国际大数据研究主题的可视化分析*

王一博[1]，郭鑫[2]，王继民[2]
（1. 北方工业大学信息工程学院计算机系，北京 100144；2. 北京大学信息管理系，北京 100871）

摘要：随着大数据时代的来临，有关大数据的理论、技术、方法与应用的研究已成为当前产、学、研的研究热点。以Web of Science数据库的文献信息为数据源，对大数据领域的论文关键词进行共词分析，构建高频关键词共现关系网络，可视化地展示该网络的“核心-边缘结构”，通过聚类分析将这一领域的研究内容划分为14个类团，并利用战略坐标图揭示该领域的各个研究主题及其发展趋势，以期为相关研究提供参考。

关键词：大数据；数据可视化；社会网络分析；聚类分析；战略坐标图
中图分类号：G350
DOI：10.3772/j.issn.1673—2286.2014.07.009

随着移动互联网、物联网、云计算技术的快速发展，以及视频监控、智能终端、应用商店的普及，全球数据量出现了爆炸式增长。Gartner提出，目前半结构和非结构化的数据，诸如文档、表格、网页、音频、图像和视频等占全球网络数据量的85%左右，大数据隐含着巨大的社会、经济和科研价值，并且整个互联网络体系架构也将面临革命性的改变[1]。

麦肯锡将大数据定义为：无法在一定时间内用传统数据库软件工具对其内容进行抓取、管理和处理的数据集合[2]。大数据的特点通常用4个V来概括，即Volume（规模性）、Velocity（高速性）、Variety（多样性）和Value（价值密度低），这被认为是大数据有别于传统数据集的特征。在大量数据的背后，有复杂的定位、访问、检索、存取、交换等活动，现有的网络环境、存储及搜索条件，都难以适应这种新的变化。随着大数据的快速发展与进步，人类的生产生活方式正在发生根本性的变革[3]。

2012年3月，美国联邦政府宣布投入2亿多美元启动大数据的研发任务，并把大数据定义为和历史上的互联网、超级计算同等重要的国家战略[4]。我国也对大数据的理论与应用开展了深入系统的研究，目前大数据已经渗透到社会经济各个层面，受到了各个学科的高度关注。因此，对国际范围内大数据领域的研究现状进行分析具有重要意义。本文基于Web of Science数据库（WoS）中以“big data”为主题的论文题录信息，通过聚类分析、社会网络分析等方法，对国际范围内的大数据领域的研究现状和研究热点进行可视化分析，以期为我国学者对大数据领域的深入研究提供参考和借鉴。

1 数据准备

1.1 数据收集

本文以美国科学情报研究所（Institute for Scientific Information，ISI）出版的国际权威引文数据库WoS（Web of Science with Conference Proceedings: SCI-EXPANDED；SSCI；A&HCI；CPCI-S；CPCI-SSH）为文献信息源，以“主题”为检索项，检索式为“big data”，时间跨度为2009年-2014年，即近6年的文献数据，因为2009年之前论文较少（仅有52篇），且与目前意义的“大数据”含义不完全相同。检索得到文献1330篇。之后对数据进行筛选，除去错检项后，得到期刊文献（包括Article，Article；Book Chapter，Article； Proceedings Paper）1113 篇，占文献总数的83.68%。

1.2 数据清洗

关键词是为了文献标引工作从报告、论文中选取出

* 本研究得到北京市科技计划项目“科学知识图谱方法在新兴产业发展态势分析中的应用研究”（编号：ZI21108002212058）资助。

来的用以表示全文主题内容信息的单词或术语。Web of Science论文的关键词分两种，一种是作者关键词（DE字段），另一种是增补关键词（ID字段）。增补关键词是ISI根据参考文献的标题中摘取的主题词[5]。为了全面反映大数据领域主题的研究，本文采用了将二者相结合的办法，即将两种关键词进行合并，并删除重复的内容。在本次研究中，有819篇文献给出了作者关键词，有531篇给出了增补关键词，有373篇既有作者关键词也有增补关键词。对数据进行处理后，得到关键词3,061个，累计频次为3,975次。

随后，笔者利用自己编写的计算机程序，对下载的WoS题录信息中的关键词进行词频统计。人工去除主题过于宽泛的通用词汇，如big data（大数据）、component（组件）、algorithms（算法）等。此外，还需要对同义词进行合并处理。主要通过词频统计表人工制定了一些映射规则，并利用计算机程序将该规则应用于原题录信息中的关键词替换。部分映射规则如表1所示。

