

科学知识网络中的链路预测研究述评^{*}

张 斌 马费成

摘 要 本文以“科学知识网络中的链路预测”为主要对象,对链路预测的类型、研究思路和方法等相关理论进行了回顾,将知识网络划分为同质网络和异质网络两种类型,从合作网络、引证网络和二分网络三个方面对同质网络的研究进行梳理,并介绍了一些异质网络中的链路预测方法。认为:针对这方面的研究近年来有成为图书情报学领域研究热点的趋势;已有研究多是描述各种链路预测指标在不同类型知识网络中的预测效果,未来应当利用链路预测量化和评价演化模型,识别和分析异常链路,以发现知识热点和创新趋势,将知识网络的研究提升到应用层次。图 5。表 2。参考文献 68。

关键词 知识网络 链路预测 同质网络 异质网络 合作网络 引证网络 二分网络

分类号 G350

A Review on Link Prediction of Scientific Knowledge Network

ZHANG Bin & MA Feicheng

ABSTRACT

Currently, the research about link prediction of knowledge network is scattered in the fields of Statistical Physics, Computer Science, Complex Networks and Library and Information Science. From the perspective of Library and Information Science, this paper mainly reviews researches on link prediction of knowledge network and systematically analyzes previous researches.

This paper used “link prediction” as topic to retrieve the literatures in WoS, and got 300 records. Using CitNetExplorer to analyze the direct citation relationships among the 300 records, we obtain the highly cited interrelated literatures to review on current link prediction types and research ideas.

Link prediction can be classified into two types: the static and the dynamic, and the corresponding dataset partition methods are different. The former uses random sampling, while the latter needs to consider the temporal state. In the field of Library and Information Science, knowledge networks vary in size and scale. So the link prediction of such knowledge networks uses the similarity-based algorithms in order to reduce computing complexity, but also introduce certain semantic and attribute information to ensure the accuracy

^{*} 本文系国家自然科学基金面上项目“知识网络的形成机制及演化规律研究”(编号:71173249)和国家自然科学基金重点国际(地区)合作研究项目“大数据环境下的知识组织与服务创新研究”(编号:71420107026)的研究成果之一。(This article is an outcome of the general project “Research on formation mechanism and evolution laws of knowledge networks” (No. 71173249) and the major international (regional) joint research project “Research on knowledge organization and service innovation in the big data environments” (No. 71420107026) supported by National Natural Science Foundation of China.)

通信作者:张斌, Email: zb0205@126.com, ORCID: 0000-0002-5591-7874 (Correspondence should be addressed to ZHANG Bin, Email: zb0205@126.com, ORCID: 0000-0002-5591-7874)

of prediction.

This paper divides the knowledge network into homogeneous and heterogeneous networks. For the link prediction of homogeneous network, this paper reviews on the research progress from the aspects of co-authorship network, citation network and bipartite network. The co-authorship network can be viewed as an undirected network, and is the easiest way to describe the real network system. This paper summarizes the predictors and steps in link prediction of co-authorship network and examines its prediction effect from author, institution, and country level. The citation networks can be viewed as a directed network, and is the first proposed knowledge network. Compared to the co-authorship network, the citation network not only has the structure information of basic data, but also involves the external information, such as authors, journals and content of articles. Therefore, the citation network not only can be predicted based on the local structure information, but also by the external information, or by the integration of structural and external information to perform machine learning to improve the accuracy of link prediction. The bipartite network is a special kind of homogeneous network, which can be used to observe structural characteristics and evolution process between subject and object. Using link prediction in bipartite network can solve some problems of recommendation systems. For the link prediction of heterogeneous network, this paper reviews on some link prediction methods, such as forecasting model based on meta-path.

This paper finds out that link prediction of knowledge network in Library and Information Science has become a hot issue in recent years. Most of current studies are empirical studies describing the prediction effect in different types of knowledge networks based on a variety of predictors in order to determine the application scope of link prediction. This paper proposes that in the future research, the link prediction should be used to quantify and evaluate the evolution models, to identify and analyze anomalous links, and to discover knowledge hot spots and innovation trends. Ten years of development shows that knowledge network can be used as an ideal carrier for link prediction, and link prediction is a powerful tool for analyzing knowledge network. On this basis, carrying out applied research from link prediction perspective based on the structure and evolution of knowledge network will be the future research direction of knowledge network in the field of Library and Information Science. 5 figs. 2 tabs. 68 refs.

KEY WORDS

Knowledge network. Link prediction. Homogeneous network. Heterogeneous network. Co-authorship network. Citation network. Bipartite network.

长久以来,人们一直在研究文献单元、信息单元、知识单元之间的内在联系,从而可以更好地组织、导航、检索和利用文献信息资源。这一研究路线主要是对文献信息等显性知识进行组织、管理和优化,并形成了以文献计量学、科学计量学和信息计量学为核心的研究领域。近些年,科学知识网络(以下简称知识网络)逐渐成为计算机科学、管理学、教育学、心理学、图书情报学等诸多领域共同关注的热点问题。在具体

研究中,以 Price^[1]为代表的统计物理方法和以 Brookes^[2]为代表的认知地图构想反映和奠定了当前的两种主流研究思路^[3],并且在知识网络的结构和演化问题上取得了很多研究成果^[4]。

研究知识网络的结构和演化是厘清知识发展脉络,对知识的发展趋势和演化动态进行预测的基础。然而,刻画网络结构特征的统计量非常多,不同的演化机制和模型也很难判断孰优孰劣。最近几年,复杂网络中的链路预测方

法受到越来越多的关注,链路预测在网络重构、网络演化模型评价、推荐系统等方面有着重要应用^[5]。知识网络是复杂网络的一种,本文在回顾链路预测相关理论的基础上,从图书情报学的角度梳理和总结知识网络中链路预测方面的研究成果,分析当前研究的不足,并提出未来的研究重点。

1 理论回顾

网络中的链路预测是指通过已知的网络节点和网络结构等信息来预测网络中尚未产生连边的两个节点之间产生连接的可能性^[5-7]。在WoS中精确检索主题为“link prediction”的文献,得到300条记录(检索时间为2015年1月2日)。使用CitNetExplorer^[8]对这300篇文献间的直接引证关系进行分析,其中被引得分前三的文献作者及内容是:Liben-Nowell和Kleinberg^[9]提出基于网络拓扑结构的相似性定义方法,将指标分为基于节点和基于路径两类,并分析了若干指标在社会合作网络链路预测中的效果;吕琳媛和周涛^[5]关于复杂网络中链路预测问题的综述;周涛等^[10]分析10种基于节点局部信息的相似性指标在6个真实网络中进行链路预测的表现。本文以“知识网络中的链路

