://simplymeasured.com/blog/. [Accessed: 18-Jul-2016].