



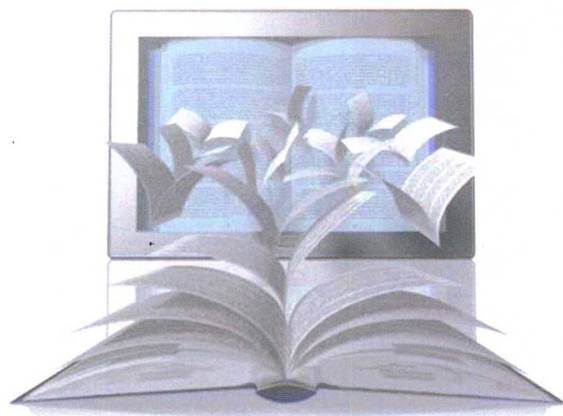
大数据时代 图书情报服务研究

DASHUJU SHIDAI

TUSHU QINGBAO

FUWU YANJIU

主编 谢 威



北京邮电大学出版社
www.buptpress.com

大数据时代图书情报服务研究

主 编 谢 威



北京邮电大学出版社
www.buptpress.com

内 容 简 介

学术交流是学会的立会之本,是学会工作的主旋律。北京科学技术情报学会紧密围绕科技情报服务,积极搭建不同形式的学术交流平台,把学术交流与科技发展中的重大问题紧密结合起来。2015年北京科技情报学会学术年会主题为“科技情报服务趋势与创新”。论文集内容包括:对当前科技发展形势下科技情报服务的趋势与发展空间研究;大数据环境下情报信息服务模式研究;图书馆服务创新与实践;科技情报工作创新经验交流;科技情报服务对重大问题应急(快速)反应能力决策咨询。充分反映了近两年我国科技情报服务领域科技工作者的最新理论方法研究,以及在实际工作中的实践与创新经验。

图书在版编目(CIP)数据

大数据时代图书情报服务研究 / 谢威主编. -- 北京:北京邮电大学出版社, 2015.12

ISBN 978-7-5635-4623-7

I. ①大… II. ①谢… III. ①情报服务—文集 IV. ①G358-53

中国版本图书馆 CIP 数据核字(2015)第 293056 号

书 名: 大数据时代图书情报服务研究

著作责任者: 谢 威 主编

责任编辑: 张珊珊

出版发行: 北京邮电大学出版社

社 址: 北京市海淀区西土城路 10 号(邮编:100876)

发行部: 电话: 010-62282185 传真: 010-62283578

E-mail: publish@bupt.edu.cn

经 销: 各地新华书店

印 刷: 北京九州迅驰传媒文化有限公司

开 本: 787 mm×1 092 mm 1/16

印 张: 22.25

字 数: 571 千字

版 次: 2015 年 12 月第 1 版 2015 年 12 月第 1 次印刷

ISBN 978-7-5635-4623-7

定 价: 48.00 元

· 如有印装质量问题,请与北京邮电大学出版社发行部联系 ·

编 委 会

主任委员：谢 威

副主任委员：吴晨生 刘景利 孟宪学 夏勇其

委 员：（按姓氏笔画）

初景利 代根兴 高淑萍 刘素清 刘彦君

李 莉 吕华侨 李长芹 李志男 苗润莲

吴 循 王 强 孙素芬 肖 雯 扈铁梅

赵世华 熊 丽

前言

2015年是全面深化改革关键之年,也是“十二五”规划的收官之年。随着全球科技革命及产业变革的新形势和新格局,我国经济发展进入了新常态,国家将创新驱动发展战略摆在了更加突出的位置,强调推动以科技创新为核心的全面创新。创新是推动一个国家和民族向前发展的重要力量,也是推动整个人类社会向前发展的重要力量。

在科技服务层面,国务院于2014年10月印发了《关于加快科技服务业发展的若干意见》,该指导意见指出要重点发展研究开发、技术转移、检验检测认证、创业孵化、知识产权、科技咨询、科技金融、科学技术普及等专业技术服务和综合科技服务,提升科技服务业对科技创新和产业发展的支撑能力。围绕科技创新全链条的科技服务体系将在指导意见之下,大大地提高科技创新能力,提升国际竞争力,使科技服务成为促进科技经济结合的关键环节和经济提质增效升级的重要引擎。