预测”为主要对象,基于链路预测的高被引文献回顾所涉及的预测类型和研究思路。

1.1 链路预测的类型

链路预测可以分为两类,未知链路预测和未来链路预测。未知链路(missing links)是指网络中实际存在但尚未被探测到的链路,这实际上是数据挖掘的过程,是静态的。未来链路(future links)是指网络中目前不存在,但应该存在或将来很可能存在的链路,这与网络演化密切相关,是动态的。两者对应的数据集划分方法也有所不同,前者多采用随机抽样,后者需要考虑时序状态。这两种类型的链路预测都很重要,可以补充或修正我们对事物的认知,并指导实践。

在图1展示的静态链路预测中,子图a表示研究使用的样本网络,或者称为观察网络,它经数据集划分后得到子图b,子图b称为测试网络,其中的实线表示训练集,虚线表示测试集。在图2展示的动态链路预测中,子图c表示在时间(段) t 下的网络,可作为训练集使用,称为训练网络,子图d表示由链路预测算法得到的节点和连边组成的网络,称为预测网络,子图e表示在时间(段) t' 下的网络,可作为测试集使用,也称为测试网络。

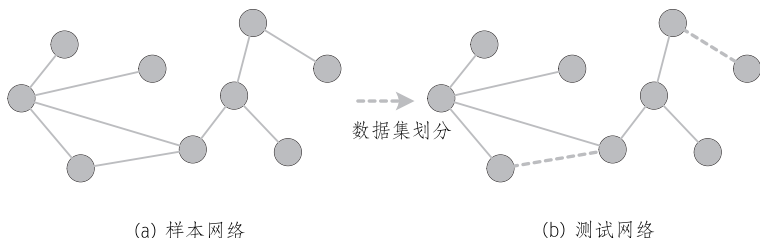


图1 静态链路预测(未知链路)

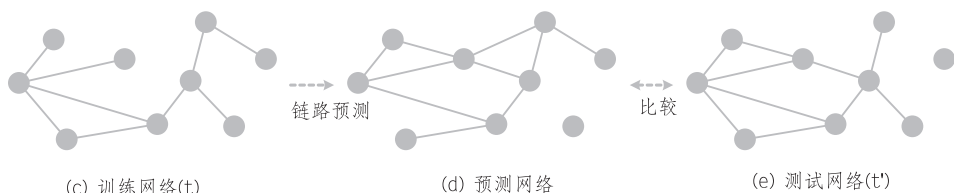


图2 动态链路预测(未来链路)

1.2 研究思路和方法

早期关于链路预测的研究思路和方法主要基于马尔可夫链和机器学习^[11-12],这在计算机科学领域已有较深的研究。在吕琳媛和周涛^[5-6]关于复杂网络链路预测的综述中,总结了基于网络结构的链路预测的三种研究思路。

第一种研究思路是利用节点间的相似性进行链路预测。这种相似性表达的是一种接近程度(proximity),研究的前提假设是两个节点之间的相似性越大,它们之间存在链路的可能性就越大。刻画相似性的方法有很多,其中最简单的是利用节点的属性。在社会网络中,如果两个人具有相同的年龄、性别、职业、兴趣、爱好等,就说明他们很相似;而在知识网络中,如果两篇论文具有相似的标题、关键词、主题词、摘要等,也可以说明它们很相似。基于节点自身属性来判断相似性的确可以得到很好的预测效果,比如论文相似性检测,但在很多情况下获取这些信息并不容易。因此,讨论更多的是基于网络结构信息的链路预测方法,其中最简单的相似性指标是共同邻居,即两个节点如果有更多的共同邻居就更可能产生连边。在社会网络中,如果两个人拥有很多相同的朋友,那么这两个人很可能也是认识的^[13];而在知识网络中,有更多共同合作者的两个科学家在未来合作的可能性较高^[14]。

第二种研究思路是基于最大似然估计进行链路预测。Clauset等^[15]提出了一种利用网络的层次结构进行链路预测的方法,在具有明显层次结构的网络中,如恐怖袭击网络、草原食物链等,预测的精确度较高。Guimerà和Sales-Pardo^[16]提出了一种基于随机分块模型(stochastic block model)^[17]来预测网络缺失连边和识别错误链路的方法。随机分块模型是一种基于最大似然估计的方法,其基本思想是网络中的节点可以分为若干模块,而两个节点之间存在连边的概率只由它们所在的模块决定。

第三种研究思路是利用概率模型进行链路预测。其基本思路是建立一个含有一组可调参

数的模型,之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,网络中两个节点产生连边的概率就等于在该组最优参数下它们之间产生连边的条件概率。

上述三种研究思路的立足点和研究对象的不同,导致其研究方法和应用范围存在差异。目前,关于第一种研究思路的探讨最多,应用范围也最广,其相似性指标也有不同的划分标准,如:①基于邻居节点和基于路径的^[9],②基于局部信息、基于路径和基于随机游走的^[6],③局部、全局和类局部的^[5]。对于具体预测指标和方法可参考吕琳媛等的研究^[5-6],在此不做详细展开。三种研究思路各有所长:基于结构相似性的链路预测只涉及网络的结构信息,相似性指标计算起来比较简单,但不同指标在不同网络中的预测能力却不一致,其预测的精确度取决于对网络结构特征刻画的好坏;基于最大似然估计的链路预测由于针对的是整个网络结构,计算复杂性较高,不太适合在规模较大的网络中应用;而基于概率模型的链路预测,不仅使用了网络结构信息,还会涉及节点属性信息,计算复杂度最高,且需要数据库的支持,但其优势在于具有较高的预测精确度。

图书情报学领域充满了各式各样的知识网络(见表1)。在知识网络中,知识节点是指知识单元的存储载体,也可以称为实体。文献计量学将论文作为基本实体使用已经很久了,可以用来测量期刊、作者、机构、国家/地区的学术影响,这些可以称为评价实体;而在科技论文中,知识实体是知识单元的微观载体,包括了关键词、主题、学科分类、数据集、关键方法、关键理论和领域实体(例如生物医学实体,包括基因、药物和疾病)^[18]。知识实体可以直接用来挖掘和分析知识的使用和迁移,最终促进知识发现。这些知识网络规模不一,有的规模还极其庞大,因此,在链路预测问题上主要采用的是第一种思路以降低计算的复杂性,同时也会引入一定的语义和属性信息来保证预测的精度。

表 1 知识网络类型和典型代表

网络类型		典型代表
同质网络	无权有向网络	引证网络
	无权无向网络	各种网络的二值特殊形态
	加权有向网络	作者互引网络、期刊互引网络等
	加权无向网络	合作网络、共词网络、耦合网络、共被引网络等
	二分网络	科学家—论文二分网络
异质网络		作者多重交互网络

