

专利研究文献的可视化分析*

Analysis through Visualizing the Patent Research Articles

高继平 丁堃

(大连理工大学 21 世纪发展研究中心暨 WISE 实验室 大连 116024)

摘要 伴随全民创新意识的增强,专利作为跟踪高新技术发展以及社会进步的重要信息源已被众多学者所认同,但是以往对专利的分析研究多聚焦于专利申请量的变化或具体专利的研究,而缺少从研究专利的文献中挖掘一些重要的信息。针对这一不足,本文将以 SCI-E 数据库中检索到的专利研究文献作为研究对象,借助分析工具 CiteSpaceII,通过文献共引和共词分析的方法,探讨专利研究文献中潜在的一些知识,同时对未来专利研究的热点做了一定的预测。

关键词 专利研究 CiteSpaceII 文献共引 共词分析 隐性知识

中图分类号 G353

文献标识码 A

文章编号 1002-1965(2009)07-0012-05

随着社会的进步、国家的发展,作为跟踪高新技术发展以及社会进步的重要信息源——专利,越来越受到专家们的重视。故而专利的研究和分析也逐渐升温,从开始专利的研究,多关注于专利拥有者、具体专利涉及的技术等,到开始关注专利的深层意义,即它与科学、技术、创新等的内在联系,再到以专利作为衡量经济的指标^[1]等。当前聚类、关联规则挖掘等文本挖掘的方法得到了广泛的研究应用,丁堃等人利用 EM 文本聚类算法和隐马尔科夫链模型对我国知识管理领域的热点发展趋势做了一定的预测^[2];林鸿飞等人利用文本挖掘中的有序聚类方法对搜索引擎的发展阶段进行了分析^[3];朱东华等人利用文本挖掘的技术预测产品技术成熟度^[4]等。

随着专利的重要性得到广泛认识,专利研究的相关文献也在迅速增长,而文献数据库中的数据也愈来愈庞大复杂,这就为专利研究的科研工作者和政策制定者如何从浩瀚的专利研究文献中获得相关领域的核心文献和如何紧跟专利研究的热点提出了一定的挑战。针对上面这些问题,本文提出下面的分析方法,以期专利分析研究者提供一定的裨益。

1 专利研究文献分析

1.1 研究数据和分析工具 在 Web of Science 中进行主题检索(“patent analysis”OR“patent process*”OR

“patent research*” OR“patent min*”),检索时间段为 1994 年至 2008 年 1 月 1 日,检索的数据库是 SCI-E (Science Citation Index Expanded),文档类型为“Article”,一共获得 923 条文献记录。之后,鉴于“patent”在医学领域中是一个重要的词,进一步经过去噪处理,最终获得 763 篇文献。

信息可视化是将抽象数据用可视的形式表示出来,以利于分析数据、发现规律和支持决策等。其中引文分析可视化是信息可视化的一个重要分支,其首先处理海量的引文数据,之后利用信息可视化技术使人们更容易地观察、浏览和理解信息,进而找到数据中隐藏的规律和模式^[5]。不过当前对此分析的作者,多应用统计学中的一些工具,如 SPSS、Pajek 等,但是其可视化的效果不仅单调,而且分析解读比较烦琐。本文使用的引文分析可视化工具是基于 JAVA 平台的 CiteSpaceII 应用软件。它是一种适于多元、分时、动态的复杂网络分析的新一代信息可视化技术^[6],同时具有监测科学文献中出现的热点和研究演化的功能。该软件通过 3 个不同的时段(开始、中端和终端)的 3 个阈值(c, cc, ccv)确定可视化的节点量、节点间的连线量和限定节点间连线长度的一个系数值(见图左上端),而具体阈值的确定则需要根据具体分析时间段内的数据量和作者的需要加以确定。

