Hans Journal of Data Mining 数据挖掘, 2016, 6(3), 125-137
Published Online July 2016 in Hans. http://www.hanspub.org/journal/hjdm
http://dx.doi.org/10.12677/hjdm.2016.63015

Hans汉斯

Big Data Research Analysis Based on Bibliometrics

Yuzhu Kuang, Mei Hong, Jiayan Zeng

School of Computer Science, Sichuan University, Chengdu Sichuan
Email: kuangyuzhu1992@foxmail.com

Received: Aug. 3rd, 2016; accepted: Aug. 25th, 2016; published: Aug. 30th, 2016

Abstract

With the development of Internet, the time of big data has come, and much attention has been paid on the research of big data all over the world. This paper analyzes the current research state and development trend of big data on the base of domestic and overseas academic research in this area. We select the 2000-2015 domestic and foreign research literature in the field of big data by the combination of automatic and manual way of literature retrieval, then analyze the paper growth and distribution and the distribution of journal, conference and cooperation of authors, etc., using the method of bibliometrics and visualization software of literature analysis, especially the hot spots and trend analysis of big data research, and provide a valuable basis and reference for the further study work of researchers.

Keywords

Big Data, Bibliometrics, Visualization, Systematic Literature Review, Data Analysis

基于文献计量的大数据研究现状分析

况俞竹，洪　玫，曾嘉彦

四川大学计算机学院，四川 成都
Email: kuangyuzhu1992@foxmail.com

收稿日期：2016年8月3日；录用日期：2016年8月25日；发布日期：2016年8月30日

文章引用: 况俞竹, 洪玫, 曾嘉彦. 基于文献计量的大数据研究现状分析[J]. 数据挖掘, 2016, 6(3): 125-137.
http://dx.doi.org/10.12677/hjdm.2016.63015

摘　要

随着互联网技术的发展，大数据时代已经来临，对大数据的研究受到世界范围的关注。本文针对国内外学术界计算领域对大数据的研究，分析该领域的研究现状与发展趋势。本文采用文献计量方法和可视化文献分析软件，通过自动和手动相结合的文献检索方式，筛选了2000~2015年国内外大数据领域的研究文献，对论文增长与分布、期刊和会议分布、作者合作等进行分析，特别分析了大数据研究的热点和趋势，为研究者的进一步研究工作提供了有价值的依据和参考。

关键词

大数据，文献计量，可视化分析，文献综述，数据分析

1. 引言

大数据是继云计算、物联网之后 IT 产业又一次颠覆性的技术革命[1]。大数据是指无法在一定时间内用常规软件工具对其内容进行抓取、管理和处理的数据集合。大数据一般是指“海量数据”+“复杂类型的数据”，大数据的特性包括 4 个“V”(Volume, Variety, Velocity, Value) [2]，即数据量大，PB 级以上数据；种类多，包括文档、视频、图片、音频、数据库数据等；速度快，数据生产、处理和 I/O 速度快；价值大，对国民经济和社会发展有重大影响。

早在 1980 年，著名未来学家托夫勒在其所著的《第三次浪潮》[3]中就热情地将“大数据”称颂为“第三次浪潮的华彩乐章”。美国的麦肯锡公司 2011 年 6 月在其报告《大数据：创新、竞争和生产力的下一个新领域》[4]中提到“数据，已经渗透到当今每一个行业和业务职能领域，成为重要的生产因素。人们对于海量数据的挖掘和运用，预示着新一波生产率增长和消费者盈余浪潮的到来。”2012 年 3 月美国政府发布《大数据研究与开发计划》[5]，投资 2 亿美元发展大数据，旨在提高从大型复杂数字数据中抽取知识与观点的能力，以帮助解决国家在科学与工程中最紧迫的诸多挑战问题。

数据量的指数级增长不但改变了人们的生活方式、企业的运营模式，而且改变了科研范式。2007 年，已故的图灵奖得主 Jim Gray 在他最后一次演讲描绘了数据密集型科研“第四范式”The Fourth Paradigm)的愿景[6]。2008 年 9 月《自然》杂志[7]出版了一期专刊——“Big Data”，2011 年 2 月，《科学》[8]期刊联合其姊妹刊推出了一期关于数据处理的专刊——“Dealing With Data”，从互联网技术、互联网经济学、超级计算、环境科学、生物医药等多个方面介绍了海量数据所带来的技术挑战。

大数据的发展已经得到了世界范围内的广泛关注，如何将巨大的原始数据进行处理、存储、分析和利用，并转化为知识和价值，成为国内外共同关注的重要研究课题。鉴于此，本文针对国内外学术界计算领域对大数据的研究，分析该领域的研究现状与发展趋势。本文采用文献计量方法，借助可视化文献分析软件，通过自动和手动相结合的文献检索方式，筛选了 2000~2015 年国内外大数据领域的研究文献，对论文增长与分布、期刊和会议分布、作者分布等进行分析，特别分析了大数据研究的热点和趋势，为研究者的进一步研究工作提供了有价值的依据和参考。

