链路预测的方法与发展综述

张月霞，冯译萱
（北京信息科技大学 信息与通信工程学院，北京 100101）

摘要：链路预测对网络结构特征的演化趋势进行挖掘有着不可磨灭的促进作用。为了对网络的未来结构变化进行预测，学者们提出了许多算法。综述了 4 类较为常见的链路预测方法，分别是基于节点属性、基于网络拓扑结构、基于机器学习以及基于最大似然的方法，比较了 4 类预测方法的优劣，并概述了几种常见的衡量链路预测算法精确度标准。最后总结并展望了链路预测的未来研究方向和发展前景。

关键词：链路预测；复杂网络；相似性指标；网络结构
中图分类号：TP393　　**文献标识码**：A　　**文章编号**：1000－8829(2019)02－0008－05
doi：10.19708/j.ckjs.2019.02.003

A Summary of the Methods and Development of Link Prediction

ZHANG Yue-xia, FENG Yi-xuan
(School of Information and Communication Engineering, Beijing Information Science & Technology University, Beijing 100101, China)

Abstract: Link prediction can promote the mining of evolutionary trends in network structure features. In order to predict links in complex networks, many algorithms have been proposed. Four types of more common link prediction methods are reviewed, including node attributes, network topology, machine learning and maximum likelihood. The advantages and disadvantages of the four types of prediction methods are compared, and several common accuracy metrics for link prediction algorithms are summarized. Finally, the future research directions and development prospects of link prediction are summarized and forecasted.

Key words: link prediction; complex networks; similarity index; network structure

复杂网络能够很好地描述社会科学、自然科学、管理科学以及工程技术等领域的相互关联的复杂模型，应用范围广泛。其以复杂系统为研究目标，利用数学、统计学、计算机等科学工具，分析和研究事物的本质结构和规律。复杂网络具有规模庞大、连接结构复杂、节点种类多样以及网络演化过程复杂的特点，使得其研究充满了挑战。但是，掌握了复杂网络的演化规律可以帮助人们更好地掌控网络结构的变化趋势，因此，吸引了越来越多的人去探索复杂网络已存在的结构特征及其发展趋势。

链路预测是研究复杂网络的核心内容之一。在复杂网络中，个体被称为节点，节点与节点间的关系称为链接。链接的内在机制、形式和结构在不同系统中的演变过程揭示了人类在社会活动中的行为和趋势[1]。研究链接的产生有助于塑造和提高对人类行为和社会网络的理解，也有助于推进对其他众多领域的研究，比如社交软件的好友推荐系统、网络超链接的预测以及股市走向等。准确的链路预测为复杂网络的进化研究工作提供了有力帮助。因此，对复杂网络中的结构和链路进行研究和建模是十分必要的。

本文对节点属性、网络拓扑结构、机器学习、最大似然等 4 类链路预测方法进行总结与概括，介绍了各类方法中学者们提出的一些经典算法，对这 4 种方法的优缺点进行了阐述。对几种较为常见的评价链路预测算法精确度的方法进行了介绍，最后对链路预测的未来研究方向和发展前景进行了总结和展望。

1 网络模型介绍

在线社交网络中的用户与用户交杂出庞大的社会网络体系，网络中的节点代表用户，网络中的连边代表

收稿日期：2018－07－19
基金项目：国家自然科学基金(51334003，61473039)
作者简介：张月霞(1978—)，女，博士，副教授，主要研究方向为移动通信、卫星通信和移动互联网；冯译萱(1994—)，女，硕士研究生，主要研究方向为复杂网络、舆情传播。

用户之间关注与被关注的关系。设定 $G(V,E)$ 为一个社交网络，其中 V 为节点的集合，E 为两点间连边的集合，网络中共有 N 个节点，M 条边。那么集合中有 $\frac{N(N-1)}{2}$ 条边，即全集 U。而网络中不可能所有节点之间都存在连边，所以 $\frac{N(N-1)}{2} > M$。所有有关链路预测的研究中提出的方法，都会给节点之间的连边计算出一个评分数值 S_{ij}，并把所有节点的评分数值按高低排序，排在最前面的评分数值最高，说明两个节点链接概率越大。网络中各符号的定义如表 1 所示。