1.2 实验结果与分析

收稿日期:2009-02-23

修回日期:2009-04-20

基金项目:国家社会科学基金项目“学科知识测度体系及其应用研究”(编号:08BTQ025)。

作者简介:高继平,男,1983 年生,硕士研究生,研究方向为专利计量;丁堃,女,1962 年生,博士,教授,研究方向为学科知识计量与创新管理。

1.2.1 核心文献分析。下面本文通过文献共被引网络分析的方法,分析当前专利研究所涉及的主要文献,继而应用中介中心度和文献引文量两种衡量标准确定了专利研究中的核心文献。其中视图1中左上角所示的则是 CiteSpaceII 运行时的一些基本信息,分别是当前运行的 CiteSpaceII 版本 2.1(V. 2.1)、运行时间 2007 年 12 月 22 日下午 2 点、运行数据保存目录(C:\citespace\实验\patent1\data\patent)、所处理的数据时间范围(1994-2007)、时间段的划分标准(2 年一个时间段,即 Slice Length=2)、网络图输出选择的阈值(3,2,20)、(4,3,20)和(5,3,20)和输出网络图中所涵盖的节点数和连线数(N=34,E=51)。

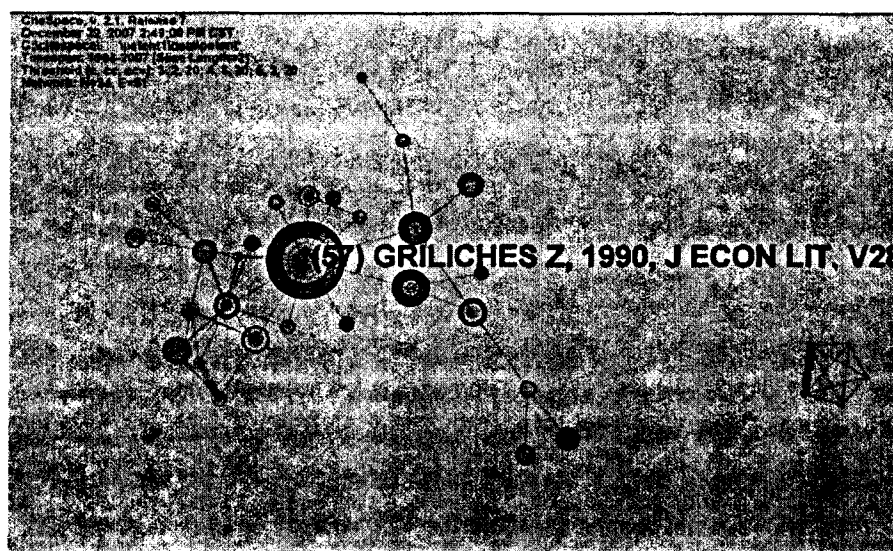


图1 文献共引网络图谱

此文献共引网络图谱中有 34 篇文献和 51 条文献共引连线。灰色圆圈所圈的节点是中介中心度(Betweenness Centrality)较高的节点(鉴于稿件出版时是黑白色,故其在期刊中的紫色是不可见的)。中介中心度主要是用来度量网络中节点的重要程度的,其直观含义是指该节点对网络图中两两节点连通的重要程度,具体的计算方法是通过计算网络中所有两两节点之间的最短路径经过该节点的概率来计算的^[6]。

表1 文献以中介中心度排序列表

中介中心度	被引次数	文献作者	出版时间
1.02	57	GRILICHES Z	1990
0.44	23	MEYER M	2000
0.30	21	JAFFE AB	1993
0.30	21	BASBERG BL	1987
0.25	29	NARIN F	1976
0.24	17	SCHMOCH U	1993
0.20	12	GIBBONS M	1994
0.11	22	NARIN F	1987
0.10	10	JAFFE AB	2002

如表1所示,中介中心度最高的文献是 Griliches 于 1990 年在 Journal of Economic Literature 上发表的

Patent Statistics as Economic Indicators: A survey。该文献主要是强调了专利统计数据的重要性,指出将其应用于经济绩效的评价中^[1]。之后中介中心度较高的其它几个关键点所代表的研究文献,进一步深化了对专利的研究。例如:中介中心度为 0.25 的文献 The Increasing Linkage between US Technology and Public Science 是 Narin 在 1976 年发表于 Research Policy 上的一篇专利研究文献,它强调了通过跟踪快速增长的科学文献与美国专利数据之间的引用情况,来进一步证实公共科学对工业技术发展所做出的贡献^[7]。而后中心度为 0.10 的是 2002 年 Jaffe 出版的一本书,尽管其在此 763 篇专利研究文献中共被引用了 10 次,但其在

Google Scholar 中的引用量却高达 286 次。此书不仅将专利、创新和技术变革有机地联系了起来,同时提出将专利文件(patent document)中所引用的参考文献(包括专利文件和科学文献)作为技术革新的源头和创新型经济研究的强有力的工具^[8]。