2. 数据来源和研究方法

2.1. 数据来源

大数据研究领域涉及面广，为了能够完整体现计算领域对大数据的研究现状，论文选取了 CNKI (中

国期刊全文数据库)作为中文文献资料来源，并选取了 ACM Digital Library (美国计算机学会电子图书馆)和 IEEE Digital Library (国际电气、电子工程师协会数字图书馆)作为外文文献资料来源。

通过对国内外文献的阅读发现，与大数据主题相关的文献往往使用“big data”、“large data”、“大数据”等关键词。因此本次检索的国内外文献检索词策略分别为：国内篇名 = “大数据”或者关键词 = “大数据”；国外篇名 = “big data”or“large data”或者关键词 = “big data”or“large data”进行精确搜索，时间跨度为 2000~2015，在 CNKI、ACM、IEEE 数据库分别进行了检索。

为了保证分析结果的有效性，对获得的论文相关数据进行人工的筛选。删除书籍、报道、会议通知、会议纪要等数据，仅保留期刊论文和会议论文；去除内容重复或内容不相关的论文，以及短文(篇幅在 5 页以下的)；对于中文论文，考虑国内部分期刊的质量问题，我们参考“计算机类核心期刊排名”，筛选刊登在 Top30 中文期刊的论文。最终共获得国外研究文献 3773 篇，国内研究文献 1041 篇，作为本文的分析数据。

2.2. 研究方法

研究方法采用文献计量学的方法。文献计量学[9]是借助文献的各种特征参数，采用数学与统计学方法来描述、评价和预测科学技术的现状与发展趋势的图书情报学分支学科。文献是贯穿于整个科研过程且反映科研能力的重要因素。文献计量则是一种具有成本和效率优势，以及准确性和客观性的定量评价方法。利用文献计量学的方法，可以对科研主体的研究工作，以及学术出版物，包括著作、期刊、论文、专利等作出较为科学的评价。

词频分析法[10]是常用的文献计量学方法之一，利用某一研究领域文献中的关键词频次高低，来确定该研究领域的文献主题概念的自然语言词汇，能够简单、直接、较为全面地概括论文的核心研究内容。

共词分析法[11]是利用某一研究领域文献中关键词共同出现的情况，来确定该研究领域中各关键词之间的关系。一对关键词在文献中出现的频次越高，它们之间的关系越紧密。在文献分析过程中，将应用 EndNote 软件对文献数据进行管理，应用 BibExcel、Pajek、VOSViewer 和 CiteSpace 软件对文献数据进行处理和分析。

3. 大数据研究文献的增长与分布情况分析

3.1. 大数据研究文献的增长情况

对采集到的论文的年份、数量特征进行统计分析，可以在一定程度上根据文献的增长情况得到国内外关于“大数据”研究的总体研究水平和发展速度，如图 1 所示。

在国内的 1041 篇文献中，2000~2005 年论文发表有缓慢上升的趋势，从 11 篇增长到 50 篇；2005~2012 年保持比较稳定的水平。2012 年后开始有明显上升趋势，论文增至 243 篇，约占总数的 23.3%。在国外的 3773 篇文献中，2000~2011 年论文发表一直处于较低水平，2011 年后开始逐渐上升，并在 2012 年迎来第一次快速增长，此后迅猛增长，仅 2014 年一年发表的文献数就达到 1147 篇，2014、2015 两年发表的文献占总数的 68%。相比之下，国内的大数据研究论文发表数量比较国外相对较多(相对于人口)，说明在这一领域研究工作较多，但在近两年的发展速度不及国外的高。

3.2. 大数据研究文献在主要期刊和会议上的分布情况

对刊载文献数量排名较前的期刊种类和刊载文献数进行分析，得出了大数据研究领域重要的刊物。统计国外大数据研究论文，期刊上发表 615 篇，国际学术会议上发表 3158 篇。国内大数据研究论文在主要核心期刊上发表 1041 篇。选取文献数排在前 15 位的期刊，对国内外刊载大数据研究文献的期刊进行统计，得到如表 1 所示结果。

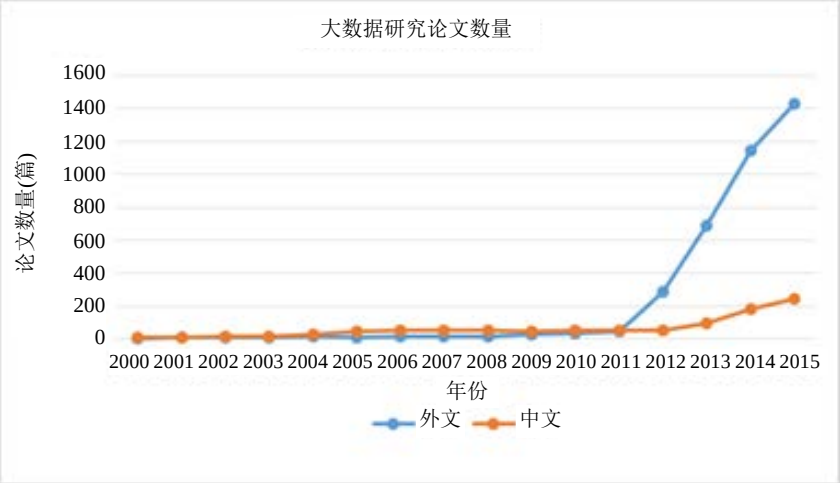

Figure 1. Variations in big data research papers
图 1. 大数据研究论文数量变化图

Table 1. Foreign top 15 conferences publishing big data research paper
表 1. 国内外刊载大数据研究文献排名前 15 的期刊情况

