

常用科学知识图谱工具实例对比^{*}

岳晓旭 袁军鹏 高继平 翟丽华 潘云涛

(中国科学技术信息研究所, 北京 100038)

摘要: 以Web of Science中近10年“三计学”领域的文献为实例, 遵循科学知识图谱绘制的一般流程, 对国内常用科学知识图谱工具Bibexcel、CiteSpace、SPSS、TDA和Ucinet在共词分析、作者合作和文献共被引等方面的异同进行对比, 总结它们的优点和不足, 为使用者提供参考。

关键词: 科学知识图谱; Bibexcel; CiteSpace; TDA; Ucinet

中图分类号: G301

DOI: 10.3772/j.issn.1673—2286.2014.05.011

1 引言

近年来, 科学知识图谱已成为图书情报学领域的一大研究热点。科学知识图谱是在对科学知识进行科学计量分析的基础上, 由相关绘图工具绘制而成, 可揭示科学知识的发展进程和结构关系^[1]。具体到操作层面, 就是对特定的文献数据集进行词频统计、文献共被引、作者合作、关键词共现以及时间序列等相关分析, 并将分析结果以恰当的图形展示出来。

由于科学知识图谱具有直观展示科学知识结构的特点, 自20世纪80年代出现以来就得到了广泛的关注^[2]。目前图书情报领域应用比较多的图谱绘制工具可分为两大类^[3]: 通用软件和专用软件。通用软件中, 如统计软件SPSS, 图谱绘制中常用到多维尺度分析、因子分析和聚类分析; 社会网络分析软件Ucinet和Pajek, 常用来展示知识结构; 此外还有GIS等相关可视化软件。专用软件包括文献计量软件Bibexcel和Histcite、图谱绘制分析工具CiteSpace和VOSviewer、文本挖掘软件TDA、Loet Leydesdorff系列软件等。在众多的知识图谱绘制工具中, 选取合适的工具, 绘制符合相关要求的知识图谱, 对于科研人员来说尤为重要。

本文基于科学计量分析的视角, 以“三计学”(科学计量学、信息计量学和文献计量学)领域的文献为实例, 对国内常用的知识图谱绘制工具Bibexcel、

CiteSpace、SPSS、TDA和Ucinet进行了应用层面的比较研究, 主要比较了各工具在对文献进行共词分析、作者合作和文献共被引等方面的异同, 以期能为科研人员在选择图谱绘制工具时起到一定的帮助。

2 工具简介与数据来源

科学知识图谱的绘制流程一般包括工具选择、数据来源选择、数据采集、数据处理、数据导入、参数设置、知识图谱绘制和知识图谱分析^[1]。首先将介绍绘制工具的类型、基本功能和数据来源。

2.1 工具类型和基本功能

结合文献[2-9]、文献[14-15]和具体使用, 对五种工具的类型和基本功能进行了介绍, 见表1。

2.2 数据来源

本文以Web of Science(WOS)的SCI-EXPANDED数据库作为数据源, 以“主题=Scientometrics OR Informetrics OR Bibliometrics”进行检索, 时间跨度为2004-2013, 文献类型为“ARTICLE OR REVIEW OR PROCEEDINGS

^{*} 本研究得到国家自然科学基金项目“论文零被引的时间演化规律、影响因素及其应用研究”(编号: 71373252)、国家软科学研究计划项目“国外科技人才发展趋势比较研究”(编号: 2012GXSSD127)资助。

表1 五种工具的类型和基本功能

功能 \ 软件		Bibexcel	CiteSpace	SPSS	TDA	Ucinet
类型		专用	专用	通用	专用	通用
数据处理功能	导入数据格式	文本、数据关联	文本、数据关联	数据关联	文本	数据关联
	数据清洗	无	有，但不易操作	无	有	无
数据分析功能	分析对象	论文、专利	论文、专利	关联数据	论文、专利	关联数据
	分析方法	计量分析、共词分析、共被引分析、引文耦合分析等	作者合作、共词分析、机构合作、共被引分析、引文耦合分析、时间序列、突发监测等	相关分析、因子分析、主成分分析、多维尺度分析、聚类分析等	聚类分析、关联分析、时间序列、突发监测等	中心性分析、子群分析、角色分析、聚类分析等
可视化展示	聚类展示	有，但比较弱	有	有	无	有
	图谱类型	节点链接图	节点链接图	树状图、饼状图	节点链接图	节点链接图

