

LiterMiner^{*}

——可视化多维文献分析工具

□ 赵斌 吴斌 / 北京邮电大学智能通信软件与多媒体北京市重点实验室 北京 100876

摘要: 随着复杂网络研究的深入, 网络分析方法被引入包括文献分析在内的多个领域。随着网络分析方法的逐步成熟, 涌现出一批优秀的文献可视化分析工具。文章在分析这些工具特点的基础上介绍了融合网络分析与联机分析处理的文献分析工具LiterMiner, 分析了该工具包括数据清理、社团发现与演化、GraphOLAP、系统架构在内的一系列系统特点, 并通过一个完整的分析流程展示了这些特性。

关键词: 文献分析, GraphOLAP, 复杂网络, 社团

DOI: 10.3772/j.issn.1673-2286.2010.08.002

1 前言

随着存储设备容量的不断增大, 人类进入了一个海量数据存储的时代。面对海量的科技文献数据, 如何对它们进行高效的处理、分析、展示, 对于进一步观察当前科学发展态势、分析科学研究特性、揭示热点研究领域以及为科研管理机构决策人员的战略决策提供依据有着重要的意义。

由于缺乏对大规模多关系型数据的处理能力, 传统文献分析工具往往局限于对数据进行局部分析, 忽视了数据间的联系。例如: 当用户搜索某个关键词时, 传统分析工具系统往往只能列出与该关键词相关的作者和文章, 而不能展示出该关键词与其他关键词、学术热点和研究领域的关系。这种分析方式不利于对文献信息进行全面、深入的了解。

近年来, 随着复杂网络研究的兴起, 网络分析方法逐步被应用到包括文献分析在内的多个领域。利用网络分析文献数据, 可以借助网络的可视化特点结合机器的特长与人的优势, 帮助分析人员发现数据的隐藏特征。这也使得结合网络分析的文献可视化分析工具成为当前文献分析工具的主流。本文在分析总结这类工具特点的基础上, 将重点介绍由北京邮电大学智能通信软件与多媒体北京市重点实验室开发的文献分

析工具LiterMiner。本文其他部分安排如下: 在第二部分相关工作中介绍当前流行的文献可视化分析工具; 第三部分介绍LiterMiner主要的系统特点, 包括数据清理、社团发现等; 第四部分通过一个完整的分析流程展示这些特点; 第五部分对全文进行总结。

2 相关工作

随着网络分析方法的逐步成熟, 基于网络可视的分析工具也越来越多。HistCite^[1]是由SCI创始人尤金·加菲尔德博士(Dr. Eugene Garfield)主导研发的一款分析引文网络的专用工具, 可以从文献数据中发现各个学科的发展规律和未来趋势。该工具的主要缺点是处理多领域数据时生成的编年图太复杂, 导致用户难以辨别关键文献及其之间的关系。CiteSpace是用来分析和展示引文网络的Java应用程序, 通过对比引文网络不同时段的多视图发现学科中的新趋势和重大事件^[2]。RefViz是一款通过文章聚类反映某一研究领域发展和动态的引文分析工具^[3], 通过连接和对比某个领域在不同时段的多视图来帮助识别知识领域中的新趋势。Jigsaw是一款多视图的文档间实体关系分析系统^[4], 通过多视图关联的方式向分析人员展示文本信息里存在的各种实体间的关系。由清华大学开发的ArnetMiner^[5]

^{*} 本文得到国家自然科学基金项目(90924029)和国家“十一五”科技支撑计划项目(2006BAH03B05)资助。

是一款文献分析工具,可以很好地找出领域专家、作者从事的领域、合作团体等,但偏重于对单个作者信息的检索和挖掘。D-Dupe是交互式的实体解析软件^[6],可以根据用户导入的相似函数和相似权重合并具有不同表述形式的同一实体,有效地处理数据中的不一致信息。在更广泛的复杂网络数据分析方面,NetworkX是一款专为复杂网络分析设计的免费软件^[7],可以进行最短路径计算、同构分析、个体网络分析、差异性分析、中心性分析等。但该软件是通过命令行来进行操作的,增加了操作难度。UCINET是一款菜单驱动的网络分析软件,界面友好^[8]。该软件不仅包括常用的网络静态特征分析程序,还包括众多的基于过程的分析程序,如多维量表(multidimensionalscaling)、二模标度(奇异值分解、因子分析和对应分析)、角色和地位分析(结构、角色和正则对等性)等以及多种统计程序。但该软件支持的数据规模较小,一般图的节点在万一级规模。