篇数	国外期刊名称	篇数	国内期刊名称
78	The Very Large Databases Journal	108	计算机应用
37	Computer	106	计算机研究与发展
30	Communications of the ACM	93	计算机工程与应用
25	ACM Crossroads	92	计算机科学
23	Computer Graphics and Applications, IEEE	90	计算机工程
20	ACM SIGMETRICS Performance Evaluation Review，	74	计算机应用研究
16	Emerging Topics in Computing, IEEE Transactions on	63	计算机工程与设计
13	IBM Journal of Research and Development	58	小型微型计算机系统
12	Intelligent Systems, IEEE	54	软件学报
12	Knowledge and Data Engineering, IEEE Transactions on	53	计算机应用与软件
12	IT Professional	47	计算机学报
11	Journal of Computing Sciences in Colleges	42	微电子学与计算机
11	Parallel and Distributed Systems, IEEE Transactions on	41	微计算机信息
10	ACM SIGKDD Explorations Newsletter	23	中国图像图形学报
10	Signal Processing Magazine, IEEE	19	系统仿真学报

可以看出，在大数据研究方面，国外在顶级期刊上发表的论文数较多，而国内高水平的研究有限，在国内顶级期刊上发表的大数据研究论文不到 10%。

为了更全面地反映大数据领域的研究情况，对刊载大数据研究论文的国际学术会议进行统计，但因国内的学术会议数量较少，没有对国内的学术会议进行统计，得到表 2。

从表 2 可以看出，大数据领域的国际学术会议，与云计算、网格计算、数据挖掘、社会网络等主题相关，说明大数据的发展与云计算、数据挖掘、社会网络等技术之间有一定的相关性。

Table 2. Foreign top 15 conferences publishing big data research paper
表 2. 国外刊载大数据研究论文排名前 15 的会议情况

排名	会议名称	篇数
1	Big Data (Big Data)	159
2	Proc. VLDB Endow.	84
3	Big Data (BigData Congress)	67
4	Computational Science and Engineering (CSE)	46
5	System Sciences (HICSS)	45
6	Trust, Security and Privacy in Computing and Communications (TrustCom)	32
7	Utility and Cloud Computing (UCC)	28
8	Cloud Computing Technology and Science (CloudCom)	19
9	Cluster, Cloud and Grid Computing (CCGrid)	18
10	Data Mining Workshop (ICDMW)	18
11	Services (SERVICES)	17
12	High Performance Computing, Networking, Storage and Analysis (SCC)	16
13	Cloud Computing (CLOUD)	16
14	Data Mining (ICDM)	15
15	Big Data and Cloud Computing (BdCloud)	14

3.3. 大数据研究文献的作者分析

通过对高产作者的统计，能够帮助研究者快速获取该领域的核心研究人员信息，便于跟踪该领域的研究进展，表 3 展示了国内外大数据研究论文发表数量排名前 15 的作者。我们发现，在国外研究作者排名前 15 中有 6 位是中国人。

同时，为了了解作者之间的研究合作关系，用合著分析方法，分析在学术研究中作者合著的情况，包括研究人员分布、结构关系等。合著的作者被认为是在地域上或学科研究上比较熟悉的人员[12]。

首先，采用 CiteSpace 文献分析工具[13]对国内的研究作者关系进行分析，生成作者合作关系图。如图 2 所示，其中共有 442 个网络节点，167 条连接线，网络密度为 0.0017。从图 2 中可以看出，国内作者分布比较分散，合作范围较小，没有形成明显的“聚焦点”。这表明随着大数据研究的进一步深入，加强国内研究人员之间的合作，形成较大的研究团队，将是大数据研究需要解决的一个关键问题，也将是大数据研究发展的必然趋势。

其次，使用 BibExcel 和 Pajek 文献分析工具，对国外文献作者合作关系进行分析。选取发文数量大于或等于 3 篇的 193 位作者，进行合作分析，得到合作关系图，截取合作关系图的一部分，如图 3 所示。图中的圆点大小代表作者发文数量的权重，箭头代表合作关系，箭头方向代表作者的主次关系。

对国外的高产作者的合作分析可以发现，国外的研究者的合作关系比较密切，形成相当多的合作团队，比如以 Carey，Michael J 等为主的研究团队，以 Jinjun，Chen 等为主的研究团队，以 Kepner，J 为主的研究团队等，在研究上有较多的合作关系，形成强大的研究能力，这是值得我们国内研究学者学习的。

4. 大数据研究的热点及其发展趋势分析

关键词代表论文的研究主题，通过提取论文的关键词，构建关键词共现网络，可以揭示该领域的研究热点。本文结合词频分析法和共词分析法，并使用文献可视化分析软件 CiteSpaceIII、BibExcel 和 VOSViewer，分别对国内、国外大数据文献的研究热点及其发展趋势进行可视化分析。

Table 3. Top 15 authors publishing big data research papers
表 3. 国内外大数据研究论文发表数量排名前 15 的作者情况

