



基于科学知识图谱分析的知识演化研究 ——以生物医学为例

马费成, 陈潇俊, 刘 向

(武汉大学 信息资源研究中心, 湖北 武汉 430072)

摘要:通过文献计量学方法绘图和引文网络分析,采用大型复杂网络分析软件Pajek和Citespace来进行数据处理,探讨了知识演化与创新的过程。以生物医学研究领域为例,分析了该学科的知识网络特征,着重分析了网络中的主题聚类、关键路径并识别了网络中权威和核心文献,为科学知识的演化与发展提供了一个较好的理解。

关键词:知识网络;权威性;中心性;知识演化;生物医学

中图分类号:G350 **文献标识码:**A **文章编号:**1007-7634(2012)01-01-07

Study on the Knowledge Evolution Based on Mapping Scientific Domain —A Case of the Biomedicine Field

MA Fei-cheng, CHEN Xiao-jun, LIU Xiang

(Center for Studies of Information Resources of Wuhan University, Wuhan 430072, China)

Abstract:In this paper, we combine bibliometric mapping and citation network analysis to discuss the process of knowledge evolution and innovation in the Biomedical field, with the use of large network analysis software Pajek and Citespace for data processing. We analysis the characteristics of disciplinary knowledge network, mainly talk about topic clustering, the critical path and the identification of authorities and hubs documents of the network. Useful conclusions in this paper can provide a better understanding for scientific knowledge evolution and development.

Keywords:knowledge network; authority; hub; knowledge evolution; biomedicine

科学知识可以假定为一个随着时间推移而沿着一条“光滑路径”增长的量,其中,那些介绍了重要思想或提出新颖观点的文献会得到引用,直到有新的研究结果对它们进行了改进。随着新知识的不断产生,知识网络结构将会发生变化,可以说此时发生了知识网络演化。知识活动中的引用行为隐含了一个“知识构建”的过程,从这些文献形成的知识网络中,我们可以看出知识的演化和创新。研究知识网络图谱的结构和特征,揭示知识演化和创新的过程,对于理清知识的发展脉络,把握学科领域发展全局,探测

和追踪创新领域及知识发展的趋势有着重要意义。

上世纪50年代,SCI创始人Garfield意识到科学引证网络可以反映科学知识之间传承、发展的关系,并尝试利用引证网络研究科学知识发展的历史、脉络和结构,这一想法在后来相关学者的研究中得到了证实^[1]。1981年,英国情报学家Brookes提出了认知地图的构想,他认为某些学科领域之间存在固有联系,可以建立一个概念联结网络,并绘制出了认知地图的雏形。对于此,马费成在其著作中进行了高度评价,他认为认知地图如果可以成功构建,便能

收稿日期:2011-10-16

基金项目:国家自然科学基金项目(71173249)

作者简介:马费成(1947-),男,贵州人,教授,博士生导师,主要从事情报学理论与方法,信息经济与信息资源管理研究。

有效地遏制人类的情报危机,同时也使情报学研究取得突破性进展^[2];赵蓉英、邱均平等认为这一构想指明了未来知识组织的最高目标^[3]。在前人思想的基础上,学者们纷纷展开了对知识网络的研究,知识网络图的形式异彩纷呈,如知识地图、概念地图、科学知识图谱、合作网络等。同时,学者们将计算机技术应用到该领域中来,Pajek、Citespace、SPSS、Netminer等分析处理软件得到了越来越广泛的运用。

本文以生物医学领域为例,结合引文分析和知识网络图谱的绘制,利用软件Pajek和Citespace来对知识的演化与创新进行系统地分析,探讨该领域研究概念的变化和关键路径的形成等,并提出未来的研究重点。

1 研究模型

要把握学科领域研究的发展,了解其中知识创新与演化的过程,我们需要着重解决以下几个问题:

①在一个学科领域的期刊文献中,哪些论文和理论是研究的主流? ②研究主题是如何随时间的变化而变化的? ③哪些文献是权威和核心的?

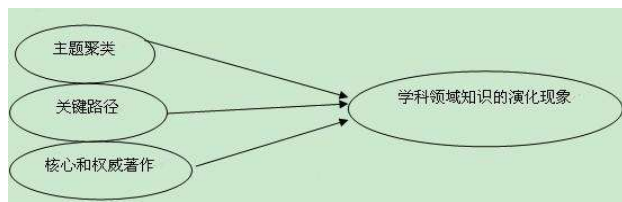


图1 基于知识网络图谱分析的知识演化研究模型

每一个学科领域,总有某些主题构成该领域的研究基础。通过对这些主题进行分析并对文献聚类,可以将某一领域划分为几个主要的子领域,并识别出这些子领域的研究随时间推移而发展的情况。通常,在知识的演化过程中,某些较为重要的文献会组成这个学科领域的“关键路径”,它可以展现该学科领域知识的演进历程,为学科领域的知识演化研究提供很好的依据(参见图1)。此外,通过核心和权威著作的分析,可以识别出该领域中那些起着重要作用的文献并探讨这些文献之间存在的关系,从而为知识的演化提供一个好的思考方向。

