氏　　名	Lankeshwara Munasinghe
学位(専攻分野)	博士(情報学)
学 位 記 番 号	総研大甲第 1633 号
学位授与の日付	平成25年9月27日
学位授与の要件	複合科学研究科　情報学専攻 学位規則第6条第1項該当
学 位 論 文 題 目	Time-aware methods for Link Prediction in Social Networks
論 文 審 査 委 員	主　査　　准教授　市瀬 龍太郎 教　授　佐藤 健 准教授　北本 朝展 教　授　相澤 彰子　国立情報学研究所 准教授　小山 聡　北海道大学

(別紙様式 2)
(Separate Form 2)

論文内容の要旨

Summary of thesis contents

We consider the classical problem of link prediction where we are given a snapshot of a social network at time t, and we seek to accurately predict the edges that will be added to the network during the interval from time t to a given future time t +1. More concretely, we are given a large network at time t and for each user we would like to predict what new edges that user will create between t and some future time t + 1. There have been numerous attempts to address the problem of link prediction through diverse approaches. Research presented in this thesis aimed the impact of temporal behavior of nodes/links for future link evolution. Thus, we introduced novel features incorporating the aspect of time which to treat the temporality in rapidly changing social networks. We have presented details of our approach in the thesis as follows:

Chapter 1: This chapter discuss about social networks, their aspects and link prediction in social networks. Online social network services has become one of the most influential and key source of service providing, information/ knowledge sharing and many other Internet base activities. Rapid growth of social networks shows the increasing popularity of these services among the users. The growth of social networks occurs as a result of adding new users and adding new links between the existing users. The emergence of new links has primacy in the study of social network evolution. Link prediction has many applications and, it offers many benefits to the users of social networking services such as providing fast and accurate recommendations or suggestions to the users. However, highly structured massive real-world networks involving heterogeneous entities with complex associations have added new challenges to link prediction research due to different factors such as sparsity, complexity, size, time-dependent nature of the networks. There have been numerous attempts to address the problem of link prediction through diverse approaches. Most common way is to measure the closeness/similarity of nodes to each other in terms of different social aspect. Vast majority them have been used static features or attributes of nodes, links and topological features to predict the future links. Only few research have been considered the temporal behavior of nodes and links. This fact motivated us to introduce time-related features which can treat the temporality in rapidly changing social networks.

Chapter 2: In this chapter, we review some of the state-of-the-art link prediction research focused on link prediction for social network which is a fundamental data mining task in various application domains, including social network analysis, information retrieval, recommendation systems, record linkage, marketing and bioinformatics. Link prediction research has been attracted great deal of attention

with the surge of online social networking services. We summarized recent progress about link prediction algorithms, emphasizing the contributions from different perspectives and approaches, such as graph theoretic approaches, similarity-based approaches, probabilistic approaches, etc. Those methods/algorithms have been used to extract knowledge regarding the evolution mechanisms of social networks which then can be used to infer the future potential links. Finally, we outline the incompetency of handling the dynamic/temporal behavior of networks by many prediction methods discussed in this chapter.

Chapter 3: This chapter presents our problem definition and discusses supervised and unsupervised learning methods, and their usage for link prediction. Machine learning methods have been extensively used in link prediction research. It has shown that machine learning methods are extremely reliable and easy to use tool for the binary classification task of existence or non-existence of links using set of features. In our approach, we used supervised learning method for link prediction. Further, we presented the set of baseline features we used in our experiments combining supervised machine learning methods.

Chapter 4: This chapter is devoted to the novel time-aware feature which is referred as Time score, computes a score for common neighbors in terms of link strength and link weights. Number of research works has been introduced time-related features and methods to deal with temporal behavior of node and links. Those features or methods have been defined using social scientific aspects such as strength of social links. Strength of social links associated with various factors depending on the network. In most cases, the strength is strongly correlated with time-related factors but in some others are not. However, generally, strength of social links strongly correlated with time-related factors. A simplest yet vital factor is link age. Link age can be interpreted in two ways: (1) elapsed time since the creation of link (2) elapsed time since the last interaction, with respect to the current time. According to our perception, the second factor is strongly correlated with link activeness. Interactions between nodes are very important for link evolution. If transactions or interactions happen frequently and regularly the links become active and strong. Active links are very important for link evolution. Thus, we started from this point and introduced a robust feature to incorporate the effectiveness of common neighbors and their temporality using the activeness of links. In the context of social networks, the effectiveness of the common neighbors depends not only on the link weights, or number of common neighbors, but also on how long the neighbors have been in contact. The time stamps of the interactions are useful in finding such information. This information provides a far better view of the importance of common neighbors than considering only the number of common neighbors. To this end, we designed a new

feature based on the following concepts.
(1) If a node pair interacted with each other recently with respect to the current time, then the link between them becomes active.
(2) If a node interacted with its neighbors within a closer proximity of time, the neighbors are more likely to become linked.
Compiling the above considerations, we introduced a new feature, Time score (TS) which can treat the temporal behavior of common neighbors. Basically, Time score assign a score or weight for a node pair based on the activeness of its common neighbors. The score or weight used as a feature in conjunction with supervised machine learning methods in order to predict links in network data sets. We applied Time score to two social network data sets, namely, Facebook friendship network data set and a coauthorship network data set. The results revealed a significant improvement in predicting future links.