在决策支撑层面,中共中央办公厅、国务院办公厅于2015年1月印发了《关于加强中国特色新型智库建设的意见》。该指导意见就发展我国智库的重大意义、指导思想、基本原则、总体目标进行了阐述,并对构建中国特色的新型智库发展新格局提出发展意见。该指导意见对健全科技情报决策支撑的制度保障体系、加强决策支撑的组织领导、深化管理体制改改革做出了明确的指导意见,为科技情报服务体系建设提供了软环境的支撑条件。

在国家战略层面,中国政府网于2015年3月公布了《中共中央国务院关于深化体制机制改革加快实施创新驱动发展战略的若干意见》。该《意见》从国家层面对总体思路和主要目标提出了若干指导意见。面对全球新一轮科技革命与产业变革的重大机遇和挑战,面对经济发展新常态下的趋势变化和特点,为推动以科技创新为核心的全面创新,必须把中国特色的新型科技情报服务作为一项重大而紧迫的任务,从而加速以科技创新为核心的全面创新。

在这些意见的引导之下,我们需要认识、适应、引领科技情报创新服务的新常态,围绕这些指导意见,对近几十年来北京科技情报服务取得的成就与存在的问题、实施情报服务创新驱动发展的瓶颈与对策展开深入研究和广泛交流。

在这个科技情报服务科技创新、科技创新引领万众创新创业的黄金时代中,为促进北京科技情报服务发挥创新引领作用,推进科技情报的创新型研究,尤其是促进科技情报在各种新兴产业中的应用,北京科学技术情报学会在科技创新中心城市北京举办主题为“科技情报服务趋

势与创新”的论坛。作为中国科技情报界最国际化、最重要的专业交流平台之一,本次论坛围绕科技情报的发展趋势、科技情报服务创新、情报信息服务实践工作等范畴,为大家提供一个专业化的学习、交流、讨论与分享平台。该论坛将借助北京科学技术情报学会作为北京市科技情报服务平台的多年服务经验与优势,吸引全国优秀的情报专家、企业相关人士、情报从业者、产业与市场研究人士以及学者。这场“科技情报服务趋势与创新”的交流盛宴将为各政府部门、企业人士及创业者提供千载难逢的学习、提升和拓展视野的机会,让与会者真正享受科技情报业界的思想大餐。本次论坛共收到论文 70 余篇,这些论文涵盖了科技情报理论与实践的多个方面,包括情报信息、计算机网络、图书馆、综合服务等,具有较高的学术和实践价值,体现了我市科技情报工作在这一年来的最新进展。

这是一个创新驱动发展的时代。创新的过程充满了艰辛,创新之道,情报先行。任何一种创新活动都离不开情报信息的服务,不知道谁是创新的启动者、哪里是科技的发展方向、如何在技术转移的过程中创新?北京科学技术情报学会将努力发现那些能够服务创新乃至引领创新的情报服务理论、方法、实践、利益相关者和资源。我们在服务政府的同时,为政府与企业之间、科技成果与市场之间搭起了一座良好的沟通桥梁,积极开展多种形式的科技活动,加速以情报服务为支撑的科技创新,推动以科技创新为核心的全面创新,最终形成情报服务创新、创新引领发展的新常态。

北京市科学技术情报学会理事长

目 录

第一篇 情报信息类

大数据分析技术与工具在情报学中的应用研究

..... 李梦茹 董 洁 孙若丹 申红艳 扈铁梅(3)

基于文献计量的北京市农业科学领域科技创新能力测度计算及其影响分析

..... 刘 如 刘彦君 李荣(11)

基于专利信息的北京科技竞争优势分析 ... 侯元元 李楠欣 张 源 王需虹 吴晨生(17)

基于发明专利的北京生物技术研究现状分析 黄裕荣 侯元元 蔚晓川 高子涵(27)

开启情报研究的“全脑”——对知识管理应用于情报研究工作的思考 ... 赵 莉 王 灿(38)

基于文献分析的大数据时代情报学变革 ... 李梦茹 董 洁 孙若丹 申红艳 扈铁梅(44)

预测性情报服务——以科技信息风险交流研究为例 李小燕(49)

从大数据分析中获取高价值情报的基本方法和步骤研究 李志男 邵 颖(54)

大数据背景下科技情报创新工作策略研究 刘恩涛 张赛男(60)

国内情报研究机构核心情报能力建设案例研究——兼谈北京市科学技术情报研究所的情报

能力建设 刘彦君 李 荣 刘 如 王 康(67)