2 数据与方法

2.1 资料与数据

本文数据来源选择了ISI,并以期刊文献为研究

对象。在选择学科领域时综合考虑了三个因素,即研究比较热、在ISI中的文献多、学科地位比较重要等。我们选择了生物医学领域,并对Cell杂志上的文献进行分析。Cell Press期刊在生物医学研究中享誉甚高,它的诸多子期刊均被SCI收录。本文选取了其中2009年影响因子分别排在前两名的子期刊——《Cell》和《Immunity》来研究,设定时间跨度为1998–2010年。搜集和处理原始数据的步骤如下:首先,以各子期刊的出版物名称作为关键词,选择SCI和SSCI两个数据库进行搜索,文献类型选择articles和proceeding papers,下载论文的基本信息,包括作者、发表时间、关键词等。然后,将数据格式进行转换(方便pajek处理),并提取这些文献的基本特征,如关键词、作者列表、同被引列表等等。最终得到了《Cell》和《Immunity》截止至2010年3月18日被ISI收录的所有论文的题录数据,共计4258条,其中最先发表的一篇文章是在1969年。这些文献基本上能代表生物医学研究领域的科研成果。

2.2 方法与步骤

第一步,建立整体的科学知识图谱。这一步将识别出哪些子领域可以吸引较多的文献,哪些发展迅速,哪些的研究热点逐渐降低等,使我们对该领域有初步的了解。通过聚类分析,可以得到一些以主题为标志的子领域。这一步采用的处理软件为citespace。由于采取的数据本来便是从两个子期刊上搜集的,一开始便有一定的分区域性,但是由于作者在投稿时,往往并未严格按照所属领域来投,同时由于某些子领域的交叉性,作者在投稿时并不能那么理性地选择期刊,而最终被该期刊收录了,所以也可以考虑将得到的子领域与现有期刊各自的内容范围进行对比分析,以对期刊特点和学科领域的划分有更好地理解。

在这一步的处理中考虑了两种途径。第一种,通过对关键字进行共词分析得到主题集群。首先,收集这些文献所有的关键词,通过软件阈值设定进行关键词去重和筛选,得到主流的关键词并计算它们在文献中共同出现的次数。然后,采用“分层凝聚、完整联系”的方法进行聚类分析,从而确定若干个主题或关键字集群,并把它们当作子领域。利用这些主题,将各篇文献“分派”到指定的子领域中去。此外,需要识别集群之间的“重叠”现象,即指文献在多个子领域存在的情况。第二种,通过文献同被引来进行聚类。与第一种不同的是,所聚类的对

象不是文献关键词,而是文献本身。得到的聚类也可以看成子领域,并用文献的标题来对这些子领域进行标签。

第二步,构建引文网络。引文网络的分析始于 Garfield 和 Torpie 对阿西莫夫 DNA 序列历史研究^[4]。通过引文网络分析,探索围绕该领域的理论和概念发展的过程。高被引文献往往具有很重要的内容,一篇经常被引用的论文可作为一个“概念的象征”,表示该论文的作者对于一个领域的科学家或一个研究主题、方法的观点^[5]。在知识的演化时期,新方法的探究与以往的成果通常是共存的。引文网络能够使我们适时地从两个方面来研究数据。一方面是纵向网络的分析(如利用“关键路径”来进行分析),它能使我们了解研究流中的收敛与发散情况,显示网络在时间上的连通性变化。另一方面是从横向的角度来看知识网络,例如通过基于“集线器”和“权威”的分析来定义重要文献等。

首先介绍关键路径。如果说知识通过引文来进行流动,那么经常用到的引文,远比几乎发挥不了作用的引文更为重要^[6]。在所有可能的引文“链”中,人们经常利用网络算法来计算最为重要的路径并定义了最为关键的路径,即“关键路径”。这种方法并不包括对所有引文文献进行完整的计数,而是通过计算整个数据集中所有可能的路径,并从中选择可以代表该领域主干理论演进的关键路径^[7]。

关键路径的起点一般是发表较早的论文,而终点则是发表较晚的论文,这样突出了“关键路径”在时间上的演进特点。在知识网络研究中,一条路径的重要特征是:在起点和终点之间不存在出现超过一次的节点或边。本文采用了“中心度”(centrality)的概念来寻找组成关键路径的节点。“中心度”是一个很重要的指标,它主要由 Freeman 提出^[8]。关于它的算法有许多种,国内吴思竹等学者将其划分为 5 种,并详细介绍了其中的中介中心度算法——它是基于最短路径的一种经典算法,能够用于揭示网络中具有连接桥作用的节点,从而发现网络连接中的关键点或脆弱点^[9]。本文便是采用“中介中心度”来进行分析,在时间轴上依次选取每一年(或者某几年)具有较高中介中心度的节点来组成该领域的关键路径。我们采用了 citespace 来进行分析。

