Digital Library Forum 数字图书馆论坛
http: // www.dlf.net.cn

基于Web的科技文献分析工具综述*

□ 田宏桥 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：随着电子引文信息的爆炸式增长，为了帮助科研工作者从海量文献数据中发现研究热点、了解领域发展趋势，科研机构和商业集团开发了一系列基于Web的文献检索工具。近年来，伴随着数据挖掘技术和信息可视化技术在知识发现领域中的迅速发展，融合上述两种技术的文献分析工具已经被研发出来并获得了很好的用户反馈。文章首先阐述了传统的文献检索工具的功能及其存在的局限性，调研了融合可视分析技术的文献分析工具并总结其功能和特点以及为文献分析带来的新颖视角，介绍北京邮电大学通信软件工程中心研发的可视分析组件——VisLib及其实用场景，最后展望了科技文献分析工具的进一步发展方向。

关键字：Web技术，文献分析，数据挖掘，信息可视化
DOI：10.3772/j.issn.1673—2286.2010.08.003

1 背景

通过引文索引与Web技术的结合，诞生了扩展性强，强大灵活的文献检索工具，为科研工作者借鉴前人经验、跟踪领域最新进展提供了极大的便利，同时在文献统计基础上对文献利用行为进行描述[1]。近年来，数据挖掘技术和信息可视化技术在知识发现过程中起到越来越重要的作用，也为传统的基于文献计量学的文献分析带来了新的动力。如何将这两种技术同传统的文献分析工具相结合，为用户提供多侧面、多维度的分析方法，更直观地展现出数据之间的联系，已经成为情报分析领域和人机交互领域（HCI）的研究热点。一些信息科技公司和大学科研机构已经率先开发出相应的产品并提供服务，其新颖独特的分析展现方式得到了用户良好的反馈，标志着文献分析工具新的发展方向。

本文第二节介绍了传统文献分析的特点并列举了具有代表性的几种工具；第三节讨论了数据挖掘技术带给文献分析领域的新视角、新思维，着重介绍ArnetMiner和微软学术搜索的技术特点和分析挖掘功能；第四节介绍北京邮电大学通信软件工程中心研发的可视分析组件——VisLib及其实用场景；最后展望了基于Web技术的文献分析工具的发展趋势。

2 基于统计分析的文献分析工具介绍

在传统的文献分析工具中，通过对期刊、作者、语种、时间分布等引文指标进行统计分析，可以获取当前的领域研究热点、实力雄厚的科研机构以及领域专家等结果，为初次接触该领域的科研人员提供了信息导向，也为科研评价提供了相应的依据。在当今的文献分析工具领域中，几乎都提供相同或与之相近的服务。稍显遗憾的是，这些统计分析值更多停留在表层上对文献数据进行分析，忽略了文献元数据之间的链接关系，其结果往往存在局限性。下面将按照所面向领域的不同和数据获取方式的差异对这些工具进行介绍。

2.1 面向全领域

- 三大检索

科学引文索引（Science Citation Index，SCI）、工程索引（Engineering Village2，EI）和科学技术会议录索引（Index to Scientific &Technical Proceedings，ISTP）是

* 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。

著名的三大科技文献检索系统，也是进行科学统计和科技评价的主要检索工具。《科学引文索引》的引文数据库是覆盖生命科学、临床医学、物理化学、农业、生物、兽医学、工程技术等方面的综合性检索刊物，尤其能反映自然科学研究的学术水平，在学术界占有重要地位。许多国家和地区均以被SCI收录及引证的论文情况作为评价学术水平的一个重要指标。《工程索引》由美国工程信息公司出版，报道工程技术各学科的期刊、会议论文、科技报告等文献。《科技会议录索引》由美国科学情报研究所编辑出版，其涉及学科基本与SCI相同。

- ScienceDirect，Scopus，Scirus

作为国际化多媒体出版集团Elsevier的旗舰电子产品，ScienceDirect（www.sciencedirect.com）是全球最著名的科技医学全文数据库之一，其直观友好的使用界面，使研究人员可以迅速链接到Elsevier出版社丰富的电子资源，包括期刊全文、单行本电子书、参考工具书、手册以及图书系列等。Scopus（www.scopus.com）是目前全球规模最大的文摘和引文数据库，涵盖了由5000多家出版商出版发行的科技、医学和社会科学方面的18,000多种期刊，内容全面、学科广泛，特别是在获取欧洲及亚太地区的文献方面，用户可检索出更多的文献数量。Scirus（www.scirus.com）是一个免费的专为科学家、研究人员和学生开发的网络检索引擎，可以使得每位想要检索科学信息的人员快捷精准地查找到所需信息——包括专家评审刊物，发明专利信息，作者主页以及大学网站等等。

- SpringerLink

SpringerLink是国际著名科技出版集团Springer的网络版全文文献服务系统。通过SpringerLink可提供全文服务的文献包括Springer出版的478种科技、医学等学术期刊，20余种世界知名科技丛书和权威的Landolt-Börnstein数值与事实型工具书。

- 中国学术期刊全文数据库

中国学术期刊全文数据库由清华大学中国学术期刊（光盘版）电子杂志社主办，是目前全世界最大的中文期刊文献全文数据库。数据库内容丰富，涵盖自然科学、工程技术、人文社科等各个学科领域。

- 万方数字资源

万方数字资源汇集中国学位论文文摘、中国数字化期刊群、会议论文文摘、科技成果、专利技术、标准法规、各类科技文献、科技机构、科技名人等近百个数据库，收录包括期刊、会议、书目、题录、报告、论文、标准、专利、工具书等各种类型文献。