“大数据”环境下科技情报服务新模式探析 孟海滨 王玉峰 阳沛湘 赵东升(75)

大数据时代企业竞争情报系统构建 ——以中国化工信息中心竞争情报平台为例

..... 文淑美 石立杰(80)

大数据下的情报信息服务:需求与挑战..... 王玉峰(88)

大数据环境下中医药科技查新的发展与创新 张素娟(94)

大数据背景下科技查新工作面临的机遇、挑战与对策研究
..... 张 炜 杜 贇 凡庆涛 陈 茜(101)

第二篇 图书馆类

着眼我校重点学科建设,提升我馆学科化服务保障能力 陈 平(109)

高校图书馆与大学生心理需要的契合..... 高淑芳 赵世华(114)

以 JCR 为代表的传统引文分析与 Altmetrics 在期刊评价中的关联性分析 李 峰(118)

大数据环境下科研机构图书馆信息服务研究 李铁虎 夏治强 席果蕊(124)

高校图书馆精细化阅读推广探讨..... 秦疏影(129)

关于军事理论科学数字图书馆数据存储备份的几点思考..... 司文委(139)

高校图书馆服务于大学生创业活动的探索研究..... 王开颜(145)

服务仍然是信息时代图书馆主题——论信息时代图书馆服务的困境与对策..... 王黎珩(152)

大数据时代图书馆信息服务模式研究..... 王丽敏(157)

基于 SCI 论文的植物转基因育种领域发展态势分析
..... 颜志辉 郑怀国 赵静娟 串丽敏 张晓静 谭翠萍 孙素芬(163)

万众创新环境下高校图书馆对企业的科技情报服务探索..... 张丰智(172)

基于特色专题数据库建设的信息推送服务探析——以国防大学图书馆为例..... 智 慧(177)

第三篇 综合类

北京地区生物产业领域院士群体关注热点现状调研——基于网络内容分析法
..... 李 荣 董晓晴 赵俊超(183)

北京市计算技术发展趋势分析..... 高子涵 乔 婧 侯元元 黄裕荣(191)

基于 Internet 的传染病监测系统研究进展..... 尉景辉 王松俊 赵东升(200)

基于模糊综合评判方法的术语关系识别·····	殷希红 李 辉 张运良(209)
我国空气质量标准执行现状及与国外标准比较研究 ·····	董 洁 李梦茹 孙若丹 申红艳 扈铁梅(217)
区域一体化背景下京津冀旅游业低碳化发展对策 ·····	李 梅 苗润莲 张惠娜 徐铭鸿(227)
美、英、韩主要城市发展低碳交通的经验对于北京的借鉴 ·····	刘平平 凡庆涛 陈 茜 王倩倩(235)
基于 DMC 的电动汽车电池技术竞争态势分析 ·····	刘志芳 贺楚涵(241)
北京市测量与测试技术现状浅析·····	乔 婧 高子涵 侯元元(254)
北京市科研机构内外部环境分析——以 S 研究所为例 ·····	王 楠 贺 杰(263)
北京研发产业在京津冀协同发展中的机遇与路径·····	邵 颖 李志男 潘锐焕(269)
以色列产业结构特点及对建设全国科技创新中心的启示 ·····	申红艳 扈铁梅 李梦茹 董 洁 孙若丹(275)
东京人口集聚过程及原因分析·····	申峥峥 于怡鑫 吕华侨(282)
北京建设国际科技创新中心的战略思考 ·····	孙若丹 扈铁梅 董 洁 李梦茹 申红艳(292)
“大数据”时代下首都新型智库建设的策略研究·····	靳晓宏 王 强 付 宏(297)
公共危机事件的网络舆情预警机制研究·····	燕 娜 张云倩 刘平平(301)
北京市青少年科技创新人才培育现状研究·····	杨濛濛 叶浅草(309)
北京地区科普创作队伍现状调查及发展研究报告·····	叶浅草 杨濛濛 贺楚涵(315)
重视颠覆性技术 促进企业创新 ·····	扈铁梅 孙若丹(328)
大众参与创新创业模式的研究进展——以大学生为例 ·····	张玉娟 张 微 郭 昱 侯立宏(334)
论技术预见对科技创新系统发展的作用·····	周京艳 王世雯 杨 萍(339)

A decorative border featuring stylized flowers, leaves, and swirling scrollwork, framing the central text box.