与上述工具相比,LiterMiner具有以下特点:

- (1) 通过使用网络解析方法,在清理数据时无需用户导入参数;
- (2) 提供了包括作者在内的多种实体的信息检索和挖掘,通过社团发现和社团演化,可以进一步分析作者所在的学术团队;
- (3) 以OLAP的方式更完整、直观地分析多类实体间的关系;
- (4) 分层展示引文网络,在处理涉及大学科的数据时,避免由于图中节点过多而导致用户难以辨别关键信息,此外LiterMiner采用更直观的方式展示引文网络的变化,无需连接和反复对比多个视图;
- (5) 提供了友好的图形界面与菜单操作使用户易于使用;
- (6) 支持较大的数据规模,目前已能处理节点数为十万级的网络。

3 多维文献分析工具LiterMiner

LiterMiner不仅提供了多类型检索、统计、报表生成等传统文献分析系统具有的基本功能,还通过抽取和分析文献数据包含的多种网络,以可视化的方式展示相关作者、机构、会议、刊物间的多种关系,为科研评价和决策提供更多依据。LiterMiner包含了数据导入、清理、分析、展示等完整的数据分析流程需要的

各项功能,本节将重点介绍和这些功能相关的系统特性。

3.1 数据清理

当前众多的文献分析工具和检索平台几乎都有专用的数据格式,LiterMiner支持导入多种主流检索平台的数据和私有数据,帮助分析人员摆脱需要选择不同软件分析不同数据带来的不便。LiterMiner目前已支持导入来自Web of Science^[9]、EI Village^[10]、万方数据知识服务平台^[11]的数据和Microsoft Excel文件、规则化的平面文件。

来源不同的集成数据中存在大量不一致的信息,例如同一作者的名字有多种不同的表述方式,而重名的作者可能因为其他属性的缺失被误认为是同一作者,这在中英文混合数据中尤为严重。在正式分析前,很有必要对数据进行清理,消除不一致信息,以提高后续分析的精度和性能。高质量的决策依赖于高质量的数据,在数据分析前检测数据异常、尽早地调整并归约待分析的数据,将在决策过程得到高回报。

传统的文献分析工具中,数据清理工作往往由分析人员完成,这个过程通常是重复、繁琐和易错的。当前流行的分析工具以判断实体属性相似度的方法消除不一致数据,在很大程度上提高了数据清理的自动化程度。但基于属性相似度的清理方法需要用户设定相似度阈值和各项属性的相似度权值,需要用户了解更多的背景知识。另外这种清理方法对由属性缺失引起的实体重名处理效果较差。LiterMiner通过分析和比对实体间的关系,使用链接分析、协同分析^[12]等方法,可以更精确地规范机构、期刊、会议和作者等实体的名称,无需用户设定参数。此外,LiterMiner提供了友好的界面供用户对清理结果进行干预,进一步保证清理的准确性。

3.2 网络分析与社团发现、社团演化

近年来在网络分析领域,在对网络结构所具有的小世界、幂率分布等静态特性的进一步分析之外,大量研究开始关注网络结构中个体的组织特征。由这些个体组成的子图中,个体间有着更高的连接特征,而与其他子单元间的个体连接则相对稀疏。这种子单元通常被称为社团。社团发现及分析对网络组织结构和

社会特征的研究有着重要意义。将社团发现方法应用到文献分析中,可以挖掘各学科领域的特征及关联关系、科研机构内部的学术团队等。LiterMiner集成了多种社团发现算法,既包括GN^[13]、CNM^[14]等经典算法,也包括LPA^[15]等时间复杂度接近线性的新算法。通过社团发现,LiterMiner可以从文献共被引网络中提取相关的研究基础、学术前沿和研究领域,也可以从机构内部作者合作关系中挖掘出相关的学术团队,从更高的层次挖掘科研机构的发展情况。