表1 映射规则（部分）

原高频词	替换后高频词
data analytics	data analysis
big data visualization	data visualization
prediction	predictive analytics
cloud	cloud computing

1.3 高频关键词

完成替换后，再次进行词频统计，并按照词频降序排列，取频次大于4的50个词作为高频关键词，部分关键词的词频列表如表2所示。

1.4 共词矩阵

共词矩阵呈现的是词与词之间的共现次数，根据表2所示的高频关键词列表，笔者利用计算机程序统计得到高频关键词的共现矩阵，部分结果如表3所示。

表3 高频词共现矩阵（部分）

高频词	词1	词2	词3	词4
词1	93	25	15	16
词2	25	79	4	21
词3	15	4	61	6
词4	16	21	6	49

在共词矩阵中，对角线上的数值即为该词出现的总频次。

1.5 相关矩阵和相异矩阵

在实际共词分析过程中，关键词共现频次受到各

表2 高频关键词表（部分）

关键词/（简称）	词频	关键词/（简称）	词频
cloud computing/（词1）	93	predictive big data analysis/（词12）	12
MapReduce/（词2）	79	NoSQL/（词13）	12
big data analysis/（词3）	61	distributed computing/（词14）	11
Hadoop/（词4）	49	social networks/（词15）	10
data visualization/（词5）	33	education/（词16）	10
data mining/（词6）	32	Clustering/（词17）	10
privacy/（词7）	19	GPU/（词18）	9
Performance/（词8）	19	Optimization/（词19）	8
machine learning/（词9）	18	Genomics/（词20）	8
social media/（词10）	14	Bioinformatics/（词21）	8
Twitter/（词11）	13	Internet of Things/（词22）	7

自词频大小的影响。为了消除初始共词矩阵绝对值差异的影响，准确揭示关键词之间的共现关系，本文利用Ochiia系数将共词矩阵转换为相关矩阵，结果如表4 所示。Ochiia系数的计算公式如下：

$$W_1,W_2\text{两词的Ochiia系数}=\frac{W_1,W_2\text{共现次数}}{\sqrt{W_1\text{出现频次}*W_2\text{出现频次}}}$$

为方便处理，用“1”与相关矩阵中的各数值相减，得到表示两词之间相异程度的相异矩阵，结果如表5所示。

表4 相关矩阵（部分）

高频词	词1	词2	词3	词4
词1	1	0.2917	0.1992	0.237
词2	0.2917	1	0.0576	0.3375
词3	0.1992	0.0576	1	0.1097
词4	0.237	0.3375	0.1097	1

表5 相异矩阵（部分）

高频词	词1	词2	词3	词4
词1	0	0.7083	0.8008	0.763
词2	0.7083	0	0.9424	0.6625
词3	0.8008	0.9424	0	0.8903
词4	0.763	0.6625	0.8903	0

在相异矩阵中，数值接近1表示相异程度较高，数值接近0则表示相异程度较低。

2 可视化分析

2.1 关键词共现的核心—边缘结构

关键词共现矩阵可转化为高频关键词之间的共现关系网络，在该网络中，结点表示高频关键词，边及其权值为关键词的共现次数[6]。核心—边缘结构分析是根据网络中结点之间联系的紧密程度，将网络中的结点分为两个区域：核心区域和边缘区域。处于核心区域的结点在网络中占有比较重要的地位，核心—边缘结构分析的目的是研究社会网络中哪些结点处于核心地位， 哪些结点处于边缘位置，它是对网络“位置”结构进行量化的分析[7,8]。基于表3得到的关键词共现关系网络，利用社会网络分析软件UCINET和Pajek进行核心—边缘网络结构的计算和呈现，结果如图1所示。

图1显示：有10个关键词处于核心位置，包括big data analysis（大数据分析）、Hadoop、MapReduce、data visualization（数据可视化）、cloud computing（云计算）、storage（存储）、clustering（聚类）、performance（性能）、data mining（数据挖掘）、privacy（隐私）。在这些核心关键词中，结点度值最大且相连边的权值最大的关键词都是cloud computing（云计算）。统计显示：平均每个核心关键词出现在63.3篇文献中。可以说，这10个核心关键词在大数据研究领域占有比较重要的位置。

