DOI:10.16353/j.cnki.1000-7490.2014.07.011

● 邱均平[1,2]，李小涛[1,2]
（1. 武汉大学　信息管理学院，湖北　武汉　430072；2. 武汉大学　中国科学评价研究中心，湖北　武汉　430072）

基于引文网络挖掘和时序分析的知识扩散研究*

摘　要： 在构建国内知识图谱研究论文的引文网络的基础上，挖掘网络中每篇文献的期刊、机构、作者、关键词信息，从上述4个层面来整合、细化引文网络，并引入时间维度进行分析。研究发现国内知识图谱研究由科技管理领域扩散到图书情报领域，进而推广应用于教育学等学科领域。我国的知识图谱研究沿着理论介绍—方法研究—应用推广的路径不断发展深化。研究表明引文网络挖掘和时序分析可有效揭示知识的扩散与演进过程。

关键词： 知识图谱；引文网络；知识扩散；网络挖掘；时序分析

Abstract: Based on the construction of citation network of research papers on knowledge mapping in China, this paper extracts journals, institutions, authors and keywords of the references for each paper in network. Then the paper integrates and refines citation network from the 4 aspects mentioned above, and introduces time dimension for analysis. The research finds that the study of knowledge mapping in China has diffused from the field of scientific and technological management to the field of library and information science, and then applies in the field of education and other disciplines. The studies of knowledge mapping in China is developing and deepening along with the path of theory introduction-method research-application and extension. The research shows that citation Web mining and timing analysis can effectively reveal the diffusion and evolution process of knowledge.

Keywords: knowledge mapping; citation network; knowledge diffusion; Web mining; timing analysis

文献是知识的载体，文献引用与被引用的过程伴随着知识的扩散与转移、继承与创新。特定领域的文献及其引用关系经过一定时间的积累，就会自发形成复杂的引文关系网络，网络的结构和特征可以在一定程度上反映该领域知识的扩散情况。因此，引文网络是一种天然适合观察知识扩散的自组织网络，已有多位学者利用引文网络对特定学科领域的知识扩散进行了研究。岳洪江收集管理科学期刊间的引证数据，采用社会网络分析方法，研究了管理科学知识的流动扩散网络的结构特性[1]；赵星等人构建了我国82个文科领域的引文网络，定量刻画了我国文科领域的知识扩散情况[2]；高霞等人构建了一个科学知识扩散的网络模型，并以国外h指数的研究论文为例进行了实证研究[3]。

以往的引文网络分析的局限在于，网络中每个节点代表一篇论文，只能从论文层面来揭示知识的结构和关联，而对于每篇论文的作者、机构、发表刊物等因素未能深入揭示。各种期刊在知识扩散中发挥着怎样的作用？多少机构参与了知识的交流？哪些作者提供的知识得到了广泛的认可和吸收？知识在扩散的过程中又发生了怎样的演化？这些问题都未能得到很好的回答。若能在引文网络的基础上对上述信息进行挖掘，进而对引文网络进行整合与细化，将能更加深入和具体地反映知识的扩散和演进过程。

本文以国内知识图谱研究领域的论文为例，在构建引文网络的基础上，挖掘每个节点的期刊、作者等信息，并从多个方面对引文网络进行重构，然后引入时间维度进行分析，以探索引文网络中知识扩散和演进的具体过程。

1　数据与方法

1.1　数据来源

以南京大学中国社会科学引文数据库（CSSCI）为数据来源，选择来源文献的关键词为检索途径，用“知识图谱”作为检索词，时间范围不限，共检索到213篇研究论文，检索时间为2013年6月。将论文的标题、作者、机构、刊名、年份、参考文献等字段导入到Excel数据库中。

1.2　研究方法

1.2.1　构建引文网络　目前常见的引文网络主要有两种，一种是基于文献间的直接引用关系构建的引文网络，是一种有向网络，例如HisCite构建的引文网络，方向从被引文献指向施引文献[4]；另一种是基于文献间共被引关系构

* 本文为国家社会科学基金重大项目“基于语义的馆藏资源深度聚合与可视化展示研究”的成果，项目编号：11&ZD152。

建的引文网络，是一种无向网络，例如 Citespace 构建的引文网络，网络中引文间的连线没有方向[5]。前者是原始的引文网络，后者是在前者的基础上提炼出共被引关系后而形成，丢失了原始引文网络中的方向性，而文献引用关系的方向反映着知识流动的方向。本文关注的是知识的扩散和演进过程，因此需要构建原始引文网络。

