Link Prediction in Twitter Using Similarity Metrics

Aditya Shirode

avshirod

Abstract

Given a graph at time t, Link Prediction aims to predict formation of new links at time $t^{'}$ ($t^{'} > t$). How this can be used to predict links in dynamic graph such as Twitter? Based on previous research done on Link Prediction in Dynamic Graphs, this project tries to apply a weighted-sum of similarity metrics such as Jaccard Coefficient, Adamic-Adar, Salton's cosine index, etc. to predict if there will be a newly created link between two users or a user and a topic. The results suggest that augmenting similarity metrics with an advanced method, such as machine learning, SVM or LDA would yield better results.

Keywords: Link Prediction, Graph Data Mining, Twitter, Dynamic Graphs

1. Research Question

Given a graph at time t, Link Prediction aims to predict formation of new links at time t ($t > t$). The research question this project tackles is how this can be used to predict links in dynamic graph such as Twitter?

The premise is that, given a snapshot of subset of Twitter network with $'u'$ users and $'h'$ topics the given users tweeted about, we can predict how the future interactions of the users might go, based on their current tweeting habits (recent topics and friend interactions).
This can be used to improve the 'Whom to follow' recommendation, 'Suggested Topics' (which currently works on geo-location data and current trends), and even targeted advertising in future (showing ads about things that you might be interested in future should give better results than ads based only on past information.)

There has been a lot of research on topic of Link Prediction *(LP)* in recent years. Starting from similarity metrics, which predict links based on *'Proximity'* measures, to implementing recent Machine Learning developments to Link Prediction, there are various methods that have been tried for LP. Their accuracy has not always been spectacular, but they have shown that there is latent information in the structure of the network than can be used for Link Prediction.

A quick literature survey would show a number of papers published on Link Prediction in Social Networks, such as Facebook, Collaboration network, etc. This project aims to see how these methods perform on a constantly changing complex graph of Twitter network.

The interactions in Twitter network are uni-directional, and they also hold a temporal significance. Majority of the work on Link Prediction in Twitter focuses on improving User Recommendations, predicting closeness in User profiles. This project argues that the same logic can be applied to topics a user tweets/ follows about. Based on logic that a user would generally tweet about -

- Topics that the user has *previously* tweeted about
- Topics that the users *ego-net* (friend circle) is tweeting about
- Topics that are currently *trending*

All of these properties are based on network structure, and hence can be predicted about using properties inherent to the network topology.

2. Literature Survey

(1) The paper **The Link Prediction Problem for Social Networks** by *Liben-Nowell and Kleinberg* [1] in 2004 (and later updated in 2007) has been the founding paper for Link Prediction domain. This paper formalized the Link Prediction problem and presented methods intrinsic to a network which can help accuracy in Link Prediction. They analyzed the co-authorship network of *arXiv.org* for five specific sub-domains by using prediction methods purely based on graph structure. A combination of methods based on node neighborhoods and path distances was used on Rank matrices of the original graph. They found out that accuracy of a method varies based on the intrinsic property it scores and how strong that property is in the graph.

Katz, low-rank inner product, and Adamic/Adar were found to be quite similar in their predictions; and random walk hitting time methods performed poorly overall. Their future work suggested on improving the accuracy of these methods (which was low at 16% at the time) by tracking changes over multiple timestamps, improving via feedback, and by applying various machine learning techniques for predicting the links.

This paper is a very good introduction to the topic of Link Prediction. Our project implements some of the metrics used in this paper to score the possible links in the Twitter network, although they might not be enough and need to be augmented with other techniques.

(2) The research survey **Link Prediction in Social Networks: The State-of-the-Art**, by *Peng, et.al* [2], is the recent and most cited survey on Link Prediction. The paper gives a great overview of link prediction in social networks covering both classical and latest link prediction techniques, link prediction problems, link prediction applications, and active research groups. It is comprehensive enough for a beginner who wants to learn the link prediction systematically. The paper establishes a general link prediction framework, and categorizes link prediction problems and techniques. It summarizes various metrics, their relevance in terms of a graph, their performance and time complexity. Tabular comparison of papers published on a certain topic gives a quick summary of what has been tried and tested in the field.