2.2 聚类结果分析

根据数据对象的特征属性，聚类分析可将数据对象集合划分为若干个不同的类团或簇，使得同一类团中的数据对象具有较大的相似性，不同类团中的数据对象具有较大的相异性[9]。将相异矩阵导入统计分析软件SPSS中进行层次聚类，得到聚类结果。根据聚类树状图，在阈值为9.5处进行划分，可将这50个高频词分成14个词团，个别英文词过长无法在SPSS聚类图中显示，故删去这些词后的少许字母，聚类结果如图2所示。

为使结果呈现更加直观，笔者将部分关键词译为中文，具体如下文所示：

K1：性能，高性能计算；

K2：元数据、数据安全、存储、HDFS（Hadoop分

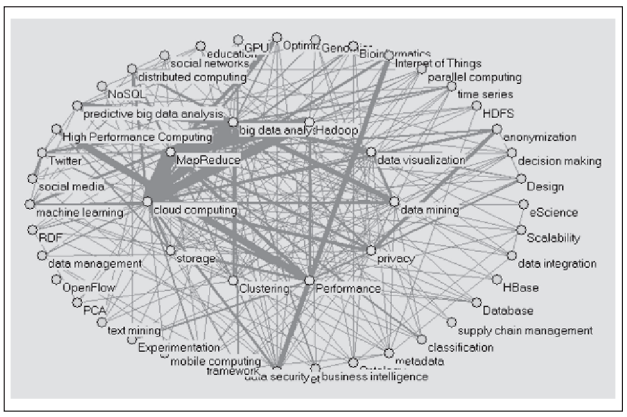

图1 核心-边缘结构图

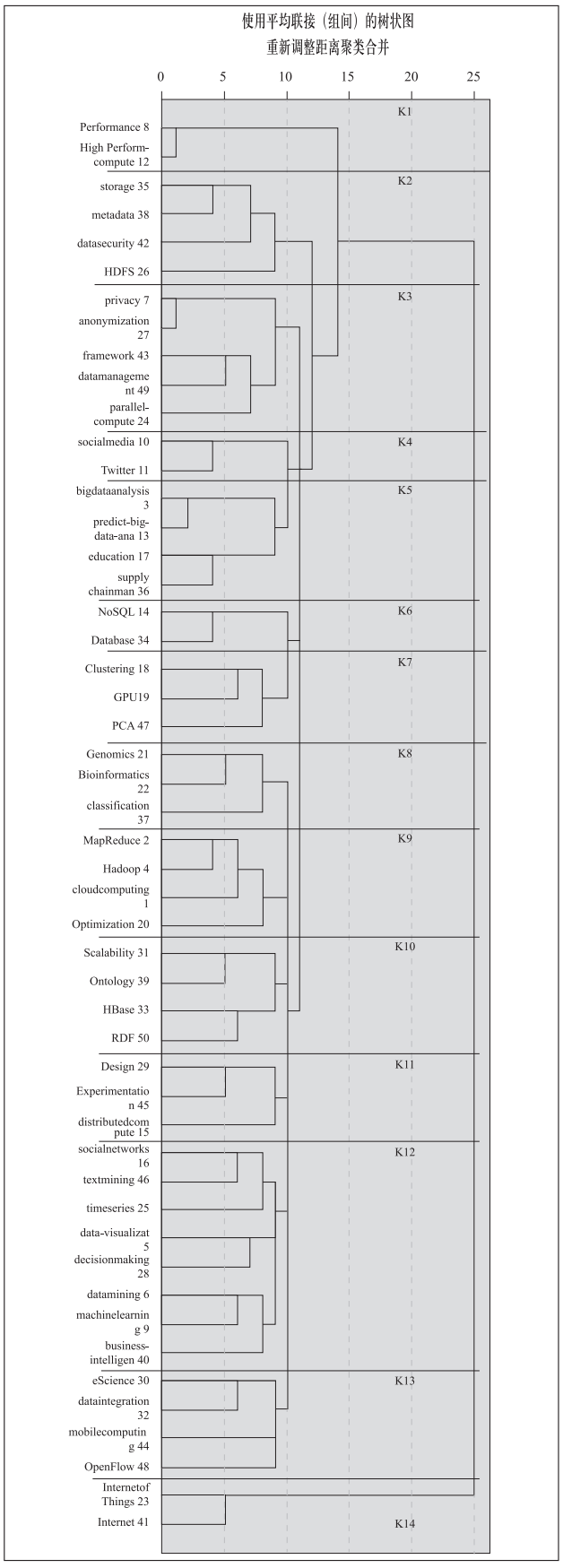

图2 聚类结果图

布式文件系统）；

K3：隐私、匿名化、并行计算、架构、数据管理；

K4：社会媒体、Twitter；

K5：大数据分析、大数据预测、教育、供应链管理；

K6：数据库、NoSQL（非关系型的数据库）；

K7：聚类、GPU（图形处理器）、PCA（主成分分析）；

K8：基因组学、生物信息学、分类；

K9：云计算、Hadoop（分布式系统基础架构）、MapReduce（一种编程模型）、最优化；

K10：可伸缩性、HBase（分布式存储系统）、本体、RDF（资源描述框架）；

K11：分布式计算、设计、实验；

K12：数据可视化、数据挖掘、机器学习、社会网络、时间序列、决策、商业智能、文本挖掘；

K13：eScience、移动计算、数据集、OpenFlow（一种新型网络交换模型）；

K14：互联网、物联网。

2.3 战略坐标图

战略坐标图可以概括地展现一个领域的结构，它把每一个研究主题放置到一个坐标系的四个象限中，进而描述各主题内部的联系情况和各主题间的相互影响的情况。该坐标系的横轴表示向心度，纵轴表示密度[9]，所有的主题词团都将划分到四个象限中。