以往的研究多是借助 HistCite 软件来构建引文网络。李运景等人利用 HistCite 构建了杂交水稻研究文献的引文编年图，梳理出杂交水稻研究的发展规律，证明了利用引文时序网络挖掘专业学科发展规律的可靠性和准确性[6]；董克、韩毅等人采用 HisCite 分别绘制了网络计量学发展的引文编年图和知识管理引文网络的主路径[7-8]。也有学者自行收集引文数据构建中文文献的引文网络，刘蓓等人构建了我国情报学领域研究论文的引文网络，从网络结构的角度分析了国内情报学研究的特点[9]。上述研究对引文网络的构建与分析均停留在文献或作者层面，尚未见有论文在构建引文网络的基础上，从刊名、机构、作者和关键词等多个角度对引文网络进行深入挖掘，并引入时间维度进行分析。

本研究将知识图谱研究论文间的相互引用视为该领域知识扩散的主要渠道，只考虑 213 篇论文间的相互引用关系。经过数据清理，对每篇论文的参考文献进行核查，对于同一篇引文的不同标注格式进行合并，从 213 篇知识图谱研究论文中得到 1871 条引文。将所有论文与引文进行比对，发现 213 篇论文中共有 104 篇论文相互之间存在着引用或被引用关系。依据这 104 篇文献的引用关系构建引文网络，用节点来表示文献，节点之间的箭头表示论文间的引用关系，箭头方向从被引文献指向施引文献。

Gephi 是由法国 M. Bastian 等人合作研发的一款网络分析领域的数据可视化软件[10]，对社会网络有很好的可视化效果，本研究采用 Gephi 对引文网络进行可视化展示。

1.2.2 引文网络挖掘 在引文网络的基础上，提取网络中每篇文献的详细信息，分别以刊名、第一作者机构、第一作者和关键词作为节点，节点间的弧（箭头）依据论文引用关系进行累加合并，并使节点在网络中按时间顺序进行排列，依次构建期刊时序网络、机构时序网络、作者时序网络和关键词时序网络，以探索知识图谱研究在上述层面的扩散和演进过程。

1）期刊时序网络。将引文网络中的文献按期刊进行合并，形成反映期刊间引用关系的期刊时序网络。属于同一期刊的文献合并为 1 个点，同一期刊的文献被引次数累加即为该刊被引次数，节点大小与期刊被引次数成正比；将文献间的引用关系按期刊进行累加，计算出期刊间的引用关系，引用次数越多，期刊间的弧（箭头）越粗。提取每个期刊刊载第一篇知识图谱研究论文的时间，将各期刊节点按该时间排列。期刊时序网络可以反映知识在期刊间的传播过程，节点大小（被引频次）和在网络中的位置可以共同反映期刊在知识网络中的重要性。

2）机构时序网络。将引文网络中的文献按第一作者所属机构进行合并，形成反映机构间引用关系的机构时序网络，节点和连线的处理方法与期刊时序网络类似。提取每个机构首次发表知识图谱研究论文的时间，将各机构节点按该时间排列，可以直观地显示机构在知识网络中的重要性。

3）作者时序网络。将引文网络中的文献按第一作者进行合并，形成反映作者间引用关系的作者时序网络，节点和连线的处理方法与期刊时序网络类似。提取每个作者首次发表知识图谱研究论文的时间，将各作者节点按该时间排列，可以直观地显示作者在知识网络中的重要性。

4）关键词时序网络。统计引文网络中所有文献的关键词频次，分析词频≥2 的关键词间的共现关系，构建关键词共现网络。将关键词按其首次出现的年份进行排列，形成关键词时序网络，可以反映知识图谱领域的研究热点

图 1 知识图谱研究论文的引文网络

和研究内容的演变。

期刊、机构、作者都是知识的载体，关键词可看作具体的知识单元，引入时间维度后，从这4个层面对引文网络进行挖掘和分析，可以从宏观到微观深入揭示知识图谱研究的知识网络扩散与演进。各个网络中的节点大小及其在网络中的位置可以共同反映该节点在知识网络中的重要性。Pajek是大型复杂网络分析的有力工具[11]，在国内外社会网络研究中得到了广泛应用，因此本研究采用Pajek来构建时序网络。