然后来看“集线器”和“权威”节点。在文献计量学、超文本链接、万维网的研究中,许多学者认为识别网络中的重要节点是非常重要的。现阶段,学术界比较推崇的 SCI 论文“影响因子”,由加菲尔德于

1972 年提出。紧接着, Pinski、Narin 和 Geller 等学者提出了一个衡量期刊的算法,他们认为如果一个期刊收录了很多来自其他期刊的文献,而这些期刊同时是有很高地位的,那么这个期刊可以认为也是很有地位的期刊^[10]。1999 年,国外学者 Brina、Page 与 Kleinberg 等采用了“集线器”和“权威”的概念,来计算提高网络引擎搜索效率的方法^[11]。Kleinberg 开发了一种用于计算“集线器”和“权威性”的迭代算法,指出这两者之间存在一个相辅相成的关系:一个良好的“权威”出版物,是被许多很好的“集线器”出版物引用的,而一个良好的“集线器”出版物,则引用了许多很好的“权威”出版物。其研究表明,该算法可以帮助过滤掉不相关的或质量差的文献,并确定高品质的文献。

我们认为对“集线器”和“权威”出版物的定义是了解引文网络中重要文献作用的一个很好途径。网络中的节点如果指向许多好的权威节点,则被认为是很好的“集线器”。如果它被许多好的集线器点连接,那么它就是一个好的权威。这一步采用 Pajek 来进行分析。

3 实验分析

3.1 聚类与归类

3.1.1 以关键词为节点进行的聚类

上文已说明,本文尝试了两种途径进行聚类分析。第一种是基于关键词的聚类分析。在 citespace 中,为控制节点和边的数量在一个合适的范围内——既不过大(如超过 1000)也不过小(如低于 500),经过多次阈值调节的实验,最终选取了一组较为合适的阈值(下同)。选取节点类型为 Keyword,将阈值设为(3,2,15)、(3,2,15)、(3,2,15),采用 Pathfinder 算法提取出其中较为重要的关键词,并将主要的文献进行归类。经 Citespace 分析得出的结果,共有 591 个节点与 2050 条边,进行自动聚类结果如图 2 所示。

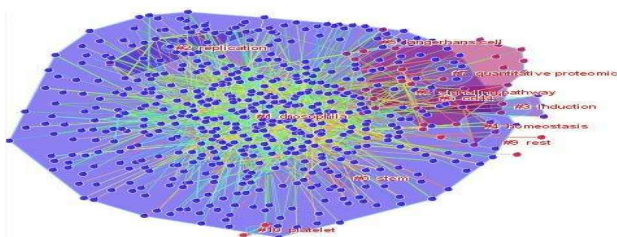


图2 利用关键词进行的聚类情况

可以看出,citespace将这些keywords分成了11类,且类与类之间存在重叠现象,这是符合客观规律的。然而我们仔细分析了这11类的详细分类情况,发现并不科学,因此放弃了此种途径的分析。

3.1.2 以文献为节点基于文献同被引进行的聚类

第二种途径是以文献本身为研究对象,通过同被引来进行聚类。结果表明,此种途径是更好的选择。将阈值设定为(6,2,20)、(5,3,18)、(4,4,15),节点类型选为“cited reference”,处理结果得到了785个节点(被引文献)和6349条边(文献之间的引用关系),通过tf*idf(词频-逆文档频率)聚类,用title terms来对各个类进行标签,结果如图3所示。

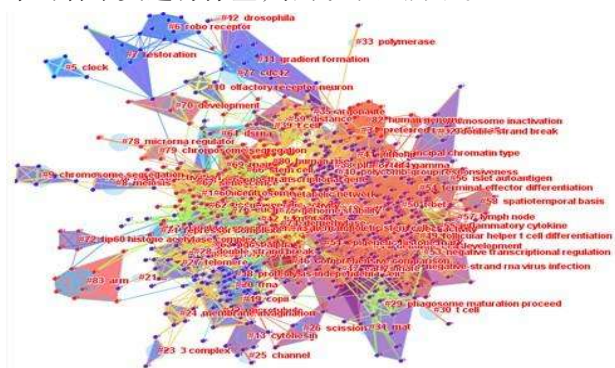


图3 基于文献共被引所得到的节点聚类情况

经过citespace自动聚类,这785篇文献分成了84类,每个子类都代表着生物医学领域的一个小的分支,如子类RNA、virus infection(病毒传染)、human genome(人类基因组)等,它们中间有交叉,但又不完全相同。为研究这些由文献组成的类随时间的变化而出现频率变化的情况,利用citespace得到了这些类在时间上的分布图,如图4所示。图中类对应的横向粗线代表了该类在时间轴上出现的分布状况,粗线越长,所跨的时间就越长。