这些关键节点所代表的文献,其重要性并不在于它的引用量很高,而在于它揭示了与专利研究密切相关的一些研究主题,例如:与科学、经济、创新、研发等的联系。

伴随着科学技术的进步,大量科学家的研究成果总是建立于其他同行已做的工作之上的,而一篇文献被领域内同行所引用的数量,不仅可以体现该作者在该领域的认可程度,又可充分体现出该文献在该领域的重要性的影响力(质量)。故针对专利文献现在的研究进展,有必要对专利研究工作中有影响力的文献进行一定的分析。

表2 文献以被引次数排序列表

被引次数	中介中心度	作者	出版时间
57	1.02	GRILICHES Z	1990
29	0.25	NARIN F	1997
23	0	ALBERT MB	1991
23	0.44	MEYER M	2000
22	0.11	NARIN F	1987
21	0.3	JAFFE AB	1993
21	0.3	BASBERG BL	1987
21	0	TRAJTENBERG M	1990
18	0	COHEN WM	1990
18	0	LEVIN RC	1987

如上表2所示,影响力最大的文献依然是在表1中提到的 Griliches 于 1990 年在 Journal of Economic

Literature 上发表的 Patent Statistics as Economic Indicators: A survey。进一步通过 Google Scholar 检索,发现它的被引量高达 1442 次,这进一步证实了它在专利研究分析中的重要作用。排在第三位的是 Albert 于 1991 年发表的 Direct Validation of Citation Counts as Indicators of Industrially Important Patents,它从多个角度指出了将专利引用量作为行业重要专利判别指标的直接有效性^[9]。此表中尤其值得关注的一篇文章是被引次数为 21 的 Trajtenberg 的 Geographic Localization of Knowledge Spillovers as Evidenced by Patent Citation,该作者于 1990 年在此文献中提出了“引用次数”(citations)的概念,就像商品有优劣之分一样,不同的专利之间也有质量的差别。那么,质量是靠什么来衡量的呢?即引用次数^[10],同时该文献进一步将专利与知识溢出(knowledge spillover)联系了起来,使得知识溢出一个抽象的概念,得以通过专利引用加以量化判别分析。同时此文献在 Google Scholar 中的被引量也高达 2 000 次,该文献为后来研究知识溢出提供了独到的见解。

1.2.2 专利研究热点变迁。科学文献在发表之后,随着时间的推移,相对于科学技术的迅猛发展,其内容会逐渐变得陈旧过时,而研究内容的陈旧过时,具体体现在代表该研究内容的词汇或短语出现的次数的变化。本文采用具体文献中的关键词表征该文献的研究内容,同时鉴于 SCI-E 数据库的特殊之处,下面所提到的关键词俱来源于该数据库中的两部分 identifier(标识符)和 descriptor(描述符)。针对 CiteSpaceII 软件所具有的时区可视化法,下面将跟踪专利研究自 1994 年以来各个时间段的研究热点,并且结合其自带的凸显检测算法(burst detection algorithm),综合判断当前专利研究的前沿。

如表 3 所示,与专利最密切相关的是知识产权(intellectual property)。知识产权即政府授予知识创新者在一定期限内的一种排他性的权利,其目的是为创新活动提供足够的激励。众所周知,知识产权主要包括专利权、商标权、版权和商业秘密等,其中专利权是知识产权的重要组成部分。排在其后边的有专利申请、知识产权法、专利保护、专利数据等等与专利研究

相关的词项。

表 3 热点分析以词频列表

频数	对应词汇	凸显值	首次出现时间
35	intellectual property		1996
30	patent applications		1994
25	intellectual property rights		1994
21	patent protection		1996
20	patent data		1998
19	patent documents	4.1	2006
18	patent system		1996
16	patent analysis		1996
14	patent citations		2000
14	patent law		1996
13	patent claims		2002
13	patent information		1998
13	patent application		1998
12	trademark office		2004
11	patent office		2000

表 3 中,最值得关注的一个词是凸显值为 4.1 的专利文件(patent document)。所谓的凸显词项指的是—些在过去的一段时间里突然激增凸显出来的词或短语^[6]。通过查找数据库中涉及到专利文件的 19 篇文献,发现它所涉及的是专利研究分析的研究对象。

