基于文献计量的“大数据”研究

杨　绎

(中国科学技术信息研究所　上海图书馆上海科学技术情报研究所　200031)

文　摘　随着计算机技术以及互联网的发展,人们早已从“信息匮乏”发展到了“信息爆炸”的时代。作为信息基础的数据无论从量上还是自身特点上也在不断发展,“海量数据”的说法由来已久,而在近两年,“大数据”(big data) 逐渐成为一个热词,本文从文献计量的角度出发,以国内文献为研究基础,以关键词分析为主要方法,对“大数据”进行了研究。

关键词　大数据　海量数据　文献计量　关键词分析

1　关于大数据

仅从字面来看,“大数据”(big data) 显然不是一个新词。而以前,人们习惯将大量数据称为“海量数据”。早在 1998 年,《科学》杂志上关于一种计算机软件 HiQ 的介绍就应用了《大数据的管理者》(A Handler for Big Data) 这一说法。

不过在学术期刊上,与近年来的“大数据”含义最为接近的可能是 2008 年 9 月《自然》杂志上的“big data”专刊,该专刊的多篇文章分别从互联网经济、超级计算、生物医药等多方面介绍了“大数据”带来的技术挑战、现有技术以及未来的发展方向,编辑总结文章《大数据: PB 级数据时代的科学》成为热点。这可能也是“大数据”一词逐步得到业界肯定和接受的开端。

从现有文献资料来看,对“大数据”时代进行背景描述往往会提到 IDC,通过进一步研究可以发现,这些描述往往来自于 IDC 接受 EMC 资助,自 2007 年开始发布的“数字宇宙”年度专题系列报告,该报告提出庞大的“数字宇宙”中包括大量半结构化、非结构化的数据。2011 年该报告题名为“从混乱中挖掘价值”。因此,虽然在该系列报告中,直到 2011 年才正式出现“big data”这个词语,但却无可否认它在揭示“大数据”现象、推动“大数据”技术发展的过程中所起到的重要作用。

不过,“大数据”一词真正成为热点还是在 2011 年。当年 5 月,EMC 在 EMC World 2011 大会上正式抛出了“大数据”概念,事实上,该届大会主题即为“云计算相遇大数据”; 差不多同时,McKinsey 发布的报告《大数据: 创新、竞争和生产力的下一个新领域》在网上引起巨大反响; 8 月,Gartner 在其著名的新兴技术成熟度曲线报告中,将“大数据”加入其中。2012 年 3 月,奥巴马政府发布《大数据研究和发展倡议》,并将为此注资 2 亿美元,这意味着美国政府对“大数据”的关注和重视。

2　文献资料收集以及分析

关于“大数据”的讨论在国外如火如荼,那么国内对其有多少关注,接受程度又有多少。为此,本文以国内文献为基础,以关键词分析为主要方法,应用文献计量方法进行研究。

2.1　文献资料的收集

一般而言,主要的文献来源包括图书、期刊、报纸以及网络资源。由于“大数据”仅在近几年成为热点,而图书的出版周期较长并不适合; 网络资源丰富而且反应速度最快,但信息噪音较强。因此本文主要以期刊和报纸资源作为文献资料来源。

本文选择中国知网(cnki) 的中国学术期刊网络出版总库以及中国重要报纸全文数据库作为检索数据库。经过前期的阅读和研究,发现目前在“海量数据”与“大数据”之间并无明显的界限; 另外,由于该研究对象属于一种社会现象,因此并不限于某一学科。最终对于期刊的检索策略定为篇名 = “大数据”or“big data”or 关键词 = “大数据”or“big data”or 篇名 = “海量数据”or 关键词 = “海量数据”进行“精确”检索,检索日期为 2012 年 4 月 11 日; 对于报纸,由于其对于新热点较期刊更敏感,因此检索策略定为题名 = “大数据”or“big data”or 关键词 = “大数据”or“big data”,检索日期为 2012 年 4 月 17 日。检索结果为期刊 1434 条记录,经去重、筛选等数据清

洗过程后,保留 1348 条记录;报纸检索得 149 条记录,经数据清洗后保留 94 条记录作为研究对象。利用参考文献[1]中提到的方法,可以将检索结果文献的题录导入 Excel,并进一步得到关键词表。