排名	篇数	作者	排名	篇数	作者
1	10	Jinjun, Chen	1	9	孟小峰
2	10	Ranjan, R.	2	7	于戈
3	10	Xuyun, Zhang	3	7	王士同
4	9	Yong, Chen	4	6	李龙澍
5	9	Carey, Michael J.	5	6	王珊
6	8	Lomotey, R. K.	6	6	葛浩
7	8	Markl, Volker	7	6	杨传健
8	8	Wanchun, Dou	8	6	李岩
9	8	Deters, R.	9	5	王伟
10	8	Chang, Liu	10	5	乐嘉锦
11	7	Borkar, Vinayak	11	4	程学旗
12	7	Xhafa, F.	12	4	高宏
13	7	Cuzzocrea, Alfredo	13	4	杨涛
14	7	Li, Chen	14	4	李建中
15	6	Yang, Li	15	4	史忠植

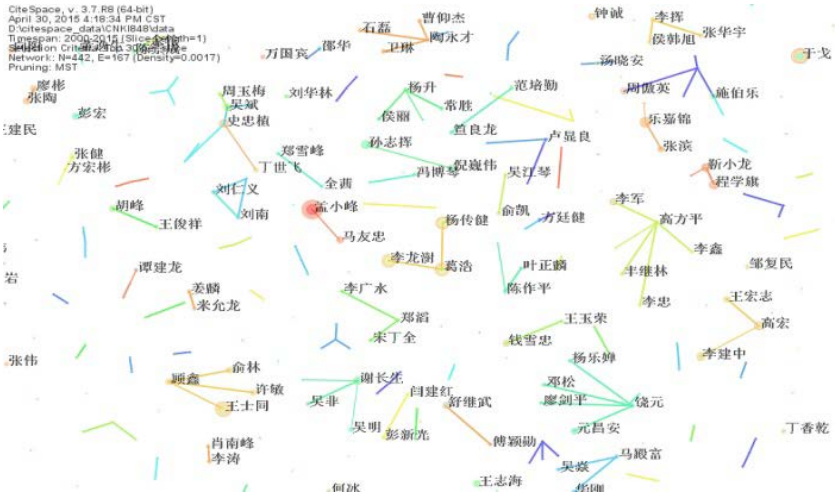

Figure 2. Domestic author cooperation relationship diagram
图 2. 国内高产作者合作关系图(由于数据量大，为了使可视化图片显示完整，因此字体显示较小，无法提供更清晰的图片)

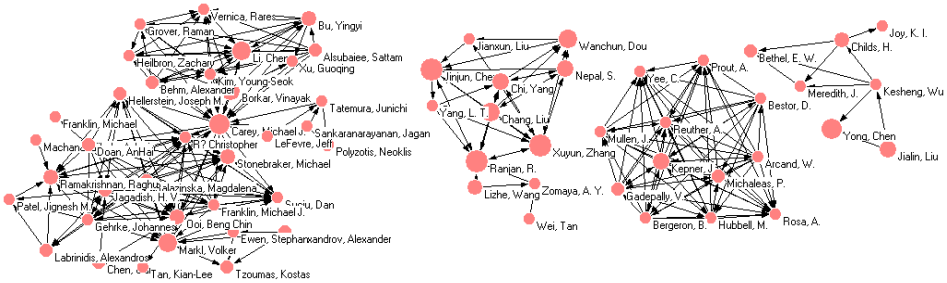

Figure 3. Foreign authors cooperation relationship diagram
图 3. 国外高产作者合作关系图

4.1. 国内大数据研究的热点

将从 CNKI 获得的 1041 篇文献数据导入 CiteSpaceIII；时间跨度选择 2000~2015 年，单个时间片长度取 1 年；节点选择为“关键词”，并选取频次高于 6 的关键词；将时间片阀值设置为 30。由 CiteSpaceIII 软件形成“关键词”关系图，显示出当前大数据研究的热门词汇，也就是热门研究主题，如图 4 所示。图中共有 388 个网络节点、250 条连接线，网络密度为 0.0033。其中每个圆形节点代表一个关键词，节点的大小代表该关键词在研究领域内出现的频率，节点的颜色年轮代表该关键词出现的年份，即持续的时间。

通过对论文的关键词进行统计，可得关键词一共 2627 个，根据关键词出现频次的高低，本文选取排名前 30 的关键词进行统计分析，如表 4 所示。

统计结果显示，我国在大数据研究领域主要涉及的热点有数据挖掘、MapReduce、云计算、Hadoop、聚类等。从这 30 个关键词可以看出，目前在大数据领域的研究，主要还是归结在大数据的获取和存储，大数据的交流(使用)，大数据的管理和内容的处理，大数据的分析，以及大数据的可视化[14]等研究方向。云计算成为大数据存储和计算的主要支撑技术，与大数据的研究密切相关。现有的 Hadoop、MapReduce 平台因其分布式并行计算能力，成为大数据研究的主要途径和工具[15]。此外，数据挖掘在关键词网络中也占据重要的地位，是数据分析的重要方法，涉及关联规则、聚类、分类、支持向量机、机器学习等。

4.2. 国内大数据研究的发展趋势

将关键词共现网络视图按时区方式(time-zone)显示，得到关键词的时区视图，可以显示国内大数据图 5 国内大数据研究趋势变化图的发展与演化趋势，如图 5 所示。

