移动社交网络中链路预测方法

吕开明
金华市高亚天然气有限公司　浙江金华　321000

摘　要：随着时代发展，网络技术日益成熟，社交网络与大众的生活愈加密不可分，为大众的生活、学习及工作过程等带来更大的便利。与此同时，链路预测成为社交网络中一项重要任务，所谓的社交网络链路预测主要是指通过对现有的网络节点和网络结构进行预测，从而判断目前尚未建立连边的两个节点之间成功产生连接可能性的一项活动。从中可知，链路预测对于研究移动社会网络结构演化有着重要意义。在此结合当前社会发展背景，充分分析了移动网络的具体内涵和特点，以及现有的链路预测方法，从而以上述信息为依据，选取最常见的四种链路预测方法进行对比分析，通过实验结果，查找出最优质的链路预测方法，以期为移动社交网络发展提供更多助力。

关键词：移动社交网络；链路预测；相似性

引言：

在科技的发展下，网络技术得到广泛应用，与此同时个人无线设备也不断普及，例如现今几乎人手一部的智能手机、GPS设备等，这些移动终端设备的普及为大众生活带来更加令人满意的服务。基于此，在社会需求的指引下，移动社交网络得到深度探索，发挥了更为重要的作用。结合实际情况来看，分析移动社交网络时，可将其当成一张具有无数个点的图画，而用户便是其中的节点，人与人之间的联系便用图中边的形式表示。在移动社交网络中，联系和实体往往随着时间变化出现，这就导致社交网络呈现复杂性和高度的动态性，更新速度极为迅速。因此研究移动社交网络过程中，链路预测方法有不可替代的作用，通过此种方法可有效超预测未来两个节点建立联系的可能性，结合模型对预测结果进行对比，通过获得的结论，可在一定程度上来促进节点建立联系。

一、移动社交网络概述

移动社交网络（Mobile Social Network）是基于网络发展而逐渐形成的一种新型社交网络，以移动终端为基础，借助网络技术的便捷性和快速性等对群体行为进行，以及结合先进技术从移动设备终端的位置信息中获得群体活动规律等的一种社交网络。例如随着科技发展，人们基于互联网基础，借助移动终端，人们的社交方式逐渐实现线上和线下交互，比如当前的智能软件——画说，该产品是一个较为真实的朋友分享社区，借助该软件，可通过智能终端随时随地的观看发布在附近地点的图片，并于发布者建立好友关系，同时也可用与现实好友通过该产品以照片的方式分先彼此精彩发现，另外该产品具备评论、私信和转发等功能，为各用户进行流畅沟通提供了有力渠道。简单来说，便是在科技的辅助下，用户将自身开展社交活动的媒介从传统的PC网页为中心转移到了以移动APP为中心，看似转移极其简单，但这代表着科技迈入新的台阶，具有不可替代的意义。将社交活动的媒介从传统的PC网页转移到移动APP上，其深层次内涵便是，在社交网络服务形成初期，人们将线下生活相关信息数据转移到线上，借助网络技术实现低成本管理，从而发展形成具有一定规模的虚拟社交活动，到随着科技发展，移动APP得到广泛应用，在其帮助下，真实生活和虚拟社会形成更深层次的交织的过程[1]。

二、移动社交网络的特点

移动社交网络基于互联网技术，实现移动通信网络与互联网深度融合，也就是用户通过使用先进的移动终端设备，例如智能手机、平板电脑等，通过移动通信网络访问互联网，借助网络技术从而更为方便地开展社交活动。当前在各种技术和科学理论的支持下，移动通信网络不断发展，从最初的2G逐渐进步，形成了3G、4G、5G等，另外目前的Wi-Fi，GPRS也得到广泛应用。同时，移动APP由于其自身的优势愈加受大众喜爱，越来越多的用户应用APP访问互联网，具体来看，其主要行为包括以下几方面，其一是移动信息搜索、移动网页信息浏览，其二是移动应用程序下载和在线应用，例如各种在线游戏等，其三在线进行电子书阅读，其四进行移动音频播放、下载及视频播放等，其五移动社交网络服务行为，例如移动邮件、微博等，其六是移动电子商务，例如网购，其七是移动网络办公，借助互联网和移动通信网络实现远程会议等[2]。结合上文进行深度分析发现，