2.2 面向专业领域

- ACM Digital Library

ACM（the Association for Computing Machinery），美国计算机学会，创立于1947年，是全球历史最悠久和最大的教育科学计算机学会。目前，ACM的成员有来自100多个国家的工业、学术界和政府的8万多位计算机专业人士。1999年起，ACM开始提供电子数据库服务——ACM Digital Library数据库（http://portal.acm.org/portal.cfm）。ACM全文数据库内容包括：全文期刊、杂志和汇刊共44种、SIG定期简讯、超过220种会议录，同时涵盖了大量的计算机学科里的核心出版文献的书目引文资料和会议文献论文资料等。

- DBLP（Digital Bibliography & Library Project）

DBLP是德国特里尔大学搭建的计算机科学文献检索网站（http://www.informatik.uni-trier.de/~ley/db/）。到目前为止，DBLP收录了超过1,300,000篇科技文献，其中涵盖了计算机学术会议、期刊、报告、书籍在内的海量文献记录，为科研工作者查询文献信息提供帮助。

- PubMed

PubMed是一个免费的搜寻引擎，提供生物医学方面的论文搜寻以及摘要（http://www.ncbi.nlm.nih.gov/PubMed/）。其数据库来源为MEDLINE，核心主题为医学，但也包括其他与医学相关的领域。PubMed的资讯并不包括期刊论文的全文，但可能提供指向全文提供者（付费或免费）的链接。

- APS Journals

APS（The American Physical Society，美国物理学会）成立于1899年，是世界上最具声望的物理学专业学会之一。APS出版的物理评论系列期刊*Physical Review*、*Physical Review Letters*、*Reviews of Modern Physics*，分别是各专业领域最受尊重、被引用次数最多的科技期刊之一，在全球物理学界及相关学科领域的研究者中具有极高的声望。

- IEEExplore

IEEE/IEE Electronic Library（IEL）数据库，收录美国电气电子工程师学会（IEEE）和英国电气

工程师学会（IEE）出版的242种期刊、8706余种会议录和近1706种标准的全文信息。读者通过检索可以浏览、下载或打印与原出版物版面完全相同的文字、图表、图像和照片的全文信息。IEL数据库所涉及的学科有计算机、电气电子、信息科学、物理学。

- CA

CA（Chemical Abstracts，化学文摘）1907年创刊，由美国化学会所属化学文摘服务社（CAS）编辑出版，现为世界上收录化学化工及其相关学科文献最全面、应用最广泛的一种文献检索工具。

- LexisNexis

LexisNexis（律商联讯）是世界领先的法律和商业资讯提供商，可提供世界范围内的报纸、杂志、商业期刊、行业新闻、税务和财会信息、金融数据、公共记录、立法档案、企业及其管理者的信息。

- IIPA

IIPA（International Index of Performing Arts，国际表演艺术期刊全文数据库）是网上仅有的关于表演艺术的资源之一，有35万篇左右的记录，主题涉及舞蹈、电影、电视节目、艺术表演等。

2.3 网络数据源

- CiteSeerX

CiteSeerX是CiteSeer的升级产品（http://citeseerx.ist.psu.edu/）。CiteSeer（又名ResearchIndex），是NEC研究院在自动引文索引（Autonomous Citation Indexing，ACI）机制的基础上建设的一个学术论文数字图书馆。这个引文索引系统提供了一种通过引文链接的检索文献方式，目标是从多个方面促进学术文献的传播和反馈。在CiteSeer提供服务的十年间，研发人员结合当前Web技术的发展趋势和用户的反馈信息，针对原系统无法应对大规模数据的问题，对系统进行了重构，扩展了原有的数据模型并增加了用户服务功能[2]，系统更名为CiteSeerX。

CiteSeerX检索WEB上的PostScript和PDF两种格式的学术论文。目前，在CiteSeerX数据库中可检索超过81万篇论文，这些论文涉及的内容主要是计算机领域。这个系统能够在网上提供完全免费的服务（包括下载PostScript或PDF格式的论文的全文）。该系统的主要功能有：①检索相关文献，浏览并下载论文全文；②查看某一具体文献的“引用”与“被引”情况；③查看某一篇论文的相关文献；④图表显示某一主题文献（或某一作者、机构所发表的文献）的时间分布。

- Google Scholar

Google Scholar（GS）是Google公司于2004年底推出的专门面向学术资源的免费搜索工具（http://scholar.google.com），能够帮助用户查找包括期刊论文、学位论文、书籍、预印本、文摘和技术报告在内的学术文献，内容涵盖自然科学、人文科学、社会科学等多种学科。Google Scholar不仅仅从Google收集的上百亿个网页面中筛选出具有学术价值的内容，而且最主要的方式是通过与传统资源出版商的合作来获取足够的有学术价值的文献资源。目前，Google公司与许多科学和学术出版商进行了合作。他们已经与学术、科技和技术出版商，如ACM、Nature、IEEE、OCLC进行了广泛的合作。这种合作使用户能够检索特定的学术文献，通过Google Scholar从学术出版者、专业团体、预印本库、大学范围内以及从网络上获得学术文献，包括来自所有研究领域的同级评审论文、学位论文、图书、预印本、摘要和技术报告。

3 融合数据挖掘与可视分析技术的文献分析工具介绍

数据挖掘技术帮助人们发现隐藏在海量数据背后的知识，其在文献分析领域起到的作用日益显著。对领域内高被引文献进行聚类分析，可以揭示学科结构、研究热点及发展方向。通过挖掘引证关系，可以追踪学科研究的脉络和演变过程。随着分析挖掘的深度和广度不断增加，越来越多的隐藏在引文数据背后的知识将被发现出来。

