

基于随机游走的带有属性网络的链接预测

陈永祥¹ 陈 峻^{1,2}

(扬州大学信息工程学院计算机系 扬州 225009)¹

(南京大学计算机软件新技术国家重点实验室 南京 210093)²

摘要 链接预测的问题是复杂网络分析中的一个重要研究领域,已经在社会学、生物信息学、信息科学以及计算机科学等领域得到了广泛的应用。提出了一个顶点具有属性的网络链接预测的随机游走算法。在此算法中,根据顶点和属性的链接相似度定义了每一条边上的传播概率。并将顶点的属性相似度作为顶点间的相似度的初值,然后根据传输概率在网络中以随机游走的方式进行传播和更新,最终得到顶点间的相似度作为链接预测的结果得分。实验结果显示,提出的算法在顶点带属性的网络中取得了比其他算法更精确的预测结果。

关键词 链接预测,属性,随机游走,复杂网络

中图分类号 TP391

文献标识码 A

DOI 10.11896/j.issn.1002-137X.2016.6.040

Link Prediction in Networks with Node Attributes Based on Random Walks Algorithm

CHEN Yong-xiang¹ CHEN Ling^{1,2}

(Department of Computer Science, College of Information Engineering, Yangzhou University, Yangzhou 225009, China)¹

(State Key Laboratory of Novel Software Technology, Nanjing University, Nanjing 210093, China)²

Abstract Link prediction is an important area in complex network analysis. It is widely used in various domains such as sociology, anthropology, information science and computer science. A link prediction algorithm was proposed based on link similarity score propagation by a random walk in networks with node attributes. In the algorithm, each link in the network is assigned a transmission probability according to the similarity of the attributes on the nodes connected by taking a link. The similarity of the attributes on the nodes is also treated as the initial value of their link similarity. The link similarity between the nodes is then propagated by the random walk in the network according to their transmission probability. At last, the link similarity is just the final score of link prediction. Our experimental results show that the proposed algorithm can obtain more accurate results than other algorithms in the networks with node attributes.

Keywords Link prediction, Attributes, Random walks, Complex networks

1 引言

现实世界中存在的实体之间的关系可以被抽象成各种复杂的网络,例如在线社交网络、蛋白质相互作用网络、神经网络、电力网络、航空网络、用户-商品网络等。这些网络中包含了成千上万的节点以及节点之间的连边。复杂网络作为复杂系统的一种拓扑近似,在构建过程中,由于时间和空间或者实验条件的限制,难免有错误或冗余的链接出现,还有不少潜在的链接并未探测到。再者,复杂网络往往是随时间动态演化的,其链接会不断地添加或去除。因此,需要根据已知的网络信息对缺失的链接以及未来的链接进行预测,这就是网络链接预测的问题^[1-3]。

链路预测问题有重大的实际应用价值。如在生物领域研究中,蛋白质相互作用网络和新陈代谢网络^[4]节点之间存在链接,即存在相互作用关系。揭示该类网络中隐而未现的相

互作用关系需要耗费高额的实验成本,而链路预测方法的结果可以指导实验,提高实验的成功率,从而降低实验成本。对疾病-基因网络的丢失和可疑链接预测的研究,有助于探索疾病的发生机制,预测和评价相应的治疗手段,同时还可以寻找新的药物靶标,为新药研发开辟新的途径^[5]。

在社会网络分析中,链路预测同样可以作为准确分析社会网络结构的有力的辅助工具。如近几年在线社交网络发展非常迅速,链路预测可以揭示用户潜在的朋友推荐给用户^[6]。在社会关系分析中,可以发现人际之间潜在的联系^[7,8]。链路预测的思想和方法还可以用于学术网络中判断一篇学术论文的类型以及合著者^[9]。链路预测的方法也可直接用于信息的推荐,在电子商务中可以用于对客户商品推荐^[10],还可以用于电子邮件的预测^[11],在无线通讯网络中判断一个手机用户是否产生了切换运营商的倾向。在对犯罪分子组成的网络的监控中,需要利用链接预测来发现犯罪分子间隐藏的联

到稿日期:2015-03-19 返修日期:2015-06-26 本文受国家自然科学基金(61379066,61070047,61379064,61472344),国家973项目(2012CB316003),江苏省自然科学基金(BK20130452,BK2012672,BK2012128,BK20140492,BK2010318,BK201010134),江苏省教育部门自然科学基金(12KJB520019,13KJB520026,09KJB20013),江苏省研究生培养创新工程项目(CXZZ13_0172)资助。

陈永祥(1991-),男,硕士生,主要研究方向为数据挖掘、人工智能,E-mail:yxchencom@163.com;陈 峻(1967-),男,教授,主要研究方向为数据挖掘、并行与分布式处理、人工智能。

系,以防止犯罪或恐怖活动的发生。

链接预测研究不仅具有较高的实际应用价值,也具有重要的理论研究意义,例如,链路预测的研究也可以从理论上帮助人们认识复杂网络演化的机制^[12]。由于刻画网络结构特征的统计量非常多,很难比较不同的机制孰优孰劣。链路预测可以为演化网络机制提供一个简单、统一且较为公平的比较平台,从而推动复杂网络演化模型的理论研究。

