数据挖掘领域研究现状与趋势的可视化分析

■ 杨良斌

[摘　要] 数据挖掘技术已成为计算机领域的一个新的研究热点，其应用也渗透到了其他各大领域。以2004－2013年SCI数据库中收录的2 263篇以“数据挖掘”为主题的文献为研究对象，使用可视化软件CiteSpace绘制关键词共现图谱、期刊共引图谱、机构合作图谱等科学知识图谱，分析数据挖掘领域的热点以及发展趋势。分析结果表明，2004－2013年数据挖掘领域研究处于稳步发展时期。最后获得该领域各研究方向的现状和演化趋势。

[关键词] 数据挖掘　知识图谱　信息可视化　共现分析

[分类号] G250

1　引言

互联网技术的出现和发展带给了人们太多的便利，在网上互换信息和合作变得越来越容易，计算机不断地提高对各种类数据和信息的收集存储和处理能力，数据库技术的成熟和普及带来的结果是所积累的信息量以指数方式暴涨[1]。伴随着急剧增长的数据量和对数据处理方面的各种需求的增加，传统的数据分析工具已经不能承载对那些海量数据的操作处理了，人们需要一个将广博的数据转换成知识的技术，数据挖掘(data mining)便在这个背景下应运而生。

数据挖掘有多种定义，其中比较有代表性的一个即是“从数据中汲取出包含着的过往不被知道的有利用价值的潜在信息”。作为近年来新兴起的学科，数据挖掘在学术界赢得了极高的关注度，在产业界赢得了赞誉。早先的数据挖掘领域经历了电子邮件阶段和信息发布阶段，而如今这项技术已步入电子商务阶段并逐步走向当下最新的全程电子商务阶段，其应用横跨各个领域并为不同领域提供联系与数据支持的基础。在新世纪信息产业与网络互联持续发展、数据激增的背景下，数据挖掘领域一直不断融入新的知识和技术方法，并不断以多角度多元化发展，其学科框架已遍及多个领域。数据挖掘相关技术如今已被各大领域大力应用，如生物学研究中用数据挖掘技术对DNA进行分析[2]；市场中可以利用数据挖掘技术对顾客的购买行为模式进行识别和区分，并能对商业上频繁出现的诈骗行为予以防备[3-4]。数据挖掘的多学科化使学术界和产业界的研究人员们面临诸多挑战，因此探究数据挖掘领域的研究热点和发展趋势对于把握该领域的研究现状和发展方向具有重要意义和参考价值[5-6]。

本文所使用的研究方法为信息可视化研究法，目前常用的科学知识图谱主要有共词分析、共引分析、多元统计分析、词频分析和社会网络分析。其分析的数据单元涵盖作者、关键词、标题、引文、摘要和作者地址等，通常采用Ucinet、CiteSpace、VOSviewer等可视化分析软件来绘制。本文所选用的CiteSpace软件，是由美国德雷克赛尔大学信息科学与技术学院的陈超美教授于2004年开发的信息可视化软件，该软件近年来在信息可视化分析领域有着不小的影响力，其关键节点测量、时间年轮等特色功能可以方便研究者们对某个领域当前的热点与发展趋势进行研究[7]。许多学者利用该软件研究了战略管理领域的智力结构，绘制了共引图谱，并可视化科学知识的结构、关系与演化过程[8-10]。本文依据来自SCI数据库的数据挖掘领域相关文献，绘制关键词共现图谱、期刊共引图谱、国家及机构合作图谱和时区视图这4种类型的科学知识图谱，进行可视化分析并探讨数据挖掘领域研究趋势和热点，以便于这一领域的相关研究人员们对数据挖掘研究的现状从总体上有一个大致的了解，并且对其今后的进一步研究有所引导和帮助，从而促进数据挖掘领域研究的深入发展。