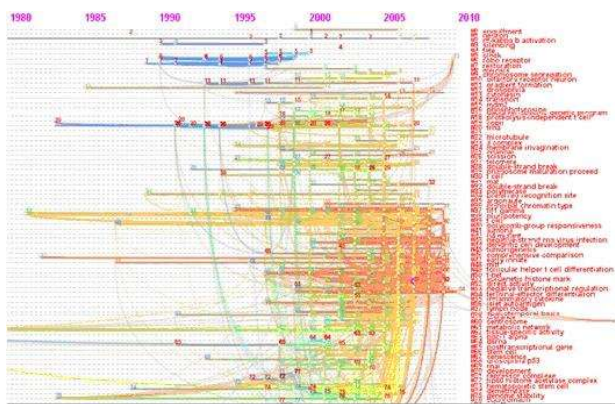


图4 各个类在时间轴上的分布情况

本文选取了其中节点数排名前3的类来进一步分析(这三类所拥有的节点数之和占据了所有节点

的三分之一)。将数据直接导出到Excel,排名前三的类分别为类80(gene)、类20(human risc)和类44(cell development),其节点数分别为95、90、64。可以说这三大类占据了大部分的节点,同时,这三类之间也有“重叠”现象。通过对文献发表的时间进行分析,得出了这三大类在时间轴上的引文分布。类80(gene)的论文涵盖的时间区域较为广泛,基本涵盖了所研究的时间区域,说明一直以来都是研究的热点。而类20(human risc)所覆盖的时间区域相对较小,大多论文都是在最近几年内,但此类中,许多论文的Burst值较高,并且论文总数在2005年达到高峰。类44(cell development)则主要集中在最近几年的时间范围内。

为验证分类的合理性,我们结合了关键词的出现频数来进一步确定生物医学领域的研究热点。利用在尝试第一种途径中得到的关键词频率数据,选出了出现次数大于70次的8个关键词术语,如表1所示。

表1 出现频数排列前八的关键词列表

序号	出现频数	英文术语	中文术语
1	546	protein	蛋白质
2	284	gene-expression	基因表达
3	248	crystal-structure	晶体结构
4	276	activation	激活素
5	148	cell	细胞
6	108	messenger-rna	信使RNA
7	85	mammalian-cells	哺乳动物细胞
8	74	Transcription actor	转录因子

可见,这些高频词基本上能归到这三大类中去,protein、gene-expression、messenger-rna、transcription factor可以归在类“gene”里,cell、mammalian-cells在类“cell development”里,crystal-structure、activation便属于类“human risc”(人源RISC复合物,一种化学分子)的范畴了。值得指出的是,由于利用软件进行自动聚类尚且存在一些缺陷,在选用标题词来给各个类进行标签的过程中,所选取的标题只能在一定程度上代表该类,但不能说完全涵盖了该领域的主题。

3.2 关键路径的描绘

我们利用citespace构建了引文网络。选定节点类型为“cited reference”和“author”,阈值同样设为(6,2,20)、(5,3,18)、(4,4,15),得到了中心性大于或者等于0.14的7篇关键节点文献的详细信息(见表2),并绘出了生物医学领域文献共引网络节点的时区视图(见图5)。

结合上面的图和表格,可以对这七篇文献进行

表2 生物医学研究的7篇关键文献(按时间排序)

文献英文标题	中文标题	被引频率	中心度	作者	年份
The genetics of <i>Caenorhabditis elegans</i>	线虫的基因遗传	4998	0.21	BRENNER S	1974
Basic local alignment tool	DNA 序列基本分析工具	29987	0.14	ALTSCHUL SF	1990
Rho GTPases and the actin cytoskeleton	Rho GTPases 和肌动蛋白细胞骨架之间的关系	3618	0.16	HALL A	1998
The language of covalent histone modifications	共价组蛋白改性的启示	2839	0.26	STRAHL BD	2000
A system for stable expression of short interfering RNAs in mammalian cells	用于哺乳动物细胞中干扰 RNA 进行稳定表达的系统	2626	0.15	BRUMMELKAMP TR	2002
Asymmetry in the assembly of the RNAi enzyme complex	RNA 干扰酶混合物分配的不对称性	1405	0.13	SCHWARZ DS	2003
Chromatin modification and their function	核染色质的改变及其功能	153	0.21	KOUZARIDES T	2007