移动互联网与传统互联网的主要区别在于以下三点，分别是用户、接入网络和终端，移动互联网在这三方面占据更大优势，其更加具有便捷性、移动性、上下文感知敏锐度以及终端个人化等固有属性。同时移动互联网可为用户标记更为明确、真实的用户标识，并可方便地从概貌层面对移动用户进行刻画。而传统互联网与之相比，其虽然也具备便捷性，但移动性不够灵活，也难以为用户提供个人化服务，另外传统互联网中用户的标识则相对较为模糊，难以保证信息真实度。例如移动用户的人口统计学数据是根据用户在注册APP时真实填写数据计入，或者是借助现今科技，根据机器学习或移动数据挖掘技术推理获得，例如在移动用户授权许可范围内，结合身份证信息获取籍贯、获取生日日期等；根据用户购买记录预测其收入情况，便于针对性地进行推荐；根据用户的社交网络信息挖掘用户的工作背景及教育背景等。移动互联网中数据收集多数由更为科学的依据，相比之下，其较传统互联网获取的数据更为真实可靠。除此之外，移动互联网还可以通过其他方式获取用户的相关信息，例如借助当前应用较为广泛的GPS全球定位系统（global positioning system）获取移动用户的地理位置信息数据或者查看用户的运动轨迹等，也可以通过机器学习和数据挖掘技术分析用户行为从而获得更为深层次的信息数据，了解更多关于用户的属性特征[3]。综上，结合上文论述可知现阶段的移动社交网络主要具备以下两个特点，概括如下。

1.移动性

用户借助移动APP及移动通信网络可在任何时间、任何地点在满足服务标准的基础上访问应用系统，位置和时间与用户之间的联系更为紧密。

2.瞬时性

用户借助移动终端，可在特定时间和特点地点利用移动通信网络进入应用程序，从而满足自身需求，这使得移动社交网络具有瞬时性，可及时更新信息数据。

三、链路预测现有方法

社交网络中的链路预测主要是指如何通过已知的网络结构等信息，预测网络中尚未产生连接的两个节点之间产生连接的可能性。具体来看，在社交网络中，网络顶点代表用户，边代表用户关系，链路预测即是对用户未来发展关系的分析。当前社会网络链路预测模型可大致分为3类，其一是基于监督学习为主的分类模型，例如决策树、神经网络、SVM、KNN及集成方法中的bagging、boossting和随机森林等。其二是概率模型，此种模型在应用时，需要事先建立一组可调参数的模型，然后应用优化策略寻找最优参数值，从而保证模型达到最佳状态，这时两个未连边的节点对的概率就是它们产生连边的条件概率。概率模型的构建方法有贝叶斯网络模型和马尔科夫网络关系模型等。其三为线性代数方法，此方法是通过降阶相似矩阵来计算网络中节点之间的相似性。Kuegis等人利用图的邻接矩阵，并定义一个函数F使得两个时刻的邻接矩阵的差异性最小，这样就将链路预测问题转换成线性代数优化问题，之后再通过矩阵变换和降维的方法将问题转换为一维的最小二乘曲线拟合问题[4]。

结合上文内容，当前现有的链路预测方法主要有以下三类，第一是以机器学习为基础的方法。Hasan等人结合前人的研究，经过总结指出监督式学习和预测，在社交网络中的两个潜在连接节点是否为一个正的或者负的示例，他们结合实际情况，尝试了多个分类算法模型，但其工作成果较时代发展滞后，与复杂网络的研究进展存在差距。其中第二种方法以最大似然估计为基础。Clauset，Moore和Newman经过研究提出了一种方法，他们从数据网络分层结构入手进行思考，并进行深度分析推断。但此种算法相对而言存在明显的缺点，其运行速度相对较慢，与当前飞速发展的网络难以匹配。其三类为基于节点邻居为基础的测量方法。结合相关研究可知，拥有更多的共同邻居数的节点，其在未来拥有连边的可能性越大[5]。

四、移动社会网络预测方法具体阐述

针对上文的描述，可将移动社交网络看做一张图片，其中个人或者实体便是其中的节点，之间的边便是两个节点之间存在联系。因此，在特殊时间t时，可将图看成G（V，E），其中V代表节点的集合，而E表示集合V上的无向边集。所谓的链路预测算法便是在未来时间 t' $\left(t'\rangle t\right)$ 中预测节点建立新的链接或者重新建立已丢失的链接的可能性。对于没有建立链接节点的 $x,y\in V$ ，然后为其分配一个相似性分数 $sim\left(x,y\right)$ ，这就是节点 x,y 之间的相似性。然后结合这些分数降序排列节点，相似性得分越高的节点，其在未来建立链接的可能性越大。