2 同质网络中的链路预测

如前所述,知识网络的类型多种多样。从知识节点上看,作者和论文是最常见的两种节点类型;从知识关联上看,共现关系和引证关系是最重要的两种关联关系,其他诸如耦合、共被引、互引网络等体现的依然是共现关系和引证关系,只是构建方式稍显特殊。目前,探讨较多的有合作网络、引证网络 and 二分网络,其链路预测主要是在同质网络中进行的,这里就其研究成果进行梳理。

2.1 合作网络

针对知识创造主体间的合作关系而构建的合作网络,可被视为无向网络,这是描述真实网络系统最简单的一种方式。合作关系隐含着知识在某种社会关系之间的交流、转移、共享,把知识研究的对象扩展到了拥有知识的人、项目和组织,基本可以列入共现关系网络的类别。科学合作可以分为作者、机构、国家/地区三个层面,由于作者从属于他所服务的机构,机构在地理上又位于其所处的国家/地区,因此,机构之间和国家/地区之间的合作网络,实质上都是作者之间的合作网络。在具体研究中,合作网络可以通过合著网络来直观呈现。

2.1.1 预测指标与评价步骤

学者们已经针对合作网络的结构问题取得了大量成果,接下来的问题就是:在时间 t 下,一

个给定合作网络在未来时间 t' 时会演化成什么样子?为了捕捉这种动态变化特征,学者们从宏观视角上对网络增长和合作行为^[19-21]进行统计和建模分析,之后又聚焦到个人的合作行为^[22-24]研究上。利用链路预测可以进一步推断和分析合作网络的演化过程。Liben-Nowell 和 Kleinberg^[9]在 arXiv 上获取了物理学中五个领域的论文预印本,并从中抽取了作者合作信息。研究中,假定合作网络为 $G=\langle V,E \rangle$,连边 $e=\langle u,v \rangle \in E$ 表示节点作者 u 和 v 在某一特定的时间 t (e) 时完成了一次合作。令 1994—1996 年间的子网络为训练集,1997—1999 年间的子网络为测试集,将前者表示为 $G_{collab}=\langle A,E_{old} \rangle$, E_{new} 则表示仅在测试集中存在,在训练集中不存在的连边 $\langle u,v \rangle$ 集合,即新产生的合作关系,其中 $u,v \in A$ 。进一步,还将合作次数在 3 次及其以上的作者确定为核心作者。研究者以随机预测、图—距离预测、共同邻居预测为基准,对多种预测指标进行评估,发现在这些预测指标中,不存在单一明确的最优指标,但是大多数指标都显著优于随机预测,这表明网络拓扑结构中确实包含了非常有用的信息。

后续研究基本遵循上述研究思路^[25],主要包括两个步骤:①将一些预测指标应用在某一训练网络上,预测可能产生的新的连接;②通过和一个实际的后续快照,即测试网络,进行比较,来评价预测指标。需要注意的是,链路预测只能预测训练网络中节点对间产生连边的概率,而不能预测连接到新增节点的概率。

对链路预测进行评价可以被视为类信息检索评价^[26],如表2所示。这就意味着,对每一种预测指标都可以计算准确率和召回率。在信息检索评价中,准确率和召回率会相互制约,

因此,通常会根据需要为“检索策略”选择一个合适的度,不能太紧也不能太松,寻求在准确率和召回率之间的一个平衡点。在链路预测中,同样如此。

表2 预测网络和测试网络中连边之间可能存在的四种关系

		测试网络	
		相关的	不相关的
预测网络	预测到的	$A = E_{pred} \cap E_{test}$	$B = E_{pred} \cap \neg E_{test}$
	未预测到的	$C = \neg E_{pred} \cap E_{test}$	$D = \neg E_{pred} \cap \neg E_{test}$

注:准确率 = $A/(A+B)$,召回率 = $A/(A+C)$

在表2中,情形A是正常状态的链路预测,情形B是异常间隙发现(Anomalous Gap Discovery, AGD),情形C是异常链路发现(Anomalous Link Discovery, ALD)^[27]。对于情形B而言,异常间隙是指预测网络中那些出现连接的可能性高,但没有在测试网络中形成实际的连边的节点对。这意味着,该链路可能会在更远的未来的某个时间点出现,或者节点间形成了某种边界。针对合作网络而言,如果作者a和作者b之前合作过多次,那么在测试网络中再次合作的可能性也很高;但如果合作没有发生,则有可能是其中某个作者的研究课题发生了改变。对于情形C而言,异常链路是指在测试网络中真实存在,但通过预测方法认为其存在可能性很低的连边。从另一个角度来看,这些异常链路可能是“不稳定的”或者在未来很可能消失。当然,我们有理由怀疑这些连边可能是虚假连接,但如果它们的确存在,那么势必在网络中具有特定功能和作用。实际上,异常间隙比异常链路具有更多的不确定性,而异常链路会对保持网络连通性有着特别重要的贡献,因此,链路预测的一个重要应用是对识别出的异常链路进行分析。

2.1.2 不同层面的预测效果

Liben-Nowell 和 Kleinberg^[9]的研究针对单一学科领域的作者合作网络。其中,在基于路径的方法中,Katz 指标^[28]及其变体和低秩逼近的预测效果很好;而在基于节点的方法中,仅考

虑节点共同邻居的方法和 Adamic/Adar 指标^[29]的预测效果也非常好。而 Guns^[26]认为,链路预测中的训练网络并不一定是测试网络的一个旧的快照,也可以将两个或者多个网络一同训练。对此,Liben-Nowell 和 Kleinberg^[9]提出的一个思路是,可以将论文标题、作者所在机构和地理位置信息加入到计算中,对预测结果进行微调。Guns 在具体实施时,以安特卫普大学为例,训练集由三个子网络组成:第一个网络是合作网络,由该校学者在2001—2003年间发表论文时形成的作者合作关系构成;第二个网络是部门网络,当作者归属于同一个部门时,由作者节点间形成连接关系构成;第三个网络是物理位置网络,当作者工作地点相互接近时,由作者节点间形成连接关系构成。需要说明的是,部门网络和物理位置网络的密度比测试网络(2004—2006)高得多。因此,作为单一训练网络来进行预测,其准确率非常低,而和合作网络(2001—2003)一起,则有助于提升准确率和召回率。

Guns 和 Rousseau^[30]对非洲、中东和南亚的城市间科研合作进行研究,构建了在三个连续时间段(1997—2001、2002—2006、2007—2011)内疟疾和肺结核研究领域中的合作网络,通过链路预测和机器学习发现潜在的富有成效的合作。在具体操作中,先构建加权合作网络,之后根据不同预测指标计算出节点对之间的相似性,这些数值表示给定节点对在未来产生连接