2　数据来源与整理

本文选取的文献数据来源于美国《科学引文索引》数据库，以2004－2013年共10年为时间跨度，以“data mining”为主题词进行检索，得到包括作者、标

[作者简介] 杨良斌，国际关系学院信息科技学院副主任，副教授，E-mail: yangliangbin@tsinghua.org.cn。

题、参考文献等项的 2 263 条文献记录，被引频次总计 18 727 次，去除自引的被引频次总计 17 612 次，施引文献 15 701 篇，去除自引的施引文献 15 072 篇，每项题录平均引用次数为 8.28 次，检索时间为 2014 年 5 月 3 日。统计得到 2004 – 2013 年数据挖掘领域每年出版文献量及论文被引情况分布图（见图 1、图 2）。从图中可以明显看到，数据挖掘领域的相关研究近 10 年来一直保持较高热度，每年的文献出版量都保持在 200 篇以上，且于 2012、2013 年分别突破 250 篇；文献被引频次逐年增加，2013 年更是达到了近 3 500 次，说明数据挖掘在近一两年的影响力逐步攀升，甚至在计算机等相关领域的用途越来越广，作用越来越不能被忽视。也许，真正属于数据挖掘技术的时代才刚刚开始。

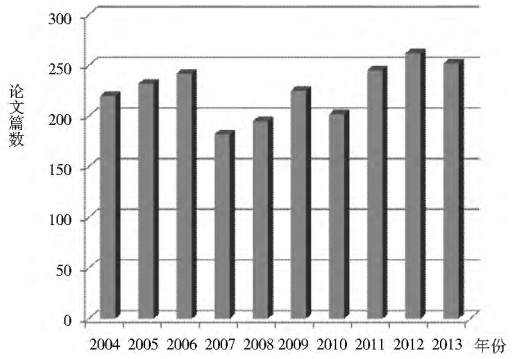

图 1 2004 – 2013 年数据挖掘领域论文发表数量的年度分布

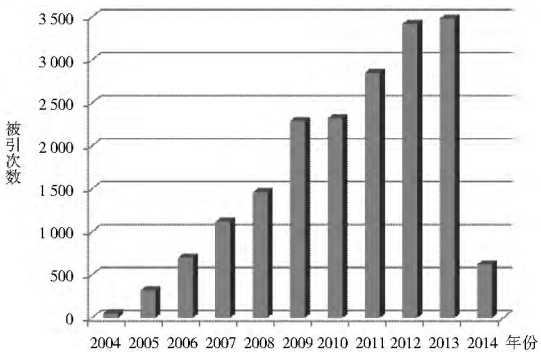

图 2 2004 – 2013 年数据挖掘领域论文被引频次的年度分布

3 各图谱的生成及分析

3.1 关键词共现图谱

共现指的是文献当中相同或不同特征项共同出现的现象，比如多篇文献当中会共同出现一些合作者、合作机构和关键词，等等。共现分析，是以揭示信息的内容关联和特征项所隐含的寓意为目的，将各式各样的信息载体中的共现信息进行定量分析的方法。其中，关键词共现分析可以通过文献集中关键词的出现频率、中心度以及这些关键词构成的共现网络的聚类和节点，来分析其内在的关联及隐含的意义，揭示某个学科、某个领域的基本研究结构，确定当前的研究热点和主题。

利用 CiteSpace 对来自 SCI 的 2 263 篇文献的关键词进行分析得到：有效的关键词 150 个，关键词出现总数 3 038 次，平均每个关键词出现 20.25 次。从中提取出频次最大的前 50 个关键词，并剔除 system、management 等一般性较宽泛的词汇，通过 CiteSpace 绘制数据挖掘领域关键词共现知识图谱（关键词频次阈值为 2）得到图 3，高频次关键词的统计情况见表 1。