在社团发现研究兴起的同时,对网络结构动态特性的分析也开始得到广泛关注。分析网络各项属性的演变过程,有助于进一步分析网络的演化过程,发现网络中的重大事件,这对网络行为预测、网络演化建模有着重要的意义。LiterMiner可以计算网络多项属性,并将演化过程以图表的形式展示出来。此外,LiterMiner还能以可视方式追踪网络中社团结构的演化过程,为进一步分析研究领域中的重大事件和学术热点的迁移、科研团队的发展与团队成员的关系提供了有力的支持。

3.3 多维分析与GraphOLAP

使用网络展示文献数据,可以直观地反映相关实体的多维特性,如果能进一步对网络作多维分析,则有利于分析人员深入理解各类实体间的关系。联机分析处理(Online Analytical Processing)系统常用来分析多维数据,以数据立方体的形式提供数据的多维视图,并通过上卷和下钻操作提供数据的多层次视图。但传统的联机分析处理系统并不支持图结构,现有的主流分析工具也都不支持对图进行OLAP操作。

近来,Chen Chen、Xifeng Yan和Jiawei Han等提出了支持图结构的OLAP框架GraphOLAP^[16]。该框架根据图的特点重新定义了传统OLAP中的相应概念和操作,如图的数据立方体、图的上卷与下钻等,并在此基础上为图结构规划了完整的多维分析流程。GraphOLAP框架到目前尚未实现,实现该框架存在诸多难点,例如图的识别、立方体的物化策略等。LiterMiner通过多种方式解决了这些难点。LiterMiner以综合考虑图的拓扑结构和多维信息的方法,代替时间复杂度较高、完全基于拓扑结构的传统算法来识别图,提高了图的存取速度,同时解决了纯拓扑方法无法辨别具有不同信息的同构图的问题。针对不同的场景,LiterMiner采用

了不同的立方体部分物化策略保证数据量较大情况下OLAP系统的响应速度,目前该工具已能分析记录数为十万级的文献数据。此外,针对从文献数据抽取的网络中包含大量异构网络的特点,LiterMiner在实现GraphOLAP框架的基础上进一步扩展该框架,定义并实现了异构图的OLAP,支持展示和进一步分析多类实体间的多种关系。

3.4 系统架构

LiterMiner系统采用SOA架构。与传统系统架构相比,SOA的特点在于:基于标准,松散耦合,共享服务和粗粒度等。网络分析展示集成工具使用SOA体系架构,其优势可表现为:

(1) 敏捷:在研究进展日新月异的环境下,系统能不断集成新的算法和子系统非常重要。一方面数据挖掘算法和可视化布点算法的发展很快,另一方面分析算法的应用涉及到了各个学科领域,对于特定学科可能存在着专用的算法,而系统并未涉及。这样就需要系统可以嵌入用户提交自行实现的算法。无论用何种语言实现,只要满足接口标准,就可以动态地嵌入到系统中,作为系统的一部分。

(2) 服务透明性:服务透明性使软件的各层之间的逻辑关联和部署关联都降到最低,完全可以做到并行开发。而平台和编程语言的无关性,有利于系统维护和版本升级。

系统概念构架可以分为三层结构,如图1所示,各层的功能如下:

(1) 数据持久层:处理各种不同格式的数据源,

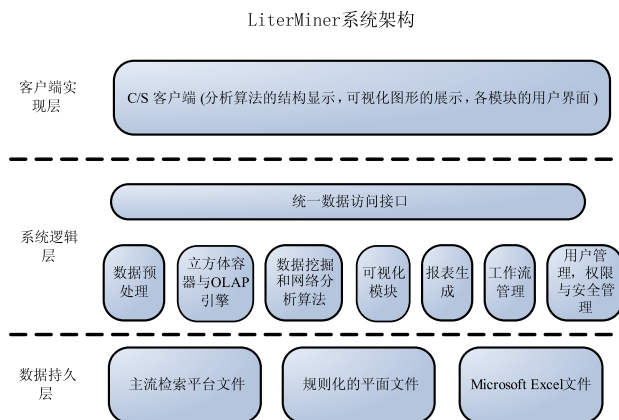


图1 LiterMiner系统架构

可以把各种数据源转化为统一格式的数据,传递给系统逻辑层,向上层屏蔽底层数据的位置和结构。

(2) 系统逻辑层:根据SOA概念设计的架构,分别把各个模块包装成通用的服务。SOA松耦合的架构利于功能的扩展和复用。本层通过统一的数据访问接口获取图形数据,可以提供如下服务:

数据的预处理,对数据进行解析,清洗和过滤,提取网络结构;

集成了数据挖掘和复杂网络分析的算法,如网络特征值计算,基本几何属性统计,社团结构挖掘等操作;将得到的结果信息传入下一层;

立方体容器和OLAP引擎,根据提取的网络结构建立数据立方体,并根据相关的立方体物化策略物化和存储部分立方体,为OLAP操作提供访问支持;

可视化结构层,将从数据持久层得到的数据,根据用户的要求进行过滤,结合图形显示信息(如布局算法、渲染效果等),将可以直接用于显示的图形数据传入下一层;

报表生成,根据用户的需求生成各类统计和可视化分析结果的表单;

workflow管理,将用户的分析需求分为数据导入、数据清理、数据分析和生成报表等过程进行管理。提供图形界面的 workflow管理模块,使用户对整个分析的流程可以动态地配置和改变;

用户管理、权限与安全,根据相应权限管理使用系统服务的各类用户。

(3) 客户端实现层:出于对执行效率的考虑,客户端为C/S结构,属于胖客户端。

4 分析实例

4.1 数据准备

分析所用数据为从Web of Science^[9]、EI Village^[10]和万方数据知识服务平台^[11]收集的北京邮电大学于2005年-2008年发表的论文(去重后共计14000余篇,由于从EI Village^[10]和万方检索平台^[11]收集的论文数据没有引文信息,引文分析所用数据为从Web of Science收集的北京邮电大学截止到2009年发表的论文)。

4.2 科研团队分析

数据导入后首先进行数据清理,包括会议、期刊、机构和作者名称的规范以及重复文章的清理。数据清理完成后,用户可以在主界面根据多个字段检索关注的记录,也可以选择分析各类实体,如图2所示。这里选择机构分析,机构分析界面如图3所示,右上方为所有机构相关信息列表,列表上方的工具栏提供了包括可视化、生成报表在内的多种功能;左上方显示了列表中选中机构包括地址、发文情况等详细信息;左下方的过滤与统计方式模块可以使用户提取和统计感兴趣的数据子集;右下方为机构发文情况的三种统计图(其他实体分析的界面与功能与机构分析类似)。