通过数据挖掘技术，计算机可以支持对大规模数据的统计与分析，但却很难提高用户对发现特征的理解能力[3]。如何更加有效地分析数据及展示分析结果，从而更有效地提高人们对分析数据的认知能力，正受到学者们越来越多的关注。可视化分析与信息可视化技术在知识发现的过程中有着不可或缺的地位，在机器输出结果与用户的交互过程中起了极为重要的桥梁作用。信息可视化最大的优势在于能够展示文本无法表现的大量信息[4]。

下面从数据挖掘技术的应用以及信息可视化两个方面，挑选国内外具有代表性的工具介绍。

3.1 ArnetMiner

ArnetMiner是由清华大学计算

机系软件研究所知识工程研究室开发的针对于科技文献领域的分析工具（www.arnetminer.org），在提供检索功能的同时，融入了数据挖掘和可视分析技术，为学术社区的分析和评价提供了一个全新的视角。ArnetMiner与传统文献分析工具相比，提供了以下几个新颖功能：

1) 领域专家发现：输入领域检索词，如data mining，系统将会返回这个领域的知名学者。同时，该领域内的顶级会议和重要文献信息也将呈献给用户。

2) 会议分析：输入会议名称，如KDD，系统将会返回这个会议中最活跃的学者，以及重要的文献。

3) 学术排名：系统定义了8个指标来衡量一个学者的学术成就（http://arnetminer.org/Academic Statistics）。用户可以限定一个领域查看排名信息，如数据挖掘领域。

4) 课程检索：输入课程的名称，如machine learning，系统将会返回教授这门课程的学者以及课程的具体信息，如课程编号，所开设的学校等。

下面以领域专家发现和学者重名处理为例介绍其挖掘功能。

领域专家发现的目的是找出某领域内最具影响力的科研工作者。ArnetMiner的发现算法[5]不仅考虑科研人员的履历信息，同时将科研人员之间的合作关系也作为算法的输入。其排名结果优于只考虑链接关系的PageRank算法[6]和仅考虑作者履历信息的算法。

图1展示了在数据挖掘领域的研究专家。

学者的重名现象会导致统计分析结果的不准确，为此ArnetMiner引入了基于隐马尔科夫条件随机场算法[7]，对重名学者进行了区分，取得了很好的效果。图2为对姓名为Wei Wang的学者进行重名处理的结果。

图1 数据挖掘领域研究专家

图2 机器学习领域研究专家

在信息可视化方面，ArnetMiner结合了社会网络分析中的个人中心网络可视化概念，将一个作者的合作者以及合作信息通过拓扑图展现出来。

在作者详细信息页面中，用户可以多侧面、多维度地分析该作者的合作网络：

- 按领域分析：用户可以选择领域信息，展示作者在不同领域的合作者；
- 按角色分析：展示作者与合作者的角色关系，如教师-学生关系，普通合作关系。不同角色使用不同颜色渲染；
- 按合作次数分析：可根据作者之间的合作强度展示相应的合作网络。

图3是韩家炜的合作网络图，在图中可以直观地看出哪些人是他的学生，谁曾经指导过他的研究以及他的普通合作者有哪些。

ArnetMiner于2006年5月首次发布，在完成了科研作者信息抽取的基础上，提供了作者、文献和会议的检索功能；在2007年的2.0版本中，提供了关联搜索、研究兴趣发现等功能；在2008年的4.0版本中，提供了图搜索和领域挖掘功能并获得了国家自然科学基金支持；在2009年发布的版本中，提供了学术统计、用户反馈和开放资源下载等功能。

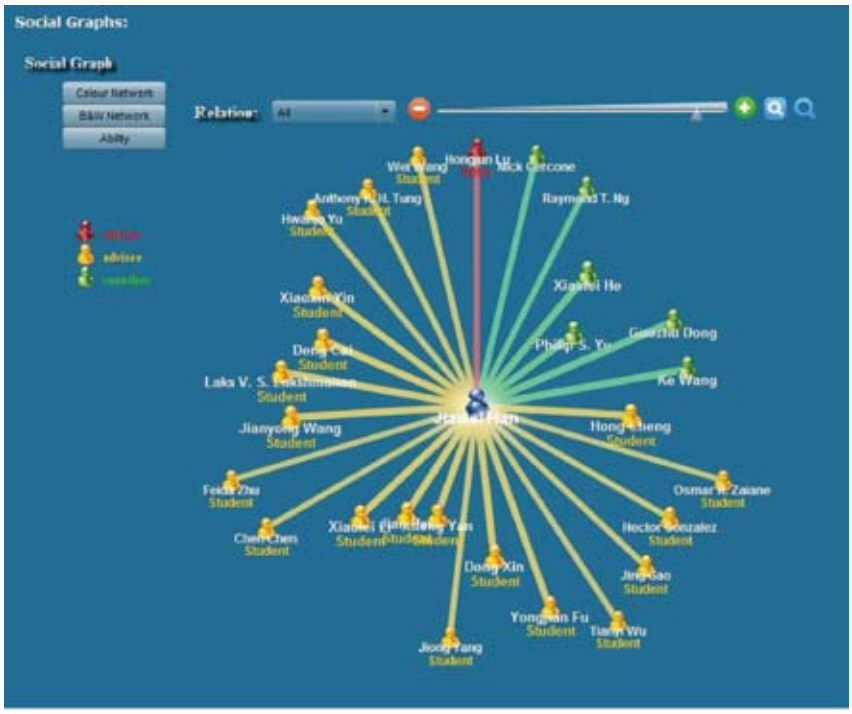

图3 Arnetminer中韩家炜合作关系图

3.2 微软学术搜索

微软学术搜索是微软亚洲研究院开发的在线免费学术搜索引擎（http://libra.msra.cn/）。它为研究员、学生、图书馆员和其他用户查找学术论文、国际会议、权威期刊、作者和学术兴趣圈提供了更加智能、新颖的获取方式。与传统的页面级搜索引擎比较，微软学术搜索使用对象级别的信息查询方式[8]，它可以帮助用户找到一个学术领域内的顶尖科学家、会议和期刊；获得一个学术兴趣圈兴起与发展的详细信息；使用户准确找到感兴趣的学术论文；发现某个领域经典、热点的学术论文和正在崛起的学术新星。目前该搜索引擎专注于计算机科学和信息科学范围内的搜索，未来有可能将搜索范围扩展到其他学科领域。

