国际生物医学大数据研究可视化分析

李凌英 刘建炜 李 俊

（中南大学信息安全与大数据研究院 长沙 410016）

〔摘要〕 以 WOS 数据库为数据源，全面采集生物医学大数据相关文献，以 CiteSpace 和 VOSViewer 软件绘制科学知识图谱，对当前生物医学大数据研究力量、研究热点及演进趋势进行分析。

〔关键词〕 大数据；生物医学；可视化分析

〔中图分类号〕R－056 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2017.07.002

Visualization Analysis of International Biomedical Big Data Research *LI Ling－ying, LIU Jian－wei, LI Jun, Information Security and Big Data Institute, Central South University, Changsha 410006, China*

〔**Abstract**〕 The literatures of biomedical big data are collected with the database of WOS as the data sources. CiteSpace and VOSViewer are used to draw scientific knowledge maps to analyze the research strength, research hotspot and evolution trend of current biomedical big data research in the paper.

〔**Keywords**〕 Big data; Biomedical; Visualization analysis

1 引言

生物医学是综合医学、生命科学和生物学的理论和方法而发展起来的前沿交叉科学、基本任务是运用生物学及工程技术手段研究和解决生命科学，特别是医学中的有关问题。近年来，随着生物科学技术、医疗技术以及计算机科学技术的快速发展，生物医学领域产生了大量的数据，如随着新一代测序技术的发展，TB、PB 数量级的基因组学、蛋白质组学、代谢组学等各类组学数据产出日渐增加，随着医院电子病历的普及，电子病历记录和医学影像等临床数据迅速积累，生物医学已进入大数据时代[1]。在大数据时代，庞大繁杂的数据以及对数据的研究对社会、科技、经济的发展将发挥支撑促进作用。大数据本身是一种潜在的战略性资源，具有小规模数据无法匹及的趋势预测能力，通过对生物医学大数据的分析和应用能够辅助医生进行临床决策、疾病诊断与个性化药物治疗等[2]。但是要想将这些资源的效益真正释放出来还面临着各类挑战，因为生物医学数据具有异质性、非结构化的特点，数据标准化与存储传输以及安全性问题都需要解决[3]。本文将研究视角扩展到国际上，以 WOS 数据库为数据源，以 CiteSpace 和 VOSViewer 可视化软件为工具，对当前生物医学大数据研究力量、研究热点及演进趋势进行可视化分析，以期促进国内生物医学大数据的研究发展。

2 数据来源与研究方法

2.1 数据来源

选取 Web of Science™ 核心合集作为检索数据

〔修回日期〕 2017－04－13

〔作者简介〕 李凌英，本科生；通讯作者：李俊，讲师。

库，检索期限设定为 1985 – 2016 年，出版类型设定为 Article，检索词为主题：（big data）AND 主题：（medical or med * or biology or bio *），检索时间为 2016 年 12 月 19 日。共计检索到 6 909 条符合条件的数据记录，下载的方式设定为全纪录（包含引用的参考文献）。

2.2 研究工具和方法

引文分析可视化是信息可视化的重要分支，在处理完海量引文数据后，利用信息可视化技术使人们更直观地观察浏览和理解信息，进而找到数据中隐藏的规律和模式。CiteSpace 软件和 VOSViewer 对文献信息的可视化，能够直观地反映学科领域的发展轨迹、知识基础、研究前沿与热点等。

3 研究力量分析

3.1 国家

将数据导入到 CiteSpace 软件中，Node Types 设置为 Country，阈值设置为 T50，选择 Pathfinder 算法，其余选用默认值，得出生物医学大数据研究领域的研究力量——国家分布，见图 1，其中圆形节点代表国家，国家发文量越多，节点越大。由图 1 可知，在生物医学大数据研究领域，各国研究力量不尽相同，主要集中在美国、中国、英国、德国、意大利等国家。美国的发文量最多，为 2 178 篇，中心性最高为 0.31，在生物医学大数据领域遥遥领先。中国的发文量位居第 2，为 868 篇，但中心性不高，为 0.01，说明中国需要加强对生物医学大数据研究的重视，多发表高质量文章。值得注意的是，法国发文量为 262 篇，位居第 10，但是中心性为 0.29，位居第 2，说明法国虽然发文量不多，但在生物医学大数据研究领域占据重要地位。此外，墨西哥、土耳其和以色列的 Burst 值分别为 5.55、5.15 和 5.06，说明近年来墨西哥、土耳其和以色列对生物医学大数据的研究予以重视。