2 结果与分析

2.1 原始引文网络

图1是国内知识图谱研究论文的引文网络。每个节点用文献的第一作者、刊名和发表年份进行标识，每个节点的大小与出度（被引频次）成正比，节点间的箭头从被引文献指向施引文献。图1中显示出了一大一小两个子网，大的子网由102个节点相连构成，是引文网络的主体部分；小的子网只由2个节点构成，位于图的左下方。图1中共包含104个节点（文献），151条边（引用关系）。图的密度（网络中实际存在的边数与最大可能的边数之比）为0.014，说明节点间联系不够密切，文献之间引用关系非常松散。

由于引文网络是有向网络，节点的点度中心度包括出度和入度两种，入度为引用其他论文的频次，出度即为该论文的被引频次。表1为引文网络中出度最大的8篇论文，其中5篇均来自《科学学研究》，可见该刊在知识图谱研究领域的重要地位。陈悦的《悄然兴起的科学知识图谱》被引27次，是目前国内论文中被引次数最多的论文，另外7篇论文的出度从5到11不等。这些论文在引文网络中起到了知识枢纽的作用，为其他论文提供了知识来源和研究基础。需要说明的是，本文的被引频次仅指被213篇知识图谱研究论文引用的频次，被其他论文引用的频次未统计在内。

2.2 期刊时序网络

提取图1所示引文网络中每篇论文所发表的期刊，将属于同一期刊的引文合并成一个节点，用刊名对节点进行标识，节点大小与该期刊所有知识图谱研究论文的累计被引频次成正比，箭头从被引期刊指向施引期刊，箭头粗细与期刊被另一期刊引用的频次成正比，按期刊首次刊载知识图谱研究论文的年份进行排列，生成期刊时序网络（见图2），共包含39种期刊。

表1 引文网络中出度（被引频次）>4的论文

第一作者	论文标题	刊名	年份	出度
陈悦	悄然兴起的科学知识图谱	科学学研究	2005	27
侯海燕	当代国际科学学研究热点演进趋势知识图谱	科研管理	2006	11
秦长江	知识图谱——信息管理与知识管理的新领域	大学图书馆学报	2009	8
陈 悦	科学知识图谱的发展历程	科学学研究	2008	6
栾春娟	国际科技政策研究热点与前沿的可视化分析	科学学研究	2009	6
刘则渊	现代工程前沿图谱与中国自主创新策略	科学学研究	2007	5
赵玉鹏	基于知识图谱的美国《科学哲学》研究前沿和热点探讨	科学学研究	2008	5
侯海燕	基于知识图谱的国际科学计量学研究前沿计量分析	科研管理	2009	5

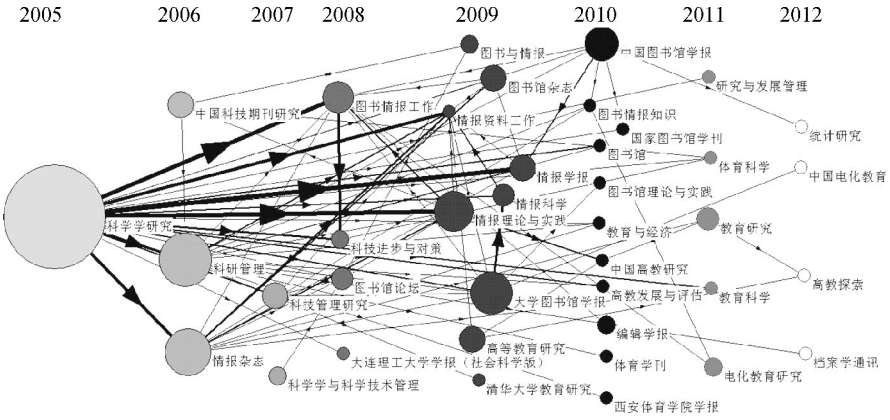

图2 期刊时序网络

从期刊所属学科领域和研究内容来看，可将期刊分为3种类型：①科技管理类。以《科学学研究》、《科研管理》、《科技管理研究》等期刊为代表，其研究内容侧重于知识图谱在科学学、科技政策、科技管理等方面的应用。这类期刊最早关注知识图谱研究，为知识图谱研究在我国的兴起奠定了基础。②图书情报类。以《情报杂志》、《大学图书馆学报》、《情报理论与实践》等期刊为代表，其研究内容侧重于知识图谱的理论、方法与应用研究。这类期刊数量众多，是当前知识图谱研究的主要阵地。③综合应用类。以《高等教育研究》、《教育与经济》、《体育学刊》等期刊为代表，其研究内容侧重于知识图谱在教育学、体育学、统计学等其他学科的推广