的可能性。为了确定每个预测指标的相对价值,对旧的数据进行随机森林分类训练,而同样的分类器用在新的数据上可以产生预测结果,最终将排名前 n 的预测结果作为科研合作建议进行推荐。

晏尔伽和 Guns^[31]使用图书情报学于2001—2010年间发表在59种期刊上的论文数据集研究作者、机构、国家/地区层面的合作网络。数据集被分为两个时间段(2001—2005、2006—2010)。所构成的网络中,作者合作网络和机构合作网络都很稀疏,而国家/地区合作网络则很稠密。研究中使用了四种评价指标:准确率和召回率、ROC曲线、AUC、nDCG,对三个层面的合作网络在八种预测指标下的预测结果进行了评价。在基于邻居信息的预测指标中,择优连接(Preferential Attachment, PA)相似性预测结果表现较好;而在基于拓扑结构的预测指标中, Katz 指标预测结果表现较好。研究结果显示针对该特定的数据集,国家/地区合作网络的预测结果要优于作者合作网络和机构合作网络,同时还给出了所预测的三个层面下排名前十的合作组合。

2.2 引证网络

表现知识前后承接关系的引证网络,是最早被提出的一种知识网络形态,它的主要研究对象是学术论文。论文间的引证关系一经发表一般不会出现变动,绝大部分都会服从时间先后关系,且大部分论文都由相关主题的其他文章引用,很少出现相互引用,网络中没有环状结构。理论上讲,这种引证关系一定要满足时间先后关系,且不出现在相互引用。但在实际情况中,由于存在优先出版等导致出版周期错乱的因素,使得引证网络中存在极少的不符合理论情形的连边。这些单向关系无法使用无向网络中的相互关系或共现关系进行刻画,因此需要借助有向网络的形式。

2.2.1 网络结构的形成机制

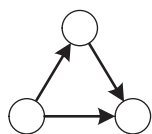
引证网络是一种无标度网络。Barabási 和

Albert^[32]于1999年提出了通过引入增长和择优连接机制构建增长网络模型,简称为BA模型,来解释真实网络无标度特性出现的内在机理。事实上,这一思想可以追溯到 Price^[1,33]和 Simon^[34]的研究,只不过当时使用的是累积优势和马太效应来描述这一过程。当然,还有其他因素影响引证网络的演化过程。比如:时间效应会在一定程度上平抑度择优所导致的马太效应的负面影响^[35];知识学习和知识引证通常还是基于局部知识领域的^[36-37]。因此,真实的引证网络往往是由多种机制混合作用而成。

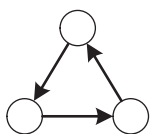
同时,学术论文中的引证与被引证还体现了同质性(homophily),使得大量文献得以分群。在一些文献中,聚类效应被当作一种特殊的同质性,即拥有共同邻居的节点被看成拥有相似的网络环境。同质性体现了节点间的相似性及相容性,能对它们之间是否形成连边产生重要影响^[38]。显然,论文间的引证关系倾向于产生在具有相似内容的文献之间^[39]。在文献计量学中,通常会使用文献耦合和共被引两种方法来揭示论文的主题相似性,以及相互之间的作用和联系。文献耦合和共被引都建立在相似的假设基础之上,即具有耦合关系和共被引关系的论文可以认为它们在学科内容上存在某种相关性。

引证网络的形成机制非常复杂,而同质性的存在使得可以利用链路预测来探究网络结构的形成机制。无向网络中3个节点的相互关系非常简单,在已经拥有共同邻居的两个节点之间产生连边,只能形成一种新结构,称为三角形结构^[40]。而在有向网络中,同样考虑3个节点,情况就复杂得多^[41]。在不考虑互惠连接机制和具体节点间连边的情况下,引证网络的三角形结构可分为前馈回路和反馈回路两种。Milo等^[42]引入模体(motif)来描述网络中反复出现的相互作用基本模式。相比3节点结构而言,4节点结构更加复杂。Milo等发现在很多网络中,4节点的双风扇结构和双平行结构是非常显著的。实际上,可以将这些识别出来的模体看

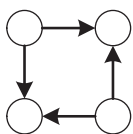
成网络生成和演化过程中在特定限制条件下形成的特殊局部结构。于是,在引证网络中3个节



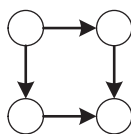
3节点前馈回路
(3-FFL)



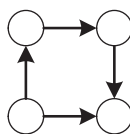
3节点反馈回路
(3-Loop)



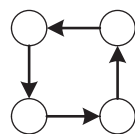
双风扇结构
(Bi-fan)



双平行结构
(Bi-parallel)



4节点前馈回路
(4-FFL)



4节点反馈回路
(4-Loop)

图3 六种含有回路的最小子图

张千明等^[43]提出的势理论(Potential Theory)为强化上述假设和描述引证网络演化提供了一种可能解释。按照势理论,只有双风扇结构和双平行结构是可定义势的。又根据同质性的理论,双风扇结构比双平行结构具有更强的同质性。不妨假设:如果一条边的添加能产生更多的双风扇结构,那么这条边存在的可能性就越大。这一假设得到了链路预测实验结果的强有力支持,即通过双风扇结构进行预测会得到准确的预测效果。需要说明的是,这一假设是一种局部驱动机制,它暗示着一个局部的层级结构,而节点会倾向于在局部以层次结构的方式进行自组织。

上述实验结果还表明,双风扇结构的一个显著优势在于它具有高度的预测稳定性,对网络结构不敏感。对双风扇结构进行分解,结果如图4所示。显然,不管以哪条虚线方式进行划分,都一定可以找到节点间的耦合关系或共被引关系。结合这两点,可以得到一条简单的推论:双风扇结构保证了对耦合关系和共被引关系分析的有效性。当然,这一推论还需要进一步验证。同时,通过双风扇结构对引证网络进行预测的效果还需要大量实证研究加以检验。

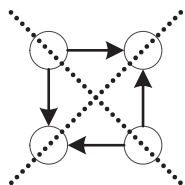


图4 双风扇结构分解

点和4个节点的情况下可得六种子图结构,如图3所示。

2.2.2 引入外部信息的预测

相比合作网络,引证网络除了基础数据的结构信息,还会涉及作者、期刊以及文章内容等外部信息。因此,引证网络不仅可以基于局部结构进行预测,还可以巧妙地利用外部信息,以及将外部信息和结构信息融合起来进行机器学习,从而提高链路预测的精确性。Popescul和Ungar^[44-45]提出了一个统计关系学习模型并应用于引证网络,对科学文献的引证关系进行预测。O'Madadhain等^[46]利用网络的拓扑结构信息以及节点的属性建立了一个局部的条件概率来进行预测。