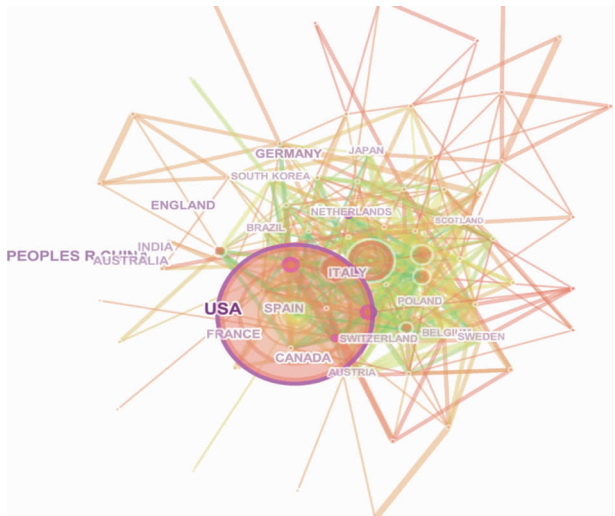

图 1 生物医学大数据研究领域的研究力量——国家分布

3.2 机构

将数据导入到 CiteSpace 软件中，Node Types 设置为 Instituition，阈值设置为 T50，选择 Pathfinder 算法，其余选用默认值，得出生物医学大数据研究领域的研究力量——机构分布，见图 2。其中圆形节点代表机构，机构发文量越多，节点越大。由图 2 可知，在生物医学大数据研究领域，各国机构研究力量不尽相同，主要集中在中国科学院、哈佛大学、密西根大学、华盛顿大学、加州大学洛杉矶分校等教育机构。中国科学院的发文量最多，为 126 篇，突显值最大为 17.45，说明中国科学院近几年对生物医学大数据研究领域非常重视。哈佛大学发文量位居第 2，为 78 篇，突显值为 13.51，位居第 4。值得注意的是，斯坦福大学的发文量为 50 篇，并不是很高，但是其突显值达到 13.89，仅次于中国科学院，说明斯坦福大学近年来对生物医学大数据研究领域高度重视。另外，图 2 也反映了在生物医学大数据研究领域各个机构的合作情况，大致可以分为 5 个聚类：以牛津大学为中心的 A 集群，以密歇根大学、华盛顿大学、剑桥大学为中心的 B 集群，以哈佛大学、加州大学、斯坦福大学为中心的

C 集群，以中科院为中心的 D 集群，以及以浙江大学为中心的 E 集群。这种机构间的合作集群情况存在地域因素的影响，如 A 集群和 C 集群，但是也存在跨地域之间的合作，如 B 集群中的剑桥大学与密歇根大学合作较为紧密，而斯坦福大学作为 C 集群和 D 集群之间的桥接与 D 集群的中国科学院合作紧密。