表 1　网络符号及定义

符号	定义
$k(x)$	节点 x 的度
$\Gamma(x)$	节点 x 的邻居节点的集合
$k_{in}(x)$	节点 x 的入度
$k_{out}(x)$	节点 x 的出度
$\Gamma_{in}(x)$	节点 x 的入度邻居节点的集合
$\Gamma_{out}(x)$	节点 x 的出度邻居节点的集合

链路预测的定义为：对选定的真实网络数据集进行处理，将连边集合 E 分成训练集 E^T 和测试集 E^P 两部分，且训练集和测试集并无交集。一般随机选择边作为测试集的边，时序链路预测中，会将最近时间内产生的边作为测试集。利用链路预测方法，计算出训练集中的给定节点对 (x,y) 的评分数值 S_{xy}，并将评分数值 S_{xy} 按大小进行排序，如果节点对的评分数值越大，那么节点间产生链接的可能性越大，将得到的预测结果与测试集进行对比得出评价结果，从而得出真实网络的演化模型并对其进行预测和分析。其具体过程如图 1 所示。

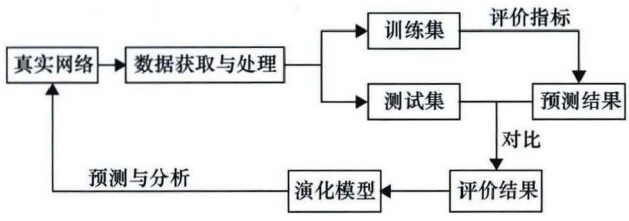

图 1　链路预测过程描述

链路预测的方法基本分为 4 类：以节点属性为基础、以网络拓扑结构为基础、以机器学习为基础以及以最大似然为基础。基于节点属性的链路预测方法有助于提高预测准确度，但是节点属性，如姓名、爱好、地理位置等信息难以获取，而且已获取信息的时效性和真实性不能确定。与该方法相比较，基于网络拓扑结构的方法容易获取网络结构信息，通过网络结构，更容易发现哪些节点更倾向于连接或断开连接。部分学者利用基于机器学习或最大似然等技术更新的链路预测方法，与节点属性、网络结构等信息相结合，综合性地对连接关系进行预测，大大提高了链路预测的准确程度。

2　链路预测方法分类

2.1　基于节点属性的方法

在早期对链路预测的研究中，多数学者采用的是基于节点属性的链路预测方法。两个节点的属性，如兴趣爱好等越相似，就越容易产生连接。在社交网络中，最简单的获得节点属性的方法就是使用标签。卜心怡等人[2]利用微博做出了实证分析，用粉丝数量、微博数量、关注人数、转发微博数量等作为节点属性标签，与基于共同邻居的算法相结合进行节点连接的预测。Jure 等人[3]利用包含超过 300 亿次对话、1.8 亿节点与 13 亿链接的数据集进行了仿真，证明人们更倾向于和相同地区、共同兴趣、爱好相似的人建立新连接。利用节点属性等信息可以提高的预测准确性，但网络用户更倾向于保护自己的隐私，信息的获取变得难度很大，即使成功获取信息也无法保证其准确性。此外，在获取信息后，如何提取有用信息也是较为烦琐的工程。因此，基于节点属性的链路预测方法在实现上具有一定的难度，更多学者倾向于使用以网络拓扑结构为基础的方法进行链路预测。

2.2　基于网络拓扑结构的方法

基于网络拓扑结构的链路预测方法的原理为：以网络结构的相似性为基础，判别节点的相似性，两个节点的结构越相似，那么它们之间越可能产生连接。以网络拓扑结构为研究基础的链路预测方法，可以分为 3 种研究方向：以局部信息、路径和随机游走为基础。

解决链接预测问题最有效的方法是建立相似性指标。不同的链路预测方法主要区别之一即是相似性指标的不同。相似性指标是定义网络节点之间相似性的评分函数，根据该函数对网络中所有的节点，两两计算相应的评分数值。

2.2.1　基于局部信息的方法

在基于局部信息的链路预测方法中，最基本的是共同邻居(Common Neighbor，CN)指标。CN 评价指标代表了两个节点间共同邻居的多少。如果其数目越多，这两个节点越可能进行连接。除此之外，CN 评价指标还有很多基于邻居其他属性的评价方式：Salton 指标[4]、Jaccard 指标、Sorensen 指标、LHN-I 指标等，如表 2 所示，分别利用了共同邻居、邻居的并集以及节点的度属性等。另外，Adamic 和 Adar 还提出了 AA 指标，其主要考虑节点对共同邻居的重要程度不同，从而