由图 5 可以看出，数据挖掘早在 2001 年就有相关文献的研究，如今成为大数据分析领域中的主要技术之一。此后陆续出现相关的技术与方法，如“聚类”、“粗糙集”、“支持向量机”、“机器学习”、“关联规则”等。从图还可以看出，数据挖掘与今后各年的研究都存在连接情况，表明数据挖掘不仅是当时的研究热点，还将是未来的的主流研究方向之一。

云计算、Hadoop、MapReduce 等关键词在 2007 年以后出现，并从 2011 年开始显著增加。从关键词共现的角度来看，大数据与云计算的关系十分密切，研究者注重大数据与云计算的结合。而大数据这个关键词是从 2012 年开始大量出现，成为计算领域的研究热点。

此外，通过 CiteSpace 软件查找 2000~2015 年期间国内大数据研究领域出现的突变词，这些突变词能够较准确地反映“大数据”研究领域中的研究趋势。结果显示，2010 年以前国内大数据出现的突变词有“支持向量机”、“聚类”、“大数据量”、“属性约简”等。2010 年以后对于大数据的研究兴趣不断增加，此阶段出现的突变词分别是大数据、MapReduce、云计算、Hadoop、云存储等。这表明国内研究者更加关注算法和技术方面的研究。除此之外，大数据的突变率高达 36.97，这一方面表明研究的延续性不够，另一方面表明新的研究正在不断兴起。

4.3. 国外大数据研究的热点

使用 BibExcel 软件对 3773 篇外文文献的关键词进行共词分析，取词频排名前 75 的关键词(词频数大于或等于 44)，应用可视化工具 VOSViewer，绘制热点主题知识图谱，并用颜色区分热度，如图 6 所示。红色代表关键词出现的频次很高，黄色代表频次较高，绿色代表平均水平，蓝色代表平均水平以下。

对国外研究论文的关键词进行统计，共 17,882 个，抽取关键词词频排名前 30 的关键词，如表 5 所示。结合图 6 和表 5 分析可知，在国外大数据研究中，呈现出的研究热点主要有数据分析(Data Analysis)，数据挖掘(Data Mining)，云计算(Cloud Computing)，数据处理(Data Handling)，数据模型(Data Models)，

Table 4. Domestic paper keywords frequency top 30
表 4. 国内论文关键词词频排名前 30

排名	词频	关键词	排名	词频	关键词
1	123	大数据	16	10	分布式
2	47	数据挖掘	17	9	无线传感器网络
3	34	MapReduce	18	9	负载均衡
4	30	云计算	19	9	遗传算法
5	29	Hadoop	20	9	大数据集
6	25	聚类	21	9	数据采集
7	22	粗糙集	22	9	数据分析
8	16	支持向量机	23	9	分类
9	16	属性约简	24	8	模糊聚类
10	14	并行算法	25	8	入侵检测
11	12	机器学习	26	8	数据库
12	11	并行计算	27	7	聚类分析
13	11	关联规则	28	7	云存储
14	10	可视化	29	7	物联网
15	10	大数据量	30	7	MapReduce 模型

Table 5. Hot spot analysis of foreign research paper
表 5. 国外论文关键词词频排名前 30

词频	关键词	词频	关键词
1373	Big data	128	Algorithm design and analysis
383	data analysis	122	learning (artificial intelligence)
343	data mining	116	Data visualization
311	cloud computing	114	Servers
272	data handling	112	pattern clustering
226	Data models	103	Clustering algorithms
210	Information management	100	Big Data Analytics
184	MapReduce	96	Computer architecture
173	Data storage systems	91	machine learning
147	Hadoop	89	Real-time systems
141	parallel processing	89	data visualisation
139	Computational modeling	88	query processing
139	Internet	85	storage management
138	Databases	79	Security
136	Distributed databases	76	social networking (online)