应用。

从被引频次来看，节点较大的期刊主要来自科技管理类和图书情报类，可能是由于这两类期刊的知识图谱研究起步早，且大多对知识图谱的理论与方法进行了创新，故被引频次高；教育学、体育学的期刊的知识图谱研究起步较晚，且主要是应用性的论文，创新性不及前二者，给其他研究者提供的参考价值相对较低，因此被引较少。

从发文时间来看，《科学学研究》是图 2 中最早出现的节点，该刊是从 2005 年开始刊载知识图谱研究论文，该刊的论文得到了国内同行的广泛引用。2006 年，知识图谱研究扩散到《科研管理》、《情报杂志》和《中国科技期刊研究》，2007 和 2008 年又陆续有几种科技管理类和图书情报类的期刊参与进来。2009 年和 2010 年是知识图谱研究在国内快速扩散的阶段，大量期刊开始发表知识图谱的研究论文，尤其许多图书情报类期刊在这一阶段开始涌现，如《情报理论与实践》、《大学图书馆学报》、《中国图书馆学报》等，这些期刊的论文为之后的知识图谱研究提供了重要基础，得到了广泛引用和关注。2011 年和 2012 年新出现的期刊大部分是教育学领域的期刊，另外还有个别统计学、体育学的期刊，说明知识图谱研究开始在教育学等领域得到广泛应用。

整体来看，科技管理类期刊是国内知识图谱研究的先驱，率先引进了国外的知识图谱研究；图书情报类期刊是国内知识图谱研究的主体，推动了知识图谱理论、方法和应用研究的发展和深化；教育学、体育学、统计学等领域的期刊是知识图谱研究的推广，标志着知识图谱研究逐渐发展成熟，在其他学科领域得到了广泛的认可和应用。知识图谱研究中知识扩散的方向是从科技管理类期刊流向图书情报类期刊，然后再扩散到教育学、体育学等领域的期刊，在扩散的过程中又不断创新、发展和深化。

2.3 机构时序网络

提取引文网络中每篇论文的第一作者所属机构，将属于同一机构的文献合并成一个节点，用机构名称对节点进行标识，节点大小与该机构的论文总被引频次成正比，箭头从被引机构指向施引机构，箭头粗细与机构被另一机构引用的频次成正比，按机构首次发表知识图谱研究论文的年份进行排列，生成机构时序网络（见图 3）。共包含 39 个第一作者机构。