2.2 期刊文献的分析

经阅读发现,许多期刊文章在题名中直接出现"大数据"、"大数据时代"等词语,明显与"大数据"现象紧密相关,但却并未将"大数据"作为关键词,这会对基于关键词的研究造成较大影响,因此在这些文章的关键词中加入"大数据"。另外,除一些中英文关键词的统一,例如"data mining"转换为"数据挖掘"、"地理信息系统"转换为"GIS"、"甲骨文"转换为"Oracle"等之外,对关键词未作更多处理。

(1)期刊论文的年度分布。从"大数据"在关键词中出现的情况可以看出人们对它的接受情况,在所有 1348 篇期刊论文中,关键词中出现"大数据"的文献所占比例如图 1 所示:

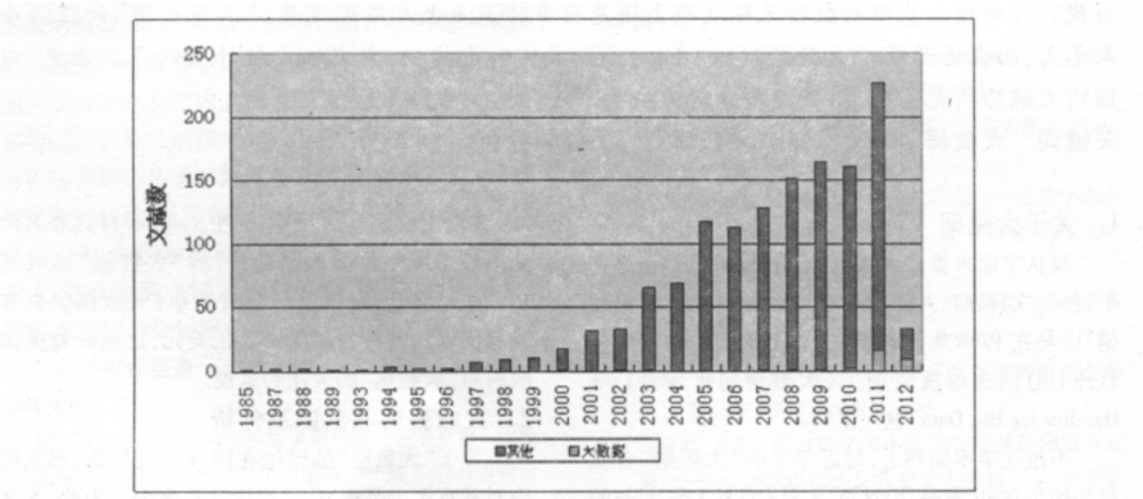

图 1 关键词包含"大数据"文献所占比例

可以看出,在国内期刊文献中,以"大数据"作为关键词的文献所占比重很小,从 2011 年开始出现,但是其所占的比重却迅速增大。2011 年为 16 篇,占全部文章的 7%,但截止到检索时间点的 2012 年前几个月中,已经出现了 9 篇文章,所占比例达到 28%,说明自 2011 年以来,"大数据"正迅速成为人们关注的焦点,不过其是否会在今后逐步取代"海量数据"的说法,虽有趋势显现,但尚无法定论。

(2)期刊论文的关键词分析。经前述的数据清洗过程后,即可进行文献计量研究。

利用 Excel 进行统计,所有期刊论文中,共出现 4116 个关键词,总计出现 7100 次。包括"海量数据"在内,词频超过 10 的高频词共有 48 个,如表 1 所示。可以看出,在"海量数据"、"大数据"、"大数据量"的基础上,"数据挖掘"、"数据处理"、"数据分析"、"数据管理"、"数据存储"、"数据采集"等是相关的研究热点,"数据仓库"、"云计算"、"Hadoop"、"MapReduce"、"分布式"、"多线程"等技术被应用于"地理信息系统(GIS)"、"商业智能"、"企业"、"数字图书馆"等领域。

表 1 期刊文献高频关键词表

关键词	词频	关键词	词频	关键词	词频
海量数据	514	数据管理	20	聚类	12
数据挖掘	178	MapReduce	19	算法	12
数据仓库	76	数据挖掘技术	19	可视化	12
数据库	72	关系数据库	19	聚类分析	11
GIS	49	企业	18	数据采集	11
云计算	43	索引	18	决策支持	11
解决方案	38	网格	18	信息技术	11
大数据	36	服务器	16	分布式计算	10
大数据量	34	应用	14	挑战	10