基于此,下面将通过一幅时区视图(time-zone)来跟踪专利研究文献中研究对象的变迁线路。

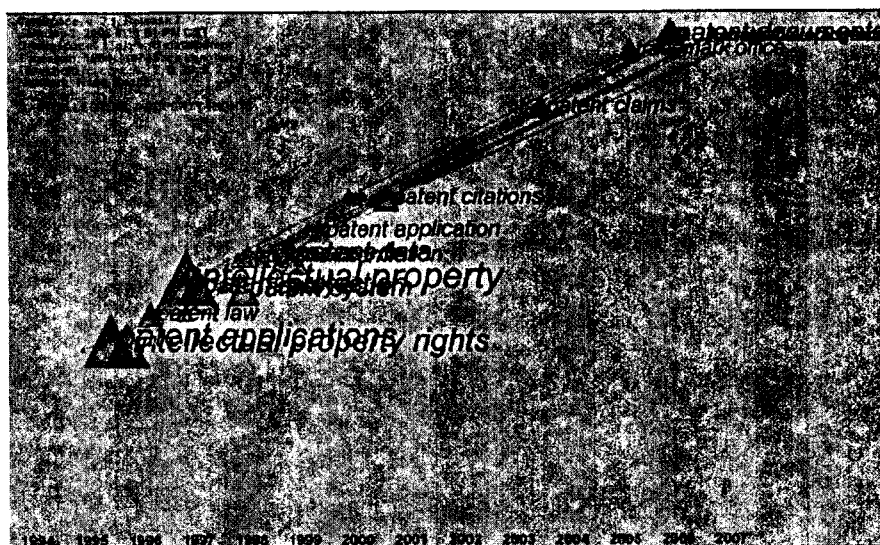


图 2 研究热点时区视图

如图 2 所示,专利研究文献中专利研究对象的變化是下面这么一个顺序。开始专利研究多以知识产权(intellectual property)以及具体专利的申请(patent application)等作为研究对象,之后开始关注于专利数据(patent data),开始以专利的量变作为研究切入点。主要是通过企业、地区间专利的量的比较,来分析企业、地区间科学技术水平的差距。之后,开始用专利引用(patent citations)来探讨科学与技术之间的联系以及核心专利的判别等。例如,Tijssen, RJW 等发表于 2000 年的文献,是以专利对科学文献的引用作为研究对象,

通过引用量来确定和比较科学技术知识对技术发展的贡献。该文献同时也检验了通过这种量化的方法来分析科学与技术之间相互作用的可行性和有效性,并且也进一步将这种计量方法引申到判别发达工业国家的技术优势和技术劣势中,例如:荷兰的技术优劣势^[11]。

之后在2003年,专利研究者开始关注于专利申请书或授权书中必不可少的权利要求书(patent claims)的重要性。这是由于专利的权利要求书是专利申请书或授权书中最核心的部分,是在整体上反映该发明的主要技术内容的文档,同时包含了该项专利的全部的必要技术特征,且清楚、完整地说明了这些特征所具有的功能及其相互之间的作用关系。Myers通过对早期激光器专利的权利要求书的仔细分析,得以推断出激光器方面的成果主要应该归功于 Townes, Shawlow 和 Bloembergen 三人^[12]。

而到了2006年、2007年,专利研究展示出一种全新的面貌,同时它的研究手法也有了新的突破。那就是词汇频数列表3中的凸显词汇“专利文件(patent document)”所示的,它表明专利研究进入了一个全新的时期。

1.2.3 专利研究新方法。专利文件中含有大量关于此发明的相关文字描述,例如:专利摘要(abstract)、权利要求书(claims)、专利描述(description)等等。而伴随着现代信息检索、文本挖掘以及人工智能等技术的发展,计算机对大规模英文文本的技术处理已经成熟。这样将这些先进的计算机技术应用于专利文件中的文本分析和文本挖掘中,就会成为一种必然和趋势。

在这次检索到的763篇专利研究文献中,共有9篇此类文献。而它们的侧重点和研究手段有所不同。其中第一篇文献是 Callaert, J 所作的 Traces of Prior Art: An analysis of non-patent references found in patent documents。文中特别强调了专利文件中的非专利引文的重要性,同时指出非专利引文中的绝大部分是期刊文献引用。因为这些期刊文献的被引用,为解答长期困扰研究者的疑难问题,即科学水平的提高到底在技术能力增强中起了多大的作用,提供了充足的信息^[13]。