详细分析,即可梳理出生物医学领域大致的演进路径。

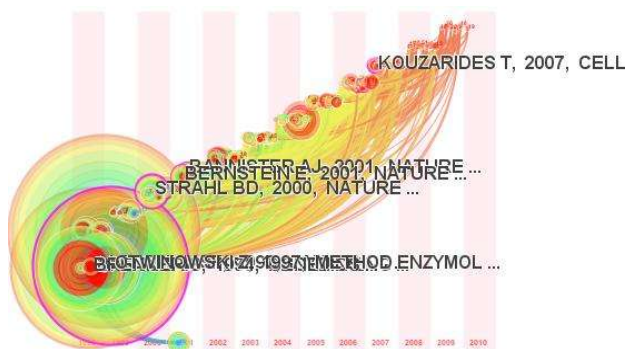


图5 生物医学领域共引网络节点的时区视图

第1篇文献是1974年,由英国剑桥大学分子生物学研究领域的BRENNER S教授发表在Genetics期刊上的《线虫的基因遗传》一文^[12]。作者主要描述了线虫各种基因变异形成的生态体,同时也定义了大约100种的基因。该文可以看作是生物医学领域的“前学科时期”,此时学者们主要研究小生物的基因并对其进行特点与表达出来的形态进行归纳与分析。

第2篇文献发表于1990年,由学者ALTSCHUL SF发表在期刊journal of molecular biology的《Basic local alignment tool》^[13]。该文主要从生物技术的角度进行研究,介绍了快速进行基因序列的对比方法,并实现了DNA和蛋白质序列数据库检索、主题检索以及基因的鉴定。它的被引频次高达2万次以上,可见这篇文章得到了学术界的广泛认可,并具有极大的利用价值,它标志着生物医学技术的萌芽,此文为生物技术的研发奠定了坚实的基础。

第3篇文献发表于1998年,是由HALL A教授发表的《Rho GTPases和肌动蛋白细胞骨架之间的关系》一文^[14]。该文主要从细胞生物学的角度出发,并兼顾健康与疾病的视觉,探讨了肌动蛋白细胞骨架的组织机理与生化机制,重点介绍了RHO(一种蛋白质,属于小G蛋白超家族的亚家族成员^[15])在调节肌动蛋白细胞骨架与控制基因转录等活动中的重要作用。相比于前面两篇文章,该文在研究对象上有

了更高层次的深入,并开始将生物学领域的知识与人体疾病和健康相结合,研究细胞骨架的组成机理,重视蛋白质在其中所起到的重要作用,该文之后,学术界掀起了一股研究蛋白质的热潮。

第4篇文献出现于2000年,是由英国著名学者STRAHL BD发表在《NATURE》期刊上题为《共价组蛋白改性的启示》一文^[16]。文中介绍了真核染色质的基本构建块——组蛋白和它们用来形成DNA的核小体,并从微观分子的角度探讨了这些蛋白质之间的相互作用关系,从而更好地揭示染色体的本质。该文继承了前人对蛋白质的研究,并进行了更深层次的探讨。从表格可以看出,本文的中介中心度达到了0.26,在所有文章当中是最高的。毋庸置疑,它在整个学科领域发展中起着最为关键的作用,不仅能够很好地连接学科网络中的各篇文献,并起着连接桥的作用,极大地推进了学科的发展。

第5篇文献为《用于哺乳类动物细胞中小干扰RNA进行稳定表达的系统》,该文由专家BRUMMELKAMP TR发表于2002年^[17]。本文中,作者提出了一个较为完整的系统,用来指导在哺乳动物细胞中小干扰RNA的合成,并分析了多种哺乳动物细胞中基因的功能特点。该文的重要意义在于对于干扰RNA的深入研究,代表了这段时期学术界研究新热点的出现。

第6篇文献是关于RNA干扰的另一篇著作——《RNA干扰酶混合物分配中的不对称性》,它是由SCHWARZ DS于2003年发表在Cell杂志上^[18]。文章中出现了一个重要的概念“human RISC”(指的是人源混合物,由某种rna和其他分子组成)。该文提出了RNA干扰(RNAi)技术中的一个重要步骤是对RISC和protein-siRNA复合物进行组装。

最后1篇文献是2007年,由英国剑桥大学的KOUZARIDES T教授发表在cell杂志上的《核染色质的改变及其功能》一文^[19]。该文重点介绍了组蛋白在人类基因组中的重要作用,并对染色质的改变机制和组蛋白甲基化的相关特性进行了阐述。我们在分析文献的过程中发现,对组蛋白的研究成为了

当前生物学领域的热点之一。通过上述分析,笔者把生物医学领域主要理论的演进路径绘制了出来,如图6所示。

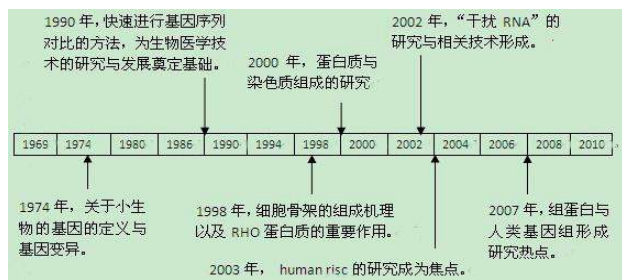


图6 七篇关键节点文献的核心观点展示的生物医学领域研究演进路径

2000年,人类基因组草图绘制完成,是生物医学领域发展的转折点,它标志着该领域从20世纪的分子生物学时代进入功能基因组时代。生物医学领域的专家罗长坤先生分析了当前生物医学的发展趋势之一为基因组学引领下的“组生物学”的发展,同时指出:与10年前相比,2010年生物医学领域基因研究的重点已经从单基因、单纯碱基测序、单蛋白质渐渐进入到在基因组和更高水平上的对多蛋白质、多基因,甚至所有基因和所有蛋白质的系统研究^[20]。这与本文得出的结果是大同小异的。