对网络链接的预测有基于节点相似性、基于最大似然估计、基于概率模型等方法。基于节点相似性的方法假设两个节点之间相似性越大,它们之间存在链接的可能性就越大。为此,有很多有关节点的相似性的定义,如基于局部信息的相似性指标主要有共同邻居(CN)指标、Salton 指标^[13]、Jaccard 指标^[14]、Sorenson 指标^[15]、大度节点不利指标(HDI)^[16]、大度节点有利指标(HPI)和 LHN-I 指标^[17]、优先链接指标(PA)^[18]、Adamic/Adar 指标^[19]和 RA 指标^[20]等。基于路径的相似性指标主要有局部路径指标(LP)^[21]、LHN-II^[17]指标和 Katz 指标^[22]等。基于随机游走的相似性指标包含平均通勤时间(Average Commute Time)^[23]、Cos+ 指标^[24]、重启的随机游走(Random Walk with Restart)^[25]、SimRank 指标^[26]等。周涛等^[27,28]提出了两种新指标:资源分配指标和局部路径指标。这两种指标具有明显好于包括 Adamic/Adar 指标在内的 9 种已知指标的预测能力。刘伟平等^[29]提出了两种基于网络局部随机游走的相似性指标,通过与其他 5 种相似性指标的比较,发现有限步的随机游走可以给出比全局收敛后的预测精度更高的结果。饶君等人^[30]针对 9 种局部信息的相似性指标,设计并实现了一种基于 MapReduce 计算模型的并行链接预测算法,使链接预测应用于大型复杂网络。东昱晓等人^[31]提出了一种基于节点引力指数的预测算法,在保持低时间复杂度的同时,提高了预测的准确率。

链路预测的另一类方法是基于最大似然估计的。该方法将网络的链接可以看作某种内在的层次结构,或者是随机分块模型结构,相应提出了层次结构模型和随机分块模型,利用最大似然估计的算法进行链接预测。层次结构模型链接预测方法在处理具有明显层次组织的网络(如恐怖袭击网络和草原食物链)时具有较好的精确度。但是,由于每次预测要生成很多个样本网络,因此其计算复杂度非常高,只能处理规模不太大的网络。基于随机分块模型的链路预测方法不仅可以预测缺失边,还可以预测网络的错误链接,例如纠正蛋白质相互作用网络中的错误链接。

还有一类链接预测的方法是基于概率模型的,该类方法的基本思路就是建立一个含有一组可调参数的模型,然后使用优化策略寻找最优的参数值,使得所得到的模型能够更好地再现真实网络的结构和关系特征,网络中两个未相连的节点间存在链接的概率就等同于在该组最优参数下它们之间产生链接的条件概率。概率模型的优势在于有较高的预测精确度,但是计算的复杂度以及非普适性的参数使其应用范围受到限制。

链接预测在社会网络分析中是一个重要的工具。它可以从未知的网络中检测出隐藏的网络,或者从现有的网络结构中预测将要发生的链接。链接预测具有很多的应用,包括预测人与人之间的朋友关系,并预测他们未来的行为,例如交流与合作。在社会安全网络中,链接预测可以用来识别隐藏的

恐怖分子和罪犯^[32]。在人类行为网络中,链接预测可以检测识别人的行为和动作^[33]。链接预测在社会网络外的其他领域同样有很多应用。在传感器网络中,链接预测可以用来检测动态的性能^[34]、确保信息的保密传输^[35],以及制定最优化路线^[36]。

现在围绕顶点属性的问题得到越来越多的关注^[37-41]。许多现实世界的网络中包含众多的顶点属性,例如用户在 Google+ 中拥有包含雇主、学校、职位和居住地址等在内的属性。在属性-推理问题中,我们旨在为丢失或有不完整属性的网络结点构造属性信息。当用户在网络中设置他们的信息为共享或在创建账号提供更多的属性信息时,这个方案会取得更好的成效。属性信息应用的重要性在隐私卷入属性的推理时会格外突出,其中包括人肉搜索、影响力分析、协同过滤^[42]和用户身份辨别^[43]等。

针对顶点带属性的网络的链接预测,已经有人提出了一些方法,例如,基于进化算法的方法^[44]、基于图核的方法^[45]、基于机器学习的方法^[46,47]、基于矩阵分解的方法^[48]、基于概率模型的方法^[49-51]等。也有人提出基于随机游走的链接预测方法^[52,53]。Backstrom 和 Leskovec^[52]提出了一个监督的随机游走算法,将网络结构和边上的属性信息结合起来,但此算法并没有完全考虑顶点上的属性信息,它仅仅包含了顶点的邻居信息。

本文提出了一个基于概率传播的顶点具有属性的网络链接预测算法。在算法中,每条边都根据与其连接的顶点的属性相似度指定了一个传输概率。顶点间的相似度得分根据边上的传输概率在网络通过随机游走进行传播和更新。实验显示,提出的算法对于顶点带属性的网络取得了很好的链接预测结果。