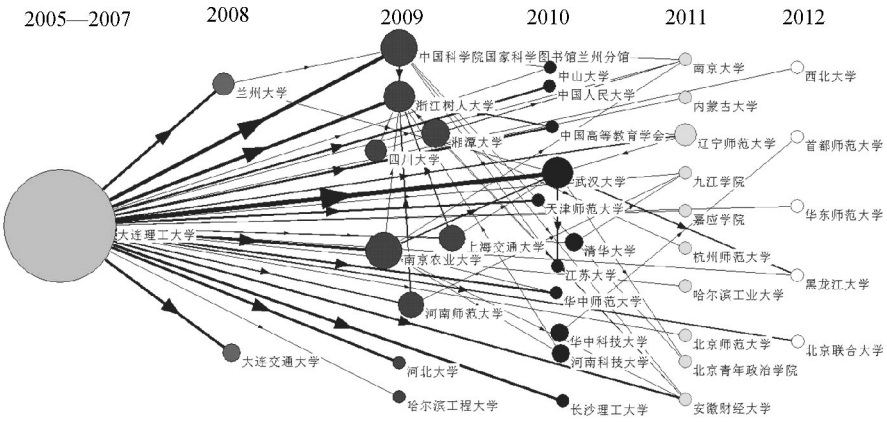

图 3 第一作者机构时序网络

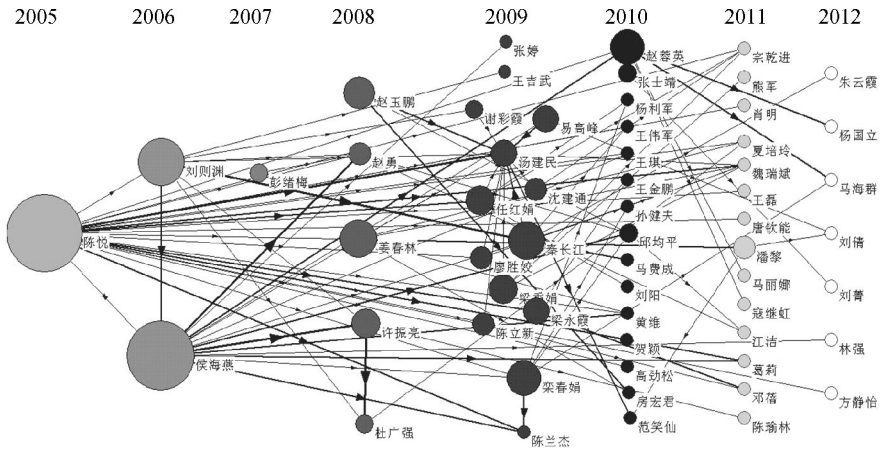

图 4 第一作者时序网络

从图 3 可以看到机构间的知识扩散现象。知识图谱研究以大连理工大学为核心直接辐射到另外 29 个机构，74.36% 的机构引用了该机构的论文。2005—2007 年全国只有大连理工大学发表了知识图谱的研究论文，2008 年兰州大学和大连交通大学开始进行知识图谱研究并引用了大连理工大学的研究成果。2009—2010 年众多机构开始发表知识图谱研究论文，如浙江树人大学、湘潭大学、南京农业大学、武汉大学等，在继承之前研究成果的基础上进行拓展创新，同样得到了众多机构的引用，成为了知识扩散网络中的知识枢纽。2011 年和 2012 年仍有不少机构开始发表知识图谱的研究成果，但它们由于起步较晚，发表论文的时间较短，较少被其他机构引用，目前处于知识扩散网络的末端或外围位置。

2.4 作者时序网络

提取引文网络中每篇论文的第一作者，将属于同一作者的文献

合并成一个节点，用作者姓名对节点进行标识，节点大小与作者累计被引频次成正比，箭头从被引作者指向施引作者，箭头粗细与作者被另一作者的引用频次成正比，按作者首次发表知识图谱研究论文的年份进行排列，生成作者时序网络（见图4），共包含59位第一作者。

从图4中可以看到作者间的知识扩散现象，知识图谱研究以陈悦、侯海燕、刘则渊等人为核心辐射到另外32位作者，被辐射作者人数占作者总数的54.24%，可见这3位作者在知识图谱研究中的重要地位。这3位作者内部也存在着知识的扩散与交流：侯海燕和陈悦互相都引用了对方的论文，因此他们之间有两个箭头方向。

2005—2007年知识图谱研究论文的第一作者只有4位：陈悦、侯海燕、刘则渊和彭绪梅，而且这4位作者均来自大连理工大学。2008年新增了5位第一作者：大连理工大学的许振亮、姜春林、赵玉鹏，兰州大学的赵勇和大连交通大学的杜广强。2009年大量有影响力的作者开始涌现，如秦长江、梁永霞、栾春娟等人，他们的研究成果同样被广泛引用，成为了知识网络中的知识枢纽。2010年新出现的作者人数最多，但只有赵蓉英等人成为了较为突出的知识枢纽。2011—2012年仍有较多的作者进入知识图谱研究领域，但由于起步较晚，主要是继承和吸收其他作者的知识，较少作为知识来源被引用，因此处于知识扩散网络的外围或末端。

2.5 关键词时序网络

提取引文网络中104篇文献的关键词，统计关键词词频，词频≥2的关键词共有59个，且它们均在2012年之前就已开始出现。对这59个词进行共现分析并绘制共词网络：每个节点表示一个关键词，节点大小与词频成正比，节点间的连线粗细与关键词共现频次成正比，将关键词按该词首次出现的年份进行排列，见图5。