Chapter 5: This chapter presents a novel algorithm for link prediction based on information flow via active links. In the previous chapter, we introduced a novel feature called Time score defined based on link activeness, which showed an impressive link prediction performances. The fundamental assumption here is if the interactions happen frequently and recently the links become active and influence other nodes to become linked. However, Time score is limited to common neighbors. Therefore, we investigated the possible ways to extending the idea of Time score to any node pair. We identified one possible way to integrate the idea of Time score to compute information flow between any node pair. Some of the recent link prediction research has introduced supervised/unsupervised random walk algorithms to compute information flow in social networks. PropFlow algorithm is one of them which uses link weights are the transition probabilities. We extended PropFlow algorithm by incorporating link activeness and introduced T_Flow algorithm. The information flow computed by T_Flow used as a feature in conjunction with supervised machine learning methods in order to predict links in network data sets. We tested T_Flow with two social network data sets, namely, a data set extracted from Facebook friendship network and a coauthorship network data set extracted from ePrint archives. We compare the link prediction performances of T_Flow with the previous version PropFlow. The results of T_Flow algorithm revealed a notable improvement in link prediction for Facebook data and significant improvement in link prediction for coauthorship data. Further, we compared T_Flow with Time score in terms of recall.

Chapter 6: In this chapter we summarize the research presented in this thesis and our contributions. We considered the classical problem of link prediction where we are given a snap shot of a social network at time t, and we seek to accurately predict the edges that will be added to the network during the interval from time t to a given

future time t + 1. Most common way is to measure the closeness/similarity of nodes to each other in terms of various social aspects. Social networks continuously evolve in response to the underlying social dynamics, and those similarities change over time due to highly dynamic behavior of nodes and links. Clearly, older events are less likely to be relevant for determining the future linkages than recent ones. Therefore, it is necessary to develop features or methods which can treat the rapidly changing network data to understand the mechanisms of network evolution. We learnt that time-related data is very important to understand underlying mechanisms of temporality. We contributed by introducing two novel time-related features which can be used with machine learning methods for link prediction in rapidly evolving social networks. We devised two novel time-related features based on link activeness. We used timestamps of interactions, which is strongly correlated with link activeness, to define our novel features. The advantage of using timestamps of interactions is they are easy to obtain across many social networks. Further, we emphasize that these methods are easily extendable to any type of network data.

(Separate Form 3)

博士論文の審査結果の要旨

Summary of the results of the doctoral thesis screening

博士論文では，ソーシャルネットワークにおいて，ネットワークのリンクの変化を予測するという課題に取り組んでいる．ここで取り組んだ課題は，ある時点のソーシャルネットワークが与えられた時に，その後のある時点までに，そのソーシャルネットワークに追加されるリンクを予測するという問題である．これまでにリンクの変化を予測するための様々な研究が行われてきたが，リンクの予測を行うためには，時間の要素が重要であるとし，時間の概念を取り入れた新たな指標を導入することで，従来よりも精度高くリンクの予測が可能であるというのが，本論文の主張である．

本論文は，全６章からなる．第１章「Introduction」では，ソーシャルネットワークやリンク予測などの本研究の背景，動機について説明している．

第２章「Related work」では，この論文に関連する研究について述べている．関連研究を「グラフ理論アプローチ」，「確率的アプローチ」，「類似性に基づくアプローチ」，「その他のアプローチ」の４種類に分けて，順に説明をしている．

第３章「Machine learning for link prediction」では，この研究の基礎となる機械学習を使ったリンク予測手法について述べている．最初に，この研究で解く問題を改めて定義している．次に，機械学習を使ったリンク予測手法を，教師無し学習と教師付学習に分けて説明している．次に，この研究で使った機械学習手法である決定木とリンク予測で使う指標について説明している．

第４章「Time score」では，リンク予測で使う，時間の概念を取り入れた新たな指標 Time Score を提案している．この指標は，リンク予測を行う対象の共通の知人に対して，（１）現在の時間から比べてどれくらい前に関係があったか，（２）どれくらい近い時間に関係があったか，を利用して構成されている．そして，この指標を機械学習手法に適用してリンク予測を行うことを試みている．この指標を２つのデータに適用して実験を行った結果，提案した指標の有効性を確認している．

第５章「T_Flow Algorithm」では，リンク予測で使う，時間の概念を取り入れた別の指標 T_Flow を提案している．第４章で述べた Time Score は，共通の知人を基に計算を行う指標であったため，リンク予測をする対象は，共通の知人を持たなければならないという制約があった．この制約を解消するために，Time Score で使われたアイデアと情報の伝搬モデルを組み合わせた新たな指標を構築した．そして，この指標を機械学習手法に適用してリンク予測を行うことを試みている．この指標を２つのデータに適用して実験を行った結果，提案した指標の有効性を確認している．さらに，Time Score と T_Flow を実験的に比較している．

第６章「Conclusion」では，博士論文で提案した Time Score と T_Flow について考察をすると共に，博士論文の結論をまとめている．

上記のように，本博士論文は，ソーシャルネットワークにおいて，リンク予測を行うという課題に対して，時間の概念を取り入れた指標を導入することで，従来までのリンク予測手法よりも高い精度でリンクの予測を行うことが可能であることを示した点で，この研究分野の発展に貢献するものである．また，この研究で取り組んだ考え方は，近年注目を

(Separate Form 3)

浴びている社会ネットワーク分析や口コミ分析などに関する分野で，基盤となる考え方であるため，基盤技術開発という観点からも意義があると認められる．さらに，博士論文の内容は，２本の査読付ジャーナル論文，１本の査読付国際会議論文，２本の国内会議論文として発表されており，社会からも評価されている．以上より，本論文は博士論文として，十分な水準であると審査委員全員一致で認められた．