2 问题的定义

一个网络可以表示成 $G=(V,E)$,其中 V 为顶点集合, E 是所有边集合。 $A=[a_{ij}]_{n \times n}$ 是网络的邻接矩阵,当 $a_{ij}=1$ 时表示顶点 v_i 和 v_j 之间有连接,若 $a_{ij}=0$ 则没有连接。网络中顶点的属性反映了对应顶点的本质特征。例如,在线网络社区中,顶点代表社区中的个人,顶点间的连接表示两者的关系,比如同事、亲属、同学等。对于每一个顶点,属性包含了每个人的年龄、兴趣、地址、职业等。两个具有同学关系的顶点表明他们具有类似的属性,比如年龄、教育情况和地址。两个具有同事关系的顶点表明他们具有地址和职业等类似的属性。在社会网络中,顶点的属性信息是链接预测的重要因素。利用属性信息相比于只考虑拓扑特征可以得到更精准的链接预测结果。两个顶点共有的属性越多,它们之间有关系并连接的可能性越高。

设网络中顶点有 m 个属性,用一个向量 $T_i=(t_{i1}, t_{i2}, \dots, t_{im})$ 来表示顶点 v_i 的属性。用向量 $T_i(i=1, 2, \dots, n)$ 作为行向量,那么 $T=[t_{ij}]_{n \times m}$ 就是网络的属性矩阵。对于顶点 $v_i, v_j \in V$,它们的相似度可以通过测量向量 T_i 和 T_j 的相似度得到,比如相关系数或者余弦相似度。同样可以通过两个向量的欧几里德距离、曼哈顿距离或马氏距离来计算相似度。若将顶点 v_i 和 v_j 之间的相似度表示为 $\text{sim}(i, j)$, $\text{Sim}=[\text{sim}(i, j)]_{n \times n}$ 就是网络的属性相似度矩阵。

我们的目标是通过邻接矩阵 A 表示的拓扑结构信息和

属性矩阵 T 表示的属性信息来预测隐藏的链接。链接预测的最终结果由一个链接相似度得分矩阵 $S=[S_{ij}]_{n \times n}$ 表示,其中 S_{ij} 表明了顶点 v_i 和 v_j 存在链接的可能性。 S_{ij} 的值越大,存在链接的可能性就越高,因此相似度矩阵 S 能够更量化地反映顶点间链接的存在。

3 随机游走和 SimRank 指标

随机游走是一种通过模拟粒子运动来求得图上概率分布的学习方法。它的基本思想是假想分散在图的一些顶点的粒子从这些顶点开始遍历这个图。在游走到某一个顶点时,粒子可能以概率 $1-p$ 游走到其邻居顶点,也可能以概率 p 跳到另一个随机选择的顶点。粒子每次游走后得出一个概率分布,该概率分布刻画了图中每一个顶点或者边被访问到的概率,用这个概率分布进行一轮游走。重复这一迭代过程,直至收敛为止,此时即可得到一个稳定的概率分布。随机游走模型广泛应用于数据挖掘和网络分析^[21,24,52,53],SimRank 指标的计算可以看作是随机游走的实例。

本文算法是基于 SimRank^[50] 指标的,这是基于链接相似度的方法经常所用的一个相似度指标。SimRank 指标基于网络上的随机游走,既符合人类的直觉,又具有坚实的理论背景。SimRank 指标专注于“在多领域兴趣度中创建对象与对象关系”^[50],它是基于“与两个很相似的对象有连接的对象之间也应该很相似”的假设来定义的。

给定一个网络 $G(V,E)$,其包含了顶点集 V 和边集 E ,用 $\Gamma(x)=\{u|u \in V, (u,v) \in E\}$ 表示顶点 v 的邻居。对于两个顶点 v_i 和 v_j ,用 S_{ij} 表示它们的链接相似度。由于在大多数情况下,任何事物都极大地与自身相似,因此指定与自身的链接相似度为 1。定义链接相似度矩阵的初值为:

$$S_{ij}^0 = \begin{cases} 1, & \text{if } (i,j) \in E \\ 0, & \text{otherwise} \end{cases}$$

然后, S 元素的值根据一个迭代公式而更新。在每次迭代中,每个顶点对 (x,y) 的链接相似度围绕着 $\Gamma(x)$ 和 $\Gamma(y)$ 更新。例如,考虑在图 1 中的顶点对 (a,b) 之间的相似度 S_{ab} ,假设图中的顶点 x 和 y 是相似的, x 与 a 连接并且 y 与 b 连接,那么 a 和 b 也会有某种程度的类似,因为它们与相似的顶点对连接。我们可以将 x 与 y 之间的相似度 S_{xy} 通过边 (x,a) 与 (y,b) 传输到 (a,b) 上,以此影响到 a 与 b 之间的相似度 S_{ab} 。