Shibata等^[47]从引证网络的结构、语义、属性3个方面的11个特征入手,构建了一个监督机器学习模型,分别对5个大规模引证网络数据集(创新、纳米生物、有机LED、太阳能电池、二次电池)进行链路预测分析。结果表明:结构方面的“基于链路的Jaccard系数”、“中介中心性差异”,语义方面的“TF-IDF向量余弦相似性”这3个特征对引证网络链路预测有较强的影响。值得说明的是,不同类型的研究领域需要不同的模型来对应,很难建立一个通用的学习模型。研究中涉及的数据集可以大致划分为两种类型,即多问题研究领域(创新、纳米生物)和单问题研究领域(有机LED、太阳能电池、二次电池)。由于引证论文一般会有局域限制,因此多问题研究领域的链路预测会存在一定障碍。此外,Shibata等还指出这一研究存在两个局限性:一是由于引证了还未正式发表的论文,会使引证网络的拓扑特性发生改变;二是虽然链路预

测可以帮助显示应该被引却被遗漏的文献,但真实的引证关系是由人的个体行为来决定的,如果将专家判断加入到学习模型中,预测结果的准确率有望得以提高。

2.3 二分网络

很多同质知识网络还可以以二分网络的形式呈现,比如:合作网络可以是以参与者和事件为两个分离集的二分网络^[48],人类疾病网络可以是以身心机能失调表现和致病基因为两个分离集的二分网络^[49]。二分网络形式众多,像科学家—论文二分网络,可以用科学家—术语、科学家—被引论文等进行替代。科学家代表了知识创造过程中的主体,论文、术语、被引论文等代表了知识创造过程中的客体。利用二分网络研究主客体间的结构特征和演化过程,有助于探讨知识创造活动的规律。在二分网络中应用链路预测,可以解决一些推荐系统的问题。

许多学者曾经考虑一个分离集为科学家的二分网络。Goldstein等^[50]较早地研究了在同一个群组中的成员更倾向于合作这一效应,并以科学家—论文二分网络为例,构建了演化模型。之后,Morris和Goldstein^[51]在此基础上又进行了改进。针对二分网络的链路预测问题,Chua和Lim^[52]对Clauset等^[15]提出的层次结构模型进行了改造,应用于二分网络,并以ICDM会议的科学家—论文二分网络,以及SIGKDD、SDM、ICDM、WSDM四个会议所得到的科学家—论文二分网络为例,发现采用二分网络层次结构模型,不仅预测精度高于简单无向网络层次结构模型,而且能更好地再现真实网络的结构特征。Guns^[53]也针对科学家—论文二分网络、科学家—术语二分网络分别进行了研究。当然,对于使用二分网络能否真正提高预测效果还存在争议^[54],因此需要大量实证研究来检验。

3 异质网络中的链路预测

当前大部分研究都聚焦于同质网络,即网

络中只存在一种类型的连边。然而,现实中大部分的知识网络都是异质网络,即网络中有很多不同类型的连边。Börner等^[55]提出过一个科学家—论文网络同步增长模型,其连边有三种类型,分别是论文间的引证关系、作者间的合作关系、作者和论文间的使用和生产关系。Börner等主要考虑的是这一异质网络的演化过程,并以1982—2001这20年间发表于PNAS上的论文数据进行验证,但并没有从链路预测的角度来考虑演化模型评价问题。

在异质网络中,节点和连边的异质特性使得链路预测变得更加困难。不同类型节点之间的连接具有不同的含义,可能具有不同的形成和演化机制,同时还会相互影响,也会导致各自在链路预测中的作用不同。因此,异质网络中的链路预测是一个新颖且富有挑战的问题。Sun和Han^[56-57]针对异质网络提出了基于元路径(meta-path)的链路预测模型。元路径是用来描述异质网络中两个节点对象之间的连接路径,用来研究两个节点对象在元层次上存在的关系。不同的元路径代表了两个节点对象之间可以通过不同类型的节点和连边建立不同的关系。显然,不同的元路径具有不同的语义来描述节点之间的接近程度。基于元路径的拓扑特征,设计监督学习模型学习不同元路径上节点对象之间的接近程度和每一条元路径的权重,来进行链路预测,这可以应用在作者合作^[58]和引证关系^[59]的预测上。目前,国内也有相关文献^[60-61]对此予以关注。

对于异质网络中的多重关系网络,比如:作者多重交互网络,通常是将节点对之间的多重关系经过加权处理,转换成单一关系^[62],之后再进行后续研究。有关多重关系网络的链路预测方法可以参考Yang等^[63]的研究。此外,不同知识实体所构成的网络也是一种异质知识网络。在现代药物发现过程中,需要对化学基因组、化合物和药物以及身体里的多种蛋白靶标和基因之间的复杂相互作用进行理解。Chen等^[64]从众多数据库中获取公开数据,并对涉及药物靶标相

互作用的数据进行整合和语义标注,知识实体关系包括了化合物—基因、药物—药物、蛋白质—蛋白质、药物副作用等,之后构建了一个异质网络。他们构建了一个统计模型,命名为语义链路关联预测 (Semantic Link Association Prediction, SLAP),其中,关联分值的计算基于邻居节点的拓扑和语义信息。研究者利用 SLAP 预测该异质网络中的未知链路,以此来识别药物靶标对和确定间接的药物靶标关联。鉴于这属于医学信息学研究范畴,这里不再赘述。

4 研究评论

通过上述回顾和梳理可知,图书情报学领域对于知识网络中链路预测问题的专门研究在近年有成为热点的趋势,成果多刊发在 *Journal of Informetrics*、*Scientometrics*、*JASIST* 等权威期刊上。已有研究多是描述各种链路预测指标在不同类型知识网络中预测效果的实

证性研究,以确定链路预测的应用范围。以合作关系研究为例,科研合作是一个很复杂的多层次社会现象,研究作者合作网络有助于理解个人的合作行为与动机,研究机构合作网络有助于审视科研机构的地位和规划学科未来,研究国家合作网络有助于确定影响科研合作的宏观性因素,比如语言、科技政策、文化等。随着时间的推移,合作关系也发生着变化。而链路预测有助于揭示这些网络的形成机制和演化过程,为合作预测提供新的思路。链路预测的结果可以为交叉学科合作趋势、研发群体科学 (the Science of Team Science, SciTS) 等提供推荐建议和决策依据,从而解决寻找合作者这一难题^[65]。由此可见,链路预测可以在一定程度上将知识网络的结构和演化研究提升到应用层次,并最终运用到预测和决策中去,有着重大的实践价值。本文总结出研究框架 (见图 5),并从研究方法和研究内容两个方面指出研究中尚需要进一步深化之处。