第一篇 情报信息类

大数据分析技术与工具在情报学中的应用研究

李梦茹 董 洁 孙若丹 申红艳 扈铁梅

(北京市科学技术情报研究所, 北京, 100048)

摘 要: 大数据时代的到来, 给情报研究带来了机遇和挑战。本文分析了大数据环境下情报研究应用的新技术, 如数据挖掘、可视化、语义处理等等, 并总结了与之相对应的新工具, 如 Weka、Sitespace 等, 探索大数据情报研究技术与工具的新发展, 寻找情报分析的新模式, 以期能为促进情报研究理论和实践的发展有所启示。

关键词: 大数据; 情报学; 大数据技术; 大数据工具

前 言

在大数据背景下, 从技术上看, 可视化、数据挖掘等计算机领域的技术, 为情报研究提供了有力的技术视角, 情报研究获得的知识反过来又给予其他技术领域的发展以引导。从工具上看, 很多大数据其他学科的分析工具, 广泛应用于军事情报、科技情报等领域, 用于情报分析的认知过程, 以指导情报分析及其工具的研发。

面对海量数据, 自动化的技术手段必不可少, 通过多种技术手段或采用不同的工具, 全方位地展示信息内容及其之间的关系, 从而避免产生信息的误读。可见, 在大数据时代, 情报研究需要多种新兴手段来加强其分析过程的科学性。

1 大数据技术在情报学中的应用

情报研究的发展走向, 决定了情报研究既不能仍然停留在定性分析上, 也不能仅仅靠简单的统计替代情报研究中的计算技术, 由此对情报研究技术提出了新的要求。美国 McKinsey Global Institute 在 2011 年 5 月发布了研究报告“大数据: 创新、竞争和生产力的下一个前沿领域”(Big data: The next frontier for innovation, competition, and productivity)。报告分六个部分, 其中第二部分讨论了大数据技术, 并围绕大数据分析技术、大数据技术和可视化三方面进行了阐述。在大数据分析技术中, 列举了 26 项适用于众多行业的分析技术, 包括聚类分析、众包、数据挖掘、自然语言处理、网络分析、预测建模、回归、可视化等。这些技术绝大部分是已有的技术, 也有部分是随着互联网的发展以及对大规模数据挖掘的需求, 在原有技术的角度发

展起来的。这些技术大致可以划分为大数据存储和处理技术、大数据查询和分析技术、大数据高级分析和可视化技术三大类。其中,前两项技术是大数据的基础技术,而大数据高级分析和可视化技术则是当前情报分析领域应用最多且应予以关注和深入研究的技术。大数据高级分析和可视化技术主要包括数据挖掘与高级分析、可视化分析与展示技术、语义处理分析。

1.1 数据挖掘与高级分析

数据挖掘(Data Mining)一般是指从大量的数据中通过算法搜索隐藏于其中信息的过程。数据挖掘通常与计算机科学有关,并通过统计学、在线分析处理、情报检索、机器学习、专家系统和模式识别等诸多方法来实现数据分析和数据库知识发现。数据挖掘的任务是从大量的数据中发现模式。根据数据挖掘的任务可分为多种类型,其中比较典型的有:关联分析、基于决策树或神经网络的分类分析、聚类分析、序列分析等。

大数据分析的理论核心就是数据挖掘算法,各种数据挖掘的算法基于不同的数据类型和格式才能更加科学的呈现出数据本身具备的特点,另外,数据挖掘的算法还能够更快速的处理大数据,也正是因为这些被全世界统计学家所公认的各种统计方法才能深入数据内部,挖掘出公认的价值。

从数据挖掘的含义看,它与情报研究有着天然的联系;从数据挖掘的方法看,有其特定的含义和实现过程,可以有效地解决情报研究的问题。但从目前的情报研究成果看,许多还仅仅停留在简单的频率统计、共词计算层次上,在知识发现的过程中,这些工作仅仅是数据挖掘的数据准备,还有待于更为深入的发掘。可见,数据挖掘能够也应该应用于情报研究领域,这不仅是数据挖掘应用扩展的结果,也是情报研究自身发展的需求。