通过共同邻居的度来突出这一特点。Zhou 等人[5]对 9 种基于局部性的指标进行了准确性对比，并提出了准确度更高的资源分配(Resource Allocation，RA)指标。大量的实验结果显示，RA 指标与 AA 指标的表现相近，但在准确性上略微高于 AA 指标。

表 2　相似性指标及其定义公式

相似性指标	定义公式
CN	$S_{xy}=\lvert\Gamma(x)\cap\Gamma(y)\rvert$
Jaccard	$S_{xy}=\frac{\lvert\Gamma(x)\cap\Gamma(y)\rvert}{\lvert\Gamma(x)\cup\Gamma(y)\rvert}$
Slaton	$S_{xy}=\frac{\lvert\Gamma(x)\cap\Gamma(y)\rvert}{\sqrt{k(x)\times k(y)}}$[4]
Sorensen	$S_{xy}=\frac{2\times\lvert\Gamma(x)\cap\Gamma(y)\rvert}{k(x)+k(y)}$
LHN-I	$S_{xy}=\frac{\lvert\Gamma(x)\cap\Gamma(y)\rvert}{k(x)\times k(y)}$
AA	$S_{xy}=\sum_{z\in\Gamma(x)\cap\Gamma(y)}\frac{1}{\log k(z)}$
RA	$S_{xy}=\sum_{z\in\Gamma(x)\cap\Gamma(y)}\frac{1}{k(z)}$[5]
LAS	$S_{xy}=\frac{\lvert\Gamma(x)\cap\Gamma(y)\rvert}{k(x)}+\frac{\lvert\Gamma(x)\cap\Gamma(y)\rvert}{k(y)}$

2.2.2　基于路径的方法

基于路径的相似性指标有 3 种不同的研究方向，局部路径(Local Paths，LP)指标、LHZ-II 指标以及 Katz 指标。

CN 指标可以看作是考虑两个节点间的二阶路径数目，LP 指标在 CN 指标的基础上进行了改进，在二阶路径的基础上再考虑三阶路径数目。当考虑的阶数越来越多，趋近于无穷时，这一指标的计算公式就相当于 Katz 指标了，与 LP 指标相比，Katz 指标的计算复杂度也成倍增高。Leicht 等人提出了基于自相似度矩阵的 LHZ-II 指标，表示如果两个节点是相似的，它们在网络中的邻居也相似，那么它们建立连接的可能性较大。LHZ-II 指标可以使用网络的邻接矩阵关系，进行多次迭代遍历网络中所有节点的相互关系，进行链路预测。

2.2.3　基于随机游走的方法

基于随机游走的指标，比如 Cos + 指标、重新开始的随机游走指标(Random Walk with Restart，RWR)以及 Sim Rank 指标等。

Fouss 等人基于马尔科夫随机游走模型，通过计算平均通勤时间等信息，为节点对之间的链接定义 Cos + 相似性指标。实验结果表明，该算法在 Mivieles 数据集上表现良好，但不能很好地应用于大型数据集。张学佩[6]定义了局部随机游走的节点相似度指标，并与其他相似性指标进行比较。结果表明，其提出的算法具有更低的计算复杂度。Jeh 等人认为，如果两个节点的邻居节点相似，则它们是相似的，并据此提出了 Sim Rank 指标。

2.3　基于机器学习的方法

现有的链路预测方法除了以相似性为研究基础，还有基于机器学习的方法，主要分为三大类：特征分类、概率模型以及矩阵分解。基于机器学习的算法有助于解决数据挖掘中的常见问题，例如类别不平衡，过度拟合等。

在基于特征分类的链路预测算法方面，很多学者使用监督学习中一些较为常用的算法，比如支持向量机、决策树、径向基网络、神经网络[7]等，对未知链接进行预测，其中预测精度最高的是支持向量机方法。Li 等人[8]将快速分类法应用于支持向量机，对分类器进行训练，把大部分预测过程中测试阶段传递到训练阶段，大大减少了预测阶段的时间复杂度。Yuan 等人[9]提取推特中用户的情感信息，通过分析情感特征来判断两个用户是否可能成为好友。实验结果表明，提出的模型优于 Logistic 分类器和随机森林。