关键词能反映论文研究的主要内容，从图5中可以看到知识图谱研究内容的扩散和演进。整体来看，出现频次最高的几个词间存在着密切的联系：知识图谱、研究热点、研究前沿、可视化，这些词可以反映我国知识图谱研究的核心——采用科学计量学的方法，对特定学科或主题领域的研究热点和前沿进行可视化分析。各年新出现的关键词体现了知识图谱研究的发展演进：2005—2006年新出现的关键词，如科学知识图谱、科学计量学、知识图谱、信息可视化等，侧重于理论研究，主要是对知识图谱的基本概念的介绍和应用的展望；2007—2009年新出现的关键词，如多元统计、因子分析、聚类分析、文献计量分析、共词分析、社会网络分析等，侧重于知识图谱绘制的方法探索和实证研究；2010—2011年新出现的关键词，如生命周期理论、知识计量学、概念图、教育经济学、体育科学等，侧重于新理论的引入和知识图谱在其他学科领域的推广应用。可见我国的知识图谱研究近年来大致上沿着理论介绍—方法研究—应用推广的路径不断发展深化。

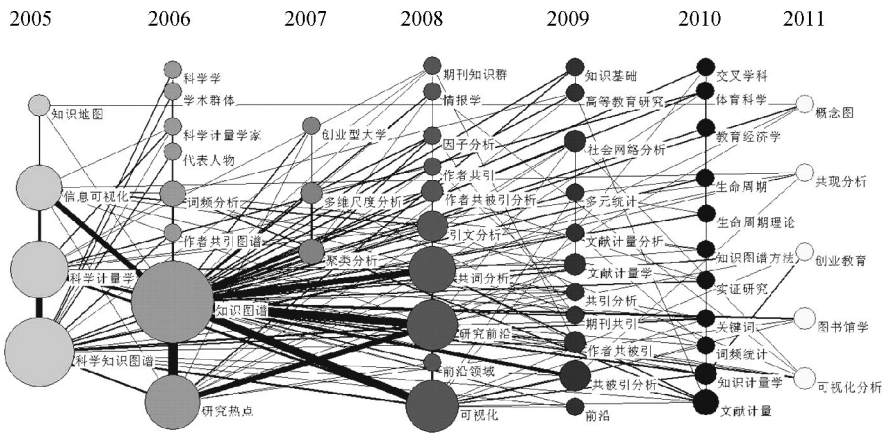

图5 关键词共现时序网络

3 分析

本文在引文网络的基础上构建了期刊时序网络、机构时序网络、作者时序网络和关键词时序网络，分别从期刊、机构、作者和关键词4个层面揭示了知识图谱研究的扩散演进过程。研究发现科技管理类期刊是国内知识图谱研究的先驱，图书情报类期刊是国内知识图谱研究的主体，目前知识图谱研究正向教育学、体育学、统计学等学科领域扩散；大连理工大学在国内率先进行知识图谱研究，其研究成果被其他机构广泛吸收和借鉴；2009—2010年以浙江树人大学、湘潭大学、南京农业大学、武汉大学等为代表的众多机构开始发表知识图谱研究论文；我国的知识图谱研究近年来大致上沿着理论介绍—方法研究—应用推广的路径不断发展深化。上述结论与杨思洛等人对知识图谱研究现状及趋势的描述相近[12]，表明引文网络挖掘结合时序分析，能有效地揭示知识的扩散与演进过程。

本研究相较以往的研究有两个特点：一是根据引文网络来提取知识单元和知识载体（关键词、作者、机构、期刊），以它们的相互引用关系构建知识网络，而且给知识网络中的每个节点都赋予了权重，以体现其在知识网络中的不同状态，能较好地反映知识结构和知识载体的分布；二是引入了时间维度从多个层面进行时序分析，可以反映知识扩散和演

进的发展过程。引文网络挖掘和时序分析相结合的方法有较好的创新性和应用价值。

1）为绘制知识图谱提供新方法，从期刊、机构等多个层面直观地展示特定研究领域的知识结构、分布及演进过程，有助于了解学术研究的发展动态和发现重要的知识节点。

2）为研究知识扩散和演进提供新视角，有助于了解网络节点之间知识的交流、传播情况，发现可能推动或阻滞知识传播的交互关系，提高对于知识网络发展的干预能力，从而促进知识管理、服务与利用。

3）为学术评价提供新思路，论文、作者、期刊、机构在知识网络中的位置可以直观反映学术水平和影响力，可以作为学术评价的依据之一，与被引频次及其他指标相结合进行综合评价可使结果更为准确。□

参考文献

[1] 岳洪江. 管理科学知识扩散网络的结构研究 [J]. 科学学研究，2008（4）：779-786.