3.3 集线器和权威的节点

上文已经指出,如果一个节点指向很多良好的“权威”节点,那么它就是一个好的“集线器”,而如果它指向很多良好的“集线器”节点,那么它就是一个好的“权威”节点。我们得到了20个既是权威又是集线器的节点。结果如图7所示。

图中标号为预先处理文献记录后得到的节点标

签。在寻找这些节点的过程中,节点对象为从1998–2010年《cell》和《immunity》杂志所出版的文献,并不包括这些文献所引用了的那些文献。从图中可以看出:这些节点之间联系紧密且都具有较高的度数,这跟之前所说的“集线器”和“权威”节点指向的节点往往较多,同时这二者之间存在一种“迭代”关系是相符合的。取出这20个节点进行分析,并对照文献列表,得到每一个节点的详细信息如下(见表3)。

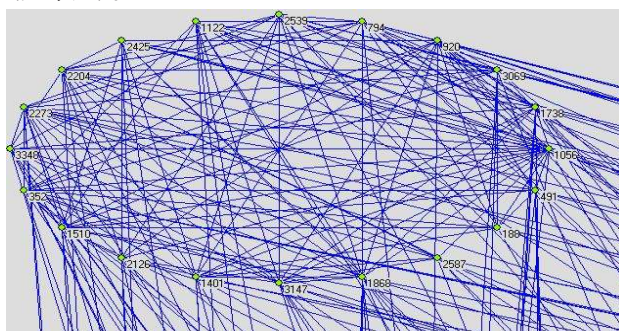


图7 生物医学领域20个权威和集线器节点文献的网络图

这20个节点组成了该领域的“集线器”和“权威”部分。结果发现,这些节点彼此之间不仅在pajek的处理结果中具有较高的联系,同时在citespace中,它们也被聚合到了一起(基本都属于第80类或者20类),可见这20个节点组成了一个较为密集的小集团。

首先,从被引频次来看,引用频次最高的是由作者ZAMORE PD于2000年发表的一文,主要对干扰RNA进行了研究。被引频次排在第二的由SCHWARZ DS发表于2003年,该文在上文得出的关键路径当中,进一步说明了它在生物医学领域中占据的重要地位。然而其他文献却没有出现在关键路

表3 权威和集线器节点的详细信息

标签	作者	年份	部分题名	被引频次
794	ZAMORE PD	2000	RNAi: Double-Stranded RNA Directs the ATP-Dependent Cleavage of mRNA	1024
1868	SCHWARZ DS	2003	Asymmetry in the assembly of the RNAi enzyme complex	919
1738	KHVOROVA A	2003	Functional siRNAs and miRNAs exhibit strand bias	797
920	GRISHOK A	2001	Genes and mechanisms related to RNA interference regulate expression	746
352	TABARA H	1999	The rde-1 gene, RNA interference, and transposon silencing in <i>C. elegans</i>	629
1401	MARTINEZ J	2002	Single-stranded antisense siRNAs guide target RNA cleavage in RNAi	595
1122	SIJEN T	2001	On the role of RNA amplification in dsRNA-triggered gene silencing	540
491	DALMAY T	2000	An RNA-dependent RNA polymerase gene in Arabidopsis	527
1056	NYKANEN A	2001	ATP requirements and small interfering RNA structure in the RNA interference	522
188	KETTING RF	1999	Mut-7 of <i>C. elegans</i> , required for transposon silencing and RNA interference	446
2126	LEE YS	2004	Distinct roles for <i>Drosophila</i> Dicer-1 and Dicer-2 in the siRNA/miRNA silencing	389
2425	GREGORY RI	2005	Human RISC couples microRNA biogenesis	316
2539	MATRANGA C	2005	Controlled confinement and release of gases in single-walled carbon nanotube	293
2587	RAND TA	2005	Argonaute2 cleaves the anti-guide strand of siRNA during RISC activation	216
1510	TABARA H	2002	The dsRNA binding protein RDE-4 interacts with RDE-1, DCR-1, and a DEXH-box	215
2204	PHAM JW	2004	A Dicer-2-dependent 80s complex cleaves targeted mRNAs during RNAi	205
2273	TOMARI Y	2004	RISC assembly defects in the <i>Drosophila</i> RNAi	177
3069	YIGIT E	2006	Analysis of the <i>C. elegans</i> Argonaute family sequentially during RNAi	127
3348	TOMARI Y	2007	Sorting of <i>Drosophila</i> small silencing RNAs	100
3147	FORSTEMANN	2007	<i>Drosophila</i> microRNAs are sorted into functionally distinct argonaute complexes	90