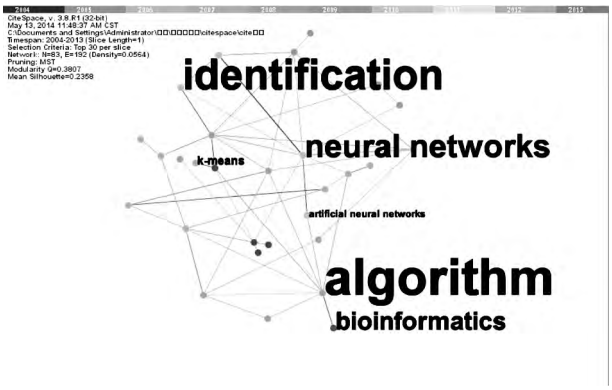

图 3 2004 – 2013 年数据挖掘领域关键词共现图谱

通过对关键词共现图谱的分析可以得到数据挖掘领域的研究热点。经计算，“classification”频次为 217，且中心度 0.24 最高，其节点最大并位于整个网络中心；其次，“algorithm”、“neural networks”、“clustering”和“knowledge discovery”也拥有较高频次，这些关键词在近 10 年的数据挖掘相关文献中出现频率较高，可视为一直以来该领域的研究热点；另外由时间年轮可以看到，这些关键词在 2010 年前后出现较多，说明这些词又是近几年数据挖掘领域的重要主题。另一方面要注意的是中心度较高的节点，中心度的大小与频次有一定关系，但并不成正比，中心度高而频次低代表着该结点可能是近年出现的枢纽结点，与其他关键词经常一起出现，起着承接的作用，在共词网络中影响力比较大，有一定的发展潜力，所以处于各个结点网络中心的“identification”、“association rules”和“artificial neural networks”等词虽然没有过高的出现频次，但是它们显然在数据挖掘研究领域引起了关注和讨论。

纵观这 10 年的高频关键词，可以发现部分数据挖掘研究方法如系统识别（system identification）、遗传算法（genetic algorithm）、特征选择（feature selection）等应

表 1　2004 – 2013 年数据挖掘领域高频次关键词统计

关键词	频次	中心度	释义和备注
classification	217	0.24	分类 – 数据挖掘领域的主要方向之一
algorithm	73	0.08	研究数据挖掘领域算法
neural-networks	72	0.14	神经元网络
clustering	70	0.03	聚类 – 数据挖掘领域的主要方向之一
decision tree	65	0.03	决策树 – 分类的常用工具之一
knowledge discovery	62	0.11	知识发现 – 当前知识挖掘的重要研究热点
identification	61	0.21	识别技术
association rules	59	0.06	关联规则
regression	57	0.05	回归拟合
optimization	53	0.01	最优化
artificial neural-networks	52	0.08	人工神经网络
Support vector machines	45	0.03	支撑向量机
genetic algorithm	40	0.04	遗传算法
machine learning	37	0	机器学习
bioinformatics	35	0.01	生物信息学
feature selection	30	0	特征选择
drug discovery	21	0.04	药物发现
k-means	16	0	k 均方聚类
artificial intelligence	16	0.01	人工智能
time-series	15	0	时间序列
self-organizing map	15	0	自组织映射
logistic regression	15	0	逻辑回归
hemodialysis	6	0	血液透析
cybernetics	5	0	控制论

用较多，涉及面较广，而其他出现频次较低或未出现的研究方法可能因针对性较强而没有大热。另外，分类(classification)、聚类(clustering)、算法(algorithm)等数据挖掘的重要研究方向一直是该领域不断突破的重点，而近几年来热门的神经网络(neural-networks)算法的研究热度高居不下，并将继续保持；同样可以看到，近一两年的研究动态偏重于数据挖掘技术探索到的新领域，如血液透析(hemodialysis)和人工智能(artificial intelligence)等。

3.2　期刊共引图谱

引文分析，指的是利用多种统计学方法和逻辑方法，对科学论文、期刊文章和作者等各种分析对象彼此之间的引用证明与被引用证明的现象进行分析，以便揭示其中的数量特征和隐含规律的一种文献计量分析方法。对 2004 – 2013 这 10 年的 SCI 数据挖掘领域期刊共引进行分析，网络节点选为期刊，时间区选择为 1 年，阈值为(15，15，15)，(15，15，15)，(15，15，20)，可视化得到图 4 的期刊共引网络。总共得到有效的期刊 150 种，其中共引文献超过 300 篇的期刊有 3 种，分别为 *MACH LEARN*、*LECT NOTES COMPUT SC*、*EXPERT SYST APPL*，可见这 3 种期刊在数据挖掘领域的影响力是足够大的，另外紧随其后的超过 200 篇共引文献出现的有 11 种期刊。图谱中颜色深浅都不相同的多个圆环组成的年轮型图案表示期刊节点，大小代表了期刊的共引频次，颜色对应着该期刊的共引年份，用彼此间连线的颜色深浅不同来区分期刊间的共引时间，从中挑选出频次超过 200 的关键节点期刊，得到表 2。