Although there have been a few surveys (Lu and Zhou [4], Hasan and Zaki [5]) on the domain of Link Prediction, we found this to be most comprehensive and easy to understand. It gave a fair review of the techniques used for Link Prediction and forms a basic for the features for scoring function chosen in this project.

(3) The paper **Network Growth and Link Prediction Through an Empirical Lens** by *Liu et.al.* [3], is the most recent paper on evaluation and accuracy of prediction methods on large scale dynamic graphs. They implement 18 link prediction algorithms on large detailed traces of Facebook, YouTube, and RenRen (China) social network graphs. The datasets contain time-labelled snapshots of the three networks taken over individual periods. The RenRen database as accumulated for upto two years after its creation, and contains about $10M$ nodes, $200M$ edges. They also propose filters that prune the set of candidate nodes and reduce the search space for edge creation

based on network dynamics. They compare the trade-off between score-based metrics (and their combinations), versus Machine Learning Classifiers (such as Decision Trees, SVMs). They conclude that no single score metric performs as consistently as SVM does across all networks. Although this paper has not been published in a top-10 journal yet, it is still very significant to the project at hand (over papers such as [6], [7], [8]), because it discusses the potential of using ranking metrics on dynamic graphs.

3. Proposed Solution

The Twitter Network is a large, dynamic, directed graph. We do not assign any weight to edges for simplicity. The nodes represent users (U) and hashtags (H). Based on findings shown in literature, we consider following similarity metrics to compute *score(x,y)* between users x and y, who are not neighbors at time t, but have a possibility to be in next time step $t^{'}$. We use a weighted sum of metrics explained in section 4 below to compute a score for every possible pair of x and y. Then we select the top 3 scores as our prediction.

4. Methodology

We use the twitter ego-net snapshot from *SNAP*, the Stanford Project [9], as our dataset. This dataset contains 81306 Nodes, with 1768149 Edges. As this dataset was created by crawling live Twitter feed, it represents the directional interactions between users. For this project, we only consider the *combined edges of all ego-nets*, representing set of users U. Since this dataset represents only one snapshot t, we modify it by removing 10% of edges at random, and consider these edges to be created in next snapshot $t^{'}$.
To add the `hashtags` to the data, we randomly choose a number between $(1, 15)$ (avg. number of topics a user generally tweets about), and assign those many topics (again) randomly picked from a list of topics (*#hashtags*) to a particular user u. Since we do not have the true future snapshot ($t^{'}$), we leave the part where we test accuracy of topic predictions. But an analysis of how user-user and user-topic interactions affect the probability of a user-topic link formation (or link-user links) will tell us the inherent (maybe hidden) connections between users and topics.

We use the following metrics to compute *score(x,y)* :

(1) **Common Neighbors (CN):** Number of nodes x and y have in common. The more friends you share, the more probability that you'll be introduced to each other.

(2) **Jaccard Coefficient (JC):** Normalizes the size of common neighbors. If most of my friends are your friends as well, then there should a higher probability of us being connected, than CN.

(3) **Sorenson Index (SI):** Besides having more CN, if we have less number of neighbors overall, we have a higher chance of being connected.

(4) **Salton Cosine Similarity (SC):** If we have similar interests (cosine between our vectors is low, and we point in same general direction), then there is good probability our paths will cross.

(5) **Hub Depressed Index (HD):** How many nodes we have common in proportion to max neighbors of either of us.

(6) **Adamic-Adar Coefficient (AA):** Weighing our CNs by number of neighbors they have. If we have common friends whose main friend circle consists of us, then there is high chance we'll be introduced.

(7) **Preferential Attachment (PA):** The more friends you have, the more friends of friends you have to possibly connect to.

(8) **Resource Allocation (RA):** Similar to AA, but punishes the contribution of higher degree CN more heavily. If we have friends who have many friends (we are one of many), we have less chance of being introduced through them than we have of being introduced through others.

You can refer to literature online for formulas for these metrics; mentioning them is not as relevant as the intuition mentioned above. We ignore complex methods of Machine Learning, and others like LDA, Evolutionary Algorithms, because of the time and resources it will take to use these algorithms on a large-scale dynamic graph like Twitter. We can use the metrics we calculated above as features for a ML algorithm, but that is out of scope given time constraints.