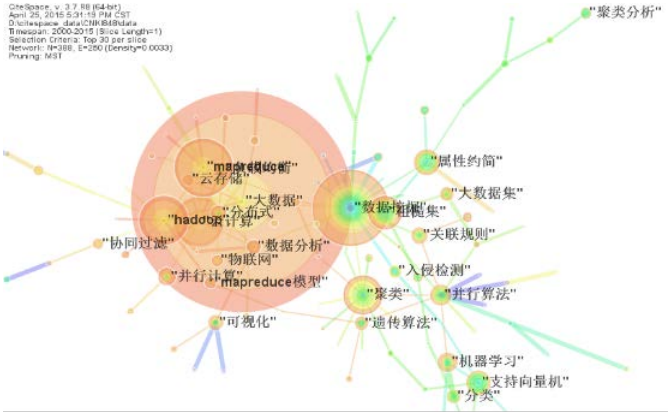

Figure 4. Domestic keywords co-occurrence diagram
图 4. 国内关键词共现关系图

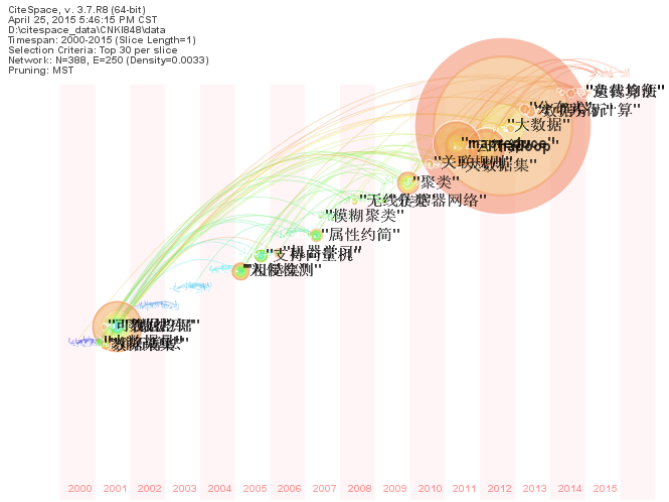

Figure 5. Domestic large data research trend variation
图 5. 国内大数据研究趋势变化图

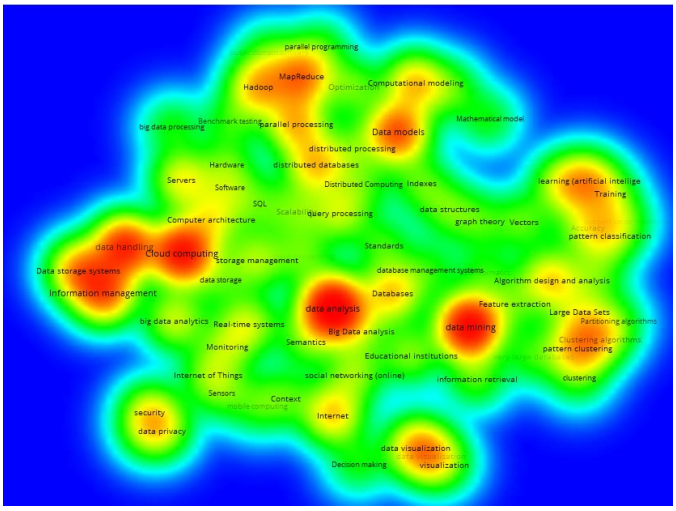

Figure 6. Hot spot analysis of foreign research paper
图 6. 国外文献关键词热点分析

以及信息管理(Information Management)等；其次是并行处理(Parallel Processing)，分布式处理(Distributed Processing)，学习、人工智能(Learning、Artificial Intelligence)，模式分类(Pattern Classification)，模式聚类(Pattern Clustering)，数据隐私(Data Privacy)，安全(Security)等。

对国外大数据领域研究论文中词频排名前 75 的关键词(词频数大于或等于 44)进行共词分析，如图 7 所示。

从图 7 可以看到，关键词之间的亲疏关系和明显的聚类现象。关系紧密的关键词表示这些主题的关联性，是该领域相互关联的技术，可以协作解决一个问题。采用 VOSViewer 工具，根据关键词共现的亲疏关系，将所有的关键词聚类，分别使用红色、蓝色、绿色、黄色、紫色标识五大类关键词，得到大数据研究领域比较热门的五个主要的研究方向，如表 6 所示。

4.4. 国外大数据研究的发展趋势

为了对国外大数据领域研究的发展趋势进行分析，将全部 3773 篇国外文献按照年份区间分组：2000~2005 年的 62 篇文献为组 1；2006~2009 年的 78 篇文献为组 2；2010~2011 年的 85 篇文献为组 3；2012~2013 年的 975 篇文献为组 4；2014 年~2015 年的 2573 篇文献为组 5。

使用文献计量分析工具 BibExcel，对五个分组的文献的关键词分别进行处理，使用可视化工具 VOSViewer 生成相应的热点分析图，将这五个热点分析图中的重要关键词提取和绘制成图 8，展示了大数据领域研究的进展和趋势。

从图 8 中可以看到，“数据挖掘(Data Mining)”一词在 2000 年至今的十几年中一直保持着很高的热度，这说明了人们对于挖掘数据中的有价值的信息一直在进行持续的探索与研究。在 2000 年至今的十几年中一直保持着较高的热度的词还有“数据分析(Data Analysis)”和“人工智能(Artificial Intelligence)”等，显然它们是与大数据研究密切相关的研究领域。在 2000 年至 2005 年期间，国外文献还对“数据库(Database)”、“划分算法(Partitioning Algorithms)”、“Clustering Algorithms(聚类算法) [16]”的关注度较高，这也反映了大数据的研究一开始是始于数据库研究的事实。同时也看到“数据可视化(Data Visualization)”技术在大数据研究的早期就成为研究的热点，显示数据可视化的重要性和必要性。

在 2006 年至 2009 年期间，研究者们对数据“聚类(Clustering) [17]”和“分类(Classification)”关注很多，这是大数据研究的重要课题。与此同时，云计算技术悄然兴起，“分布式计算(Distributed Computing)”也在逐步地引入大数据研究中，成为主要的大数据计算解决方案[18]。

在 2010 年至 2011 年期间，“云计算(Cloud Computing)”一跃成为第二热的关键词，并在 2010 年至 2015 年期间保持高热，说明了云计算技术对大数据研究的影响力。“数据模型(Data Model)”一词在 2010 年至 2015 年期间保持着高热或次热的热度，说明了数据建模是大数据信息处理的重要方法，应是未来的一个研究热点。与此同时，“MapReduce”一词首次出现在次热词列表中，而“聚类(Clustering)”和“分类(Classification)”相关技术在 2012 年之后关注度下降，回复到平均水平。