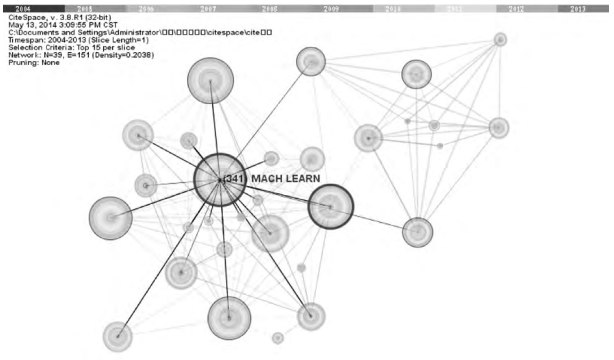

图 4　2004 – 2013 年数据挖据领域的期刊共引图谱

表 2 2004 – 2013 年数据挖掘领域高频次被引期刊

序号	共引期刊	频次	中心度
1	*MACH LEARN*	341	0. 41
2	*LECT NOTES COMPUT SC*	305	0. 18
3	*EXPERT SYST APPL*	302	0. 15
4	*DATA MINING CONCEPTS*	294	0. 1
5	*IEEE T KNOWL DATA EN*	285	0. 52
6	*DATA MINING PRACTICA*	268	0. 02
7	*COMMUN ACM*	239	0. 03
8	*LECT NOTES ARTIF INT*	232	0. 04
9	*C4 5 PROGRAMS MACHIN*	223	0
10	*BIOINFORMATICS*	214	0. 11
11	*DATA MIN KNOWL DISC*	211	0. 07
12	*P NATL ACAD SCI USA*	206	0. 1
13	*SCIENCE*	205	0. 2
14	*NATURE*	202	0. 11

图 4 中，各种不同的连线，表示相关文献的首次共被引的年份，由结点的位置可以观察出文献的核心层度，对其中的高引频文献进行分析研读，便可分析出数据挖掘领域的期刊文献的趋势和走向。从表 2 中可以看出排在被引频次前 5 位的期刊分别是 *MACH LEARN*、*LECT NOTES COMPUT SC*、*EXPERT SYST APPL*、*DATA MINING CONCEPTS*、*IEEE T KNOWL DATA EN*，因此在关注数据挖掘领域的核心期刊时，对于那些来源期刊发文量和共引频次排名前列的核心期刊，应给予着重考虑。外围的中心度最高的两个节点 *IEEE T KNOWL DATA EN*（0. 52）、*MACH LEARN*（0. 41）各自引领自己的期刊网络，说明这两种期刊正是数据挖掘领域各大期刊的共引核心期刊，并以各自为中心形成了共引期刊子网网络。利用对该领域期刊共引网络图谱中共引频次和中心度的分析，便可获知与自身科研水平相符合的数据挖掘领域核心期刊分布情况。

3.3 机构合作图谱

科学合作在很多时候已经成为科研交流中的关键所在，合作的不断拓展缘于科学发展的动态性、技术的多样性、知识的复杂性以及高度发展的技能等的专业性。合作不仅可以促进科学家相互之间的知识传播和学术交流，共享现金实验工具和实验设备；同时，在科学合作的过程中，隐性技艺和知识也被相应转化和分享。长期以来，科学计量学家对科学合作特别是国际科学合作问题本身的研究给予了极大的关注，尤其是在数据挖掘这一研究领域，数据信息应被全世界全人类所共享，各国之间以及各国的机构包括研究院、学校等机构间的合作交流更是推动该领域向前发展的关键所在，只有囊括越全面越有针对性的数据，不断加强国际交流，数据挖掘技术研究才能不断攀登新的高度。