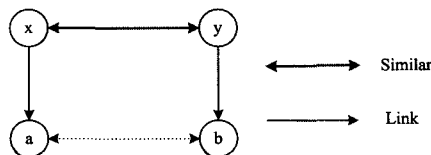


图 1 用链接相似度 S_{xy} 修正 S_{ab}

根据上述分析,对 a 与 b 之间的相似度 S_{ab} 的迭代公式为:

$$S_{ab}^{(k+1)} = \frac{c}{|\Gamma(a)| \cdot |\Gamma(b)|} \sum_{x \in \Gamma(a)} \sum_{y \in \Gamma(b)} S_{xy}^{(k)} \quad (1)$$

其中, c 为一个常数,又称为衰减系数,表明相似度在传递过程中的衰减程度。这个操作会在相似度传输和修正过程中一直被重复执行直到结果收敛。 S_{xy} 的最终结果就是顶点对 (x,y) 的链接相似度,表明了两顶点存在链接的可能性。(这样的 S_{xy} 在收敛以后也可以理解为一个随机游进程:两个随机游

走者分别以顶点 x,y 为出发点开始游走, S_{xy} 在一个确定顶点相遇的概率)

式(1)还可以表示成如下的矩阵形式:

$$S = cW^T S W + (1-c)I \quad (2)$$

其中, I 为单位矩阵, $W=[w_{ij}]$ 为传输矩阵,它的元素可由邻

接矩阵 A 计算得到: $w_{ij} = \frac{a_{ij}}{\sum_{k=1}^n a_{kj}}$ 。

4 基于随机游走的带有属性网络的链接预测

一个顶点带属性的网络 $G=(V,E)$ 可以用一个邻接矩阵 $A=[a_{ij}]_{n \times n}$ 和一个属性矩阵 $T=[t_{ij}]_{n \times m}$ 来表示。矩阵 T 的每一行是一个向量 $T_i=(t_{i1}, t_{i2}, \dots, t_{im})$,表示顶点 v_i 的所有属性。每个顶点的属性反映了对应顶点的性质, m 是属性的数量。

在顶点带属性的网络中进行链接预测,需要同时考虑邻接矩阵 A 表示的拓扑结构信息和属性矩阵 T 表示的属性信息。因此,首先要计算所有顶点对之间的属性相似度。定义顶点对 v_i 和 v_j 之间的属性相似度等同于它们的属性向量 T_i 和 T_j 之间的相似度。我们可以用余弦相似度、相关性系数和欧氏距离等方法来求相似度。在本文中,选用了余弦相似度作为属性相似度。因此对于顶点 v_i 和 v_j ,关于它们的属性向量 T_i 和 T_j 的相似度定义为:

$$sim(i,j) = \frac{\sum_{k=1}^m t_{ik} t_{jk}}{\sqrt{\sum_{k=1}^m (t_{ik})^2} \cdot \sqrt{\sum_{k=1}^m (t_{jk})^2}} \quad (3)$$

定义一个链接相似度矩阵 $S=[S_{ij}]$,元素 S_{ij} 的初值为属性相似度,即 $S_{ij} = sim(i,j)$ 。为了让链接相似度在边上传播,也要考虑顶点间的属性相似度。因此,在边 (i,j) 上指定了一个传输概率 P_{ij} 。链接两个具有较大属性相似度顶点的边上应该分配更高的传输概率。定义在边 (i,j) 上的传输概率 $P_{ij} = sim(i,j)$ 。如图 2 所示,顶点 x 连接 a ,顶点 y 连接 b ,在边 (x,a) 和 (y,b) 上的传输概率分别为 P_{xa} 和 P_{yb} 。可以将 x 与 y 之间的相似度 S_{xy} 通过边 (x,a) 与 (y,b) 传输到 (a,b) 上,以此通过传输概率 P_{xa} 和 P_{yb} 影响到 a 与 b 之间的相似度 S_{ab} 。

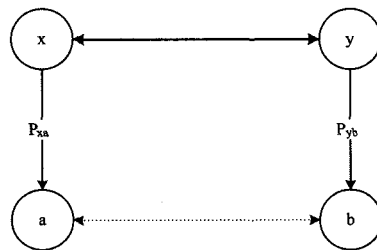


图 2 通过 P_{xa} 和 P_{yb} 传输相似度 S_{xy}

如果顶点 x 和 y 有很高的链接相似度,并且边 (x,a) 和 (y,b) 上有更高的传输概率 P_{xa} 和 P_{yb} ,那么 a 和 b 将会在某种程度上相似,因为它们连接到相似的顶点。因此,计算 S_{ab} 的迭代公式为:

$$S_{ab}^{(k+1)} = \frac{c}{P_x \cdot P_y} \sum_{x \in \Gamma(a)} \sum_{y \in \Gamma(b)} (P_{xa} + P_{yb}) \cdot S_{xy}^{(k)} / 2 \quad (4)$$

其中, c 是衰减系数, $\Gamma(x)=\{v|v \in V, (x,v) \in E\}$ 是顶点 x 的邻居集合, $P_x = \sum_{v \in \Gamma(x)} S_{xv}$ 是顶点 x 与其所有邻居相似度的总和。