PAPER”，经过人工筛选，最终获取数据1028条，以TXT形式进行下载。检索时间为2013年12月16日。

由于SPSS和Ucinet只能导入矩阵格式的数据，在分析过程中，将TXT格式保存的数据直接导入Bibexcel、CiteSpace和TDA进行处理、分析，而对SPSS和Ucinet则导入的是Bibexcel分析过程中产生的共现矩阵，分析数据字段为关键词（DE字段）、作者（AU字段）和参考文献（CR字段）。

3 科学知识图谱绘制工具应用比较

本文尝试使用五种科学知识图谱绘制工具对同源文献数据进行关键词共现、作者合作和文献共被引三个方面的分析，分析五种工具在分析数据、绘制图谱过程中的异同，总结绘制工具的优缺点，同时对五种工具的分析结果进行验证。

3.1 数据预处理

（1）清洗数据

由于选取的五种工具中，只有TDA具有清洗数据的功能，首先将TXT格式数据导入TDA进行清洗，在选

用TDA自身的叙词表清洗数据之后，进行手工清洗，清洗规则主要包括同义词合并，单复数、大小写合并，去除特殊符号（如“-”、“’”）等。

而对于Bibexcel和Citespace，则是在产生分析结果后，由人工针对性的进行数据清洗，清洗规则同上。

（2）频次统计

五种工具中，Bibexcel和TDA具有统计频次功能。基于清洗后的数据集，使用两种工具分别进行统计分析，分析结果基本一致。统计结果表明，关键词共2062个，为了更好地揭示领域的研究热点，去掉“三计学”，即Scientometrics、Bibliometrics和Informetrics，选择频次大于等于5共91个；作者共2084个，选择频次大于等于5的共53个；被引用论文数23752篇，选择被引次数大于等于15的共84篇。由于篇幅限制，此处只列举了部分，具体如表2所示。

（3）构建矩阵

Bibexcel和TDA具有构建矩阵的功能，使用Bibexcel和TDA分别构建高频词、高产作者、高被引论文共现矩阵（共现矩阵为互相关矩阵），结果一致。为了绘制图谱的方便，将复杂的引文字段在矩阵和绘制的部分图谱中用序号表示。由于篇幅限制，此处略去相关矩阵。

表2 高频词、高产作者、高被引论文列表 (前10)

高频词	频次	高产作者	频次	高被引论文	频次
citation analysis	156	Egghe L	27	HIRSCH JE, 2005, PNATL ACAD SCI USA, V102, P16569	27
impact factor	80	Abramo G	22	MOED H F, 2005, CITATION ANAL RES EV,	22
h-Index	63	D'Angelo CA	21	GARFIELD E, 2006, JAMA-JAM MED ASSOC, V295, P90	21
Research productivity	39	Ho YS	19	GARFIELD E, 1972, SCIENCE V178, P471	19
Publications	36	Kostoff RN	16	EGGHE L, 2006, SCIENTOMETRICS, V69, P131	16
Research evaluation	35	Glanzel W	15	SEGLE N PO, 1997, BRIT MED J V314, P498	15
Research trend	29	Konur O	14	GRFIELD E, 1955, SCIENCE V122, P108	14
Scientific journals	25	Prathap G	13	SMALL H, 1973, JAM SOC INFORM, V24, P265	13
Journal impact factor	25	Rousseau R	13	PRICE DJD, 1963, LITTLE SCI BIG SCI	13
Web of Science	23	Leydesdorff L	13	PRICE DJD, 1965, SCIENCE, V149, P510	13