对 2004 – 2013 这 10 年的 SCI 数据挖掘领域国家和机构合作进行分析，网络节点确定为研究机构，时间选择为 1 年，可视化得到数据挖掘领域研究机构合作图谱，见图 5。共得到 332 个研究机构的数据，可见数据挖掘领域的研究已经被大部分国家所重视，并均有针对性研究机构，数据挖掘在计算机、统计甚至电子商务等诸多市场前沿的学科方向上都进入革新发展的时代。表 3 给出了数据挖掘领域近 10 年发表研究成果前 15 名的研究机构。

图 5 2004 – 2013 年数据挖掘领域研究机构合作图谱

表 3 2004 – 2013 年数据挖掘领域高出现频次的机构

机构英文名	机构中文名	所在国家(地区)	出现频次
Univ Iowa	爱荷华大学	美国	32
Chinese Acad Sci	中国科学院	中国	30
Natl Chiao Tung Univ	台湾“国立”交通大学	中国台湾	18
Iowa State Univ	爱荷华州立大学	美国	18
Monash Univ	莫纳什大学	澳大利亚	16
Univ Illinois	伊利诺伊大学	美国	15
Natl Cheng Kung Univ	台湾成功大学	中国台湾	15
Chinese Univ Hong Kong	香港大学	中国香港	14
Univ Nebraska	内布拉斯加大学	美国	13
Zhejiang Univ	浙江大学	中国	13
Natl Univ Singapore	新加坡国立大学	新加坡	13
Univ Calabria	卡拉布里亚大学	意大利	13
Natl Tsing Hua Univ	“国立”清华大学	中国台湾	13
Shanghai Jiao Tong Univ	上海交通大学	中国	12
Univ Florida	佛罗里达大学	美国	12

从图 5 中可以了解到数据挖掘领域的文献研究中机构的分布及合作关系。大多数的合作网络都是科研机构和大学组成的，其中比较活跃的机构包括 Univ Lowa（爱荷华大学）、Chinese Acad Sci（中国科学院）、

Natl Chiao Tung Univ(台湾“国立”交通大学)、Iowa State Univ(爱荷华州立大学)和 Monash Univ(莫纳什大学),不难发现美国、中国和中国台湾的高等院校是数据挖掘领域研究文献的主要贡献者。机构的合作分布从一定程度上反映并解释了国家(地区)间的合作分布,考虑数据挖掘研究领域的文献数量,美国仍处于世界领先水平,一些近几年起步的国家及机构为该领域的研究注入了新的信息力量,如塞尔维亚、爱沙尼亚、斯诺文尼亚等欧洲国家。而由结点的时间年轮可以看出,中国的诸多研究机构虽然起步较晚,但是在近两年发展速度非常快,且在云计算实践和大数据应用等方面取得了一定的突破。

综合对关键词共现图谱、期刊共引图谱和机构合作图谱进行的可视化分析可以看到,从研究主题上来看,数据挖掘不断在主要技术方法上寻求新的突破,并不断蔓延到各个领域,以生物学和医学两个领域的贡献尤为突出,恰巧这两个突出领域也是发达国家相对于其他国家的科研优势。从机构的合作构成来看,学术强国或地区在数据挖掘领域的技术和方法上面能够自主研究和创新,如美国、中国、中国台湾和欧洲等学术强国或地区等,但大多数的研究合作还是局限于国家(地区)内部,无论从期刊的发表上还是机构的合作上来看,这样的现象十年来都未有缓解。从期刊共引上来看,主要的数据挖掘领域权威期刊集中在西方国家,东方国家和第三世界国家在该领域的力量虽不容忽视,但影响力还没有完全形成。未来,随着东方国家的数据技术的进步,其在该领域拥有话语权还是很有可能的。

4 时区视图与发展趋势分析

Time-zone(时区视图)是 CiteSpace 的特色之一,利用其突变探测功能,可根据热点关键词或者文献随时间的变化来形成时区视图,用以展示研究对象的演变路径。