1.2 可视化分析与展示技术

可视化分析(Visual Analytics)是一门通过交互的可视化界面来进行关联分析推理的科学,目的是帮助用户在大规模信息分散或数据结构不统一的复杂数据内容的基础上进行分析决策,并得出完整的分析图表。图表中包含所有事件的相关信息,也完整展示数据分析的过程和数据链走向。它不同于信息可视化,信息可视化关注计算机自动生成信息的交互式图形表示,以及图形表示的设计、开发及其应用,而可视化分析在此基础上加入了知识发现过程,关注自动分析方法及其选择,以及如何将最佳的自动分析算法与适当的可视化技术相结合,以达到辅助决策的目的。

可视化技术应用在情报分析中,可以大大提高情报分析的效果,是情报学领域一个较新的研究热点。信息可视化技术应用在情报研究中可以弥补传统方法的一些缺陷,对信息从一个全新的角度进行观察分析,发现以往的方法所不能发现的隐藏情报,并对其进行分析解释,得出有价值的结论得出对决策都有用的情报,从而大大提高情报分析的效率和效果。

大数据分析的使用者有大数据分析专家,同时还有普通用户,但是他们二者对于大数据分析最基本的要求就是可视化分析,因为可视化分析能够直观的呈现大数据特点,同时能够非常容易被读者所接受,就如同看图说话一样简单明了。

1.3 语义处理分析

语义是关于意义(Meaning)的科学。语义分析是审查源程序有无语义错误,为代码生成阶段收集类型信息。它的任务是对结构上正确的源程序进行上下文有关性质的审查,进行类型审查。语义技术提供了机器可理解或是更好处理的数据描述和程序,整合了自然语言处理、信息抽取、数据库技术等技术方法,旨在让计算机更好地支持处理、整合、重用结构化和非结构化信息。

核心语义技术包括语义标注、知识抽取、检索、建模、推理等。语义技术可以为信息的深层挖掘打好基础,即通过对各类信息的语义处理,在获取的富有语义的结构化数据上使用各种数据挖掘算法来发现其中的潜在模式。

2 大数据工具在情报学中的应用

对于大数据来说,最重要的还是对于数据的分析,从里面寻找有价值的数据帮助做出更好的决策。因此,基于以上技术和平台,将应用于情报学的大数据工具分为数据挖掘工具、可视化分析工具、语义引擎工具三大类。

2.1 数据挖掘工具

由于较少有专门针对情报研究领域研发的挖掘工具,现有情报分析通常借助于其他工具,不同工具的功能不同,这就导致常常同时使用好几个分析工具。这带来的问题是,因为工具的分割性,分析缺乏完整性,就导致潜在模式的丢失。这里主要介绍两种常用的数据挖掘工具 Weka 和 RapidMiner。

(1) Weka 数据挖掘的开源软件

WEKA 全名怀卡托智能分析环境(Waikato Environment for Knowledge Analysis),是一款免费的、非商业化的、基于 JAVA 环境下开源的机器学习以及数据挖掘软件。WEKA 作为一个公开的数据挖掘工作平台,集合了大量能承担数据挖掘任务的机器学习算法,包括对数据进行预处理,分类,回归、聚类、关联规则以及在新的交互式界面上的可视化。

2005 年 8 月,在第 11 届 ACM SIGKDD 国际会议上,怀卡托大学的 Weka 小组荣获了数据挖掘和知识探索领域的最高服务奖,Weka 系统得到了广泛的认可,被誉为数据挖掘和机器学习历史上的里程碑,是现今最完备的数据挖掘工具之一,并且 Weka 的每月下载次数已超过万次。

(2) RapidMiner 数据和文本挖掘的开源软件

RapidMiner 是世界领先的数据挖掘解决方案,在一个非常大的程度上有着先进技术。它数据挖掘任务涉及范围广泛,包括各种数据艺术,能简化数据挖掘过程的设计和评价。

RapidMiner 免费提供数据挖掘技术和库,100%用 Java 代码(可运行在大部分操作系统上),数据挖掘过程简单,强大和直观,可以用简单脚本语言自动进行大规模进程,多层次的数据视图,确保有效和透明的数据,有强大的可视化引擎,许多尖端的高维数据的可视化建模,并