从年份变化图 1 知道，2012~2013 年国外大数据领域研究增长迅速，期间“数据处理(Data Handling)”成为大数据研究领域最火热的关注点；“信息管理(Information Management)”和“数据存储系统(Data Storage System)”也由于数据的爆发式增长得到密切关注。与此同时，边缘词中出现了“隐私(Privacy)”和“安全(Security)”等词[19]，说明随着计算机技术的发展和时代的进步，数据的安全性和保护数据隐私成为人们关注的问题。

大规模数据集的并行运算[20]能有效地提高大数据处理速度，因此在 2010 年以后，“MapReduce”、“Hadoop”[21] [22]迅速成为大数据处理的主要解决方案。

图 1 显示大数据领域研究在 2014、2015 年得到迅猛发展，“数据分析”、“云计算”、“数据挖掘”、

Table 6. Foreign keywords clustering
表 6. 国外论文关键词聚类情况

分类	关键词	研究方向
红色类	Data mining, pattern clustering, large data sets, partitioning algorithms, feature extraction…	数据挖掘相关的技术和算法
绿色类	Cloud computing, Hadoop, MapReduce, parellel processing, distributed processing…	云计算、分布式计算相关技术和平台
蓝色类	Data analysis, Internet, social networking, internet of thing, security, data privacy…	大数据分析相关的网络和技术
黄色类	Data model, Computational modeling, data visualization, decision making…	数据建模与可视化相关的方法和技术
紫色类	Data handling, data storage systems, information management…	数据存储和管理、处理相关的方法和技术

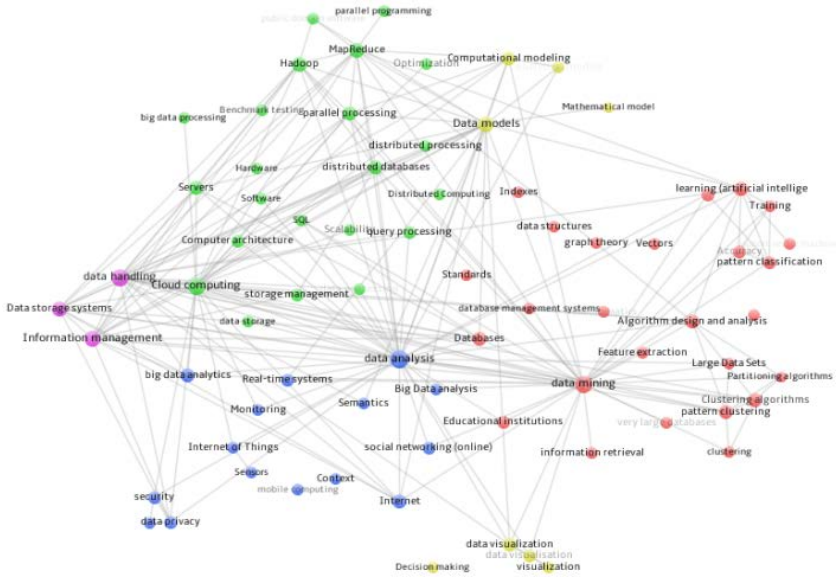

Figure 7. Keywords co-word analysis of foreign research paper
图 7. 国外文献关键词共词分析

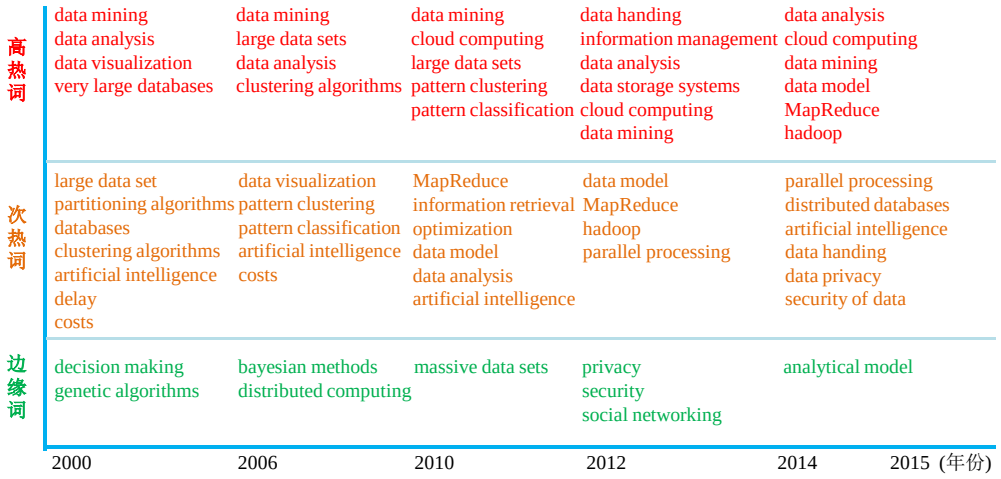

Figure 8. Change of hot spot on foreign big data research
图 8. 国外文献大数据研究热点变化

“数据模型”、“MapReduce”、“Hadoop”等继续成为大数据研究的热点，“并行处理”[23]的热度得到提升，“数据隐私”“数据安全”越来越得到人们的关注，并且数据的安全和隐私问题永远不会过时，将是未来的研究热点之一。