在机构列表中双击北京邮电大学,分析该机构

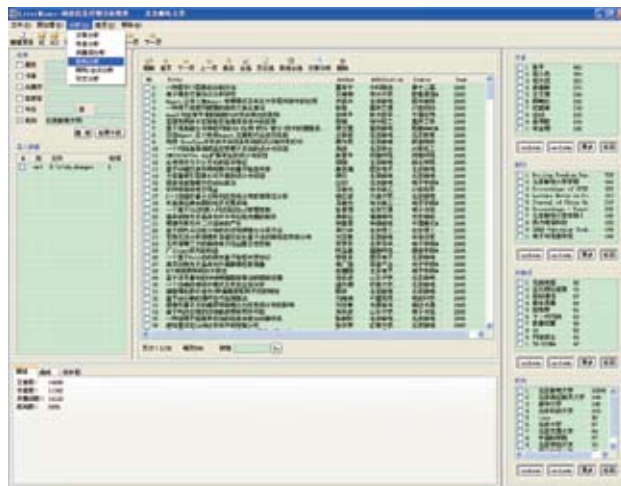


图2 LiterMiner主界面

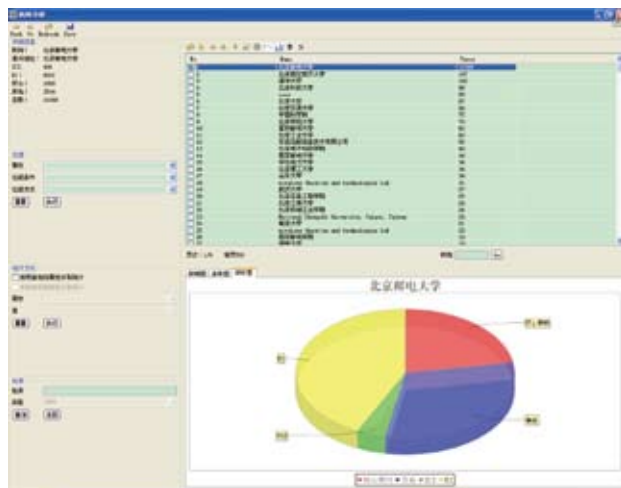


图3 机构分析界面

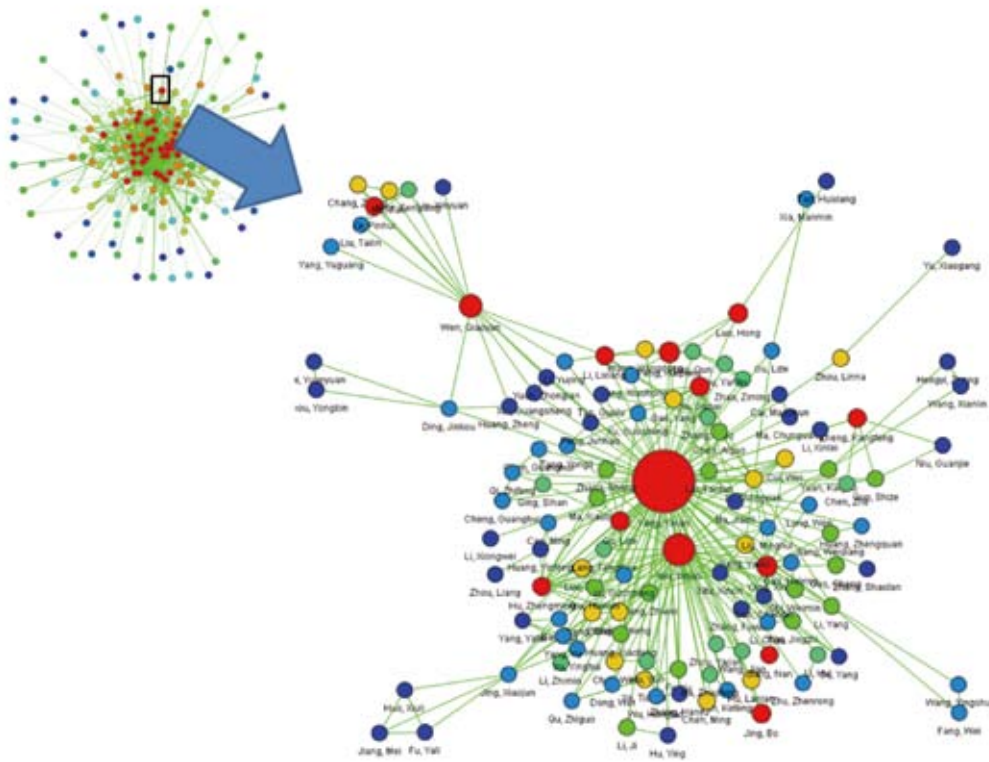


图4 机构内部团队合作关系

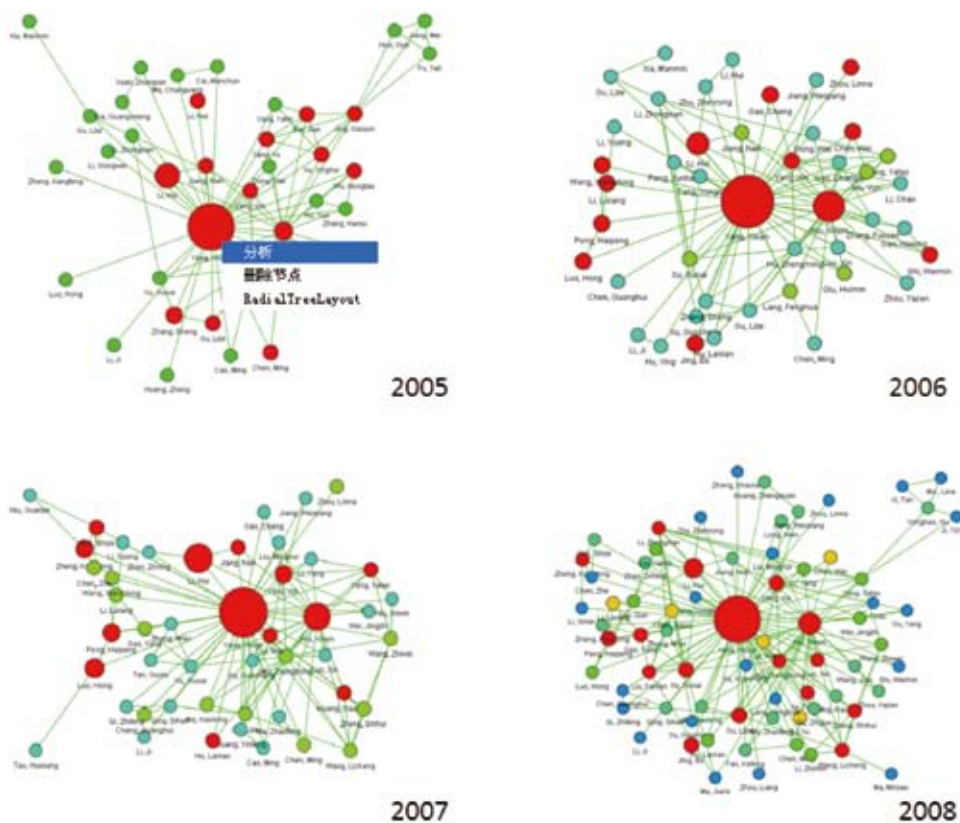


图5 团队演化过程展示

内部团队，各团队合作关系的可视化效果如图4左上所示。可以对图中指定团队下钻进一步分析团队内部的合作关系。图4右下展示了对某一团队下钻后取得的成员合作网络。该图中节点的大小反映了作者的发文量，节点颜色反映了作者的活跃程度。拖动GraphOLAP面板中时间维的滑块条，图中节点的大小和颜色根据年份不同而变化，可以直观地查看整个团

队的演化过程（演化过程如图5所示）。

LiterMiner支持以异构网OLAP的方式观察各类实体间的关系，在GraphOLAP面板的实体维选择相应的实体便可观察多类实体间的关系。此外LiterMiner也支持以Jigsaw中关联分析的方式分析各类实体间的联系，如图5所示，右键点击该团队核心作者A的重要合作者B，在菜单中选择分析，各类实体分析界面将会列出与