数据处理	26	支持向量机	13	优化	10
数据中心	26	数字图书馆	13	元数据	10
Hadoop	25	海量数据存储	13	联机分析处理	10
数据存储	25	计算机	13	存储系统	10
关联规则	25	决策树	13	数据仓库系统	10
商业智能	21	分布式	12	内存映射文件	10
数据分析	21	多线程	12	信息化建设	10

高频关键词从各方面揭示出了对“大数据”及“海量数据”的研究热点，但仅从关键词表不能看出各个关键词之间的关系，为此需要对关键词进行共词分析并在此基础上进行进一步的研究，并以可视化的方式进行呈现。为此，首先利用 Excel 的“数据透视表”功能得到共词矩阵，如表 2 所示。

表 2　关键词共词矩阵(局部)

	GIS	Hadoop	MapReduce	存储系统	大数据	大数据量	多线程	分布式	分布式计算
GIS	32	0	0	0	0	0	1	0	0
Hadoop	0	43	10	0	1	0	0	0	3
MapReduce	0	10	37	0	2	0	0	0	1
存储系统	0	0	0	23	0	0	0	1	0
大数据	0	1	2	0	52	0	0	1	0
大数据量	0	0	0	0	0	5	0	0	0
多线程	1	0	0	0	0	0	8	0	0
分布式	0	0	0	1	1	0	0	18	0
分布式计算	0	3	1	0	0	0	0	0	17
服务器	0	0	0	0	1	0	0	2	0
关联规则	0	0	1	0	0	0	0	0	0
关系数据库	0	1	1	0	2	0	0	0	0

矩阵中对角线代表相应与其他高频关键词有共现关系的高频关键词出现的总次数，其他各行、列的交叉点代表相应的两个关键词共现的次数，由于关键词的共现没有方向关系，因此该矩阵是一个对称矩阵，即元素以对角线为对称轴对应相等。将该矩阵导入 Ucinet 软件，生成相应的以##h 为后缀的文件后，可以用 NetDraw 可视化软件绘制网络图，为了排除极值干扰，图中去除了“海量数据”节点，节点间以 2 为共现次数阀值，结果如图 2 所示，图中线条粗细代表了关键词的共现次数，节点大小代表该节点的中介中间性(Betweenness)，点越大代表该节点在网络中更为重要。

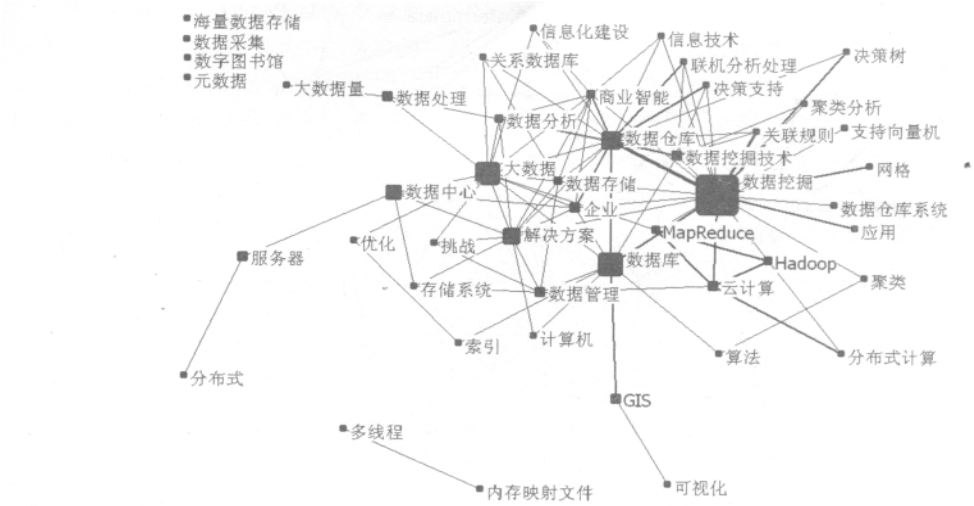

图 2　“大数据”相关期刊文献的关键词网络图

从图中可以看到,“数据挖掘”占有最为重要的地位,这说明了人们对于挖掘数据中的有用信息一直在进行探索与研究。“大数据”这个词虽然仅在近两年开始兴起,但是已经形成一个重要节点,“数据分析”、“数据中心”、“数据仓库”以及“商业智能”等关键词与之有着密切联系,另外,“大数据”与“企业”相关,似乎意味着目前企业界对于“大数据”有着更高的接受程度;另外,“云计算”、“MapReduce”以及“Hadoop”三者之间有着密切的联系,与“大数据”之间的联系并不紧密,但事实上,MapReduce 以及 Hadoop 带来的分布式、并行式计算针对的目标就是超大规模的“大数据”,甚至在业界一提到“大数据技术”便会想到以上两者,这也从侧面说明国内对于“大数据”的接受程度并不是很高,人们更愿意使用“海量数据”这个耳熟能详的词语。