基于概率模型的链路预测算法的主要思想为，首先建立一个可以调整参数的可预测模型，然后在调整参数的过程中找到使模型达到最优的参数值，可令模型的预测准确度达到最高。构建概率模型的方法主要有马尔科夫网络关系模型以及贝叶斯网络模型等。Liben 等人[10]首先利用机器学习的方法进行了链路预测，且预测准确度较高。Asil 等人[11]使用监督学习的方法进行链路预测，并证明多数情况下，加权网络比未加权网络在监督和非监督方法中都有更好的预测结果。Gupta 等人[12]将链路预测问题视为二元分类问题，应用贝叶斯分类来预测网络中缺失的链接。基于概率模型的链路预测研究可以扩展到包括社交网络分析在内的各个领域，可以用来发现生物学中蛋白质的相互作用[19]，帮助建立电子社交网络中的推荐系统[13]，还可以识别出社交网络中隐藏的连接。

在基于矩阵分解的链路预测算法，Menon 等人[15]提出了一个扩展矩阵分解的模型来解决有向网络中的链路预测问题，并在实验中证明，显式特征通常能够提供比隐式特征更好地链路预测结果，但两者结合可以更好地提高预测性能。郭丽媛[16]提出用集体矩阵分解方法将可用的链接信息从相对密集的交互网络转移到稀疏的目标网络，通过网络相似性来建立源网络和目标网络之间的对应关系。Ahmed 等人[17]针对动态图中的链路预测问题，提出了一种基于非负矩阵分解

的方法,该方法从动态网络的时间和拓扑结构中学习潜在特征,可以获得更高的预测结果。该方法应用新的迭代规则构造具有重要网络特征的矩阵因子,并证明了算法的收敛性和正确性。

2.4 基于最大似然的方法

基于最大似然的链路预测方法的基本思路是:根据目前已经观测到的网络中的链路计算网络的似然值,并假设真实网络的似然值最大,最后根据网络似然最大化,计算所有未连接节点间的连边概率。以网络结构为研究基础的最大似然估计方法较为适用于具有组合结构的网络类型。

2008 年,Clauset 等人提出一种通过网络数据推断网络层级结构的方法,证明层次结构可以用来预测网络中丢失的链接,并具有较高的精确度。刘继嘉[18]提出了一个扩展的经典随机块模型,预测了网络中的丢失链接和错误链接。田甜等人[19]提出了一种基于最大似然估计的层次随机图模型,利用脑网络数据建立层级随机图,通过改进的马尔科夫蒙特卡洛算法对树状图空间进行采样,最后计算脑网络边的平均连接概率,实验验证,该算法比传统的算法计算复杂度低。

3 衡量算法精确度

衡量链路预测算法准确度的指标最常用的有 3 种:AUC(Area Under ROC Curve,ROC 曲线下面积)、Precision 和 Ranking Score。

3.1 AUC

AUC 的评价方式为:每次从测试集 E^P 中随机选取一条边,再随机从不存在的边的集合中选取一条边,对两者的评分数值 S_{ij}进行比较,若测试集中的边分数高,则记 1 分,若两者相等,记 0.5 分。假定一共比较 n 次,其中有 n'次测试集分数高,有 n''次两者分数相同,则 AUC 计算值为

$$AUC = \frac{n' + 0.5n''}{n} \tag{1}$$

3.2 Precision

在链路预测完成后,需要对所有的边计算的评分数值由高到低排序,而 Precision 计算的是排在前 L 位的边中,预测正确的边所占比例,即如果在预测后的排序结果中,排在前 L 位的边有 m 个与测试集中的边相同,那么 Precision 的计算值为

$$Precision = \frac{m}{L} \tag{2}$$

当两个链路预测方法的 AUC 值相同时,用 Precision 比较其准确性,Precision 值越大,越倾向于把连边准确的节点对排在前面。

3.3 Ranking Score

RankingScore 是计算测试集 E^P 中的边在评分数值大小的排序里的排名。使集合 $H = U - E^T$,表示不存在的边和测试集的并集。r_i 表示边 $i(i \in E^P)$的排序的位置。边 i 的 RankingScore 值为 $RankS_i = \frac{r_i}{|H|}$,计算测试集中所有的边得到系统的 RankingScore 值为

$$RankS = \frac{1}{|E^P|}\sum_{i \in E^P} RankS_i = \frac{1}{|E^P|}\sum_{i \in E^P} \frac{r_i}{|H|} \tag{3}$$