同时也有一些专家将计算机中的一些数据挖掘技术应用于专利文件的处理中。例如, Hodrea, IB 就明确地提出将 Rose 和 Gurewitz 编写的确定性退火算法(deterministic annealing algorithm)应用于专利分类系统中^[14];而 Trappet, AJC 在 Development of a Patent Document Classification and Search Platform Using a Back-Propagation Network 一文中,详细介绍了将误

差反向传播网络(back-propagation network)应用在专利分类和专利搜索中。此方法首先通过文本挖掘技术从专利文件中抽取出关键词,然后通过关键词的频数具体地确定相应关键词的重要性。继而通过相关性分析确定所得关键词之间的相似度,从而将相似度较高的关键词分为一组,最后进一步利用误差反向传播网络算法对专利进行分类^[15]。通过比较发现,这9篇专利研究文献多关注于具体的数据挖掘算法,都是验证具体算法可以应用于专利文件的研究,即使提到具体的专利文件,也只是起到验证算法实验中的一个样本的作用,而没有系统地研究专利文件的结构,没有将算法集合于一起形成一个专利情报分析系统。

不过,有两篇专利研究文献为专利文件的分析研究提供了全面的方法和全新的发展方向。那就是作者 Yuen-Hsien Tseng 的 Text mining techniques for patent analysis 和 Suh, JH 的 A new visualization method for patent map: Application to ubiquitous computing technology。

其中 Yuen-Hsien Tseng 明确地提出将专利文件分为两个部分,即结构化部分和非结构化部分。其中结构化部分指的是专利文件中的专利入档时间(filing date)、申请时间(application date)、代理人(assignee)等,它们可以直接被存储到数据库管理系统中处理;另一部分非结构化部分就是专利文件中的专利标题(title)、专利摘要(abstract)、权利要求书(claims)、发明描述(description)等。其中非结构化部分具有非常重要的作用:首先,通过匹配分段后的非结构化部分中的片段,可以提高专利分类系统的精度,方便用户的专利搜索,减轻专利分析专家的负担;其次,关键词抽取以及之后的共词分析,不仅可以显示该专利所涉及的具体技术领域,甚至可以将该专利所涉及的技术细节暴露出来。之后进一步通过关键词聚类,专利的分类就更加精确了,最后还可以通过可视化的方法,将专利间的引用、技术进化演变更加直观的表现出来,同时也可以预测未来的技术发展方向。

之后,可以进一步“创造”技术,使技术的发展朝着人类期望的方向发展。这在 Suh, JH 的文章中有所提及,通过采用 k-means 聚类算法和基于语意理解网络,将专利文件中的结构化部分和非结构化部分通过专利地图(patent map)的方式表现出来。这样不仅将专利所涉及的技术表露无疑,也可以详细跟踪技术的发展历程,进而人类可以有目的的将技术发展朝着我们期待的方向引导^[16,17]。

1.3 实验小结 通过上面专利研究文献的分析,发现目前对专利的分析研究正处于高速发展阶段,而

其中当前专利研究的前沿是对专利文件的分析处理,不过目前对专利文件的研究,仍然是处于一个起步阶段,而所做的大部分工作也只是研究起步阶段中的一些试探性尝试。所以,在不久的将来,专利文件的研究仍然会是一个热点。

本文通过使用文献计量学中的文献共引分析和共词分析等方法,结合 CiteSpaceII 软件提供的文本挖掘手段,对专利研究文献做了初步的研究分析。不过鉴于 SCI-E 数据库中的专利研究文献仅是所有专利研究文献的一部分,且检索词的确定主观性比较强等存在的问题,之后作者将会进一步深入研究,以期克服上面所提到的一些缺陷。

2 结束语

本文通过将计量学领域的分析方法应用于专利研究文献的分析中,继而通过可视化的方法将其中隐含着的一些知识表现出来,从而不仅将与专利研究密切相关的一些领域清晰地显现了出来,同时在跟踪专利研究热点方面,也取得了比较满意的效果。这项工作不仅将计量学中的一些方法与文本挖掘中的一些方法以及可视化技术结合了起来,同时也为以后跟踪技术发展历程提供了一些开拓性的尝试。