表3 四种工具绘制图谱的算法与展示方式对比

软件	算法			图谱表示方式
	相似度	特征词权重	聚类	
CiteSpace	Cosine、Jaccard、Dice	tf*idf、互信息	光谱聚类算法 (spectrum cluster)、最大期望EM聚类	作者字段, 外国人姓在前, 名在后; 中国人名在前, 姓在后; 节点的大小代表节点项目频次的高低, 节点间连线代表项目间关系; 不同类簇由不同颜色背景表示, 节点的多少可代表类簇的大小; 可通过增加标签赋予类名; 可通过参数的设置来改变图谱。
SPSS	欧式距离、切比雪夫距离、Cosine、WARD距离等	无	二阶聚类、K-均值聚类、系统聚类	作者字段, 外国人名在前, 姓在后, 姓简写; 中国人姓在前, 名在后, 名简写; 树状图形象地展示关键词聚类过程的层级关系, 通过设定距离阈值来确定聚类结果。
TDA	Pearson系数、Cosine、Max proportional	tf*idf	未公开	作者字段, 外国人名在前, 姓在后; 中国人姓在前, 名在后; 点的大小代表项目频次的高低, 连线代表项目间存在关系, 连线的粗细、虚实代表节点间关系的强弱。
Ucinet	共现频次	频次、中心性、K-core	系统聚类	作者字段, 外国人名在前, 姓在后, 姓简写; 中国人姓在前, 名在后, 名简写; 节点的大小代表节点的中心度的大小, 连线的粗细代表节点间关系的强弱; 绘制结果可进行调整, 以便图谱更为直观明了。

3.2 科学知识图谱的绘制与分析

五种工具中, 除Bibexcel绘图功能比较弱, 其余四种工具均可绘制相应的图谱。在绘制图谱的过程中, 不同工具采用不同算法处理数据、不同方式展示图谱。在对文献进行共词分析、作者合作分析和文献共被引分析及相关图谱的展示过程中, 四种工具所采用的算法及

图谱表示方法如表3所示。

将清洗后的txt格式的数据导入CiteSpace, 选择不同的字段, 采用Cosine计算相似度、tf*idf计算权重绘制不同的知识图谱。依次将三种共现矩阵导入SPSS和Ucinet, 在SPSS中, 选用Cosine计算相似度, 进行系统聚类, 绘制树状图, 由于选取的项目较多, SPSS绘制的图谱线性排列, 由于篇幅限制, 在下面截取部分具有代

表性的图; 在Ucinet中, 选择K-core分析, 进行可视化, 绘制网络结构图谱。将在TDA中生成的共现矩阵直接可视化, 生成知识图谱。

(1) 关键词共现结果与分析

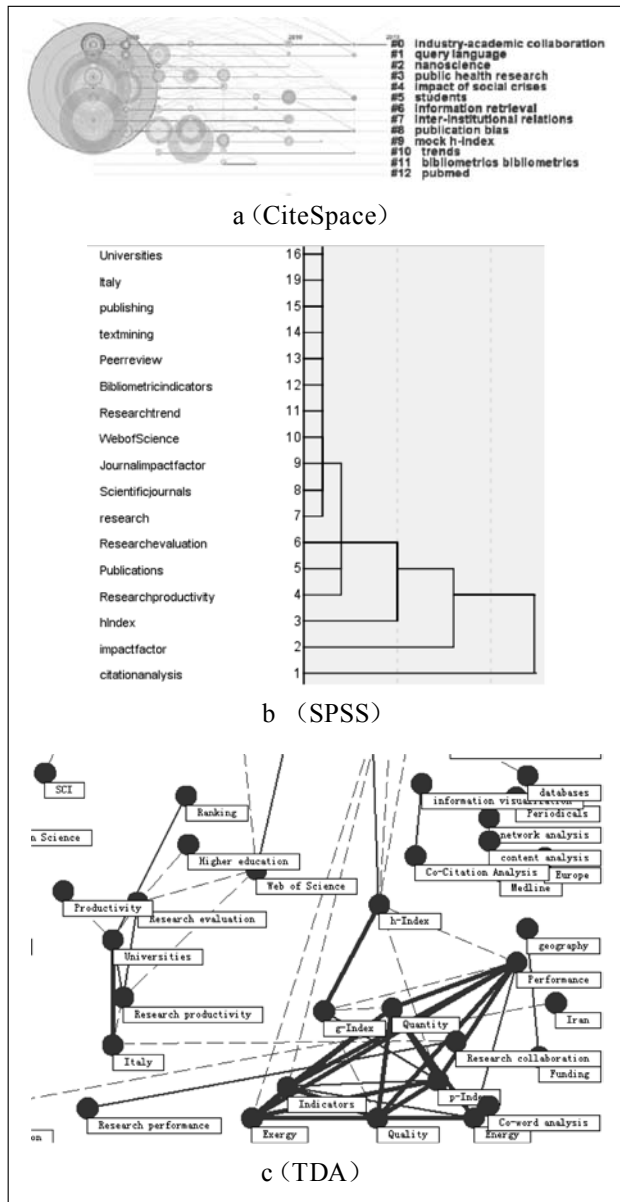


图1 共词分析可视化结果 (部分)