与传统文献分析工具不同，微软融入了垂直搜索技术[9]。通过抽取文献数据中的实体，如作者、期刊、会议等等，并构建实体之间的联系网络，通过对实体网络应用链接挖掘算法[10]，可以解决实际应用中的问题，如作者重名处理，期刊、会议排名以及相关性分析等。

在信息可视化方面，微软学术搜索同样融入了个人中心网络展示，在图中心位置的高亮节点代表分析作者，其合作者按照合作关系的强弱分布在其四周。基于SiverLight技术实现的网络图带给用户简明、清晰的感受。图4展示了韩家炜的个人中心网络，图5为与韩家炜合作关系最强的作者。

在2009年11月发布的Beta版中，视觉浏览（Visual Explorer）页面应用了人立方搜索技术，综合了与每位作者有关的合作者信息，用户可以直接在视觉浏览中点击查看其他作者信息。这帮助用户能够很快了解到某个研究领域的学术圈有哪些重要学者。同时新增加了作者的照片；点击作者图像可以显示作者的信息；点击作者之间的连接线，可以浏览合作的论文信息；在视觉浏览框内检索作者姓名可浏览他的学术关系图。

3.3 DBLP VIS

为了方便用户了解大量数据之间的链接关系，DBLP开发了可视化模块DBLP VIS（http://dblpvis.uni-trier.de/），将会议、作者和关键词之间的联系直观地展现出来。