参考文献

- Griliches Z. Patent Statistics as Economic Indicators: A survey[J]. *Journal of Economic Literature*, 1990; 1661-1707
- 丁 莹,李 鑫. 我国知识管理领域研究热点问题及发展趋势预测[J]. *情报杂志*, 2007, 26(9): 2-4
- 唐 琴,许 侃,林鸿飞. 搜索引擎发展阶段研究及热点发现[J]. *情报学报*, 2008(5): 664-669
- 刘玉琴,朱东华,吕 琳. 基于文本挖掘技术的产品技术成熟度预测[J]. *计算机集成制造系统*, 2008, 14(3): 506-510
- 李运景,侯汉青. 引文分析可视化研究[J]. *情报学报*, 2007, 26(2): 301-308
- Chaomei Chen. CiteSpaceII: Detecting and Visualizing Emerging Trends and Transient Patterns in Scientific Literature[J]. *Journal of the American Society for Information Science and Technology*, 2006, 57: 359-377
- Narin F. The Increasing Linkage between US Technology and Public Science[J]. *Research Policy*, 1976, 26(3): 317-330
- Jaffe AB. Patents, Citations, and Innovations: A Window on the Knowledge Economy[M]. The MIT Press, 2002
- Albert MB. Direct Validation of Citation counts as Indicators of Industrially Important Patents[J]. *Research Policy*, 1991, 20(3): 251-259
- Trajtenberg M. Geographic Localization of Knowledge Spillovers as Evidenced by Patent Citation[J]. *The Quarterly Journal of Economics*, 1993: 577-598
- Tijssen RJW. Technological Relevance of Science: An Assessment of Citation Linkage between Patents and Research Papers[J]. *Scientometrics*, 2000, 47(2): 35-54
- Meyers RA, Dixon RW. Who Invented the Laser: An Analysis of the Early Patents[J]. *Historical Studies in the Physical and Biological Sciences*, 2003: 115-149
- Callaert, J. Traces of Prior Art: An Analysis of Non-patent References Found in Patent Documents[J]. *Scientometrics*, 2006, 69(1): 3-20
- Hodrea, IB. The Rose - Gurewitz - Fox Approach Applied for Patents Classification[J]. *European Journal of Operational Research*, 2006, 173(3): 815-826
- Trappet, AJC. Development of a Patent Document Classification and Search Platform Using a Back - Propagation Network[J]. *Expert Systems with Applications*, 2006, 31(4): 755-765
- Yuen - Hsien Tseng. Text mining techniques for patent analysis [J]. *Information Processing and Management*, 2007, 43(5): 1216-1247
- Suh, JH. A new visualization method for patent map: Application to ubiquitous Computing Technology[J]. *Lecture Notes in Computer Science*, 2006, 4093: 566-573
- work of Scientific Collaborations[J]. *Physica A*, 2002, 311(3-4): 590-614
- Newman MEJ. Scientific Collaboration Networks I Network Construction and Fundamental Results[J]. *Physical Review E*, 2001, 64: 016131
- Newman MEJ. The Structure and Function of Complex Networks [J]. *SIAM Review*, 2003, 45(2): 167-256
- Barrat A, Barthélemy M, Pastor-Satorras R, et al. The Architecture of Complex Weighted Networks[J]. *Proceedings of the National Academy of Sciences of the United States of America*, 2004, 101(11): 3747-3752
- Barrat A, Barthélemy M, Vespignani A. Weighted Evolving Networks: Coupling Topology and Weighted Dynamics[J]. *Physical Review Letters*, 2004, 92(22): 228-701
- 张丹红,李晓辉. 复杂网络理论的情报学意义探讨[J]. *情报资料工作*, 2007(6): 12-14
- 刘 杰,陆君安. 一个小型科研合作复杂网络及其分析[J]. *复杂系统与复杂性科学*, 2004(7): 56-61
- 王福生,杨洪勇. 作者科研合作网络模型与实证研究[J]. *图书情报工作*, 2007(10): 68-71
- 杨洪勇,张嗣瀛. 作者合作复杂网络模型[J]. *情报科学*, 2008(5): 774-778
- 车宏安,顾基发. 无标度网络及其系统科学意义[J]. *系统工程理论与实践*, 2004(4): 11-16
- 汪小帆,李 翔,陈关荣. 复杂网络理论及其应用[M]. 北京:清华大学出版社, 2006: 9-13

(责编:王平军)

(责编:王平军)