图 6 是数据挖掘领域关键词共现时区视图,反映出了近 10 年该领域研究主题和热点的变化,10 年内的大量文献还是来源于 5 年之前,2004 – 2008 年这 5 年间有关分类方法、聚类分析方法和时间序列的研究达到了一个高度,并为该领域研究方法后来在其他领域的应用奠定了知识基础,提供了参考来源;可见数据挖掘领域的发展是有据可循的,每个阶段都有一定的基础和发展变化。近两三年新兴起的研究方向如基因、细胞分裂等医学和生物学方向也取得了很大的学术突破。

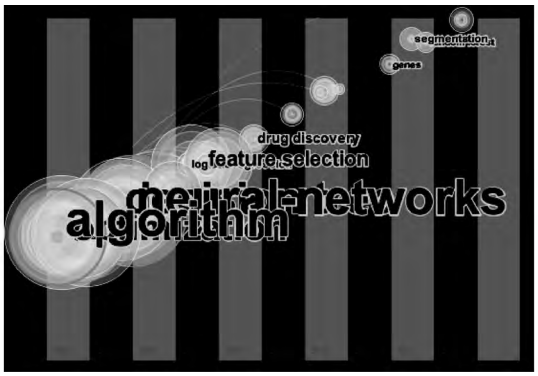

图 6 2004 – 2013 年数据挖掘领域关键词共现时区视图

表 4 2004 – 2013 数据挖掘领域文献作者出现频次

作者	出现频次
A. Kusiak	19
Tzung Hong-Pei	13
K. Rajan	9
A. Kusiak	9
Y. Okuno	8
T. Sakaeda	8
Chen Chunhao	8
K. Kadoyama	8
V. S. Tseng	8

由关键词共现时区视图可知,在数据挖掘领域中算法、系统等方向依然是研究热点。首先要提到神经元网络技术(neural-networks),该技术属于软计算领域内的一种重要方法,多年来学者们不断进行人脑神经相关的机能模拟得到的成果,已可喜地被实际应用于各大产业项目。该技术不仅应用广泛,且带有较强的普适性和可操作性,因此成为了数据挖掘领域热点方法之一。

代表决策树(decision tree)的节点在图中也占据了较显眼位置,决策树算法是目前应用最广泛的归纳推理算法之一,也是一种简单的知识表示方法,它将用例一步步分类使每一类代表不同的派别。分类规划是比较直观的,因为其相比之下更容易被人们所理解,但这种方法只局限于分类和规划任务。因此在近些年分类(classification)主题大热的趋势下,决策树得到了大量的应用,并派生了很多理论和技术方法。其他关键词中有一部分尤为突出,即与生物学研究相关的课题,如遗传算法(genetic algorithm),它是一种高效探索算法,伴随自然群体遗传演化机制形成,算法对自然群体循环操作,然后根据预定的标准对每个实验单体进行评估,便可求得满足要求的最优解。在环境和生物领域接受着严峻挑战的今天,数据挖掘相关算法无疑给

相关领域带来了福音，遗传算法也使得很多课题得到了高效的解决办法。另外“最优化”(optimizition) 这个关键词不得不引起注意，它主要分为两方面: 一方面是数据挖掘技术的最优化，使其更好地运用于实际应用当中; 另一方面指利用数据挖掘技术和相关成果来优化企业或机构的内部活动，如优化客户服务、监察管理等。这两种意义都代表着数据挖掘会向着价值化的方向发展，并且会被更多的领域所应用，这是数据挖掘领域较稳定的热点。