图 1 Weka 官网

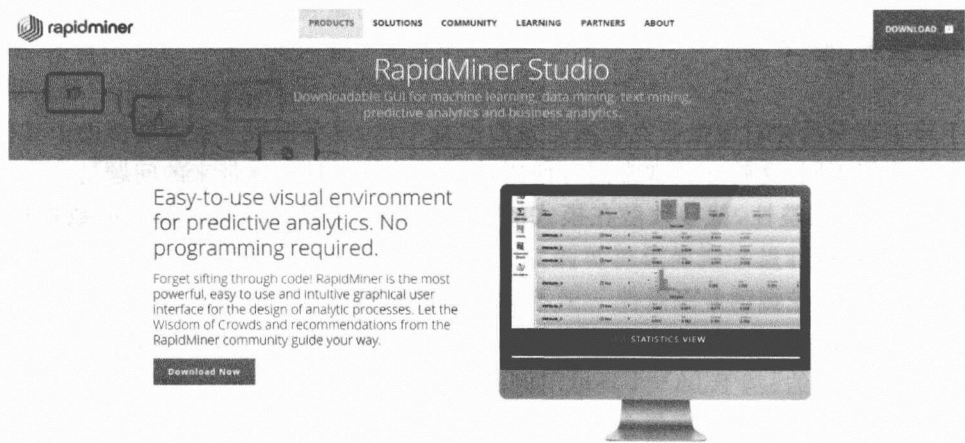


图 2 RapidMiner 官网

有 400 多个数据挖掘运营商支持。

目前耶鲁大学已成功地应用在许多不同的应用领域,包括文本挖掘,多媒体挖掘,功能设计,数据流挖掘,集成开发的方法和分布式数据挖掘。

2.2 可视化分析工具

目前的情报分析系统,虽然也提供了多种视图来揭示信息,但更多的是一种分析结果的呈现,分析人员无法了解分析方法、分析结果的局限性或者有效性。同时,现有的分析工具需要分析人员输入各种繁杂的参数,又缺乏对情报分析认知过程的支持,这就对使用人员的专业化

程度提出了较高的要求,增加了分析的难度。而可视化分析则可以较好地解决这一问题,它整合了多个领域包括采用信息分析、地理空间分析、科学分析领域的分析方法,应用数据管理和知识表示、统计分析、知识发现领域的成果进行自动分析,融入交互、认知等人的因素来协调人与机器之间的沟通,从而更好地呈现、理解、传播分析结果。这里主要介绍 Pajek、UCINET、Jigsaw 和 Citespace 这四种可视化软件。

(1) Pajek 信息可视化软件

Pajek 是一个特别为处理大数据集而设计的网络分析和可视化程序,是用于研究目前所存在的各种复杂非线性网络的有力工具。Pajek 在 Windows 环境下运行,用于带上千乃至数百万个结点大型网络的分析和可视化操作,并具有以下功能和特点。Pajek 可以同时处理多个网络,也可以处理二模网络和时间事件网络,还提供了纵向网络分析的工具。

(2) UCINET 信息可视化软件

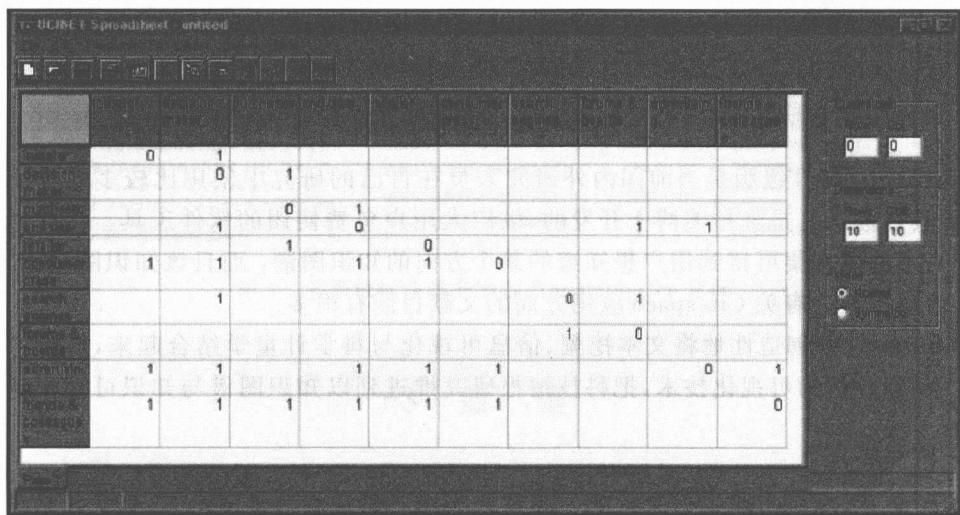


图3 UCINET 软件界面

UCINET 是目前最流行的社会网分析软件,社会网络分析法包括中心性分析、子群分析、角色分析和基于置换的统计分析等。

UCINET 是由加州大学欧文(Irvine)分校的一群网络分析者编写。UCINET 网络分析集成软件包括一维与二维数据分析的 NetDraw,还有正在发展应用的三维展示分析软件 Mage 等,同时集成了 Pajek 用于大型网络分析的 Free 应用程序。另外,该软件包有很强的矩阵分析功能,如矩阵代数 and 多元统计分析。它是目前最流行的,也是最容易上手、最适合新手的社会网络分析软件。