对比四种工具的分析结果, CiteSpace分析结果表明, 近十年“三计学”领域的研究热点集中在产学研合作 (Industry-academic collaboration)、h指数 (h-index)、信息检索 (Information retrieval) 等。CiteSpace得出的结果与其他三种工具的结果差别较大, CiteSpace是在自动聚类的基础上, 增加了自动标引功能, 即从施引文献中的标题、关键词、摘要部分通过

不同算法选择权重最大的词语, 用以表征整个聚类的含义。标签尽管是权重最大的词语, 但是依然无法完全表征整个聚类, 故建议聚类分析中增加专家干预, 选择最佳的标签词体现整个聚类。

TDA和Ucinet得出的研究热点一致, 因此只展示了TDA的结果。引文分析 (Citation Analysis)、h指数 (h-index)、影响因子 (Impact factor) 三个词, 在TDA和Ucinet的结果中成为核心词, 但在SPSS的分析结果中是孤立的, 未能与其他词聚为一类。总结原因, 以上三个词频次较高, 但与其共现的词可能较为分散。

四种工具所绘图谱具有不同特点: CiteSpace绘制的图谱中聚类标签并不具有一定的代表性, 适合做突发词的监测, 但不适合通过共词聚类来发现研究热点; SPSS绘制的图谱明显地展示了聚类过程和聚类结果; 但由于关键词集合较大且集中, 图谱不能很好地区分类别, 适合用于分析区分明显且较小的数据集; TDA和Ucinet绘制的图谱可形象地看出核心词和热点词群, 但类团区分不明显, 较适用于关键词共现。

(2) 作者合作结果与分析

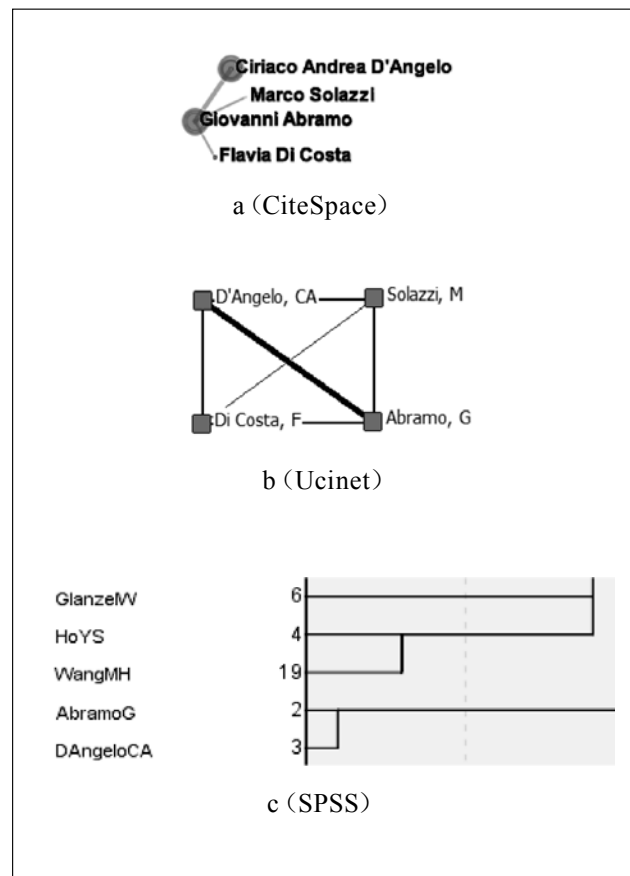


图2 作者合作可视化结果1 (部分)

