

基于小世界现象的学科信息 门户链接设计优化策略^{*}

肖 雪

(南开大学商学院信息资源管理系 天津 300071)

摘 要 从平均最短路径、集团系数、对数路径和中心节点描述了小世界网络的特性和模型,从知识组织和用户行为的角度指出学科信息网络具有更为明显的小世界现象。采用小世界度量指标对 CSDL4 个学科信息门户进行分析,发现网络链接存在的问题。据此提出学科信息门户链接设计的优化策略:采用知识链接技术发展多重链接、基于凝聚子群分析和语义网确定链接集合边界、基于数据挖掘和知识地图技术寻找捷径、运用信息计量和社会网络分析方法识别中心节点。

关键词 小世界现象 学科信息门户 网络链接 社会网络分析 知识链接

中图分类号 G254 **文献标识码** A **文章编号** 1002-1965(2011)10-0134-05

Hyperlink Design Optimization Strategies to Subject Based Information Gateways —based on Small-world Phenomenon

XIAO Xue

(Department of Information Resource Management, Business School of Nankai University, Tianjin 300071)

Abstract This paper describes the mode of small-world network from characteristic path length, clustering coefficient, short cut and central node. And it further points out the small-world phenomenon also exists in subject based information gateways from perspective of knowledge organization and user behavior. Using factors of social network analysis, the hyperlinks in four subject based information gateways of CSDL are analyzed and problems are found. Then the paper proposes optimized strategies for hyperlink design, including developing multiple links, determining link set boundaries based on cohesive subgroup and semantic web analysis, seeking shortcut with the help of data mining and knowledge map, identifying central node using informetrics and social network analysis method.

Key words small-world phenomenon subject based information gateways network hyperlink social network analysis knowledge linkage

0 引 言

1967年,美国社会心理学家 Milgram 通过著名的发信试验,发现任意两个人之间最多通过 6 个人就能取得联系,由此提出了“六度分离”理论。1970 年 White 运用模型,提出了一个修正估计值——约为 7 个中间人^[1]。尽管对于六度分离的确切数值存在分歧,但与总人口的数量级相比,无论哪一个数值都是非常小的,这就从科学的角度表明世界虽大,但也很小,“小世界现象”由此得名。2002 年哥伦比亚大学社会

学系 Watts 和 Strogatz 通过电子邮件在全球范围开展了一个“小世界研究计划”,再次重复 Milgram 的试验,结果表明邮件平均经过 5~7 步传递到目标接收者^[2],再次验证了人际网络中小世界现象的存在,也使“小世界现象”这一术语广为学术界所接受。此外,研究者们还发现在生物细胞网、脑神经网络、电力网、航线网络、互联网等多个领域中链条距离长度各有不同,但基本都表现为一个很小的常数,表明小世界现象对于刻画真实世界十分奏效。根据“大世界悖理”,世界尽管很大,总是可以缩成“小世界”,而“小世界”则能

收稿日期:2011-04-15 修回日期:2011-06-08
基金项目:南开大学校内青年项目“基于小世界效应的学科信息组织优化策略”的研究成果之一(编号:NKQ08023)。
作者简介:肖 雪(1979-),女,讲师,研究方向:信息服务与用户研究。

保障信息交流扩大进行^[3],因而小世界原理为实现从大世界到小世界的渡越、从无序繁衍走向有序控制提供了明晰的思路。

1 小世界现象的原理概述

任何网络都可以抽象为多个节点(代表网络中的个体)和各点之间的连线(代表个体之间的联系)构成的集合,存在小世界现象的网络也不例外。因此,探究小世界原理首先就要构建具有普适性的小世界网络模型,复杂网络和图论对此提供了很好的分析思路。研究者最早将小世界网络解释为规则网络,即网络中每个节点(共有 N 个节点)都遵循既定的规则,只和该节点最邻接的 K 个节点建立连接(见图 1 的左图);与规则网络相反的是随机网络,即网络中节点之间的连接是完全无规则的,每个节点都有同等的机会和其它节点建立连接,不存在高度连通节点和集聚情况(见图 1 的右图),这两种情形都与实际不相符。Watts 和 Strogatz 将规则网络上的每一条边按一定的概率 p ($p \approx 0.1$) 进行重定向,增加与其他节点之间的连接,同时保证没有重复的边和自连接的边,这时就会出现少量的快捷连接,它们会伸展到较远的节点,但由于 p 很小,网络模型总体仍大致维持规则结构(见图 1 的中图),也就是说小世界网络是具有一定随机性的一维规则网络^[4-5],这就是著名的 W-S 小世界网络模型。