[2] 赵星，谭旻，余小萍，等. 我国文科领域知识扩散之引文网络探析 [J]. 中国图书馆学报，2012（5）：59-67.

[3] 高霞，陈凯华，官建成. 科学知识扩散的网络模型 [J]. 研究与发展管理，2013，25（2）：45-54.

[4] 李运景，侯汉清. 引文编年可视化软件 HistCite 介绍与评价 [J]. 图书情报工作，2007（12）：135-138.

[5] 陈超美. CiteSpace Ⅱ：科学文献中新趋势与新动态的识别与可视化 [J]. 陈悦，等，译. 情报学报，2009，28（3）：401-421.

[6] 李运景，任银玲，何琳，等. 利用引文时序可视化挖掘专业学科发展规律 [J]. 情报学报，2010，29（5）：880-888.

[7] 董克，刘德洪，江洪，等. 基于主路径分析的结果改进研究 [J]. 情报理论与实践，2011，34（3）：113-116.

[8] 韩毅. 引文网络主路径的结构洞功能探索——以知识管理领域为例 [J]. 图书情报工作，2012，56（24）：65-70.

[9] 刘蓓，袁毅，ERIC B. 社会网络分析法在论文合作网中的应用研究 [J]. 情报学报，2007，27（3）：407-417.

[10] BASTIAM M，HEYMANN S，JACOMY M. Gephi：an open source software for exploring and manipulating networks [C] // International AAAI Conference on Weblogs and Social Media，2009：361-362.

[11] BATAGELJ V，MRVAR A. Pajek：a program for large network analysisp [J]. Connections，1998，21（2）：47-57.

[12] 杨思洛，韩瑞珍. 知识图谱研究现状及趋势的可视化分析 [J]. 情报资料工作，2012（4）：22-28.

作者简介：邱均平，男，1947 年生，教授，博士生导师。
李小涛，男，1986 年生，博士生。

收稿日期：2013－11－25

（上接第 4 页）

系，及时监控并反映我国科技系统以及相关系统安全运行的状况，捕捉各种危险因素所释放的预警信号，便于及时调控。□

参考文献

[1] 卢国琪. 对维护科技安全的思考 [J]. 今日论坛，2008（15）：106.

[2] 赵刚. 地缘科技视角下的国家科技安全研究 [D]. 武汉：华中科技大学，2007：142-143.

[3] 潘正祥，杨迎会. 全球化与国家科技安全 [J]. 中国科技论坛，2007（6）：19-23.

[4] 杨名刚. 论国家科技安全诉求的现实困境与出路 [J]. 学术交流，2011（9）：95-98.

[5] 解析创造力 [EB/OL]. [2014-03-28]. http://hi.baidu.com/cold_adonis/item/c85d4f1be29bf3406926bba4.

[6] 张殿清. 情报与反情报 [M]. 北京：世界知识出版社，1997：109.

[7] 杨春平. 国家技术安全体系研究 [D]. 大连：大连理工大学，2007：4.

[8] 沈固朝. 将情报思维纳入保密意识中 [J]. 保密工作，2011（5）：32-34.

[9] 王沙骋. 保密工作中的反情报思维 [J]. 保密工作，2012（3）：40-41.

[10] 高金虎. 试论反情报 [J]. 保密科学技术，2013（9）：22-27.

[11] MILLER J P. Millennium intelligence：understanding and conducting competitive intelligence in the digital age [M]. Medford：CyberAge Books，2000：27-30.

[12] 詹欣. 美国情报部门对中国军事的评估 [D]. 广州：华东师范大学，2007：4-10.

[13] 曾原. 从引文网络中发现美国精英对华研究的信息源规律——图情思路及方法在中美关系领域应用的尝试 [J]. 图书情报工作，2009，53（14）：10-14.

[14] 赵文绮，陈广玉. 美国对华研究引文数据库系统的建立及应用 [J]. 图书情报工作，2009，53（14）：15-19.

[15] 郑刚. 刍议反情报工作中的监测、分析与追查方法 [EB/OL]. [2014-03-28]. http://www.docin.com/p-538484281.html.

[16] 曹平. 技术创新理论模型的多维解读 [J]. 技术经济与管理研究，2010（4）：33-36.

作者简介：胡雅萍，女，1986 年生，博士生。研究方向：情报分析，科技情报。
李骁，男，1990 年生，硕士生。研究方向：情报分析。

收稿日期：2014－03－13