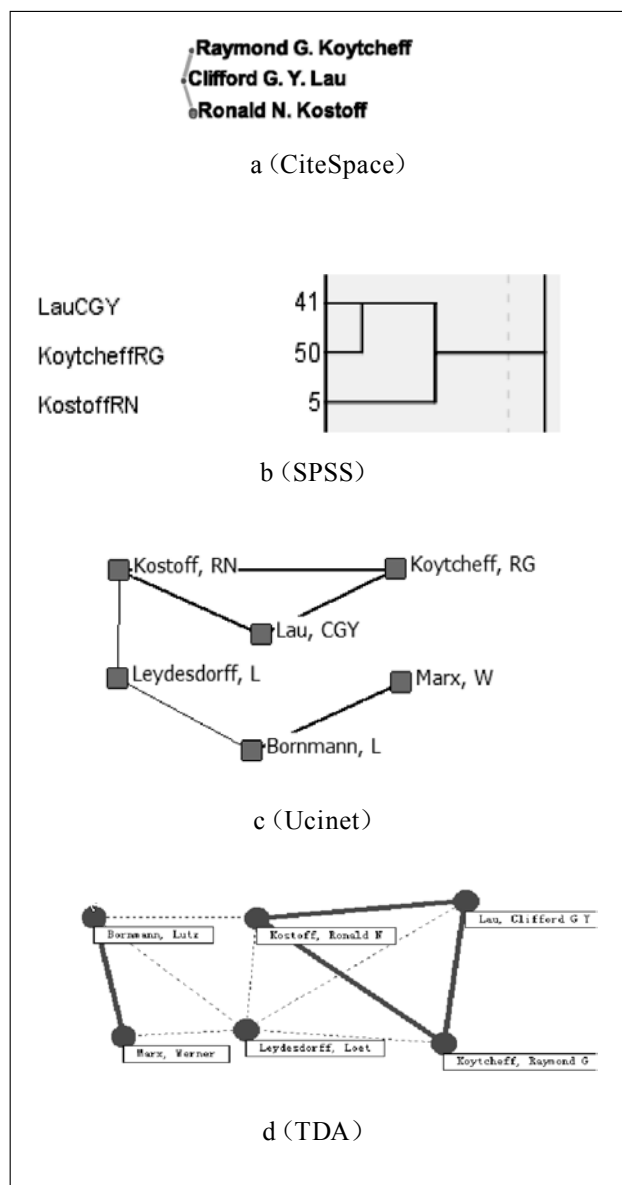


图3 作者合作可视化结果2 (部分)

由于Bibexcel、CiteSpace和TDA在分析AU(作者)字段上存在差异,因此四种工具展示的作者合作图谱存在差异。对比四种绘制工具所绘制的图谱,得出的学术团体基本一致,例如以D'Angelo,CA、Abramo,G为代表的学术团体(图2);以Lau, CGY、Kostoff, RN、Koytcheff, RG为代表的学术团体(图3)等。但一些作者在CiteSpace或SPSS中为独立作者,但在TDA和Ucinet中却和其他作者或团体联系在一起(比如作为两个不同团体的桥梁),例如Egghe, L、Leydesdorff, L、Rousseau, R、Glanzel, W等。总结原因,不同工具在参数设置和阈值的确定上可能存在不同。

在绘制的图谱中,CiteSpace绘制的图谱较好地揭示了该领域的核心作者、作者间的合作关系;SPSS绘制的图谱可揭示聚类团体;TDA绘制的图谱可形象地看出核心作者和作者群、作者间及作者群间的关系;Ucinet绘制的图谱形象地展示出了核心作者和作者间关系的强弱。CiteSpace、TDA和Ucinet都比较适合用于作者合作分析。

(3) 文献共被引结果与分析

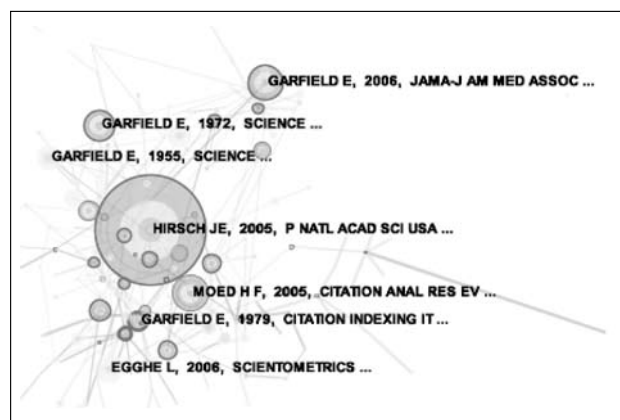


图4 CiteSpace文献共被引可视化结果

在四种工具绘制的文献共被引知识图谱中,只有CiteSpace绘制的文献共被引知识图谱效果较好;由于引文字段记录较长、节点较多,SPSS、TDA和Ucinet绘制的聚类结果不能揭示出高被引论文及高被引论文聚类结果。