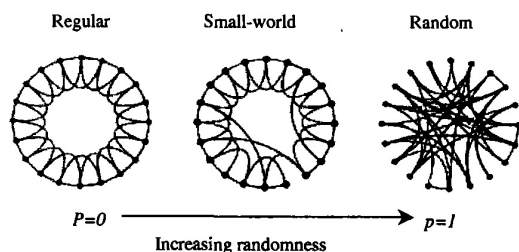


图 1 规则网络、W-S 小世界网络和随机网络的简化模型

资料来源:Watts Duncan J, Strogatz Steven H. Collective dynamics of "small-world" networks[J]. Nature, 1998, 393: 440-442. 图中显示了通过调节 p 值进行重定向,实现的从规则网络到小世界网络再到随机网络的转化,图中 $N=20, K=4$ 。

W-S 小世界模型中涉及四个重要特性,即特征路径长度(Characteristic path length L)、集聚程度(Clustering coefficient C)、捷径(Short Cut)和中心节点(Central node),这成为深入理解小世界现象形成机制的切入点。各类网络中的节点数量和位置在建立连接之前都是不确定的,因此难以形成纯粹的随机网络,但节点之间也存在随机的连接,它会有助于产生捷径。另一方面,虽然连接难以受制于某一具体规则,但节点总是围绕某一主题建立相关关系,从而会出现网络中与某个节点相连的节点间又存在彼此相连的现象,表现出很高的集团化聚类效应^[6]。这一效应不仅为其中

的各节点提供连接通道,也成为限制网络增长的重要力量,使得网络中所有节点对的平均路径长度 L 随着网络大小 N 呈对数增长(L 值较小)。此外,节点的连接常常受到优先连接机制和马太效应的影响^[7]。即网络中越是有更多连接的节点越能获得更多的连接,连接数量的累积使得整个网络并不均匀,产生出少量拥有较多连接的中心节点以及一些联系更加紧密的子团体即凝聚子群,中心节点的存在为网络中任意两个节点创造了联系途径,有利于降低网络的特征路径长度。总之,小世界现象的形成展现了网络内部结构和关联关系的建构原理,它与多种网络机制的合力作用息息相关,也会对网络内外的个体产生影响。

2 学科信息门户的小世界现象分析

学科信息门户(Subject Based Information Gateways,简称 SBIGs)通过对某一学科领域的资源进行收集、分析、鉴别、标引和组织,实现学科网络资源内容的高度组织集成,为用户提供访问某一学科资源与服务的单一入口和统一协作的学术交流环境^[8]。作为互联网的一部分,它自然具有了小世界特征,不仅如此,由于学科知识网络、用户信息行为及自身建设的一些特点,学科信息门户的小世界现象表现更为明显。

首先,学科知识网络中存在小世界现象。数学界的“艾尔德数”最早反映了数学领域合作网络的小世界现象,此后多项研究表明,在生物医学、计算机科学、物理学、生物学、图书情报与文献学等领域的研究者合作网络中也都存在小世界现象^[9-10]。文本层面的小世界现象也有发现,石晶等人证明了由文本形成的词汇共现图呈现短路径、高聚集度的特性^[11]。从专类网站的链接分析来看,Lada Adamic 分别分析了 64 826 个 web 网站和 11 000 个 .edu 网站,发现两者的集聚程度分别为 0.081 和 0.156,特征路径长度分别为 4.228 和 4.062^[12],说明后者具有更高的集聚程度和更短的特征路径长度,这与后者主题集中性更强有关。当然,由于同一学科的地区发展差异以及不同学科的发展差异,小世界现象也存在差异,如同是“艾尔德数”,欧美作者就比亚非等地作者的数值普遍偏小。