5 结语

进入 21 世纪的信息化时代以来，数据挖掘领域相关技术逐渐成为计算机领域的一大热点，无论是学术界还是产业界都对其抱以相当积极的关注。本论文通过信息可视化软件 CiteSpace，使用数据挖掘领域的相关文献，绘制了关键词共现图谱、期刊共引图谱、研究机构合作图谱和时区视图，并对其进行综合整理与分析，探明了数据挖掘领域研究的热点和未来的发展趋势。得到以下结论: (1) 当前数据挖掘技术的研究热点主要分为两个方面: 一是数据挖掘的关键技术，主要是分类算法和聚类算法等; 二是数据挖掘的具体应用，包括可视化、遗传算法和血液透析等，这是随着数据挖掘技术处理大数据由概念化转向价值化而出现的。(2) 数据挖掘领域相关研究机构以高等院校为主，国家以美国等发达国家为主，而拥有较大数据量的中国、澳大利亚和加拿大等国家在近年来后来居上，但各机构间还是疏于研究合作的。(3) 大数据研究的发展趋势主要在于以下两个方面: 一方面涉及具体应用的研究会越来越多，未来数据挖掘技术将涉足更多的领域; 另一方面是会更加注重安全和隐私的保护，很多数据挖掘相关技术还处在发展阶段，因此其应用存在不少安全隐患，随着数据挖掘领域相关技术的发展，安全方面的研究也会越来越多。

CiteSpace 虽然生成的图谱较为直观，但在很多方面如图谱种类、图谱复杂度等方面还不太成熟，比起 Ucinet 等主流可视化软件还有许多不足的地方，因此以后在利用可视化方法分析某个领域的时候，可以用多种可视化软件进行分析，再综合对比分析结果，这样得出的研究结论可能会更加全面、更有说服力。另外，本文对关键词共现图谱的分析中，省去了一些对研究结果意义不大的概念性关键词，其中对这类词的判断非常重要，而这完全取决于分析者对其分析领域的理解，所以本文对关键词共现网络的分析或多或少存在疏漏之处，今后需注意此问题。

参考文献:

[1] 陈雪刚. 数据挖掘技术在个性化 web 中的应用研究 [D]. 长沙: 湖南大学，2010.

[2] 朱扬勇等. DNA 序列数据挖掘技术 [J]. 软件学报，2007，18(11) : 2766 - 2781.

[3] Huang Li-Juan，Gan Xiao-Qing. Customer's clustering analysis and corresponding marketing strategies based on SOFM neural network in e-Supply chain [J]. Systems Engineering Society of China，2007，27(12) : 49 - 55.

[4] 郑继刚等. 数据挖掘研究的现状与发展趋势 [J]. 红河学院学报，2010，8(2) : 45 - 48.

[5] 张士靖等. 信息素养领域演进路径、研究热点与前沿的可视化分析 [J]. 大学图书馆学报，2010(5) : 101 - 106.

[6] Alberto H F. Laender，Berthier A. Ribeiro-Nrto，Altigran S. da Silva，et al. A Brief Survey of Web Data Extraction Tools. SIGMOD Record，2002. 31(2) : 84 - 93.

[7] BÊrner K，Chen C M，Boyack KW. Visualizing Knowledge Domains [J]. Annual Review of Information Science & Technology，2003，37 (5) : 179 - 255.

[8] Nerur S P，Rasheed A A，Natarajan V. The intellectual structure of the strategic management field: An author co-citation analysis [J]. Strategic Management Journal，2008，29(3) : 319 - 336.

[9] Chen C，Ibekwe-SanJuan F，Hou J. The Structure and Dynamics of Co-Citation Clusters: A Multiple-Perspective Co-Citation Analysis. Journal of the American Society for Information Science and Technology. 2010，137(14) : 144 - 307.

[10] Cobo M J，López-Herrera A G，Herrera-Viedma E，et al. Science mapping software tools: Review，analysis，and cooperative study among tools. Journal of the American Society for Information Science and Technology. 2011，62(7) : 1382 - 1402.

(上接第 170 页)

[8] 国家自然科学基金委员会编. 国家自然科学基金项目统计资料 [R/OL]. [2015 - 06 - 10]. http://www.nsfc.gov.cn/nsfc/cen/xmtj/pdf/2005_table.pdf.

[9] 长青，李东. “建校 80 周年暨合校 15 周年巡礼”科学研究工作回望 [EB/OL]. [2015 - 06 - 10]. http://news.nwsuaf.edu.cn/xnxw/43821.htm.

[10] 靳军. “盘点 2014 新亮点之一”学科建设与科学研究篇 [EB/OL]. [2015 - 06 - 10]. http://news.nwsuaf.edu.cn/xnxw/48164.htm.