径当中,这并不能说明其他节点并不重要。在关键路径的选取中,采用的是中介中心度的算法来进行节点文献的筛选——被选为关键路径当中的节点,必须有很好的连接桥作用,它们组合起来能构成该学科领域主干理论的演进轨迹。而“权威”和“集线器”节点则是从另外一个方面来衡量节点文献的重要性。值得一提的是,这些文献基本集中在两大主要子类“gene”和类“human risc”当中,符合科学发展的一般规律。

其次,从这些文献的内容来看,它们较为相似,都集中在RNA的研究,并从不同角度研究了干扰RNA和human risc在基因表达中的重要作用。在上文关键路径的分析中,得到了分别发表于2002年和2003年的两篇关于干扰RNA和human risc研究的文献,在这两年附近的一段时期内,这两个话题一直是研究热点。从表格中发现,这些节点多数是发表在2002年之后,当然在此之前也有几篇。可见,这两个分析结果是前后相呼应的。

4 结果讨论

通过结合软件citespace和pajek,并借鉴已有成果以及向相关专家人士进行咨询,顺利地得到了三个方面的实验处理结果。下面对这些结果进行进一步总结与讨论。

第一、关于聚类与归类。通过两种途径的尝试,得到了效果较好的一种聚类方法,即首先便以文献为聚类对象,基于文献共被引聚类,并采用tf*idf算法选取文献标题来对聚类进行标签。由于参与聚类的节点数目大,初步处理得到了84个子类,我们选取了其中最重要的三大子类(分别是gene、human risc、cell development)来具体分析,从而对生物医学领域的知识网络提供了一个整体的把握。

第二,关于关键路径的描述。采用了中介中心度的算法,来对众多文献按照时间顺序进行筛选,从而得到了七篇最具有代表性的文献,并绘制出了生物医学领域主干理论的演进路径,同时也指出了学科发展过程中出现的研究热点。大体来看,生物医学领域的研究经过了以下几个过程:其一是在“前学科时期”,关于小生物的基因的定义与基因变异的描述,其二是生物医学技术发展,其三是对细胞骨架的组成机理以及RHO蛋白质的研究,其四是对染色体本质的探讨,其五是对“干扰RNA”与其相关生物技术的研究,其六是human risc的研究热潮,其七是对

组蛋白在人类基因组中作用的研究。

第三,关于“集线器”和“权威”节点的寻找。我们借鉴了Web超链接分析中的HITS算法来寻找知识网路中的“hubs”和“authorities”,最终得到了20篇满足这样条件的文献。一般来说,那些“权威”的节点之间本身没有很强的联系,它们往往通过“集线器”节点来进行连接。而如果这些节点都既满足“权威”的条件,又满足“集线器”的条件,那么它们往往会形成一个紧密联系的小团体,正如实验结果所示。

值得指出的是,这三个方面的分析并不是相互独立的,相反,它们之间保持着密切的联系。关键路径中的节点文献,往往归属于聚类中的主要子类,而“集线器”和“权威”节点,部分节点会是组成关键路径的一部分,且同样归属于主要子类。

5 结 语

本文重点研究了生物医学领域发展过程中的知识演化。在处理数据的工具上,利用了多种软件和计算机辅助工具来进行研究。在内容上,借鉴了国内外较为新颖的数据处理算法与研究角度来更好地分析知识的演化机理。

(1)生物医学领域是一个体系庞大的学科领域,其研究对象纷繁复杂,但总的来说,对基因表达和细胞的研究一直以来都是热点。

(2)学科领域的知识往往是沿着某条隐形的路径来进行演化的,该路径上的节点文献发挥着关键作用。

(3)知识网络中的重要节点可以为我们提供一个对该领域研究主题的直观理解。

后续的研究,将加强对学科子领域的分析,并尝试从子领域中找出关键路径,与本实验结果进行对比分析,同时从时间轴上进行深层次地挖掘,以更好地研究知识演化随时间的变化而出现的规律。

参考文献

- 1 刘 向,马费成.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.
- 2 马费成.论布鲁克斯情报学基本理论[J].情报学报,1983,2(4):318-321.
- 3 赵蓉英,邱均平.知识网络研究(I)——知识网络概念演进之探究[J].情报学报,2007,26(2):198-209.
- 4 Ed C.M. Noyons ClaraCalero-Medina.Combining mapping and citation network analysis for a better understanding of the scientific development: The case of (下转第15页)