从对国外大数据的领域研究的分析可以看出，研究具有一定的延续性，而且在一开始就聚焦大数据研究的主要问题，研究数据挖掘、数据分析、数据建模、云计算、数据可视化等关键技术，研究线路比较明确。

5. 总结

本文通过文献计量的方法，对检索的文献数据进行可视化的分析，得出一系列有参考价值的结果。

在计算领域的大数据研究已经成为热点，而且在未来的几年将有继续增长的趋势；一些期刊和会议在大数据研究方面有较大的贡献，成为研究者们交流研究工作的主要平台；涌现出一批投身大数据研究的学者，以及以他们为核心的研究团队。

从大数据的研究热点和发展趋势来看，在向着系统化、人性化方向发展，关注解决与机器相关的问题和与人相关的问题。研究热点和趋势涉及数据挖掘、数据分析、云计算、MapReduce、Hadoop 等。国外的研究已经越来越关注大数据环境下的数据安全、数据隐私等问题。我们相信，国内外学术界对大数据的研究，将为大数据的有效利用，提供更多的有价值的解决方案。

参考文献 (References)

[1] 赛迪智库. 大数据时代需要加快布局[EB/OL]. http://www.cio360.net/index.php?m=content&c=index&a=show&catid=201&id=53375, 2012-05-17.

[2] 孟小峰, 慈祥. 大数据管理: 概念、技术与挑战[J]. 计算机研究与发展, 2013, 50(1): 146-169.

[3] 阿尔文·托勒夫, 著. 第三次浪潮[M]. 黄明坚, 译. 北京: 中信出版社, 2006: 19-25.

[4] 麦肯锡全球研究院. 大数据:创新、竞争和生产力的下一个新领域[EB/OL]. http://wenku.baidu.com/view/2e494d6d9b6648d7c1c746a7.html, 2014-05-04.

[5] Big Data Across the Federal Government. http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_fact_sheet_final_1.pdf

[6] Hey, T., Tansley, S. and Tolle, K. (2009) The Fourth Paradigm: Data-intensive Scientific Discovery. Microsoft Research, Redmond, Washington. http://research.microsoft.com/en-us/collaboration/fourthparadigm/

[7] Nature.Big Data. http://www.nature.com/news/specials/bigdata/index.html

[8] Reichman, O.J., Matthew, B., Mark, P.H., *et al.* (2011) Challenges and Opportunities of Open Data in Ecology. *Science*, **311**, 703-705. http://www.sciencemag.org/

[9] 郑文晖. 文献计量法与内容分析法的比较研究[J]. 情报杂志, 2006, 25(5): 31-33.

[10] 汤建民. 2006 年国内科学学研究的词频分析与计量研究[J]. 科学学研究, 2007, 25(s2): 518-522.

[11] 冯璐, 冷伏海. 共词分析方法理论进展[J]. 中国图书馆学报, 2006, 32(2): 88-92.

[12] 冯博, 刘佳. 大学科研团队知识共享的社会网络分析[J]. 科学学研究, 2007, 25(6): 1156-1163.

[13] 陈超美, 陈悦, 侯剑华, 梁永霞. CiteSpaceII: 科学文献中新趋势与新动态的识别与可视化[J]. 情报学报, 2009, 28(3): 401-421.

[14] 任磊, 杜一, 马帅, 张小龙, 戴国忠. 大数据可视分析综述[J]. 软件学报, 2014(9): 1909-1936.

[15] 孟小峰, 慈祥. 大数据管理: 概念、技术与挑战[J]. 计算机研究与发展, 2013, 50(1): 146-169.

[16] Yang, Y., Ni, X.H., Wang, H.J., *et al.* (2012) Parallel Implementation of Ant-Based Clustering Algorithm Based on Hadoop. In: *Proceedings of ICSI* 2012, Springer, Berlin, 190-197.

[17] Nair, S. and Mehta, J. (2011) Clustering with Apache Hadoop. In: *Proceedings of ICWET* 2011, ACM, New York, 505-509.

[18] Isard, M. and Yu, Y. (2009) Distributed Data-Parallel Computing Using a High-Level Programming Language. In: *Proceedings of SIGMOD* 2009, ACM, New York, 987-994.

[19] Agrawal, R. and Srikant, R. (2000) Privacy Preserving Data Mining. In: *Proceedings of SIGMOD* 2000, ACM, New York, 439-450.

[20] Hadoop, W.T. (2009) The Definitive Guide. 2nd Edition, O'Reilly Media, California.

[21] Dean, J. and Ghemawat, S. (2008) MapReduce: Simplied Data Processing on Large Clusters. *Communications of the ACM*, **51**, 107-113.

[22] Hadoop. http://hadoop.apache.org/index.html

[23] Chaiken, R., Jenkins, B., Larson, P., *et al*. (2008) SCOPE: Easy and Efficient Parallel Processing of Massive Data Sets. *Proceedings of the VLDB Endowment*, **1**, 1265-1276.