该方法还可以理解成一个随机游进程。假设两个游走

者分别从顶点 i 和 j 出发随机游走在网络中。通过边 (a, b) 的传输概率为 P_{ab} , S_{ij} 意味着两者在确定顶点相遇的概率。该算法框架描述如下。

```
算法 RandWalk
输入:网络邻接矩阵  $A=[a_{ij}]$ ;属性相似度矩阵  $T=[t_{ij}]$ 
输出:顶点相似度矩阵  $S=[S_{ij}]$ 
Begin
(1)For  $i=1$  to  $n$  do
     $P_i=0$ ;
    For  $j=1$  to  $n$  do
        根据式(3)计算顶点  $v_i$  和  $v_j$  的相似度  $\text{sim}(i,j)$ ;
         $S_{ij}=\text{sim}(i,j)$ ; //设置顶点  $v_i$  和  $v_j$  链接相似度初始值;
         $P_{ij}=\text{sim}(i,j) * a_{ij}$  //设置边  $(i,j)$  上的传输概率;
         $P_i=P_i+P_{ij}$ ;
    End for  $j$ 
    End for  $i$ ;
(2)While not convergence do
    For  $i=1$  to  $n$  do
        For  $j=1$  to  $n$  do
            根据式(4)更新  $S_{ij}$ ;
        End for  $j$ 
    End for  $i$ ;
    End while;
End
```

步骤(1)中,算法首先计算属性相似度 $\text{sim}(i,j)$ 以及传输概率矩阵 $P=[P_{ij}]$,同时初始化链接相似度矩阵 $S=[S_{ij}]$ 。步骤(2)模拟随机游走过程,不断迭代更新相似度矩阵 S ,直到 S 中的所有元素值收敛。最终, S_{ij} 的值就是顶点 v_i 和 v_j 之间的链接相似度,表明两顶点之间存在链接的可能性。

5 实验及分析

5.1 实验配置

为了测试和评价提出的顶点带属性的网络链接预测的算法 RandWalk,在一些数据集上做了一系列的实验。所有的实验都是在奔腾 4 处理器、Windows XP 系统、1.7GB 内存的计算机以及 VC++6.0 环境下完成的。

用 AUC(Area Under Curve)评分标准来评价结果。在算法计算出并对所有顶点对的相似度排序后,AUC 可以解释为随机选择一个存在链接的分数高于一个随机选择不存在链接的分数概率。每一次随机选择一个存在的链接和一个不存在的链接并比较它们的链接相似度得分。如果进行 n 次独立的比较,有 n' 次存在的链接有较高的分数,并且有 n'' 次有一样的分数,则 AUC 值为:

$$AUC=(n'+0.5n'')/n$$
 (5)

一般情况下,一个较高的 AUC 值表明算法有一个较好的表现。AUC 最大值为 1,同时,对于一个随机产生的评分方案,其 AUC 值为 0.5。因此,一个有效的评分方案的 AUC 值不应低于 0.5。

5.2 测试数据集

本实验中采用的数据集都是顶点具有属性的。从一个计算机科学文献网站 Digital Bibliography Library Project (DBLP) 上选取测试了 8 种具有代表性的数据集,包括:Computer Science Conference (ACM),Computers、Environment and Urban Systems(CEUS),International Conference on Information and Communication Security (ICICS), International

Journal of Computer Graphics & Animation (IJCGA), International Journal of Network Security (IJNS),Journal of Computer-Mediated Communication (JCMC),Mathematical Structures in Computer Science(MSCS),Natural Language to Data Bases (NLDB)。

在每一个网络中提取了两个矩阵:邻接矩阵和属性矩阵。在每一个网络中,顶点表示作者,属性表示作者的相关论文标题中的关键字,链接表示了作者之间的论文的合作关系。

表 1 列出了这 8 种数据集的拓扑结构特征,其中 N 表示顶点数, M 表示总连接数, NUM_c 表示连通分量的数量以及最大连通分量中的顶点数。 e 代表网络效率, C 和 r 代表聚集系数和配对系数, K 代表平均度数。

表 1 数据集的网络拓扑结构

Networks	N	M	NUMc	e	C	r	K
ACM	1465	1209	16/688	0.0014	0.3621	0.5570	1.6505
CEUS	1047	1543	40/284	0.0041	0.7092	0.4789	2.9475
ICICS	888	1398	187/208	0.0139	0.7484	0.2726	3.1486
IJCGA	940	1699	533/160	0.0748	0.5955	0.1469	3.6149
IJNS	1059	1305	58/309	0.0039	0.6394	0.1733	2.4646
JCMC	1198	1477	39/463	0.0032	0.5138	0.6972	2.4658
MSCS	870	825	160/319	0.0080	0.4282	0.2658	1.8966
NLDB	847	1211	96/225	0.0072	0.7130	0.2112	2.8595

5.3 实验结果

将提出的算法 RandWalk 在这 8 个数据集上进行测试,并且与基于 CN,Salton,Jaccard,Sorensen,HPI,HDI,LHN_I,PA,LP 和 Kaze 等指标的其他几类算法就预测结果的精度进行比较。表 2 列出了各个算法在 8 个数据集上预测结果的 AUC 值的比较。