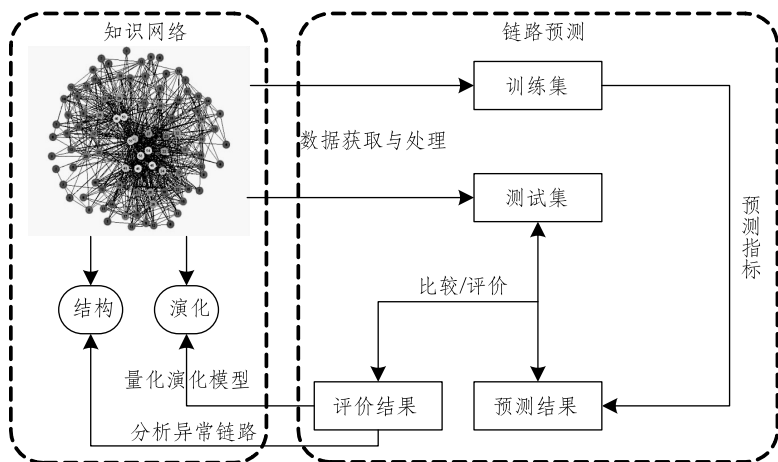


图5 研究框架

在研究方法上,Liben-Nowell 和 Kleinberg^[9] 在此方面做了开创性的论述,后续研究也基本遵循他们的研究思路展开。基于拓扑结构相似性进行链路预测的方法简单可靠,对于结构相似的网络具有一定的普适性。基于最大似然估计和概率模型的链路预测方法也取得了丰硕的

成果,虽然它们的研究框架要远远复杂于基于结构相似性的研究框架,但它们除了能提供链路预测的结果之外,还能对网络结构进行深刻的揭示,比如层次组织形态、网络分块结构等。此外,还应重视对机器学习方法的吸收、改进与整合。

在研究内容上, 知识网络是研究知识关联关系、探索发展脉络、追踪创新趋势的有效途径。知识网络的结构和演化为观察到的知识交流现象提供了解释; 而链路预测可用来检验科学理论能否对尚未观察到的现象进行预测和评价。

(1) 对于知识网络中的链路预测研究, 目前仍以预测指标的选择以及预测精度的提高为主。知识网络有很多不同类型, 不同的预测指标也有其优缺点和适用范围。因此, 首先需要找到适合相应知识网络类型的预测指标, 在此基础上, 找到一种合适的途径整合不同的预测结果。对于预测精度高低的考察, 取决于结构相似性的定义是否能够很好地抓住目标网络的结构特征。以学科合作网络为例, 不同学科合作网络的拓扑结构差别很大^[66], 而网络结构对于预测指标的选择有着重要影响^[10, 67]。目前还没有对算法性能和网络结构特征之间关系较深入的研究。就目前研究结论来看, 并不存在单一明确的最优指标。因此, 在对预测指标的评价和推广上还需要考虑具体情况。

(2) 对于知识网络中的链路预测研究, 较少涉及利用链路预测来量化和评价演化模型问题。目前, 学者们认为, 真实网络内在的演化机制决定了呈现出的纷繁有趣的网络特征^[68]。传统的针对演化机制的研究, 更多关注知识发展的描述模型和过程模型, 通过统计和仿真的手段对演化过程和演化机制进行解释。网络演化主要是节点的出现和消失、节点之间连边的改变这两个方面, 而演化模型的关键是给出两个节点之间产生连边的概率, 因此原则上一个演化模型可以对应一种链路预测方法。链路预测的本质是挖掘产生连边的原因, 这同时也是网络演化模型所关心的问题。但借助链路预测的理论框架和评价方法间接对演化模型的表现进

行定量比较的研究较少。

(3) 对于知识网络中的链路预测研究应更加侧重于识别和分析异常链路, 以发现知识热点和创新趋势。当前链路分析主要是通过基于时序词频统计、聚类分析等统计方法并结合可视化结果来跟踪主流知识与潜在知识。通过对合作网络、引证网络和共词网络等多种结构表征的比较, 发现那些被忽略的连接和潜在的联系。然而这通常是一种事后方法, 有明显的时滞, 一篇文章发表之后, 作者、参考文献、关键词这些属性信息就确定了, 断边重连机制无法应用其中。但如果通过某种方法得到一条连边存在的可信度、重要性或可能性, 那么就可以用它来预测知识的创新和发展趋势。链路预测的一个重要应用就是链路分析。针对网络中已有连接的节点对, 可以用相似性指标来评价连边存在的可信度或者重要性; 针对网络中还未产生连接的节点对, 节点对间的相似性可以用来预测产生连边的可能性。即使在实际知识网络中没有这些连边或者是连接关系较弱, 但它们对于科学家而言也是很重要的, 至少可以提醒科学家是否忽略了一些具有潜在意义的知识链路。这为开展科学研究和指导知识生产提供了重要依据, 同时也是情报和知识管理领域长久以来的一个核心问题。

当前关于知识网络中的链路预测研究零散地分布在统计物理、计算机、复杂网络、图书情报等领域的研究成果中。本文以知识网络的不同类型和典型代表为分类基础对已有研究进行初步梳理。从十多年的发展情况来看, 知识网络可以作为链路预测研究的理想载体, 而链路预测也是分析知识网络的有力工具。以此为基础, 围绕知识网络的结构和演化模型, 从链路预测角度开展应用性研究, 是今后图书情报学领域在知识网络研究中应该着力的方向。

参考文献

- [1] De Solla Price D J. Networks of scientific papers[J]. Science, 1965, 149(3683): 510-515.