对词团密度和向心度的计算有不同的方法，本文采用的计算公式为：

$$\text{密度}=\frac{2*\sum_{i,j\in K,i\neq j}E_{i,j}}{n}，\quad \text{向心度}=\frac{\sum_{i\in K,j\notin K}E_{i,j}}{n}$$

其中，E_{ij}是关键词i和关键词j共现的次数，K代表通过聚类分析得到的某一词团，n是该词团所含关键词的数目。

根据3.2节得到的聚类结果和高频词共现矩阵，利用上述计算公式，对每个词团的密度和向心度进行计算。随后，利用SPSS软件对计算结果进行Z-score规范化，并根据规范化结果绘制战略坐标图，结果如图3所示。

从战略坐标图3可以看出，第一象限包括K1、K5和K9三个词团，第二象限仅包含K14词团，其他词团均位于第三象限。

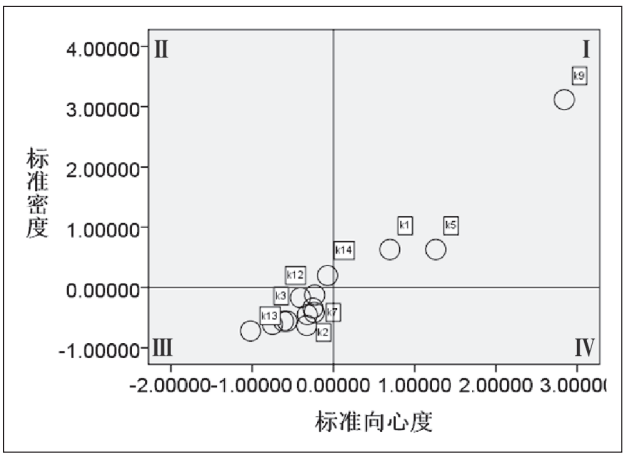

图3 战略坐标图

K1词团主要研究的是高性能计算。虽然大数据技术得到了快速的发展，但仍然面临着许多问题。比如，大数据的存储和处理都给计算机系统带来沉重的负荷，传统的计算方式已经不能适应大数据的处理。而高性能计算能够有效降低海量数据处理系统的压力，提高系统运行效率。可以说，高性能计算完美契合了大数据在运算能力、高性能存储等方面的需求。K5词团研究的是大数据的应用。业界普遍认为，“大数据”拥有“大价值”。使用大数据进行商业分析、趋势预测，以及在供应链管理方面的应用，成为许多学者关注的问题。K9词团主要研究的是大数据的处理工具与技术。云计算和大数据具有紧密的联系，云计算为大数据提供了基础架构平台，大数据的处理和应用可在这个平台上运行。基于MapReduce框架开发的Hadoop则是现今公认的处理大数据最有效的工具。