表 2 各个算法预测结果的 AUC 值比较

AUC	ACM	CEUS	ICICS	IJCGA	IJNS	JCMC	MSCS	NLDB
RandWalk	0.8280	0.9276	0.9421	0.9201	0.9017	0.8937	0.8842	0.9298
CN	0.8222	0.9319	0.9312	0.9116	0.8964	0.8781	0.8186	0.9091
Salton	0.8222	0.9321	0.9316	0.9119	0.8966	0.8782	0.8187	0.9093
Jaccard	0.4075	0.2894	0.3602	0.5551	0.3164	0.4737	0.4775	0.3398
Sorensen	0.8222	0.9321	0.9388	0.9115	0.8965	0.8782	0.8187	0.9093
HPI	0.8222	0.932	0.9388	0.9119	0.8966	0.8781	0.8186	0.9093
HDI	0.8222	0.9321	0.9387	0.9112	0.8965	0.8781	0.8186	0.9092
LHN_I	0.8222	0.932	0.9385	0.9106	0.8964	0.878	0.8186	0.9091
PA	0.5440	0.514	0.5481	0.6191	0.4676	0.6085	0.5611	0.4909
LP	0.8219	0.9377	0.9408	0.907	0.8994	0.8777	0.8175	0.916
Kaze	0.8219	0.9377	0.9388	0.8847	0.8992	0.8775	0.8139	0.9153

从上表中可以看出,在 8 个测试数据集中,除了 CEUS,本文的 RandWalk 算法都取得了最高的 AUC 值。尽管基于 Kaze 指标的算法利用了整体的路径信息,RandWalk 算法依然在 7 个数据集中取得了比 Kaze 更好的结果。RandWalk 算法取得更好的结果的原因是它同时结合了拓扑结构和属性信息,更精确地反映了顶点带属性网络链接预测的结果。这证明了本文 RandWalk 算法可以提升对顶点带属性网络链接预测的结果。

结束语 本文研究了对顶点带属性网络的链接预测问题,提出了一个基于随机游走的 RandWalk 链接预测算法。在此算法中,每一个顶点根据属性相似度分配了一个传输概率,将顶点之间的链接相似度得分根据边上的传输概率在网路内随机游走传播和更新。实验结果显示,本文 RandWalk 算法相比其它算法在顶点带属性网络可以取得更精确的链接预测结果。