近十年“三计学”领域的基础文献主要包括“HIRSCH JE, 2005”、“MOED H F, 2005”、“GARFIELD E, 2006”和“GARFIELD E, 1972”等。通过对高被引文献的关键词进行聚类,可揭示该领域的研究前沿^[14]。

4 结论与不足之处

通过以上实例分析,对五种科学知识图谱绘制工具的部分功能有了一定了解,以下对五种工具在数据处理功能和可视化功能上的差异作了总结,并绘制了折线图,具体如图5所示。

对各个工具的特点以及与其他工具相比的优势和劣势总结如下:

① Bibexcel是一款计量分析软件,它最大的优点构建共现关系矩阵,但可视化功能较弱,需借用其他可视化工

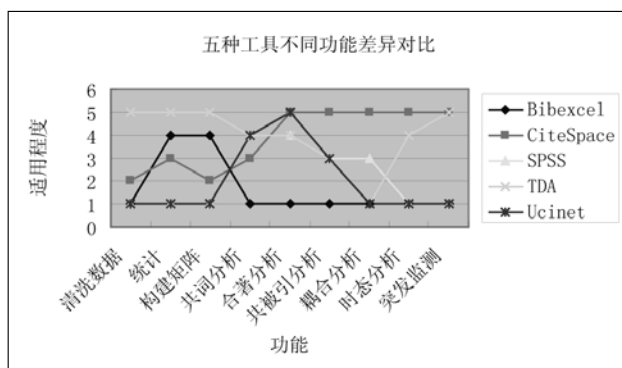


图5 五种工具不同功能差异对比

注：其中“1-5”五个数字分别代表“不可以、可以但不适合、可以在某些情况下适合、可以并且适合、可以并且非常适合”五个梯度。

具。与TDA相比，Bibexcel缺乏数据清洗功能，且构建关系矩阵步骤较繁琐，不可一步到位。因此，Bibexcel更适合用于简单的计量分析以及为一些可视化工具（如Ucinet）提供辅助，如统计、构建共现矩阵等；

② CiteSpace是主要用于计量和分析科学文献数据的信息可视化软件，该软件自身的功能比较完善，特点是时间序列分析和突发监测。在构建知识图谱的整个流程中，它各个步骤的处理都比较合理，可以突出显示核心项目。但数据清洗功能较差，聚类标签不能完全代表类别，算法还需完善，更适合用于作者合作、文献共被引、时间序列分析和突发监测；

③ SPSS具有完整的数据输入、编辑、统计分析、图形制作等功能，可进行聚类分析，它在数据的预处理和数据的规范化处理方面能力有限，且处理的数据不易过多，但SPSS聚类效果较好，所绘制的树状图可明显揭示上下位类的层级关系。SPSS适合用于分析偏小、分布均匀的数据集；

④ TDA是一款数据挖掘和可视化分析工具，它自身的功能比较完善，既可构建共现矩阵，也可绘制图谱，且绘制的图谱项目核心点突出明显。TDA最大的优势是具有清洗数据的功能，与CiteSpace相比，绘制结果的调整不是很灵活。因此，TDA更适合用于数据清洗和构建矩阵，为SPSS和Ucinet等一些可视化工具提供支持；

⑤ Ucinet是一款社会网络分析工具，它的数据分析功能强大，可分析较大的数据集，不仅可以突出项目核心点，还可通过连线的粗细揭示项目间关系的强

弱，绘制的知识图谱可调整以美化外观。Ucinet适合用于分析关键词共现和作者合作，分析核心关键词、核心作者。

由于受论文篇幅限制，本文仅使用了各工具的一种算法进行分析，分析不够全面；并且，在对五种工具的应用对比过程中，仅选取了关键词共现、作者合作和文献共被引三个方面分析，对比不够全面。