从战略坐标图中可以看到，K9这一词团无论是密度还是向心度都远高于其他词团，这也就意味着该词团内部联系紧密，且与其余各词团有广泛的联系。可以认为，大数据研究工具Hadoop和云计算处理技术是大数据领域最为核心的研究内容。根据核心—边缘的分析结果，云计算、Hadoop和MapReduce都属于核心关键词，这也印证了核心地位。除此之外，大数据的应用方式和高性能计算也是大数据领域较为核心的研究内容。

K14词团研究的是物联网。物联网是物物相连的互联网络，其用户端延伸和扩展到了任何物品与物品之间，进行信息交换和通信。同时，物联网的数据几乎都是半结构化甚至是非结构化的，并且增长率非常高。作为大数据的重要来源，物联网已经得到了许多学者的关注。从战略坐标图可以看到，K14词团密度较高，向心度略低，这意味着对于物联网本身的研究已相对成熟。

剩余几个词团的向心度和密度都比较低，处于大数据研究领域的边缘位置，尚未成熟。比如，K3词团研究的是大数据带来的隐私问题，K10研究的是大数据的信息表示问题。此外，还有一些研究主题是大数据在某些具体领域中的应用，比如K4词团研究了大数据在社交网络中的应用，K8词团研究了大数据在医疗领域中的应用，等等。

3 结语

本文基于WoS收录的有关大数据（Big data）的研究论文，利用社会网络分析、聚类分析、战略坐标图等研究方法，通过可视化手段对国际范围内大数据领域的研究现状进行分析和解读，揭示了该领域若干研究主题及其发展现状。研究结果显示，大数据的处理技术与工具如云计算技术、物联网技术、MapReduce、Hadoop、大数据分析与预测、高性能计算等研究主题是大数据领域的核心研究内容。

基于文献信息对大数据研究主题进行预测是本文下一步分析的重点内容。

参考文献

[1] 陈如明.大数据时代的挑战:价值与应对策略[J].移动通信,2012(17):14-15.

[2] 严霄凤,张德馨.大数据研究[J].计算机技术与发展,2013(4):168-172.

[3] 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(1):146-169.

[4] 李国杰,程学旗.大数据研究:未来科技及经济社会发展的重大战略领域——大数据的研究现状与科学思考[J].中国科学院院刊,2012(6):647-657.

[5] 朱庆华,彭希羡,刘璇.基于共词分析的社会计算领域的研究主题[J].情报理论与实践,2012(12):7-11.

[6] 岳洪江,刘思峰.国外管理学博士论文研究主题的可视化分析[J].科学学与科学技术管理,2008,29(3):91-94.

[7] 张世怡,刘春茂.中文网站社会网络分析方法的实证研究[J].情报科学,2011(2):246-252.

[8] 刘军.整体网分析讲义:UCINET软件实用指南[M].格致出版社,2009.

[9] HAN Jiawei, KAMBER M.数据挖掘概念与技术[M].范明,等译.北京:机械工业出版社,2007.

作者简介

王一博，男，1992年生，北方工业大学信息工程学院计算机系本科生。
郭鑫，男，1992年生，北京大学信息管理系本科生。
王继民，男，1966年生，北京大学信息管理系副教授，研究方向：文本信息处理、Web挖掘、复杂网络等，通讯作者，E-mail：wjm@pku.edu.cn。

Visualization Analysis of the Achievements in International Big Data Domain

WANG YiBo[1], GUO Xin[2], WANG JiMin[2]
(1. College of Information Engineering, North China University of Technology, Beijing 100144, China;
2. Department of Information Management, Peking University, Beijing 100871, China)

Abstract: With the explosion of Internet data, the era of big data is coming. Taking WoS (Web of Science) as data source, this paper analyzes the key words of big data domain through co-word analysis, and determines periphery structure of co-word network by using social network analysis, which visually shows co-word network. We divide the research topic of this field into 14 groups by applying clustering analysis, and reveal the core research topics of big data field by combining with strategic diagram, thus to provide some experience and reference in theory and application of big data.

Keywords: Big data; Visualization analysis; Social network analysis; Clustering analysis; Strategy coordinate diagram

（收稿日期：2014-06-17）