参考文献

- [1] Lichtenwalter R N. New precepts and method in link prediction [C]//Proceedings of ACM KDD'10. 2010;243-252
- [2] Lü Lin-yuan, Zhou Tao. Link prediction in complex networks: A survey[J]. Physica A, 2011, 390; 1150-1170
- [3] Lü Lin-yuan. Link Prediction on Complex Networks[J]. Journal of University of Electronic Science and Technology, 2010, 39 (5); 651-661 (in Chinese)
吕琳媛. 复杂网络链路预测[J]. 电子科技大学学报, 2010, 39 (5); 651-661
- [4] Zhang L, Hu K, Tang Y. Predicting Disease-related Genes by Topological Similarity in Human Protein-Protein Interaction Network[J]. Central European Journal of Physics, 2010, 8(4); 672-682
- [5] Guimera R, Sales-Pardo M. Missing and Spurious Interactions and the Reconstruction of Complex Networks[J]. Proceedings of the National Academy of Science, 2010, 106(52); 22073-22078
- [6] Papadimitriou A, Symeonidis P, Manolopoulos Y. Fast and accurate link prediction in social networking systems[J]. Journal of Systems and Software, 2012, 85(9); 2119-2132
- [7] Bao Zhi-feng, Zeng Yong, Tay Y C. sonLP: social network link prediction by principal component regression[C]//Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. 2013; 364-371
- [8] Fournet J, Barrat A. Contact patterns among high school students[J]. PLoS ONE, 2014, 9(9); e107878
- [9] Sun Y, Barbary R, Gupta M, et al. Co-Author Relationship Prediction in Heterogeneous Bibliographic Networks[C]//Proceedings of 2011 International Conference on Advances in Social Networks Analysis and Mining, 2011 (ASONAM'2011). 2011; 121-128
- [10] Li Xin, Chen H. Recommendation as link prediction in bipartite graphs: A graph kernel-based machine learning approach[J]. Decision Support Systems, 2013, 54(2); 880-890
- [11] Huang Z, Lin D K J. The time-series link prediction problem with applications in communication surveillance[J]. INFORMS J. on Computing, 2009, 21; 286-303
- [12] Liu Hong-kun, Lv Lin-yuan, Zhou Tao. Uncovering the network evolution mechanism by link prediction[J]. China Sciences: Astronomy, Physics Mechanics, 2011, 41(7); 816-823 (in Chinese)
刘宏鲲, 吕琳媛, 周涛. 利用链路预测推断网络演化机制[J]. 中国科学: 物理学力学天文学, 2011, 41(7); 816-823
- [13] Salton G, McGill M J. Introduction to modern information retrieval[M]. Auckland; McGraw-Hill, 1983
- [14] Jaccard P. Etude comparative de la distribution florale dans une portion des Alpes et des Jura[J]. Bulletin de la Société Vaudoise des Science Naturelles, 1901, 37; 547-579
- [15] Sorensen T. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons [J]. Biol Skr, 1948, 5(4); 1-34
- [16] Ravasz E, Somera A L, Mongru D A, et al. Hierarchical organization of modularity in metabolic networks[J]. Science, 2002, 297(5586); 1553-1555
- [17] Leicht E A, Holme P, Newman M E J. Vertex similarity in networks[J]. Phys Rev E, 2006, 73; 026120
- [18] Barabasi A L, Albert R. Emergence of scaling in random networks[J]. Science, 1999, 286(5439); 509-512
- [19] Adamic L A, Adar E. Friends and neighbors on the web[J]. Social Networks, 2003, 25(3); 211-230
- [20] Zhou T, Lv L, Zhang Y C. Predicting missing links via local information[J]. Eur Phys J B, 2009, 71(4); 623-630
- [21] Lü L, Jin C H, Zhou T. Similarity index based on local paths for link prediction of complex networks[J]. Phys Rev E, 2009, 80; 046122
- [22] Katz L. A new status index derived from sociometric analysis [J]. Psychometrika, 1953, 18(1); 39-43
- [23] Klein D J, Randic M. Resistance distance [J]. J Math Chem, 1993, 12(1); 81-95
- [24] Fouss F, Pirotte A, Renders J M, et al. Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation [J]. IEEE Trans Knowl Data Eng, 2007, 19(3); 355-369
- [25] Brin S, Page L. The anatomy of a large-scale hypertextual Web search engine[J]. Comput Netw & ISDN Syst, 1998, 30(1-7); 107-117
- [26] Jeh G, Widom J. SimRank; A measure of structural context similarity[C]//Proceedings of the ACM SIGKDD 2002. New York; ACM Press, 2002; 538-543
- [27] Zhou T, Lv L, Zhang Y C. Predicting missing links via local information[J]. European Physical Journal B, 2009, 71(4); 623-630
- [28] Lv L, Jin C H, Zhou T. Similarity index based on local paths for link prediction of complex networks[J]. Physical Review E-Statistical, Nonlinear, and Soft Matter Physics, 2009, 80(4); 046122
- [29] Liu W P, Lv L. Link Prediction Based on Local Random Walk [J]. European Physics Letter, 2010, 89; 58007
- [30] Rao Jun, Wu Bin, Dong Yu-xiao. Parallel Link Prediction in Complex Network under Use MapReduce [J]. Journal of Software, 2012, 23(12); 3175-3186 (in Chinese)
饶君, 吴斌, 东昱晓. MapReduce 环境下的并行复杂网络链路预测[J]. 软件学报, 2012, 23(12); 3175-3186
- [31] Dong Yu-xiao, Ke Qing, Wu Bin. Link Prediction Based on Nodes Similarity[J]. Computer Science, 2011, 38(7); 162-164 (in Chinese)
东昱晓, 柯庆, 吴斌. 基于节点相似性的链接预测[J]. 计算机科学, 2011, 38(7); 162-164
- [32] Lv L Y, Zhou T. Link prediction in complex networks: A survey [J]. Physica A, 2011, 390; 1150-1170
- [33] Ahmed H S, Faouzi B M, Caelen J. Detection and classification of the behavior of people in an intelligent building by camera [J]. International Journal on Smart Sensing and Intelligent Systems, September 2013, 6(4); 1317-1342
- [34] Yang L T, Wang S, Jiang H. Cyclic Temporal Network Density and its impact on Information Diffusion for Delay Tolerant Networks[J]. International Journal on Smart Sensing and Intelligent Systems, 2011, 4(1); 35-52
- [35] Liu Z H, Ma J F, Zeng Y. Secrecy transfer for sensor networks: from random graphs to secure random geometric graphs[J]. International Journal on Smart Sensing and Intelligent Systems, 2013, 6(1); 77-94