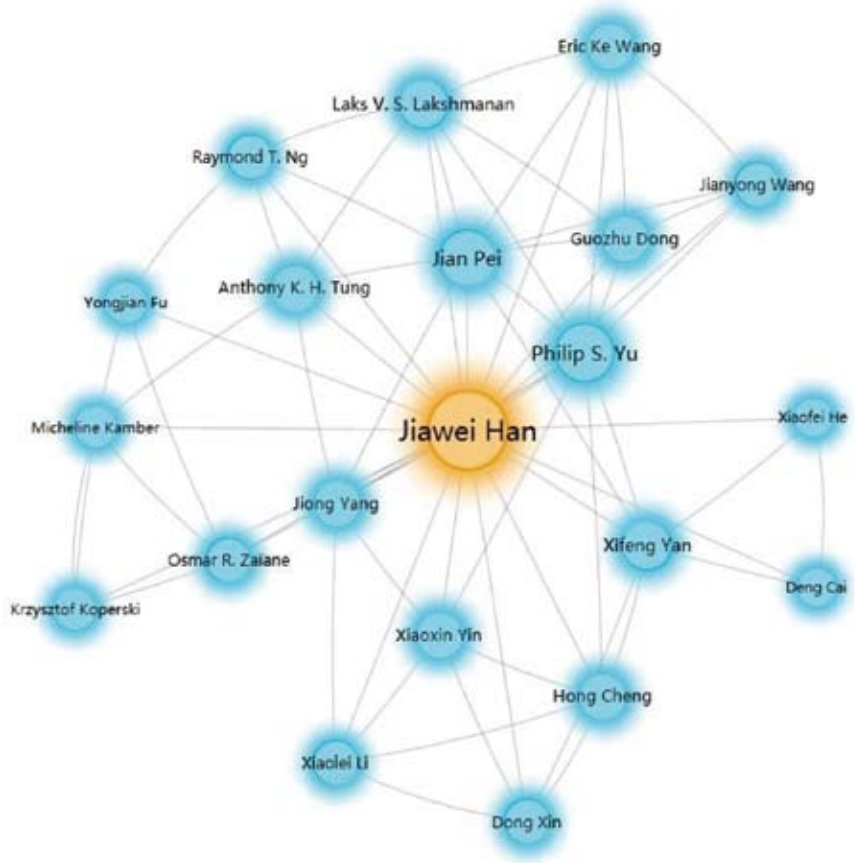

图4 微软学术搜索中韩家炜的合作关系图

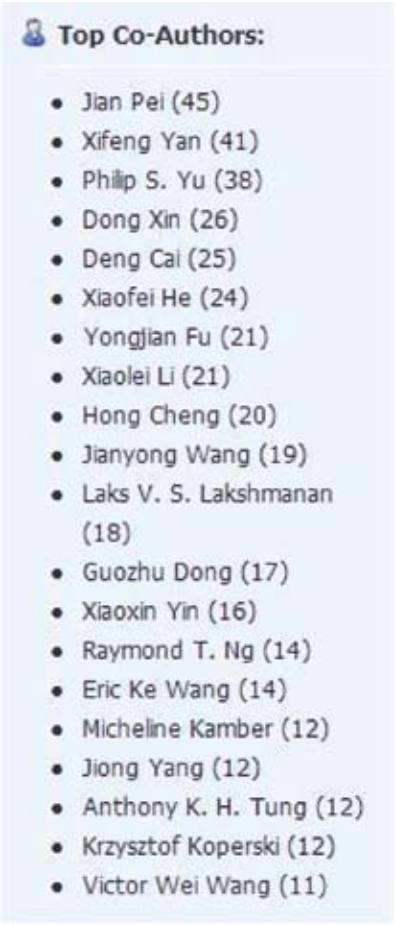

图5 韩家炜的合作者

DBLP VIS考虑三种实体类别。

1) Persons：代表一篇文献的作者。同名作者用数字后缀区分；

2) Streams：标识文献所发表的期刊或会议；

3) Keywords：从文献标题中提取的关键词。

DBLP VIS还考虑上述三种实体之间的关系，包括作者合作关系（person→person）在内的7种关系。

下面以作者合作关系为例详细说明。在作者合作关系图中，从合作时间、合作次数方面展示了核心作者与其合作者之间的关系，如图6所示。

图中的处于同心圆圆心位置的节点代表核心作者，分布在不同半径的圆周上的节点代表与核心作者共同发表文献的合作者。通过对节点进行渲染，核心节点与合作者之间的关系可以直观地表现出来。

1) 比例图：代表该作者与核心作者共同发表的文献占其总发表文献数量的比例；

2) 节点直径：代表该作者在相应时间段内与核心作者共同发表的文献数量；

3) 距核心节点的距离：代表该作者与核心作者的合作强度，合作强度越大，距离核心节点越近。合作强度由年均合作发文量表示。

在DBLP VIS的0.2版本中，节点之间的边可以根据不同时间段的合作强度渲染，用户可以直观地看出在哪一年作者的合作次数最多。同时还增加了节点过滤和搜索功能。

3.4 CDBLP

CDBLP由中国人民大学网络与移动数据管理实验室（WAMDM）开发，目标是建立一个“以人为本”，即以作者为中心来展示计算机类中文文献的集成数据库系统，从而为用户提供权威的论文数据和方便的查询服务（www.cdblp.cn）。该系统集成了国内现有权威计算机期刊、会议的中文

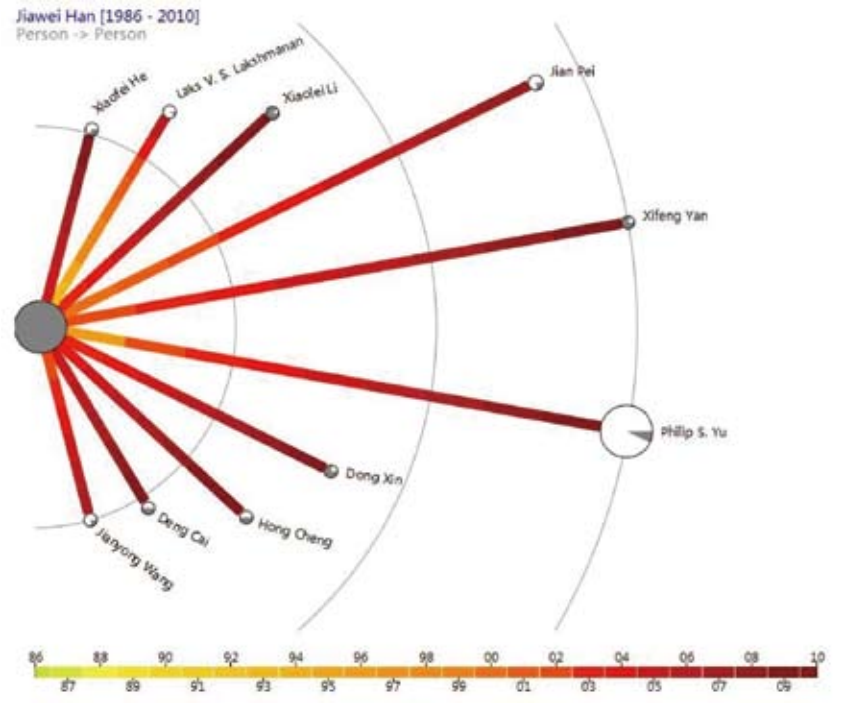

图6 DBLP VIS作者合作关系图

文献数据，为研究人员提供良好浏览的文献数据查询服务。该系统有以下特点：

1) 文献数据来源经过详细甄选，只收录国内计算机领域权威核心期刊和学术会议的数据；

2) 以作者为中心的学术成果检索，为每位作者提供集成化的检索结果，展示该作者发表的中文论文情况，并展示该作者的合作作者情况；

3) 提供基于作者名的精确匹配检索（Author Search）及基于作者名、论文题目、论文关键字、发表年份的模糊检索；

4) 基于来源的文献浏览功能，系统支持对已收录文献按期刊出处和发表会议浏览；

5) 能够精确地展示论文的基本信息，例如中英文的题目、作者、摘要、关键字等信息。

在作者详细信息展示页面中，除了按时间的发文量统计曲线和发表文献列表外，还提供了合作关系的可视化展示。其可视化功能具有如下特点（如图7所示）：

1) 在合作关系图中，只展示合作关系最紧密的21个合作者。合作者与核心作者的距离与合作次数成反比；

2) 作者详细信息展示。将鼠标悬浮在代表某一作者的节点时，将展示该作者的详细信息，如作者单位，系统收录论文数量，合作作者总数等。若用户对该作者感兴趣，则可以点击相应的链接进入其详细信息页面；