其次,从用户信息行为方面来看,也显示出可获取信息很多,但实际获取信息有限的小世界现象。从获取数量上看,Spink 在 1997-2002 年对 Excite 搜索引擎的 Web 日志统计发现,大多数用户只查看返回结果的前十条,每页 10 个记录的话,平均查看结果的数量是 2.35 页^[13]。从获取过程来看,用户当超出了一定页数仍对结果不满意,则选择其他方式途径或放弃。即使是较令人满意的检索,也常表现出“适可而止”的行

为,通常不甚追求结果的“全面无遗漏”^[14]。这就意味着个体用户会在自身可承受的知识负荷与信息获取成本的前提下,主动的在小范围内获取信息,因而用户信息获取的集聚程度很高。另外,论文引用体现着用户对信息的主动利用,研究发现其中同样存在着小世界现象^[15]。

第三,学科信息门户的建设特点有利于形成小世界现象。从资源选择的角度来看,学科信息门户中的资源都是围绕某一学科主题进行集中的,相互之间具有天然的关联性,被链接的信息之间往往又互相链接;从信息组织的角度来看,分类法和主题法提供了学科知识之间有序和多重链接的基础;从用户角度来看,学科信息门户主要针对专业用户,他们希望在信息门户中获得一站式的信息服务,因此建立多个信息点之间的链接必不可少;从技术的角度来看,超链接技术既实现了信息集成和有序组织的功能,也定义了超文本的非线性结构,可以快速实现不同网页和信息点之间的切换,提供信息获取的捷径。

综上所述,学科信息门户的小世界现象实质是学科信息、用户需求和知识组织共同作用的结果,超链接则是重要的实现手段。链接具有数量、结构、集聚度、距离、可达性等多维属性,考察链接状况可以获取整体网络的发展状况。

3 学科信息门户的链接现状分析

我国的学科信息门户建设大约始于 1999 年上海图书馆的“数字图书馆资源总汇表”和 2000 年 CALIS 组织的学科导航库^[16]。目前已有国家科学数字图书馆(CSDL)、中国科技图书文献中心(NSTL)、中国林业科学研究院、武汉大学、武汉理工大学等多个主体参与,建设的学科信息门户涉及生命科学、化学、数学物理、资源环境、图书情报、林业、交通运输等多个领域。其中,CSDL 自 2001 年启动以来先后建立了 5 个学科信息门户,建设比较规范,在国内学科信息门户中具有很强的代表性。因此,本文重点对这 5 个网站的链接情况进行调查分析。运用链接获取工具 SocSciBot 3 分别爬取 5 个网站,但因生命科学学科信息门户无法爬取,最终仅获得了 4 个学科信息门户的链接数据。随后,采用社会网络分析工具 Pajek 和 Ucinet 6 进行测量,内容包括网络密度、节点数量、链接数量、网络集聚度、特征路径长度等小世界度量常用指标(见表 1)。

分析发现,这 4 个学科信息门户的节点数量和链接数量都较多,说明门户纳入的资源数量较多,学科资源较丰富。网络的特征路径长度都是较小的常数,说明门户确实存在小世界现象。但网络中链接的关联性

并不乐观。网络集聚度的数值位于 0 和 1 之间,值越大说明节点之间越紧密,Ucinet 提供了两种集聚度^[17]:一种是基于局部密度的集聚度,有 3 个门户的数值都偏小,说明网络链接结构是较疏松的;另一种是基于传递性的加权集聚度,发现在第一种计算中值较大的图书情报学科信息门户此时也变得非常小,说明它虽然局部密度高,但可传递性差。可达网络密度等于关联度,各个点之间越相关,密度就越大,可以看出这一数值也是非常之小的,与之相对的是在网站中还存在大量不能互相抵达的节点对,其数量级甚至到亿。不可达节点对的存在有一定的合理性,但如此巨大的数量说明信息门户在关联链接和深层链接上的表现较弱。