- [2] Brookes B C. The foundations of information science part iv. Information science: the changing paradigm[J]. Journal of Information Science, 1981, 3(1): 3-12.
- [3] 刘向, 马费成, 陈潇俊, 等. 知识网络的结构与演化——概念与理论进展[J]. 情报科学, 2011, 29(6): 801-809. (Liu Xiang, Ma Feicheng, Chen Xiaojun, et al. Structure and evolution of knowledge network——concept and research review[J]. Information Science, 2011, 29(6): 801-809.)
- [4] 张金柱, 韩涛, 王小梅. 复杂网络在图情领域的研究和应用综述[J]. 情报学报, 2012, 31(9): 907-914. (Zhang Jinzhu, Han Tao, Wang Xiaomei. Overview of complex network research and its application in library and information science[J]. Journal of the China Society for Scientific and Technical Information, 2012, 31(9): 907-914.)
- [5] Lü L, Zhou T. Link prediction in complex networks: a survey[J]. Physica A: Statistical Mechanics and Its Applications, 2011, 390(6): 1150-1170.
- [6] 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. (Lü Linyuan. Link prediction on complex networks[J]. Journal of University of Electronic Science and Technology of China, 2010, 39(5): 651-661.)
- [7] Getoor L, Diehl C P. Link mining: a survey[J]. ACM SIGKDD Explorations Newsletter, 2005, 7(2): 3-12.
- [8] Van Eck N J, Waltman L. CitNetExplorer: a new software tool for analyzing and visualizing citation networks[J]. Journal of Informetrics, 2014, 8(4): 802-823.
- [9] Liben-Nowell D, Kleinberg J. The link-prediction problem for social networks[J]. Journal of the American Society for Information Science and Technology, 2007, 58(7): 1019-1031.
- [10] Zhou T, Lü L, Zhang Y C. Predicting missing links via local information[J]. The European Physical Journal B-Condensed Matter and Complex Systems, 2009, 71(4): 623-630.
- [11] Sarukkai R R. Link prediction and path analysis using markov chains[J]. Computer Networks, 2000, 33(1): 377-386.
- [12] Zhu J, Hong J, Hughes J G. Using markov chains for link prediction in adaptive web sites[M]//Soft-Ware 2002: Computing in an Imperfect World. Springer Berlin Heidelberg, 2002: 60-73.
- [13] Kossinets G. Effects of missing data in social networks[J]. Social Networks, 2006, 28(3): 247-268.
- [14] Newman M E J. Clustering and preferential attachment in growing networks[J]. Physical Review E, 2001, 64(2): 025102.
- [15] Clauset A, Moore C, Newman M E J. Hierarchical structure and the prediction of missing links in network[J]. Nature, 2008, 453(7191): 98-101.
- [16] Guimerà R, Sales-Pardo M. Missing and spurious interactions and the reconstruction of complex networks[J]. Proceedings of the National Academy of Sciences, 2009, 106(52): 22073-22078.
- [17] Holland P W, Laskey K B, Leinhardt S. Stochastic blockmodels: first steps[J]. Social networks, 1983, 5(2): 109-137.
- [18] Ding Y, Song M, Han J, et al. Entitymetrics: measuring the impact of entities[J]. PloS One, 2013, 8(8): e71416.
- [19] Barabási A L, Jeong H, Neda Z, et al. Evolution of the social network of scientific collaborations[J]. Physica A: Statistical Mechanics and Its Applications, 2002, 311(3): 590-614.
- [20] Jeong H, Neda Z, Barabási A L. Measuring preferential attachment in evolving networks[J]. Europhysics Letters, 2003, 61(4): 567-572.
- [21] Newman M E J. Coauthorship networks and patterns of scientific collaboration[J]. Proceedings of the National Academy of Sciences, 2004, 101(suppl 1): 5200-5205.

- [22] Kretschmer H. Author productivity and geodesic distance in bibliographic co-authorship networks, and visibility on the web[J]. *Scientometrics*, 2004, 60(3): 409-420.
- [23] Liu X, Bollen J, Nelson M L, et al. Co-authorship networks in the digital library research community[J]. *Information Processing & Management*, 2005, 41(6): 1462-1480.
- [24] Yan E, Ding Y. Applying centrality measures to impact analysis: a coauthorship network analysis[J]. *Journal of the American Society for Information Science and Technology*, 2009, 60(10): 2107-2118.
- [25] Guns R. Link prediction[M]//Ding Y, Rousseau R, Wolfram D. *Measuring scholarly impact*. Springer, 2014: 35-55.
- [26] Guns R. Generalizing link prediction: collaboration at the University of Antwerp as a case study[J]. *Proceedings of the American Society for Information Science and Technology*, 2009, 46(1): 1-15.
- [27] Rattigan M J, Jensen D. The case for anomalous link discovery[J]. *ACM SIGKDD Explorations Newsletter*, 2005, 7(2): 41-47.
- [28] Katz L. A new status index derived from sociometric analysis[J]. *Psychometrika*, 1953, 18(1): 39-43.
- [29] Adamic L A, Adar E. Friends and neighbors on the web[J]. *Social Networks*, 2003, 25(3): 211-230.
- [30] Guns R, Rousseau R. Recommending research collaborations using link prediction and random forest classifiers[J]. *Scientometrics*, 2014, 101(2): 1461-1473.
- [31] Yan E, Guns R. Predicting and recommending collaborations: an author-, institution-, and country-level analysis[J]. *Journal of Informetrics*, 2014, 8(2): 295-309.
- [32] Barabási A L, Albert R. Emergence of scaling in random networks[J]. *Science*, 1999, 286(5439): 509-512.
- [33] Price D S. A general theory of bibliometric and other cumulative advantage processes[J]. *Journal of the American Society for Information Science*, 1976, 27(5): 292-306.
- [34] Simon H A. On a class of skew distribution functions[J]. *Biometrika*, 1955, 42(3/4): 425-440.
- [35] 刘向, 马费成. 科学知识网络的演化与动力——基于科学引证网络的分析[J]. *管理科学学报*, 2012, 15(1): 87-94. (Liu Xiang, Ma Feicheng. Evolution and dynamics of scientific knowledge network: based on the study of scientific citation network[J]. *Journal of Management Sciences in China*, 2012, 15(1): 87-94.)
- [36] 马费成, 刘向. 科学知识网络的演化模型[J]. *系统工程理论与实践*, 2013, 33(2): 437-443. (Ma Feicheng, Liu Xiang. Evolvement model for scientific knowledge networks[J]. *Systems Engineering - Theory & Practice*, 2013, 33(2): 437-443.)
- [37] 刘向, 马费成, 王晓光. 知识网络的结构及过程模型[J]. *系统工程理论与实践*, 2013, 33(7): 1836-1844. (Liu Xiang, Ma Feicheng, Wang Xiaoguang. Formation and process model of knowledge networks[J]. *Systems Engineering - Theory & Practice*, 2013, 33(7): 1836-1844.)
- [38] McPherson M, Smith-Lovin L, Cook J M. Birds of a feather: homophily in social networks[J]. *Annual Review of Sociology*, 2001, 27: 415-444.
- [39] Cheng X Q, Ren F X, Zhou S, et al. Triangular clustering in document networks[J]. *New Journal of Physics*, 2009, 11(3). doi:10.1088/1367-2630/11/3/033019.
- [40] Rapoport A. Spread of information through a population with socio-structural bias: i. Assumption of transitivity[J]. *The Bulletin of Mathematical Biophysics*, 1953, 15(4): 523-533.
- [41] Brzozowski M J, Romero D M. Who should I follow? Recommending people in directed social networks[C]//*Proceedings of the 5th International Conference on Weblogs and Social Media (ICWSM)*. Menlo Park: the AAAI Press, 2011: 458-461.
- [42] Milo R, Shen-Orr S, Itzkovitz S, et al. Network motifs: simple building blocks of complex networks[J]. *Science*, 2002, 298(5594): 824-827.