3) 合作论文展示。单击节点之间的边时，将显示这两个作者合作撰写的论文（最多显示六篇）。

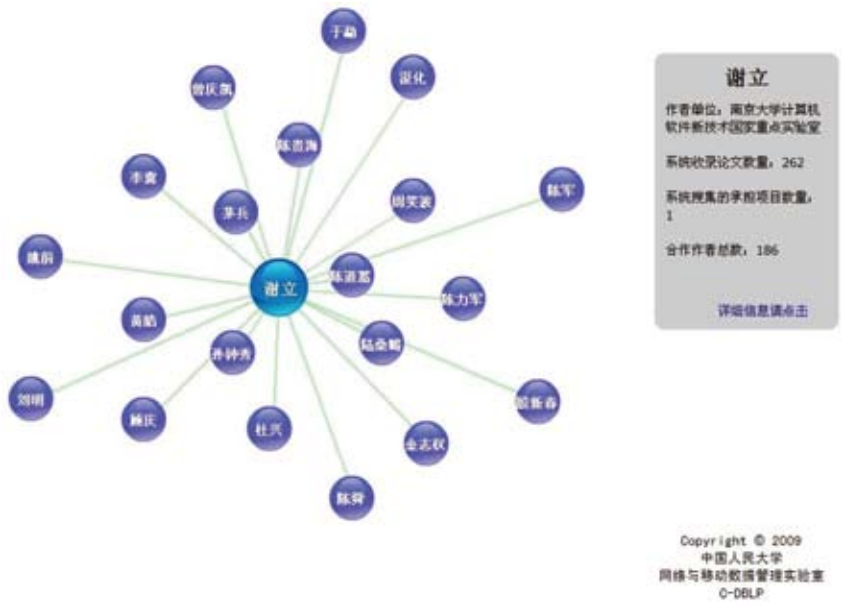

图7 CDBLP合作关系图

2009年3月，CDBLP正式推出合作关系的可视化展示，直观友好的动态图形化界面为用户提供更丰富的信息。在以后的几个月中，系统新增重名区分功能，集成了作者的相关图片并在搜索结果页面提供文献BibTex信息展示。

3.5 Web of Science

通过一篇文章的参考文献、引证文献、相关记录可以了解这篇高影响力论文的课题基础、最新发展趋势以及交叉学科的研究成

果。科学研究是一个在“继承”中“创新”的过程。引证关系图（Citation Map），以其特有的动态图形界面，揭示了科学文献间的相互继承关系。利用引证关系图，可以回溯某一研究文献的起源与历史，或者追踪其最新进展，及其对交叉学科和新学科的发展研究的重要参考价值。

图8展示了某篇文献的引证关系图。

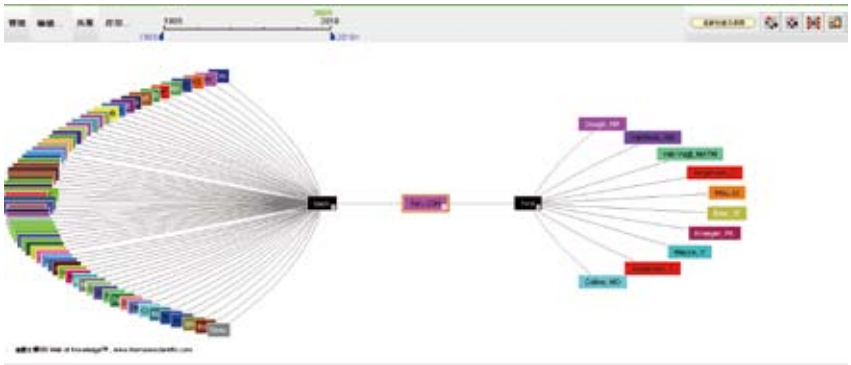

图8 引证关系图

3.6 GoPubMed

GoPubMed是基于PubMed的检索工具（http://www.gopubmed.org/web/gopubmed/），通过将用户输入的关键词提交给PubMed，并对PubMed的返回结果进行统计分析，对所有检索结果或单独对分类类目中的术语在PubMed中检索得到的文献进行统计分析，包括年代分布、核心作者、核心期刊、作者分布可视化地图等。其具体功能包含：

1) 主题词统计：统计发表文献最多的前20个主题词或关键词，可以看出主要研究领域、学科主题；

2) 作者发文量统计：统计发表文献最多的前20名作者，同时可以直接查看每名作者发表的文献，可以了解本领域的权威专家，可以重点关注其研究方向，了解该领域的研究重点；

3) 重要期刊分析：统计发表相关文献最多的前20种期刊，帮助用户了解该领域内的重要期刊；

4) 出版年代统计：统计近20年发表文献的年度分布，并以图表的形式展示；

5) 国家/地区发表文献分布图：在世界地图上用红点标示发表相关文献的国家和城市，用户可以了解哪些国家或城市对该领域投入的研发力量最多。

图9为以SARS病毒为检索词的国家/地区发表文献分布图。从图中可以看出中国和日本的科研产出最为突出。

图9 SARS Virus研究国家/地区分布图

3.7 文献分析工具对比评价

以上对传统的文献分析工具和当今流行的可视分析工具进行了简单介绍，下面将按照其提供的具体功能对软件进行对比评价，见表1。

4 可视分析应用实践

在国家“十一五”科技支撑计划“科技文献信息服务系统关键技术研究及应用示范”项目中，北京邮电大学通信软件工程中心复杂网络小组设计开发了可视分析组件——VisLib，已经应用于中国科技分析评价服务平台（http://168.160.200.61/form/help/Introduction.aspx）。该组件的目的是构建具有可视分析功能的独立软件模块，侧重于布点算法和对网络图形的渲染，为用户提供友好的访问方式和快速灵活的部署方案。

其突出的特点是：

1) 网络图形的渲染效果。相对于CDBLP和ArnetMiner提供的交互性相对较弱的网络图形展示功能，VisLib实现了基于Flex框架的网络分析与可视化模型，充分利用Flash技术的交互性和图形渲染功能，为用户提供更强大的展示效果和更友好的访问方式。例如，用户可以选择不同的布局算法从不同角度观察网络结果，通过设置限制条件，对网络图形中的节点和关系进行过滤等。