点度中心度反映了一个节点与其他节点的直接联系,各门户网站的点入度和出度均值相同,但从标准偏差差来看情况并不相同,出度的标准偏差值较小,说明点出度差异较小,而入度的标准偏差值较大,说明各节点的入度差异明显。从详细列表中能发现更多问题。在点入度列表中发现网站首页入度很高,这符合网站一般情形,但多数网页的入度很低,说明指向各节点的链接少,如果一旦去掉进入链接,那么多数页面就无法获取。

表 1 CSDL4 个学科信息门户的网络链接情况

学科信息门户名称	化学学科 信息门户	环境资源学 科信息门户	物理数学学 科信息门户	图书情报学 科信息门户	
节点数量	15850	8093	11224	5016	
链接数量	172212	75250	111661	182424	
不可达节点对数量	180703614	41231737	70087635	135405	
网络 集聚 度	基于局部密度	0.379	0.251	0.298	0.825
	基于传递性	0.025	0.014	0.008	0.014
网络密度	0.0006855	0.0011489	0.0008864	0.0072505	
特征路径长度	4.2673	3.79913	3.96152	5.89465	
最远节点距离	6	5	7	20	
点出度中心度均值	10.838	9.275	9.836	36.357	
点出度中心度标准差	18.084	14.495	12.079	7.027	
点入度中心度均值	10.838	9.275	9.836	36.357	
点入度中心度标准差	171.523	126.561	163.814	398.411	
网出度中心势	1.036%	4.508%	1.944%	4.321%	
网入度中心势	27.986%	36.927%	44.263%	98.756%	

从网的中心势来看,数值在 0 到 1 之间,环形网络的中心势为 0,星形网络的中心势为 1^[18]。4 个门户的网的出度中心势普遍较小,比较接近 0,表明网络的出链呈现出环形结构,化学学科信息门户尤其明显;图的入度中心势则表现不一,图书情报学科信息门户几乎为 1,这意味着网络的入链近乎星状结构,网络结构脆弱,一旦去除入度大的节点如首页,那么多数链接都将失效;同时也说明除了中心节点外其他节点之间的横向联系非常薄弱,可以看出节点最远距离(20)与特征

路径长度(5.89465)相比差异较大,而且这两个数值也比其他门户的相应数值大。

总之,学科信息门户通过将链接建立在对主题的属、分、参等多重关系描述和对分类全面标引的基础上,实现资源节点的横向联系,展现出基本的小世界特性。但分类法和主题法拘泥于传统等级列举式体系结构,体现的主要是一种显化的联系,也导致出现链接深度不足、链接关系不够紧密、节点对链接可达性低、节点的多重链接不充分等问题。

4 学科信息门户链接设计的优化策略

小世界现象为学科信息的链接建设提供了启示:其一,小世界就是一个关联之网,因此要充分发掘知识节点的关联,在知识概念层次增加链接数量;其二,个体用户仅追求知识的小范围扩展,进行信息推送和检索服务时就需要控制链接规模,寻找链接集合边界;其三,小世界现象说明总体的重大变化可能来自局部微小的网络变动(捷径),因此要考虑对学科信息的内部关联进行深入挖掘以发现和利用捷径;其四,中心节点的作用值得重视,要探索发现的方式。综合考虑这些启示以及学科信息门户的现存问题,笔者提出以下优化策略。

4.1 采用知识链接技术发展多重链接 目前学科信息门户的链接对象主要是网站,而对网页内容以及网站内的单项服务很少建立链接,这就导致链接的深度、数量有限,配置也主要拘泥于本地信息。知识链接技术的应用可以实施在三方面,一是对知识体的本质属性和附加属性、横向和纵向关联、同质网络和异质网络进行多维识别,设定知识之间的多重关联,在知识交汇节点处建立多个链接点,增加学科门户的交叉链接数量^[19];二是增加知识链接对象的颗粒度和表现形式,深入到被链接网站的网页及服务模块层面进行重点索引,在被链接网站的资源描述页面中增加热词标引,使知识链接更为自由和开放;三是基于开放链接标准 OpenURL 提供用户可以获得的扩展链接服务,即信息门户不仅列举和描述每一种资源,而且分析用户可以访问的资源体,如介绍某一学科数据库时自动分析用户网络环境中能否访问该数据库,可以的话就直接提供链接。