- customer churn prediction models with advanced rule induction techniques[J]. *Expert Systems with Applications*, 2011, 38(3): 2354-2364
- [12] Sun Zhi-jun, Xue Lei, Xu Yang-ming, et al. Overview of deep learning[J]. *Application Research of Computers*, 2012, 29(8): 2806-2810(in Chinese)
孙志军, 薛磊, 许阳明, 等. 深度学习研究综述[J]. *计算机应用研究*, 2012, 29(8): 2806-2810
- [13] S Jin-bo, L Xiu, L Wen-huang. The Application of AdaBoost in Customer Churn Prediction[C]// 2007 International Conference on Service Systems and Service Management. IEEE, 2007: 1-6
- [14] Lemmens A, Croux C. Bagging and boosting classification trees to predict churn[J]. *Journal of Marketing Research*, 2006, 43(2): 276-286
- [15] Datta P, Masand B R, Mani D, et al. Automated Cellular Modeling and Prediction on a Large Scale[J]. *Artificial Intelligence Review*, 2000, 14(6): 485-502
- [16] Hung S, Yen D C, Wang H. Applying data mining to telecom churn management [J]. *Expert Systems with Applications*, 2006, 31: 515-524
- [17] Burez J, Van den Poel D. Handling class imbalance in customer churn prediction[J]. *Dirk Van den Poel*, 2008, 36(3): 4626-4636
- [18] Lecun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[C]// *Proceedings of the IEEE*. 1998
- [19] Liu Jian-wei, Liu Yuan, Luo Xiong-lin. Research and development on deep learning[J]. *Application Research of Computers*, 2014, 31(7): 1921-1930(in Chinese)
刘建伟, 刘媛, 罗雄麟. 深度学习研究进展[J]. *计算机应用研究*, 2014, 31(7): 1921-1930
- [20] Page L, Brin S, Motwani R, et al. The PageRank Citation Ranking: Bringing Order to the Web[C]// *Stanford InfoLab*. 1998: 1-14
- [21] Zhu X, Ghahramani Z. Learning from labeled and unlabeled data with label propagation[R]. Technical Report CMU-CALD-02-107, Carnegie Mellon University, 2002
- [22] Rendle S. Scaling factorization machines to relational data[J]. *PVLDB*, 2013, 6(5): 337-348
- [23] Zeng J, Cheung W K, Liu J. Learning topic models by belief propagation[J]. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2013, 35(5): 1121-1134
- [24] Xu Xiang-yang. Application of x2 Test in Analysing Students' Score Difference[J]. *Journal of Changzhou Teachers College of Technology*, 2001, 7(4): 13-16(in Chinese)
徐向阳. 卡方检验在学生成绩差异性分析中的应用[J]. *常州技术师范学院学报*, 2001, 7(4): 13-16

(上接第 203 页)

- [36] Jia X, Xin F, Chuan W R. Adaptive spray routing for opportunistic networks[J]. *International Journal on Smart Sensing and Intelligent Systems*, 2013, 6(1): 95-119
- [37] Sarkar P, Chakrabarti D, Jordan M I. Nonparametric Link Prediction in Dynamic Networks[C]// *Proceedings of the 29th International Conference on Machine Learning*. Edinburgh, Scotland, UK, 2012
- [38] He Yu-lin, Liu J N K, Hu Yan-xing, et al. OWA operator based link prediction ensemble for social network[J]. *Expert Systems with Applications*, 2015, 42(1): 21-50
- [39] Barbieri N, Bonchi F, Manco G. Who to follow and why: link prediction with explanations[C]// *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2014: 1266-1275
- [40] Bliss C A, Frank M R, Danforth C M, Peter Sheridan Dodds, An evolutionary algorithm approach to link prediction in dynamic social networks[J]. *Journal of Computational Science*, 2014, 5(5): 750-764
- [41] Cao Bin, Liu Nan, Yang Qiang. Transfer Learning for Collective Link Prediction in Multiple, Heterogenous Domains[C]// *Proceedings of the International Conference on Machine Learning*. 2010: 159-166
- [42] Melville P, Sindhvani V. Recommender systems[C]// *Encyclopedia of Machine Learning*. Springer, 2010
- [43] Bartunov S, Korshunov A, Park S T, et al. Joint link-attribute user identity resolution in online social networks[C]// *Proceedings of the Workshop on Social Network Mining and Analysis (SNA-KDD)*. 2012
- [44] Bliss C A, Frank M R, Danforth C M, Peter Sheridan Dodds, An evolutionary algorithm approach to link prediction in dynamic social networks[J]. *Journal of Computational Science*, 2014, 5(5): 750-764
- [45] Li X, Chen H C. Recommendation as link prediction in bipartite graphs: A graph kernel-based machine learning approach[J]. *Decision Support Systems*, 2013, 54(2): 880-890
- [46] Zeng Z Z, Chen Ke-jia, Zhang Shao-bo, et al. A link prediction approach using semi-supervised learning in dynamic networks [C]// 2013 Sixth International Conference on Advanced Computational Intelligence (ICACI). 2013: 276-280
- [47] Backstrom L, Leskovec J. Supervised random walks: Predicting and recommending links in social networks[C]// *Proceedings of the WSDM Conference*. 2010: 635-644
- [48] Menon A K, Elkan C. Link prediction via matrix factorization[C]// *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD)*. 2011: 437-452
- [49] Hu F Y, Wong H S. Labelling of Human Motion Based on CB-GA and Probabilistic Model[J]. *International Journal on Smart Sensing and Intelligent Systems*, 2013, 6(2): 583-609
- [50] Jia X, Xin F, Chuan W R. Adaptive spray routing for opportunistic networks[J]. *International Journal on Smart Sensing and Intelligent Systems*, 2013, 6(1): 95-119
- [51] Kang C, Pugliese A, Grant J, et al. STUN: querying spatio-temporal uncertain(social) networks[J]. *Social Network Analysis & Mining*, 2014, 4(1): 1-19
- [52] Backstrom L, Leskovec J. Supervised random walks Predicting and recommending links in social networks[C]// *Proceedings of the WSDM Conference*. 2010: 635-644
- [53] Luong N T, Nguyen T T, Jung J J, et al. Discovering Co-author Relationship in Bibliographic Data Using Similarity Measures and Random Walk Model[J]. *Intelligent Information and Database Systems, Lecture Notes in Computer Science*, 2015, 9011: 127-136