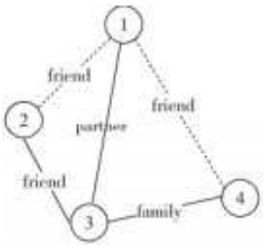

图1　移动社交网络链路预测示例

图1简单地阐述了链路预测的相关内容。在t时刻，节点2和节点3是朋友，节点1和节点3是合作伙伴，结合图示，在 t' 时刻，可能节点3将节点2介绍给节点1，同理节点4也可能会认识更多的节点。

结合上文内容，在此选择基于节点邻居为基础的测量方法进行详细阐述。在此主要对四个经典的基于局部信息的相似性指标进行比较，具体来看，分别是共同邻居CN（common neighbors），Adamic－Adar指　数（AA）、Leicht－Holme－Newman指 数（LHN-I） 和Jaccard指 数（JC）。接下来对方法涉及内容进行详细阐述。

首先CN方法主要指如果两个节点之间的共同邻居越多，则这两个节点更倾向于连边。确切地讲，此种方法的可定义为：

$$Sin_{CN}(x,y)=\left|\Gamma(x)\cap\Gamma(y)\right| \quad (1)$$

其中上述式子中，$\Gamma(x)$ 代表网络节点中 x 的邻居节点集合，本文研究中关于 x 的邻居节点定义为与节点 x 直接相连的节点，只有符合此要求才能计入计算范围。节点 x 与节点 y 的相似度主要是指它们两者共同拥有的相邻节点个数。

其次，AA指标思想主要指度小的共同邻居节点的贡献度相较于度大的共同邻居节点更大，因此根据共同邻居节点的度为每一个节点赋予相应的权重值，则赋予权重值等于该节点的度数值的对数分之一，也就是 $\frac{1}{\lg k}$。其具体定义如下：

$$sim_{AA}(x,y)=\sum_{z\in\Gamma(x)\cap\Gamma(y)}\frac{1}{\log k(z)} \quad (2)$$

再次，LHN-I指标为部分节点赋予了较高的相似性，这些节点有很多的共同邻居数，之所以出现部分节点共同邻居数增多是因为这些邻居预期数量存在，而不是由于这些邻居节点潜在最大值的存在。因此该定义式子分母较为特殊，为 $k(x)\times k(y)$。与节点 x 和节点 y 的共同邻居数的预期数目成正比。具体定义式为：

$$sim_{LHN-1}(x,y)=\frac{\left|\Gamma(x)\cap\Gamma(y)\right|}{k(x)\times k(y)} \quad (3)$$

最后的为JC指数，其是确保那些具有共享较高比例共同邻居数的节点对享有较高的预测值。

其定义式为：

$$sim_{Jaccard}(x,y)=\frac{\left|\Gamma(x)\cap\Gamma(y)\right|}{\left|\Gamma(x)\cup\Gamma(y)\right|} \quad (4)$$

五、结果分析

结合上文内容，本文采用的衡量链路预测算法精度的指标为AUC，AUC是从整体出发，衡量算法精度的一种指标。其定义为：$AUC=\frac{n'+0.5n''}{n}$。上式中 n 表示独立比较的次数。n' 表示测试集中的分数值大于不存在边的分数值次数，n'' 表示两个分数值相等的次数。

每次随机从测试集合中选取一条边与随机选取的不存在的边进行比较，判断测试集中的边地分数值，如果发现测试集中的边分数值大于不存在边的分数值，此时可加1分，如果两个分数值相等，则加0.5分。

为了保证数据集准确、可靠，在此选择的是由米兰大学（UniMi）掌上移动跟踪记录设备进行辅助，收集需要用到的相关数据。该数据集包含了米兰大学44个移动设备的运动轨迹，可尽可能地保证信息数据完整。

表1　数据描述

数据集	米兰大学数据
节点	49
链接数	11895

为了能够支持研究进行，顺利产生测试集，首先对测试网络中的所有未链接节点进行分析，确定初始值，做好前期准备工作，然后开始安排测试网络中的节点类别，将其划分为正类或者负类。其中下表为具体情况信息。