4.2 基于凝聚子群分析确定链接集合边界 凝聚子群在学科信息门户中就表现为由链接关系密切的知识节点构成的小群体,我们可以通过建立凝聚子群来确立链接的集合边界,控制用户使用时的链接延展范围。有两种凝聚子群的建立方法可供参考,一是建立在子群成员之间的可达性基础上的凝聚子群 n -派系,

它考虑的是点与点之间的距离,可以设定一个临界值 n 作为节点链接距离的最大值, n 越大,对派系成员限制的标准就越松散。二是建立在点度数基础上的凝聚子群 k -核,它指的是一个网络中所有节点的一个子集,该子集中的每个节点至少与 k 个节点相连^[20]。 k 值不同,得到的 k 核也不同,运用 k -核试探法将可以发现有意义的凝聚子群。建立任一类型的凝聚子群都需要设定数值,数值的合理设置需要全面考虑知识节点的关系。在传统信息组织方法基础上运用本体论,从语义层次对领域内的概念和概念关系进行细致深入、明确规范的表达,形成稳定的知识节点集合。引入 web2.0 和自动标引技术增添动态知识节点,经过多次检验后可纳入已有的节点集合中,形成可持续发展的语义网络,在对其结构分析的基础上设置凝聚子群的数值,就能迅速提取出有价值的链接集合。

4.3 基于数据挖掘和知识地图技术寻找捷径 捷径在相距较远的节点之间发挥了重要的沟通作用,但如何发现捷径却比较困难,因此有必要探讨如何在公开的信息中发现信息之间的深度联系。数据挖掘通过对大量的模糊的信息数据进行关联分析,能够提取出隐含的但又潜在有用的知识关联模式,对于寻找捷径是一种非常适用的技术方法。数据挖掘的关联分析涉及到多种算法和规则,其中,隐性关联知识发现规则的适用性更强。它通过共词分析法寻找频繁在一起出现的两个或多个主题词/副主题词组合,不仅可以复现学科知识中被主观弱化的知识联系,而且可以发现不同学科领域间尚未被发现的关联^[21-22]。对于前者我们可以增加对应链接,后者就可以被作为捷径用于改善网络链接。知识地图是组织中相关知识及其关系的图示,通过抽象化的编码语言来展开知识脉络,可以将毗邻的知识单元联系起来进行知识关系的双向描述。运用知识地图既能表达显性的知识联系,也能推理出潜在的知识联系,对于探索学科信息门户的捷径同样可以发挥作用。

4.4 运用信息计量和社会网络分析方法识别中心节点 信息计量学中的布拉德福定律、洛特卡定律和齐普夫定律分别从文献分布、著者分布和词汇分布三个角度指出少数文献、著者和词汇构成了学科领域的核心,它们在学科信息网络中容易形成为中心节点。这三大原理推导中采取的方法都可以用来识别中心节点,更方便的方式是词频统计分析和引文分析法。对具有实指意义的词汇进行切分和词频统计,拥有高词频密度和向心力的词汇可作为中心节点予以关注;运用引文分析法可以揭示学科领域内的核心信息源、著者和机构,也可以看作是中心节点运用到链接设计

中。但要注意的是,这两种方法在进行计算时,将所有节点视为均等,不考虑关联节点的性质和重要程度,有可能带来误判,实际操作时可以考虑为链接赋予不同的权重,通过迭代计算,最终发掘中心节点。

社会网络分析方法用“点的中心度”作为评估节点在网络中的中心程度的指标,有三种中心度:节点的“中心入度”是指该节点的入链数量,节点的“中心出度”是指节点的出链数量,节点的“中心均衡度”是指节点的入链与出链总数。通过对这三种中心度进行计算和排序就可以寻找出中心节点^[23],同时还可以绘制网络可视化图直接找出这些关键节点。在识别的基础上,可对中心节点的科学性进行考察和调整,从而改善学科信息门户的链接质量。