2.3 “大数据”密切相关文献的关键词分析

与期刊相比,报纸对于新兴热点更为敏感,这从文献数量便可看出。为了对“大数据”有进一步的了解,将关键词包括“大数据”的期刊和报纸文章集合在一起进行研究,经进一步阅读,又去除了几篇相关度较低的文章,最终对 28 篇期刊文章以及 94 篇报纸文章共计 122 篇文章进行关键词分析。由于期刊与报纸文章提供的关键词区别较大,对所有文章进行重新标引,为了最大程度减小主观判断造成的影响,主要提取文献中提到的主要机构/产品/技术以及涉及的机构/产品作为关键词,以便在今后的研究中能够选择重点研究对象。结果如表 3 所示。

用对期刊文献类似的研究方法得到关键词共现矩阵并绘制图 3 所示的网络图。可以看出,IBM 在网络中处于中心地位,在文献中它大多以“主要机构”的身份入选关键词表,这表明 IBM 是目前国内“大数据”的主要推动机构,这与其一直宣扬的“智慧城市”、“智慧地球”一脉相承;另一个中心节点“IDC”则不同,它多半以“涉及机构”的身份入选,在大多数情况下,IDC 都是因为在“数字宇宙”报告中的论断,作为介绍“大数据”的时代背景而出现的。

表 3 “大数据”密切相关文献高频关键词表

IBM	44	Gartner	11	Facebook	8
分析	31	SAP	11	Google	7
IDC	30	**处理**	11	Netezza	7
EMC	24	**数据中心**	11	Sybase	7
存储	24	Microsoft	11	**云计算**	6
Hadoop	19	BI	10	MapReduce	5
Oracle	18	Informatica	10	**丹麦维斯塔斯**	5
Teradata	15	**麦肯锡**	10	**数据仓库**	5
HP	14	**管理**	9	**沃尔玛**	5

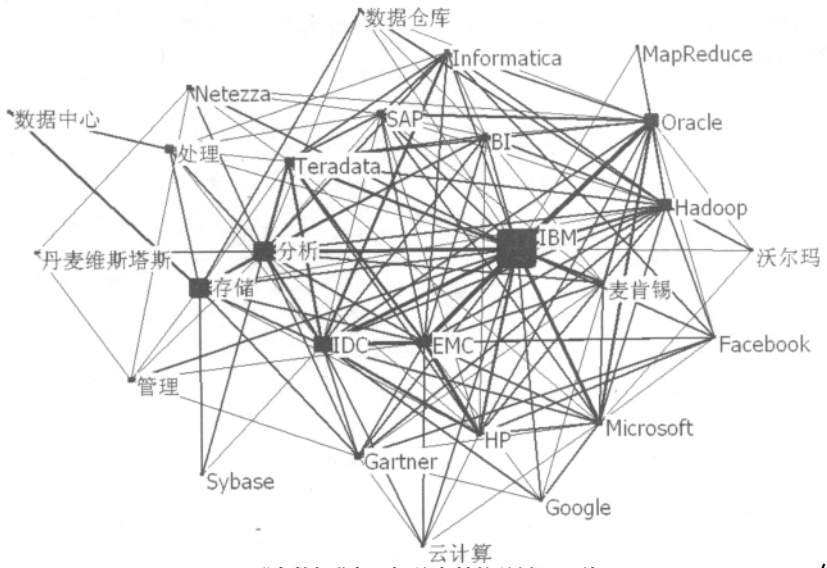

图 3 “大数据”密切相关文献的关键词网络图

(下转第 37 页)