- ments/un/unpan038851.pdf,2011-04-30.
- 6 eEurope2002 Action Plan[EB/OL].http://ec.europa.eu/information_society/eeurope/2002/action_plan/pdf/actionplan_en.pdf, 2011-04-30.
 - 7 Web Accessibility Benchmarking Cluster deliverables[EB/OL]. <http://www.wabcluster.org/deliverables.html>,2011-04-30.
 - 8 European Internet Accessibility Observatory[EB/OL].<http://eiaio.net/>,2011-04-30.
 - 9 Darrell M. West. State and federal Electronic Government in the United States, 2008. [EB/OL]. http://www.brookings.edu/~media/Files/rc/reports/2008/0826_egovernment_west/0826_egovernment_west.pdf,2011-04-29.
 - 10 Shi Y. The accessibility of Chinese local government Web sites: An exploratory study[J]. Government Information Quarterly, 2007, (24): 377-403.
 - 11 Shi Y. E-Government Web Site Accessibility in Australia and China : A Longitudinal Study[J]. Social Science Computer Review, 2006, (24): 378-385.
 - 12 Abanumy A., Al-Badi A., Mayhew P. E-Government Web-site Accessibility: In-Depth Evaluation of Saudi Arabia and Oman[J]. The Electronic Journal of e-Government, 2005,3 (3): 99-106.
 - 13 Kuzma J. M., Price C.. Analysis of UK Parliament Web Sites for Disability Accessibility [EB/OL]. [<http://eprints.worc.ac.uk/617/1/kuzmaukdisaforconferencerevised.pdf>],2011-04-19.
 - 14 J Kuzma J. M., Yen D., Oestreicher K. Global e-government Web Accessibility: An Empirical Examination of EU, Asian and African Sites[EB/OL]. http://eprints.worc.ac.uk/591/1/Global_e-government_Web_Accessibility.pdf,2011-04-19.
 - 15 Latif M. H. A., Masrek M.N. Accessibility Evaluation on Malaysian E-Government Websites[EB/OL]. <http://www.ibimapublishing.com/journals/JEGSBP/2010/935272/935272.pdf>, 2011-04-19.
 - 16 蒋淑君.网页界面设计中的残疾人用户可访问性分析[J].中国特殊教育,2004,(1):90-94.
 - 17 张翠玲.Web 站点可访问性研究概述[J].西北工业大学学报(社会科学版),2010,(3):67-70.
 - 18 郭金兰.中国政府网站可访问性现状[J].情报科学,2009,(12):1802-1805.
 - 19 Web Accessibility Checker[EB/OL].<http://achecker.ca/checker/index.php>,2011-04-27.

(责任编辑:赵立军)

(上接第 7 页)

- the absorptive capacity field[J].Journal of Informetrics, 2008, (2):272-279.
- 5 Moed HF. Citation analysis in research evaluation[M]. Dordrecht:Springer,2005:80.
 - 6 De Nooy, Mrvar, Batagelj. Exploratory social network analysis with Pajek[M].New York: Cambridge University Press,2005:334.
 - 7 Mina A, Ramlogan R, Tampubolon G, Metcalfe JS.Mapping Evolutionary trajectories: Applications to the growth and transformation of medical knowledge[J].Research Policy, 2007, (36): 789-806.
 - 8 翟伟希,陶乃航,郭吉安.社会网络分析在团体知识共享中的应用[J].图书情报工作,2010,54(4):127-130.
 - 9 吴思竹,张智雄.网络中心度计算方法研究综述[J].图书情报工作,2010,54(18): 107-110.
 - 10 Pinski G, NarinF.Citation Influence For Journal Aggregates of Scientific Publications: Theory, with Application to the Literature of Physics[J].Information Processing & Management, 1976, (12): 297-312.
 - 11 Jon M, Kleinberg.Authoritativessourcesinahyperlinked environment[J].Journal of the Association for Computing Machinery, 1999, 46(5): 604-632.
 - 12 BRENNER S. THE GENETICS OF CAENORHABDITIS ELEGANS[J].Genetics, 1974, 77(1): 71-94.
 - 13 ALTSCHUL SF. Basic local alignment tool[J].journal of molecular biology, 1990, 215(3): 403-410.
 - 14 HALL A.Rho GTPases and the actin cytoskeleton[J].science, 1998, 279(5350): 509-514.
 - 15 百度百科.RHO 蛋白[EB/OL].<http://baike.baidu.com/view/2964599.htm>,2011-07-02.
 - 16 STRAHL BD.The language of covalent histone modifications [J].nature, 2000, 403(6): 41-45.
 - 17 BRUMMELKAMP TR.A system for stable expression of short interfering RNAs in mammalian cells[J].science, 2002, 296 (5567): 550-553.
 - 18 SCHWARZ DS. Asymmetry in the assembly of the RNAi enzyme complex[J].CELL, 2003, 115(2): 199-208.
 - 19 KOUZARIDES T.Chromatin modifications and their function[J].CELL, 2007, 128(4): 693-705.
 - 20 罗长坤.当前生物医学发展趋势与特征[J].医学与哲学(人文社会医学版), 2011, 32(2): 1-4.

(责任编辑:赵立军)