5 结束语

小世界现象较好地描述了学科信息门户的链接特性,也为优化门户的链接设计提供了启示。运用社会网络分析工具,文中对 CSDL 的学科信息门户的链接情况进行了数据分析,证实了小世界现象的存在,也发现一些问题。这些问题表现为链接关联和分派的不足,实质则反映出传统信息组织方法的不足。结合信息组织和小世界特性优化链接设计是可行的,本文对此提出了一些策略性的指导,但更有意义的还是深入探讨不同学科以及同一学科链接网络的小世界现象差异,提出更具操作性的实施方式以及实践案例,这都将是今后继续探索的方向。

参考文献

- [1] [美]瓦茨著;陈禹等,译.小小世界:有序与无序之间的网络动力学[M].北京:中国人民大学出版社,2005:19
- [2] 朱亚丽.“六度分离”假说的信息学意义[J].图书情报工作,2005(6):59-61,32
- [3] 刘植惠.大世界悖理与小世界现象——情报交流研究新进展[J].重庆图情研究,2006(1):1-3,28
- [4] Watts Duncan J, Strogatz Steven H. Collective Dynamics of "Small-world" Networks[J]. Nature, 1998, 393:440-442
- [5] 司徒俊峰. Internet 的小世界网络研究[J]. 情报杂志, 2004(12):86-88
- [6] 庞景安. 网络信息资源的计量与评价[M]. 北京:科学技术文

献出版社,2007:128

- [7] 王 炼. 小世界现象形成机制与马太效应[J]. 情报学报, 2007, 26(3):477-480
- [8] 孔 敬,李广建. 学科信息门户:概念、结构与关键技术. 中国图书馆学报,2005(5):50-53,90
- [9] Newman M E J. The Structure of Scientific Collaboration Networks[J]. Proceedings of the National Academy of Sciences of the United States of America, 2001, 98(2):404-409
- [10] 晏尔伽,朱庆华. 我国图书馆、情报与文献学领域作者合作现状——基于小世界理论的分析[J]. 情报学报, 2009, 28(2):274-282
- [11] 石 晶,胡 明,戴国忠. 基于小世界模型的中文文本主题分析[J]. 中文信息学报, 2007, 21(3):69-75
- [12] Adamic Lada A. The Small World[EB/OL]. [2010-03-05]. http://www.hotiron.com/chaotics/bibliography/the_small_world_web.pdf
- [13] Spink, Amanda, Xu Jack L. Web Searching: the Excite study [EB/OL]. [2009-09-28] <http://www.shef.ac.uk/is/publications/infres/paper90.html>
- [14] 朱震远. 网络信息检索环境中知识链接的设计——基于语用和用户行为研究的视角[J]. 图书情报工作, 2010(16):130-133,81
- [15] Redner S. How Popular is Your Paper? An Empirical Study of the Citation Distribution. European Physical Journal B, 1998, 4(2):131-134
- [16] 吕慧平,陈益君,周 敏. 中国学科信息门户网站建设的现状与问题探讨[J]. 现代情报, 2006(9):137-141
- [17] 刘 军. 整体网分析讲义——UCINET 软件实用指南[M]. 上海:格致出版社,2009:169
- [18] 刘 军. 整体网分析讲义——UCINET 软件实用指南[M]. 上海:格致出版社,2009:98-99
- [19] 周晓英. 知识链接的发展阶段、发展动因和类型特征分析[J]. 图书情报工作, 2010, 54(12):36-40
- [20] 刘 军. 整体网分析讲义——UCINET 软件实用指南[M]. 上海:格致出版社,2009:117-122
- [21] 曹志杰,冷伏海. 共词分析法用于文献隐性关联知识发现研究[J]. 情报理论与实践, 2009(10):99-103
- [22] 钟伟金,李 佳. 共词分析法研究(一)——共词分析的过程与方式[J]. 情报杂志, 2008(5):70-72
- [23] [英]迈克·赛沃尔著;孙建军,李 江,张煦等译. 链接分析:信息科学的研究方法[M]. 南京:东南大学出版社,2008:185

(责编:白燕琼)