Using Users’ Activity Metrics for Link-Prediction in a Large Online Social Network

Oliver Posegga, Kai Fischbach
Department of Information Systems and Social Networks
University of Bamberg
An der Weberei 5
96052 Bamberg, Germany
oliver.posegga@uni-bamberg.de
kai.fischbach@uni-bamberg.de

Martin Donath
voola GmbH
Maastrichter Str. 23
50672 Cologne, Germany
martin.donath@voola.de

Abstract: A considerable amount of recent research has been conducted on the link-prediction problem, that is the problem of accurately predicting edges that will be established between actors of a social network in a future time period [LNK07, LZ10].

In cooperation with the provider of a German social network site (SNS), we aim to contribute to this line of research by analyzing the link-formation and interaction patterns of approximately 9.38 million members of one of the largest German online social networks (OSN). It is our goal to understand the role users’ activity levels play in link-prediction based on local structural similarity metrics. Such metrics estimate the likelihood of future link-formation between actors based on the structure of their common neighborhoods, i.e. the networks comprised of their mutual acquaintances.

Neighborhood-based metrics are usually applied to a SNS’s social graph, which is comprised of the corresponding users and their unweighted mutual connections, which are often referred to as friendships [WBS+09]. Unfortunately, the social graph is not necessarily a good predictor of strong relationships between actors [WBS+09], as it neglects much of the information usually provided by a SNS – especially activity-related information such as private and public user interaction. Furthermore, the social graph neither allows for the differentiation between recently established and long-time relationships, nor is it designed to reflect the intensity of relationships between actors.

We argue that a consideration of both the temporal nature and the intensity of relationships could be used to improve the link-prediction performance of neighborhood-based similarity metrics. Therefore, we propose applying weighted versions of well-known neighborhood-based similarity metrics (i.e. Common Neighborhood, Jaccard’s Coefficient, Adamic/Adar, Resource Allocation, Preferential Attachment) to a combination of the unweighted social graph and a weighted graph derived from actors’ recent communication activities. Furthermore, we have developed and evaluated a set of custom metrics to capture the activity of actors’ common neighborhoods. To
evaluate the performance of the proposed metrics, we have tracked the activities and relationships of 9.38 million users of a large German SNS over a period of 60 days.

Analyzing a random sample of the resulting dataset, we have found that a small fraction of the network causes most of the observed activity: 42.64% of the network’s population account for all observed interactions and 25.33% are responsible for all private communication. We have also established that the degree of recent interaction is positively correlated with imminent link-formation – active users are more likely to establish new friendships. The evaluation of our link-prediction approach yields results which are consistent with comparable studies [LZ10, DB04]. Classical metrics seem to outperform more activity-enhanced metrics. More explicitly, Adamic/Adar [Ada03] and Resource Allocation [LZ11] seem to perform best, closely followed by Common Neighbourhood and one of our own common neighbourhood activity indices. We conclude that weighted metrics tend to predict strong ties, whereas users of SNS establish strong and weak ties. Our findings indicate that members of SNS prefer quantity over quality in terms of establishing new connections. In our case, this causes the most simple metrics to perform best.

References