(3) Jigsaw 文档可视化分析系统

佐治亚理工学院的 John Stasko 等人应用 Pirolli 等人提出的情报分析概念模型,建立了一个名为 Jigsaw 的可视化分析系统,并将其应用于学术研究领域以及研究网络文领域,也说明了将可视化分析技术应用于情报研究的可行性。

Jigsaw 是一款利用文本挖掘算法建立各种文档可视化视图的软件,可生成文档的聚类图、时间轴、词树图等。建模是系统根据任务自动完成的,并产生相对模型的可视化图表,用户可自由调节图表的属性和外观,是一个典型的可视化分析系统。

(4) Citespace 可视化工具

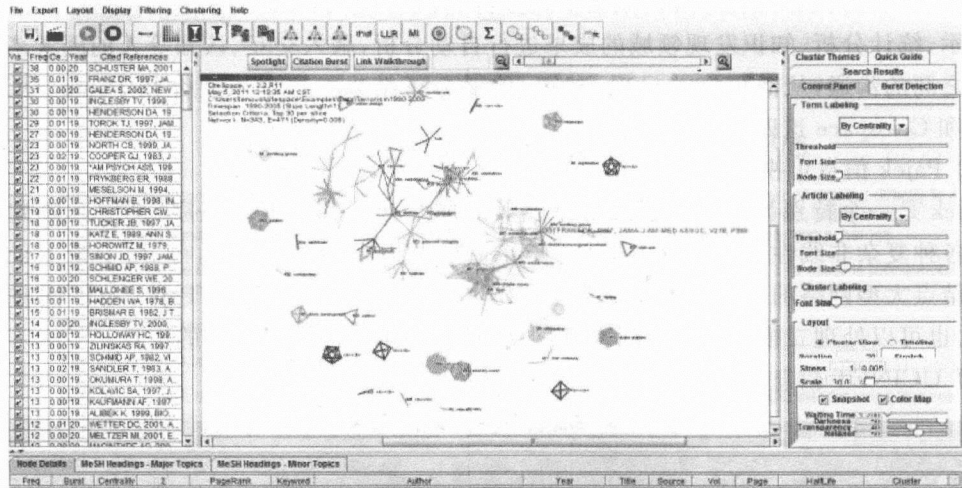


图 4 Citespace 软件界面

Citespace 及其升级版是当前国内外研究人员在自己的研究中使用比较多的科学知识图谱绘制工具。该工具是陈超美博士开发的、供广大用户免费使用的软件工具。该软件易于获取,基于一定的数据集可得到用户想知道的某个方向的知识图谱,而且该知识图谱稳定、可读性良好、信息丰富。有关 Citespace 应用方面的文献目前有很多。

CiteSpace 软件创造性地将文本挖掘、信息可视化与科学计量学结合起来,形成了适于多元、分时、动态分析的可视化技术,把科技情报研究推进到以知识图谱与知识可视化基础的新阶段。

2.3 语义引擎工具

从现有的情报研究实践和工具看,语义支持的缺失是一个普遍问题。对于传统的情报研究对象,如科技论文、专利等,有较为成熟的分析工具,但这些工具往往缺少深层次的语义支持。对于新型情报研究对象,如网络新闻、博客等,已有如动态监测科研机构等的系统工具。但总体来说还处于起步状态,目前较多的还是依赖人工筛选出所需信息,并整理成结构化的数据,同样也不利于大规模的数据分析。这些问题的存在,使得应用语义技术成为广泛需求及必然。

(1) 大规模语义计算平台 Annotation.cn

Annotation.cn 多领域标注服务平台系统主要包含词表管理模块、标注模块、用户管理模块三个模块。其中词表管理包含词表维护、类型管理、概念管理、概念查看四个功能;标注模块包含文档管理、文档标注两个功能;系统管理包含用户管理、角色管理功能。