2) 独立的部署方案和可视分析服务。VisLib实现了基于Http协议、SOAP协议和AMF协议在内的多种数据交互方式，完全独立于系统的底层数据源和系统的实现平台，如Java、C#等，用户可应

表1 各种文献分析工具功能对比评价图

功能/工具名称	文献计量学统计	重名作者处理	专家挖掘	网络可视化
SCI ISTP	☆☆☆☆☆		☆☆☆	☆☆☆
EI	☆☆☆☆☆		☆☆☆	
ScienceDirect Scopus Scirus SpringerLink	☆☆☆☆☆		☆☆☆	
中国学术期刊全文数据库 万方数字资源	☆☆☆☆		☆☆☆	
APS IEEExplore CA LexisNexis IIPA	☆☆☆☆☆		☆☆☆	
ACM Digital Library	☆☆☆☆		☆☆	
DBLP	☆☆☆	☆☆☆	☆☆	☆☆☆☆☆
PubMed	☆☆☆☆		☆☆☆	
CiteSeerX	☆☆☆		☆☆	
Google scholar	☆☆☆		☆☆	
ArnetMiner	☆☆☆	☆☆☆☆	☆☆☆☆☆	☆☆☆☆
微软学术搜索	☆☆☆☆		☆☆☆☆	☆☆☆☆☆
CDBLP	☆☆☆	☆☆☆☆	☆☆	☆☆☆☆
GoPubMed	☆☆☆☆		☆☆☆	☆☆☆

用VisLib组件快速实现可视分析功能。除此之外，VisLib也可以作为独立的可视化服务发布，用户只需实现数据访问接口便可通过VisLib提供的可视分析服务对关系数据可视化。

如图10所示，可视分析界面的上部是针对于网络中的节点、关系过滤设置，调整图形渲染属性的面板。图10展示的是某一位作者的个人中心网络，同时提供针对于网络中节点信息的展示功能，如弹出的对话框，显示了用户所选择作者的个人信息。用户可双击某一作者查看其个人中心网络。

图11展示的是某篇文献的引证树，该文献的参考和引证文献分别按时间顺序分布在其左右，并对不同学科类别的文献进行渲染，用户可以直观地了解学科间的参考引证关系。同时，通过对时间信息的调整，用户可以查看某一时间段内的参考引证关系。

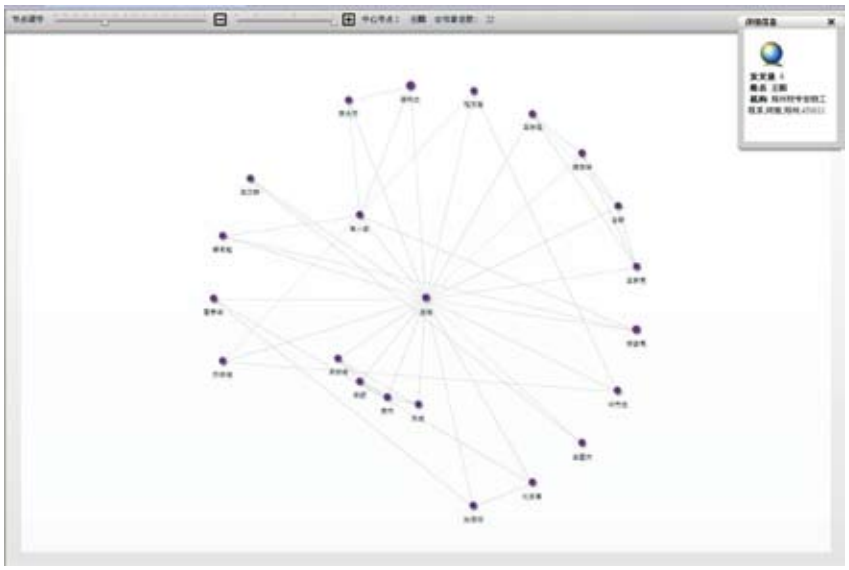

图10 作者个人中心网络

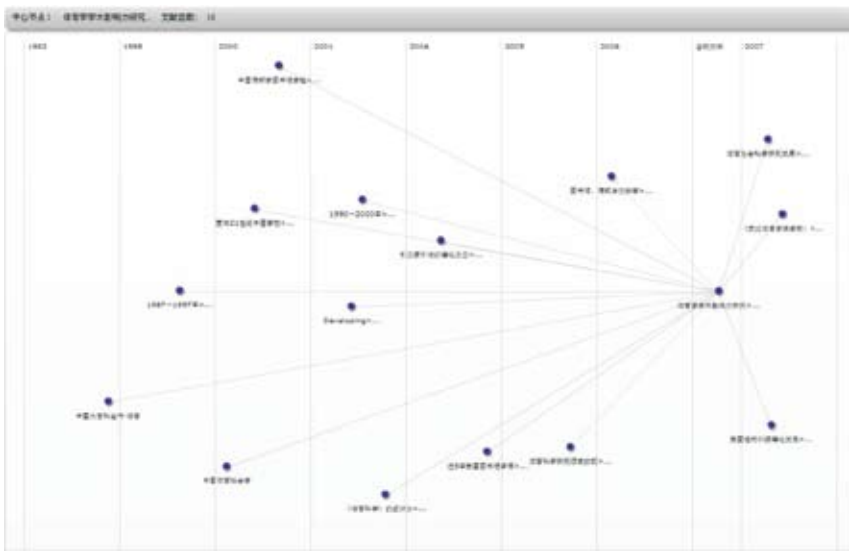

图11 文献引证树

5 文献分析工具发展趋势

通过比较分析国际上流行的文献分析工具，其功能向着两个方向发展：一种是深入探究文献领域内的实体关系，如科研作者、科技文献、科技会议等，相对于文献计量学的统计数据，结合链接关系和数据挖掘的分析结果更客观、公正。代表工具为ArnetMiner和微软学术搜索。一种是提供对关系数据的可视化展示，比起传统文献检索工具提供的表格和柱状图，网络图形更能够提高用户发现数据特征的能力。总结文献分析工具发展趋势如下：