参考文献

- [1] 胡泽文,孙建军,武夷山.国内知识图谱应用研究综述[J].图书情报工作,2013,57(2):131-137.
- [2] 朱靖.信息经济学研究的可视化分析[J].情报学报,2013,32(11):1222-1232.
- [3] 杨思洛,韩瑞珍.国外知识图谱绘制的方法与工具分析[J].图书情报知识,2012,20(6):101-109.
- [4] 刘则渊,陈悦,侯海燕,等.科学知识图谱:方法与应用[M].北京:人民出版社,2008.
- [5] CHEN C. The Centrality of Pivotal Points in the Evolution of Scientific Networks [C]// Proceedings of the International Conference on Intelligent User Interfaces, 2005.
- [6] 邓维彬,唐兴艳,等.SPSS19(中文版)统计分析实用教程[M].北京:电子工业出版社,2012.
- [7] 韩静娴,董丽娟.利用Thomson Data Analyzer专利分析工具的专利分析实例研究[J].中国现代教育装备,2012,7(19):74-76.
- [8] 刘军.整体网分析讲义:Ucinet软件使用指南[M].北京:上海人民出版社,2009.
- [9] 廖胜娇.科学知识图谱绘制工具:SPSS和TDA的比较研究[J].图书馆学研究,2011(3):46-49.
- [10] 廖胜娇.科学知识图谱绘制工具VOSviewer和CiteSpace的比较研究[J].信息工作研究,2011(7):137-139.
- [11] 张力,赵星,叶鹰.信息可视化软件CiteSpace和VOSviewer的应用比较[J].信息资源管理学报,2011(1):95-98.
- [12] 董丽平.两种信息可视化工具在学科知识领域应用的比较研究:人胚胎干细胞文献分析[D].沈阳:中国医科大学,2010.
- [13] 李淑丽.信息可视化工具的比较研究[D].哈尔滨:黑龙江大学,2006.
- [14] CHEN C. Mapping scientific frontiers [M]. Berlin: Springer, 2003.
- [15] CHEN C. 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(3): 359-377.

作者简介

岳晓旭，女，1990年生，硕士在读，研究方向：科学计量学。

袁军鹏，男，1973年生，博士，研究员，研究方向：科学计量学、科技政策。通讯作者，E-mail: junpengyuan@126.com。

高继平，男，1983年生，博士，助理研究员，研究方向：科技政策，科学计量学。

翟丽华，女，1977年生，博士，副研究员，研究方向：科学计量学，情报学。

潘云涛，女，1967年生，硕士，研究员，研究方向：科学计量学、科技评价。

The Comparative Study of Scientific Knowledge Mapping Tools Based on the Actual Case

YUE XiaoXu, YUAN JunPeng, GAO JiPing, ZHAI LiHua, PAN YunTao
(Institute of Scientific and Technical Information of China, Beijing 100038, China)

Abstract: Following the general process of scientific knowledge mapping, taking the literature of "Scientometrics, Informetrics and Bibliometrics" field from Web of Science (WOS) in recent 10 years as an example, domestic commonly used scientific knowledge mapping tools, Bibexcel, CiteSpace, SPSS, TDA and Ucinet, were analyzed and compared in similarities and differences in measurement applications, from the co-word analysis, author collaboration and co-citation etc. Finally, their advantages and disadvantages were summarized, reference for the user were also provided.

Keywords: Scientific knowledge mapping; Bibexcel; CiteSpace; TDA; Ucinet

(收稿日期: 2014-03-28)

《汉语主题词表》(工程技术版) 即将出版

《汉语主题词表》自1980年问世以来，并经1991年进行自然科学版修订，在我国图书情报界发挥了其应有的作用，曾经获得了国家科学技术进步二等奖。为了适应网络环境下知识组织与数据处理的需要，2009年由中国科学技术信息研究所主持，并联合全国图书情报界相关机构，完成了《汉语主题词表》(工程技术版)的重新编制工作，即将由科学技术文献出版社正式出版。

《汉语主题词表》(工程技术版)共收录优选词19.6万条，非优选词16.4万条，等同率0.84。在体系结构、词汇术语、词间关系等方面进行改进创新。为了方便工程技术领域不同专业用户使用，《汉语主题词表》(工程技术版)按专业分13个分册出版。同时建立《汉语主题词表》服务系统，提供在线概念检索和辅助标引服务。

《汉语主题词表》(工程技术版)作为网络环境下知识组织和数据处理的基础性工具，期待着全国图书情报界共同推进其持续更新建设，也期盼社会各界进行深层次应用和实践。