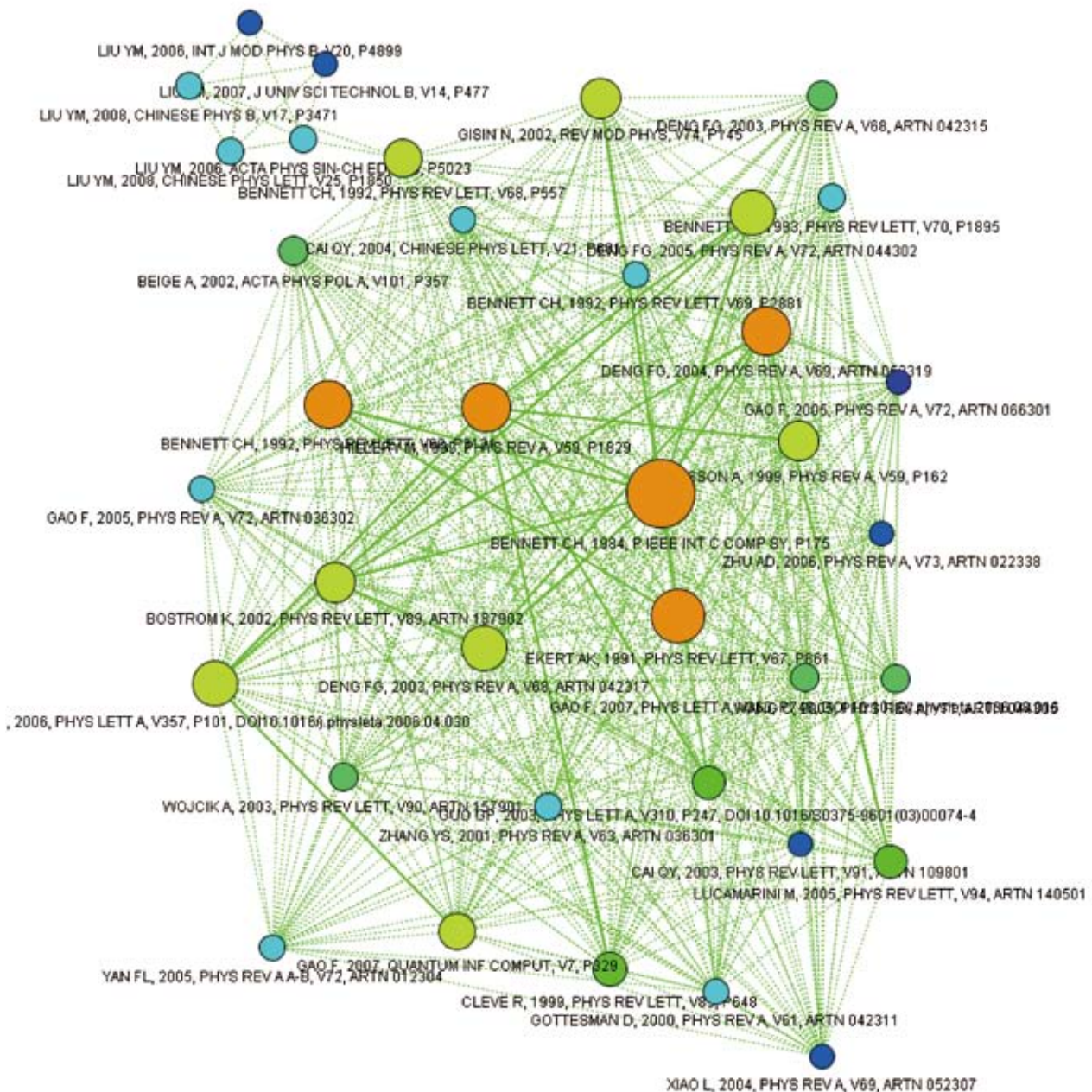


图6 光通信方向参考文献共被引关系

作者B相关的各类信息。

4.3 引文分析

LiterMiner将文献数据中的引文,根据共被引关系聚类得到研究基础,在研究基础上进一步聚类得到研究领域的划分,同时将文献根据引用相同研究基础的关系聚类得到学科前沿。在导入数据并清理后,选取引文分析观察各引文聚类(研究基础)间的关系,图6为选取其中的光通信聚类下钻后得到的论文共被引关系,节点的尺寸反映了论文的绝对被引次数,颜色反映了相对被引次数。与团队分析类似,可以通过时间维OLAP,直接观察节点大小与颜色的变化分析该研究方向的研究历程和重大事件。与HistCite、CiteSpaceII和

RefViz相比,使用LiterMiner分析引文网络演化无需连接和反复对比同一聚类多个时段的视图;由于采用了学科领域、研究基础、共被引关系的多层次视图,避免了分析多领域数据时图中节点过多而导致用户难以辨别关键文献或聚类的问题。