1) 软件服务的概念

SaaS（Software-as-a-Service）是在21世纪开始兴起的一种完全创新的软件应用模式。众多的分析工具都不仅满足于只提供给用户可运行的软件系统，而尝试将软件功能作为一种服务发布出去。在这方面，ArnetMiner已经提出了相应的解决方案。ArnetMiner将检索和挖掘都作为Web Service服务发布，使用户可以将其提供的服务无缝地融入其他的系统中。

2) 灵活的体系架构

很多工具软件的现有版本和历史版本的体系结构相差很大。比如CiteSeerX通过重构其原有的数据模型来应对文献数据的快速增长，并在新系统中融入了诸多Web 2.0技术，更加贴合用户的需求。信息技术是不断向前发展的，如何将新技术快速应用到系统中，就需要系统架构师们更多地关注模块的重用性和可扩展性，设计灵活的体系架构以应对技术和需求带来的冲击。

3) 强大的计算资源

随着电子文献信息的爆炸式增长，如何在较短时间内完成数据清洗、格式转换以及数据挖掘算法已经成为了困扰许多工程技术人员的难题。网格技术的应用以及近年来提出的云计算概念，通过整合多台普通计算机成为强大的计算资源，并设计适合并行分布式运算的计算框架，如MapReduce[11]等，来解决因数据规模庞大而带来的计算难题。在这方面，ArnetMiner以用MapReduce框架实现了大规模社会网络的影响因子计算[12]。

6 总结语

本文首先讨论了传统文献检索工具的特点及其存在的局限性，并以当今常用的检索工具为例具体介绍；接着分析了数据挖掘和可视化技术带给文献分析领域的全新视角，并列举多种当今流行的可视分析和挖掘工具，重点介绍了具有代表性的ArnetMiner和微软学术搜索，展示其提供的强大功能、具体应用场景和发展路线；随后介绍了北京邮电大学通信软件工程中心开发的可视分析组件及其相对于现有可视分析模块的特点，并展示了其在国家科技支撑项目中的实用场景；本文最后展望了基于Web的文献分析工具的进一步发展方向。

参考文献

[1] HERR B.W. Designing Highly Flexible and Usable Cyberinfrastructures for Convergence [J]. Annals of the New York Academy of Sciences, 2007(1093):161-179.
[2] LI H, COUNCILL I, LEE W, GILES C. L. CiteSeerx: an architecture and web service design for an academic document search engine [C]// Proceedings of the 15th international Conference on World Wide Web, 2006.
[3] ASSENT I, KRIEGER R, MÜLLER E, SEIDL T. VISA: visual subspace clustering analysis [J]. SIGKDD Explor. Newsl., 2007,2(9):5-12.
[4] CHEN C M. Top 10 Unsolved Information Visualization Problems [J]. IEEE Computer Graphics and Applications, 2005,25(4):12-16.
[5] ZHANG J, TANG J, LI J. Expert finding in a social networks [C]// Proceedings of Database Systems for Advanced Applications, 2007.
[6] BRIN S, PAGE L. The Anatomy of a Large-Scale Hypertextual Web Search Engine [C]// Proceedings of the 7th international conference on World Wide Web, 1998.
[7] ZHANG D, TANG J, LI J, WANG K. A Constraint-Based Probabilistic Framework for Name Disambiguation [C]// Proceedings of the Sixteenth Conference on Information and Knowledge Management, 2007.
[8] NIE Z, ZHANG Y, WEN J-R, MA W-Y. Object-Level Ranking: Bringing Order to Web Objects [C]// Proceedings of the 14th international World Wide Web conference, 2005.
[9] ZHENG J. NIE Z. Architecture and Implementation of Object-Level Vertical Search [C]// Proceedings of the international Conference on New Trends in information and Service Science, 2009.
[10] KLEINBERG J. Authoritative Sources in a Hyperlinked Environment [C]// Proceedings of the ACM-SIAM Symposium on Discrete Algorithms, 1998.
[11] DEAN J, GHEMAWAT S. MapReduce: simplified data processing on large clusters [J]. Commun. ACM, 2008,51(1):107-113.
[12] TANG J, SUN J, WANG C, YANG Z. Social Influence Analysis in Large-scale Networks [C]// Proceedings of the Fifteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2009.

作者简介

田宏桥，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。
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

Survey on Web-Based Tools for Scientific Literature Analysis

Tian Hongqiao, Wu bin / Beijing key Laboratory of intelligent Telecommunication Software and Multimedia, Beijing University of Posts and Telecommunications, Beijing, 100086

Abstract: With the electronic citation information explosion in recent years, scientific research institutions have invented a series of web-based tools to help researchers find hot topic and research trend out of mass data. Data mining and information visualization technology play an important role in the process of knowledge discovery and analysis. Tools integrated with these two features have been developed for researchers. We will first introduce the traditional analysis tools and limitations they are confronted with; after that, advanced analysis tools with their functions and features will be presented in detail; finally, we introduce the visualization module we developed and its usage scenarios, and propose the future developing directions of web-based tools.

Keywords: Web technology, Scientific literature analysis, Data mining, Information visualization

（收稿日期：2010-05-31）