4　结语

运营模式本身比较错综复杂，没有特定的单一模型能够为每个出版商的开放获取提供收益支撑的解决方案。本文通过对 OAJ 运营收入来源的分析，通过对 PLoS ONE 出版流程、论文处理费的调研，以及通过对 SCOAP3 将传统期刊转化为 OAJ 的独特运营方式的研究，提出开放获取出版在挖掘多种资金来源渠道的同时，积极探索新的增值服务，并寻求开放出版的新思路、新方法。目前，PLoS ONE 对出版流程的革新和 SCOAP3 从传统出版到开放出版的跨步在业内已经形成了两种典范。当然，随着实践的发展和 OAJ 发展的不断深入，后续还需要探索新的运营方式，需要在广泛调查研究的基础上，收集科研人员、科研团体、科研机构、资助机构、图书馆员、出版商等社会各界对 OAJ 运营的看法、建议，不仅仅是对已有案例的分析总结，而是收集一线的第一手数据资料，首先从理论上构建新的运营方式，再从实践上检验论证新运营方式的合理与否。

参考文献

1　PLoS. Progress Update(2010)[R/OL]. [2012-05-30]. http://www.plos.org/wp-content/uploads/2011/05/Progress-Update-final_with_links-070210-small.pdf

2　PLoS. PLoS Progress Report(2009)[R/OL]. [2012-05-30]. http://www.plos.org/wp-content/uploads/2011/05/PLoS_progress_report.pdf

3,5,11　Kaufman-Will Group. The facts about Open Access: A Study of the Financial and Non-financial Effects of Alternative Business Models for Scholarly Journals [R/OL]. [2012-05-30]. http://sippi.aaas.org/Open_Access/FAOAcompleteREV.pdf

4　Raym Crow. Income models for open access: an overview of current practice [R/OL]. [2012-05-30]. http://www.arl.org/sparc/bm~doc/incomemodels_v1.pdf

6　John Regazzi. The Shifting Sands of Open Access Publishing: a Publisher's View [J]. Serials Review. 2004,130(4):275-280

7　PLoS. 2010 Progress Update. [R/OL]. [2012-05-30]. http://www.plos.org/media/downloads/2011/2010_PLoS_Progress_Update_hi.pdf

8　李武. 基于开放存取的学术期刊出版模式研究(上)[J]. 数字图书馆论坛,2005(11):35-40

9　John Willinsky. Scholarly Associations and the Economic Viability of Open Access Publishing [J/OL]. Journal of Digital Information. 2004,4(2) http://journals.tdl.org/jodi/article/view/104/103

10　SPARC. Income Models for Supporting Open Access [EB/OL]. [2012-05-31]. http://www.arl.org/sparc/publisher/incomemodels/guide2-2.shtml

12　John Willinsky. The stratified economics of open access [J]. Economic Analysis & Policy. 2009,39(1):53-70

13　Adam Chesler. Open Access: A Review of an Emerging phenomenon [J]. Serials Review,2004(30):292-297

14　Caralee Adams. SPARC Innovator: PLoS ONE [EB/OL]. [2012-06-04]. http://www.arl.org/sparc/innovator/plos-one.shtml

15　The SCOAP3 Working Party. Towards Open Access Publishing in High Energy Physics [R/OL]. [2012-06-04]. http://scoap3.org/files/Scoap3WPReport.pdf

16　SCOAP3. About SCOAP3 [EB/OL]. [2012-06-05]. http://scoap3.org/about.html

17　SCOAP3. SCOAP3 tendering process has started [EB/OL]. [2012-06-05]. http://scoap3.org/news/news88.html

何莉娜　女 1987-，中国科学院国家科学图书馆，硕士研究生。

郑建程　1957-，中国科学院国家科学图书馆资源建设部副主任、研究馆员、硕士生导师。

(收稿日期：2012-06-09　编发：许桂菊)

(上接第 32 页)

3　结论

通过文献计量的方法，可以定量地对国内各界对“大数据”这一现象的接受、关注程度作出判断。可以看出，国内对于“数据”的研究一直较为重视，目前对于“大数据”这个新的名词接受程度上不算高，但在 IBM 等公司的有意炒作以及 MapReduce 等新的数据分析技术发展推动下，“大数据”正成为继“云计算”、“物联网”等之后新的关注热点。而要对“大数据”有更深的了解，可以针对 IDC 的“数字宇宙”、IBM 的“智慧地球”等进行进一步分析。

参考文献

1　郭春侠，储节旺. EXCEL 实现共词分析的方法——以国内图书情报领域知识管理研究为例 [J]. 情报杂志,2011(3):45-49

2　James Manyika, Michael Chui, Brad Brown 等. Big data: The next frontier for innovation, competition, and productivity [R]. http://www.mckinsey.com/mgi/publications/big_data/

杨　绎　上海图书馆上海科技情报研究所，读者服务中心参考馆员。

(收稿日期：2012-05-08　编发：王宗义)