5 总结

本文在介绍当前流行的文献可视化分析工具的基础上,重点介绍了北京邮电大学智能通信软件与多媒体北京市重点实验室开发的文献分析工具LiterMiner,该工具支持多种数据格式,以网络可视的方式从多个维度深入分析文献数据,可以为科研管理的评价和决策工作提供有力的支持。

参考文献

- [1] Garfield E. Histofigraphic mapping of knowledge domains literature[J]. Journal of Information Science, 2004, 30(2): 119-145
- [2] Chen C. Citespace ii: Detecting and visualizing emerging trends and transient patterns in scientific literature[J]. Journal of the American Society for Information Science and Technology, 2006, 57(3): 359-377.
- [3] RefViz[EB/OL].[2010-05-01]. <http://www.refviz.com>.
- [4] Stasko J, Gorg C, Liu Z. Jigsaw: supporting investigative analysis through interactive visualization[J]. Information Visualization, 2008, 7(2): 118-132.
- [5] Tang Jie, Hong Mingcai, Zhang Jing, et al. ArnetMiner: Toward Building and Mining Social Networks[C]. // Proceedings of the Thirteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Jose, August 12-15, 2007[C].
- [6] Mustafa Bilgic, Louis Licamele, Lise Getoor, et al. D-Dupe: An Interactive Tool for Entity Resolution in Social Networks[C]. // Proceedings of IEEE Symposium on Visual Analytics Science and Technology 2006, Maryland, October 31 - November 2, 2006[C].
- [7] Hagberg A, Schult D, Swart P. NetworkX[EB/OL]. [2010-06-11]. <http://networkx.lanl.gov/>.
- [8] Borgatti, S., M. Everett, L. Freeman. UCINET[EB/OL]. [2010-06-20]. <http://www.analytictech.com/ucinet/>.
- [9] W. of Science[DB/OL].[2010-05-01].<http://www.isiknowledge.com>.
- [10] E. Village[DB/OL].[2010-05-01].<http://www.engineeringvillage2.org.cn>.
- [11] WanFangData[DB/OL].[2010-05-01]. <http://www.wanfangdata.com.cn>.
- [12] Wu bin, Xu Chao-Qun, Wang Wen-Bin, et al. A Link-based method for Name Disambiguation and its Application[J]. Computer Science, 2007, 35(3):197-199.
- [13] Girvan M, Newman M E J. Community structure in social and biological networks[J]. Proc. Natl. Acad. Sci, 2001, 99: 7821-7826
- [14] Clauset A, Newman M E J, Moore Christopher. Finding community structure in very large networks[J]. Phys. Rev. E, 2004, 69(6).
- [15] Raghavan Usha N, Albert Reka, Kumara Soundar. Near linear time algorithm to detect community structures in large-scale networks[J]. Phys. Rev. E, 2007, 76(3).
- [16] Chen Chen, Yan Xifeng, Zhu Feida, et al. Graph OLAP: a multi-dimensional framework for graph data analysis[J]. Knowledge and Information Systems, 2009, 21(1): 41-63.

作者简介

赵斌, 硕士研究生, 研究方向为数据挖掘、信息可视化、复杂网络。通讯地址: 北京邮电大学179信箱 100876 E-mail: dh031013@gmail.com
吴斌, 副教授, 主要研究领域为数据挖掘、复杂网络及智能信息处理。通讯地址: 同上。E-mail: wubin@bupt.edu.cn

LiterMiner: A Visual Multi-dimensional Literature Analytic System

Zhao Bin, Wu Bin / Beijing Key Laboratory of Intelligent Telecommunications Software and Multimedia, Beijing University of Posts and Telecommunications, Beijing, 100876

Abstract: With the in-depth study of complex network, network analytic methods have been used in many domains, including the analysis of literature. First, this paper investigates international popular visual literature analytic tools based on network analytic technology. Then the architecture and technical features of one multi-dimensional literature analytic tool: LiterMiner are introduced in detail. Finally, we demonstrate the features of LiterMiner with a complete analytic process.

Keywords: The analysis of literature, GraphOLAP, Complex network, Community-structure

(收稿日期: 2010-06-05)