表2　训练集的节点分布

数据集	米兰大学数据
有链接的	592
无链接的	584
总共数目	5122

表3　测试集的节点分布

数据集	米兰大学数据
有链接的	360
无链接诶的	816
总共数目	1924

结合上表数据，本文主要采用对比的方法进行研究，通过数据的对比从而得出结论。将取得的数据集用上述的四种不同形式的方法进行验证，最终得到不同的数据结果。为了提高实验数据的真实性和准确度，应用MATLAB仿真工具获得真正的跟踪数据。米兰大学中的数据集是一个很小的移动社交平台，其中充分体现了移动社交网络的特点，在此网络中不同节点之间可以频繁进行互动。因此，以该学校的数据集为基础支持研究有实际参考价值。（见图2）

结合图1来看，从中可明确发现，随着时间变化，链路数目显现周期性变化，因此，很研究结合实际情况，综合考虑具体因素影响，选择400000s到800000s的数据

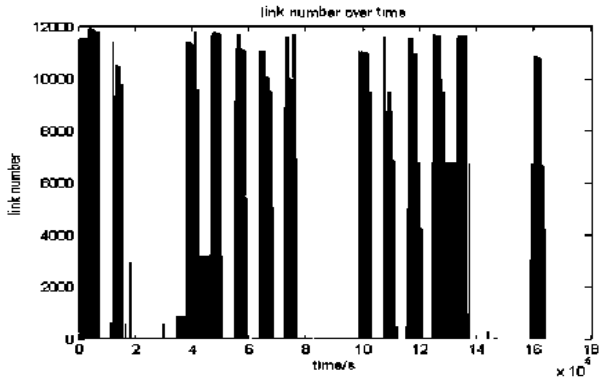

图2　链接数量随时间而变

为基础开展实验研究。

最终，结合数据信息开展了100次独立实验，实验次数较多，可确保实验不受随机性影响，具有实际价值。经过独立实验研究，从中发现CN指标是在所有指标中表现最佳。结合表4，基于共同邻居的方法（CN）的AUC值较其他算法值相对更高。基于此，结合实验研究结论，可知道CN预测效果和预测精度较其他算法更具有参考性，可以更为精准地预测节点之间的下一时刻连边时间。

表4　不同方法在400000s到800000s的AUC值

指标	CN	AA	LHN-1	JC
AUC值	0.9735	0.9730	0.9715	0.9712

六、结语

移动社交网络是以社交网络为基础进行发展而来的，社交网络是随着E-mail、BBS、博客、微博等Internet的应用而自然发展起来的反映社会交往群体的一种形态，其本质是提供一个在人群中分享兴趣、爱好、状态和活动等信息的在线平台。随着互联网发展起来的社交网络对人类社会活动的方式、效率等产生了深远影响。在科技力量的支持下，移动终端不断的发展，现今借助高科技手段，人们已经实现了随时随地进入网络访问，与此同时，如何在快速发展的移动社会网络中通过分析网络变化，分析获得人们关注的内容，成为研究重点，关于此方面的研究受到很多学者重视。本文探究了移动社交网络具体内容和其特点，然后对集中链路预测方法进行研究分析，利用AUC指标进行对比，结合结论，发现在研究的几种方法中，应用共同邻居方法（CN）取得的效果最佳。

参考文献：

[1]顾秋阳，琚春华，吴功兴.基于子图演化与改进蚁群优化算法的社交网络链路预测方法[J].通信学报，2020，41（12）：21-35.

[2]袁榕.链路预测改进算法的研究[D].南京邮电大学，2020.

[3]潘永昊，于洪涛.基于网络同步的链路预测连边机理分析研究[J].自动化学报，2020，46（12）：2607-2616.

[4]赵学磊，季新生，刘树新，赵宇.基于广义共同邻居的有向网络链路预测方法[J].网络与信息安全学报，2020，6（05）：89-100.

[5]彭逸超.异构信息网络中链路预测问题的研究[D].哈尔滨工业大学，2020.

[6]孟绪颖，张琦佳，张瀚文，张玉军，赵庆林.社交网络链路预测的个性化隐私保护方法[J].计算机研究与发展，2019，56（06）：1244-1251.