- [43] Zhang Q M, Lü L, Wang W Q, et al. Potential theory for directed networks[J]. PloS One, 2013, 8(2): e55437.
- [44] Popescul A, Ungar L H, Lawrence S, et al. Statistical relational learning for document mining[C]//Proceedings of the Third IEEE International Conference on Data Mining (ICDM 2003). University of Pennsylvania, 2003: 275-282.
- [45] Popescul A, Ungar L H. Statistical relational learning for link prediction[C]//Proceedings of the Workshop on Learning Statistical Models from Relational Data at IJCAI, 2003.
- [46] O'Madadhain J, Hutchins J, Smyth P. Prediction and ranking algorithms for event-based network data[J]. ACM SIGKDD Explorations Newsletter, 2005, 7(2): 23-30.
- [47] Shibata N, Kajikawa Y, Sakata I. Link prediction in citation networks[J]. Journal of the American society for Information Science and Technology, 2012, 63(1): 78-85.
- [48] Zhang P P, Chen K, He Y, et al. Model and empirical study on some collaboration networks[J]. Physica A: Statistical Mechanics and Its Applications, 2006, 360(2): 599-616.
- [49] Goh K I, Cusick M E, Valle D, et al. The human disease network[J]. Proceedings of the National Academy of Sciences, 2007, 104(21): 8685-8690.
- [50] Goldstein M L, Morris S A, Yen G G. Group-based Yule model for bipartite author-paper networks[J]. Physical Review E, 2005, 71(2): 026108.
- [51] Morris S A, Goldstein M L. Manifestation of research teams in journal literature: a growth model of papers, authors, collaboration, coauthorship, weak ties, and Lotka's law[J]. Journal of the American Society for Information Science and Technology, 2007, 58(12): 1764-1782.
- [52] Chua F C T, Lim E P. Modeling bipartite graphs using hierarchical structures[C]//Proceedings of the 2011 International Conference on Advances in Social Networks Analysis and Mining (ASONAM). New York: IEEE Press, 2011: 94-101.
- [53] Guns R. Missing links: predicting interactions based on a multi-relational network structure with applications in informetrics[D]. Belgium: University Antwerpen, 2012.
- [54] Guns R. Bipartite networks for link prediction: can they improve prediction performance[C]//Proceedings of the ISSI 2011 Conference, 2011: 249-260.
- [55] Börner K, Maru J T, Goldstone R L. The simultaneous evolution of author and paper networks[J]. Proceedings of the National Academy of Sciences of the United States of America, 2004, 101(Suppl 1): 5266-5273.
- [56] Sun Y, Han J. Mining heterogeneous information networks: a structural analysis approach[J]. ACM SIGKDD Explorations Newsletter, 2013, 14(2): 20-28.
- [57] Sun Y, Han J. Meta-path-based search and mining in heterogeneous information networks[J]. Tsinghua Science and Technology, 2013, 18(4): 329-338.
- [58] Sun Y, Barber R, Gupta M, et al. Co-author relationship prediction in heterogeneous bibliographic networks[C]//International Conference on Advances in Social Networks Analysis and Mining (ASONAM). IEEE, 2011: 121-128.
- [59] Sun Y, Han J, Aggarwal C C, et al. When will it happen? Relationship prediction in heterogeneous information networks[C]//Proceedings of the Fifth ACM International Conference on Web Search and Data Mining. ACM, 2012: 663-672.
- [60] 黄立威, 李德毅, 马于涛, 等. 一种基于元路径的异质信息网络链路预测模型[J]. 计算机学报, 2014, 37(4): 848-858. (Huang Liwei, Li Deyi, Ma Yutao, et al. A meta path-based link prediction model for heterogeneous information networks[J]. Chinese Journal of Computers, 2014, 37(4): 848-858.)

公 告

近期,有网站(www.zgtsgxb.cn)假借《中国图书馆学报》名义接收投稿,并向作者索取审稿费,严重侵害了本刊的声誉及作者的利益,影响恶劣。本刊已向有关管理部门投诉,并视情况进一步采取相关的法律措施。

本刊再次郑重声明:

《中国图书馆学报》唯一官方网站的网址是: <http://www.jlis.cn>, 本刊不以任何形式向作者收取审稿费、版面费等任何费用。

(《中国图书馆学报》编辑部)

- [61] 康文杰, 郑倩冰, 陈侃. 基于社会网络分析的学术合作关系研究[J]. 计算机技术与发展, 2014, 24(5): 1-5. (Kang Wenjie, Zheng Qianbing, Chen Kan. Research on relationship of academic cooperation based on social network analysis[J]. Computer Technology and Development, 2014, 24(5): 1-5.)
- [62] Liu P, Chen F L, Ma Y L, et al. Research community detection from multi-relation researcher network based on structure/attribute similarities[J]. Chinese Journal of Library and Information Science, 2013, 6(1): 14-32.
- [63] Yang Y, Chawla N, Sun Y, et al. Predicting links in multi-relational and heterogeneous networks[C]// Proceedings of the 2012 IEEE 12th International Conference on Data Mining. IEEE Computer Society, 2012: 755-764.
- [64] Chen B, Ding Y, Wild D J. Assessing drug target association using semantic linked data[J]. PLoS Computational Biology, 2012, 8(7): e1002574.
- [65] Pavlov M, Ichise R. Finding experts by link prediction in co-authorship networks[C]//Proceedings of FEWS, 2007: 42-55.
- [66] Ma F, Li Y, Chen B. Study of the collaboration in the field of the Chinese humanities and social sciences[J]. Scientometrics, 2014, 100(2): 439-458.
- [67] Feng X, Zhao J C, Xu K. Link prediction in complex networks: a clustering perspective[J]. The European Physical Journal B-Condensed Matter and Complex Systems, 2012, 85(1): 1-9.
- [68] Dorogovtsev S N, Mendes J F F. Evolution of networks[J]. Advances in Physics, 2002, 51(4): 1079-1187.

张 斌 武汉大学信息资源研究中心 2012 级管理科学与工程博士研究生。

通信地址: 武汉市武昌区八一路珞珈山。邮编: 430072。

马费成 武汉大学信息资源研究中心教授, 博士生导师。通信地址同上。

(收稿日期: 2014-10-18; 修回日期: 2015-01-07)