The logic that we have applied for finding common friends can be extended to include topics as well. If my ego-net is tweeting about a certain topic (using a certain hashtag), there is good chance that I will tweet about the same. Also, if a topic is trending at the time, or if someone famous tweets about something, I might continue the discussion. These are fair assumptions that we can use to analyze and predict links between users and topics.

Based on the intuition behind using above metrics, and the results shown in literature, we expect our model to perform fairly on the given dataset.

5. Results

Due to the randomness in removing edges, and only predicting top 3 possible connections based on the score, the accuracy is fairly low. For our given data of around 176K edges, (with about 10% removed for prediction), we get an accuracy of about 10% which is fairly consistent with the results that you can get with just using the plain metrics. (Refer [1]) (Also we have a poor check for accuracy due to randomness with which we are creating edges.)

We can use a machine learning model to implement supervised learning and use feedback to improve the weight vector. That will definitely improve the accuracy.

By observing the two output files *users_top3.txt* and *topics_top3.txt*, we can find out if the connections made by our algorithm were logical; but it hard to do in a data represented by just numbers. Accuracy can't tell us which metric performed better, because we do not know the underlying structure (real life connections) to help intuition. A real Twitter data would have helped to test the intuition behind the suggested connections.

6. Conclusion and Future Work

We have presented a very rough logic, and an initial set of test results, that indicate the latent information in a graph structure might point towards Link Prediction in Twitter topics.

For a full fledged study, a proper analysis of a Twitter Network, taken over two specific time snapshots, will give a better overview about the accuracy of these methods. Using a k-trace graph (which is a reduced state space version of actual connection graph) will give better time efficiency (and also

make high probable edges more evident.) It would be interesting to see what augmenting our methods with methods which require high computational power (Katz measure, PageRank, SimRank).

7. References

[1] Liben-Nowell DKleinberg J. The link-prediction problem for social networks. Journal of the American Society for Information Science and Technology. 2007;58(7):1019-1031. doi:10.1002/asi.20591.

[2] Wang P, Xu B, Wu Y, Zhou X. Link prediction in social networks: the state-of-the-art. Science China Information Sciences. 2014;58(1):1-38. doi:10.1007/s11432-014-5237-y.

[3] Liu Q, Tang S, Zhang X, Zhao X, Zhao B, Zheng H. Network Growth and Link Prediction Through an Empirical Lens. 2016. Available at: https://www.cs.ucsb.edu/ ravenben/publications/pdf/linkpre-imc16.pdf

[4] Linyuan L, Tao Zhou, Link prediction in complex networks: A survey, Physica A: Statistical Mechanics and its Applications, Volume 390, Issue 6, 15 March 2011, Pages 1150-1170, ISSN 0378-4371, http://dx.doi.org/10.1016/j.physa.2010.11.027.

[5] Al Hasan, M., and Zaki, M. J. (2011). A survey of link prediction in social networks. In Social network data analytics (pp. 243-275). Springer US.

[6] Yuxiao Dong, Jie Tang, Sen Wu, Jilei Tian, Nitesh V. Chawla, Jinghai Rao, and Huanhuan Cao. 2012. Link Prediction and Recommendation across Heterogeneous Social Networks. In Proceedings of the 2012 IEEE 12th International Conference on Data Mining (ICDM '12). IEEE Computer Society, Washington, DC, USA, 181-190. DOI=http://dx.doi.org/10.1109/ICDM.2012.140

[7] Pankaj Gupta, Ashish Goel, Jimmy Lin, Aneesh Sharma, Dong Wang, and Reza Zadeh. 2013. WTF: the who to follow service at Twitter. In Proceedings of the 22nd international conference on World Wide Web (WWW '13). ACM, New York, NY, USA, 505-514.

DOI: http://dx.doi.org/10.1145/2488388.2488433

[8] Bliss C, Frank M, Danforth C, Dodds P. An evolutionary algorithm approach to link prediction in dynamic social networks. Journal of Computational Science. 2014;5(5):750-764. doi:10.1016/j.jocs.2014.01.003.

[9] J. McAuley and J. Leskovec. Learning to Discover Social Circles in Ego Networks. NIPS, 2012.
Source: http://snap.stanford.edu/data/egonets-Twitter.html