4 结束语

综上所述,无论是通过节点属性、网络拓扑,还是基于概率模型,都是通过已知的数据,尽可能贴近实际情况刻画链路的连接走向,但它们各自有优缺点。基于节点属性的预测方法通过获取用户信息来确定连接关系,但无法确定用户信息的真实性和准确性,深入挖掘又涉及用户的隐私问题,单从这一方面进行预测,准确率难以保证。通过网络拓扑来进行预测,计算复杂度比较低,只通过网络结构预测链路需要获取的数据较为简单。机器学习是最近较为热门的方向,与链路预测相结合会得到更高的预测准确率。概率模型是数据挖掘的传统模型,它同时考虑了节点属性和网络结构,能够得到较好的准确度,可是计算复杂度和用户属性的获取都是它无法大规模进行采用实施的原因。

目前,链路预测广泛应用于生物系统、社交关系、推荐好友方式、股市走向等方面,已经成为一门热门的交叉学科。如何在不侵犯用户隐私、计算复杂度低的条件下设计出快速、准确的链路预测模型,从而应用于大规模复杂网络上,是接下来要继续攻克的难题。

参考文献:

[1] Mohammad Al Hasan, Vineet Chaoji, Saeed Salem, et al. Link prediction using supervised learning[C]//Proceedings of SDM Workshop on Link Analysis, Counterterrorism & Security. 2006.

[2] 卜心怡,陈美灵. 社交网络中的链路预测研究[J]. 图书馆学研究,2016(17):17 - 21.

[3] Leskovec J, Horvitz E. Planetary-scale views on an instant-messaging network[C]//International Conference on World Wide Web. 2008:915 - 924.

[4] Salton G, Mcgill M J. Introduction to modern information retrieval[M]. Auckland: MuGraw-Hill, 1983.

[5] Zhou T, Lü L, Zhang Y C. Predicting missing links via local information[J]. European Physical Journal B, 2009, 71(4): 623 - 630.

[6] 张学佩. 基于 3D 卷积神经网络的多节点间链路预测方法研究[D]. 南昌:南昌航空大学,2018.
[7] 张新良,郭晓迪,朱琳. 基于神经网络的时滞非线性系统的广义预测控制[J]. 测控技术,2017,36(2):54 - 57.
[8] Li Y,Niu K,Tian B. Link prediction in Sina Microblog using comprehensive features and improved SVM algorithm[C]// IEEE International Conference on Cloud Computing and Intelligence Systems. 2015:18 - 22.
[9] Yuan G C,Murikananiah P K,Zhang Z,et al. Exploiting sentiment homophily of link-prediction[C]//Proceedings of the 8th ACM Conference on Recommender System. 2014:17 - 24.
[10] Libent-Nowel D,Keinlberg J. The link-prediction problem of social network[M]. John Wiley & Sons,Inc. 2007.
[11] Asil A,Gürgen F. Supervised and fuzzy rule based link prediction in weighted co-authorship networks[C]//International Conference on Computer Science and Engineering. 2017:407 - 411.
[12] Gupta A K,Sardana N. Naïve Bayes approach for predicting missing links in ego networks[C]//IEEE International Symposium on Nanoelectronic and Information Systems. 2017:161 - 165.
[13] Airoldi E M,Blei D M,Fienberg S E,et al. Mixed membership stochastic block models for relational data with application to protein-protein interactions[C]//Proceedings of the International Biometrics Society Annual Meeting. 2006,9(5):1981 - 2014.
[14] Huang Z,Li X,Chen H. Link prediction approach to collaborative filtering[C]//Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries. 2007:141 - 142.
[15] Menon A K,Elkan C. Link prediction via matrix factorization[C]//European Conference on Machine Learning and Knowledge Discovery in Databases. Springer-Verlag,2011:437 - 452.
[16] 郭丽媛,王智强,梁吉业. 基于边重要度的矩阵分解链路预测算法[J]. 模式识别与人工智能,2018,31(2):150 - 157.
[17] Ahmed N M,Chen L,Wang Y,et al. DeepEye:link prediction in dynamic networks based on non-negative matrix factorization[J]. Big Data Mining & Analytics,2018,1(1):19 - 33.
[18] 刘继嘉. 基于相似性演化的动态网络链路预测算法研究[D]. 合肥:中国科学技术大学,2018.
[19] 田甜,杨艳丽,郭浩,等. 基于层次随机图模型的脑网络链路预测[J]. 计算机应用研究,2016,33(4):1066 - 1069.

□