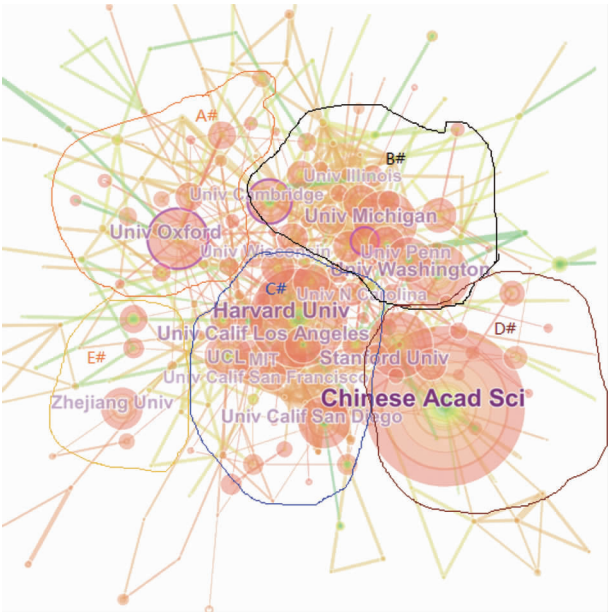

图 2 生物医学大数据研究领域的研究力量——机构分布

3.3 作者

将数据导入到 CiteSpace 软件中，Node Types 设置为 Author 和 Citedauthor，阈值设置为 T50，选择 Pathfinder 算法，其余选用默认值，得出生物医学大数据研究领域的研究力量——作者分布，见图 3，其中圆形节点代表作者，作者发文量和被引频次越大，节点越大。图 3 所示为 1985 – 2016 年综合发文量、被引频次后，在生物医学大数据研究领域的核心作者及其之间的合作情况。Dean J 的发文量和被引频次最大，频次为 111，中心性为 34.17；Dean J[4] 在 2008 年发文介绍了 MapReduce，提供了面向大型集群的简化数据处理，这篇文章在 1985 – 2016 年生物医学大数据文献中的被引频次为 66 次，位居第 1，说明 Dean J 在国际生物医学大数据研究领域的广泛国际影响力。Boyd D 的发文量和被引频次位居第 2，频次为 69，中心性为 23.69；Boyd D[5] 在 2012 年发表的文章中探讨了大数据给文化、技术、学术领域带来的挑战，这篇文章在 1985 – 2016 年的生物医学大数据文献中的被引频次为 44 次，位居第 2，说明 Boyd D 在国际生物医学大数据研究领域的广泛国际影响力。Mccrae RR 的发文被引频次位居第 3，频次为 66，中心性为 6.31。Schadt EE 和 Manyika J 的中心性较高，分别为 15.04 和 13.66；Schadt EE[6] 在 2010 年发表的文章中介绍了如何掌握不同类型的计算环境（如云和异构计算）来成功解决基因蛋白质高通量技术带来的大数据问题，这篇文章在 1985 – 2016 年生物医学大数据文献中的被引频次为 29 次，位居第 6；Manyika J[7] 在 2011 年的报告中介绍了对大数据的看法，指出要发展大数据的全部潜力，必须解决隐私、安全、知识产权等相关问题，这篇文章在 1985 – 2016 年生物医学大数据文献中的被引频次为 40 次，位居第 3。我国作者 Zhang Y，Wang Y 在生物医学大数据领域的发文量较高，但是被引频次很低，说明我国作者应加强生物医学大数据的研究，发表高质量的文章以增强我国在生物医学大数据领域的影响力。

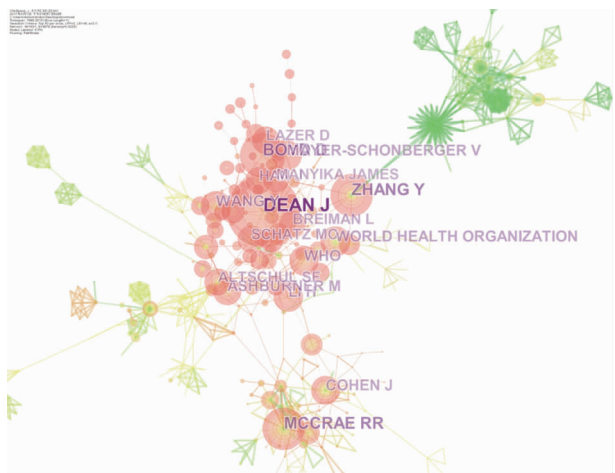

图 3 生物医学大数据研究领域的研究力量——作者分布

4 研究热点与前沿分析

4.1 VOSViewer 分析

将数据导入到 VOSViewer 软件中，共现阈值设置为 20，其余选用默认值，得出生物医学大数据研究领域的研究热点分布，见图 4。由图 4 可知，生物医学大数据研究领域的研究热点可分为 3 个子聚类：第 1 个子聚类为是蓝色部分，主要是研究方

法，包括年龄、性别等；第 2 个子聚类为绿色部分，主要是产生大数据的生物医学领域，包括基因（DNA、gene、genome、gene expression 等）、蛋白质（protein、protein - protein interaction 等）、癌症（cancer，tumor 等）以及生物标记物等；第 3 个子聚类为红色部分，主要是生物医学大数据的技术方法，包括生物医学大数据的应用（disease diagnosis、therapy、clinical study、public health、hadoop 等）、生物医学大数据的挑战（algorithm、scheme、format 等）以及基于文献的生物医学大数据研究（library、paper 等）等。

图 4 生物医学大数据研究领域的研究热点知识图谱（VOSViewer）

4.2 CiteSpace 分析

将数据导入到 CiteSpace 软件中，Node Types 设置为 Keyword + Term，选择 Burst，阈值设置为 T50，选择 Pathfinder 算法，其余选用默认值，以 Time-Zone 格式查看，得出生物医学大数据研究领域的研究前沿时区分布，见图 5，其中圆形节点代表热点词，热点词出现频次越多，节点越大。由图 5 可知可以清晰地看到生物医学大数据研究前沿的演变，研究范围不断扩大，内容不断加深，随着时间延续，生物医学大数据相关的研究领域逐年扩大，研究主题与方向更加丰富。之前的生物医学大数据研究领域主要是生物医学大数据的应用，包括生物信息学、疾病基因、风险因子的分析、癌症、医疗保健等，而后期主要围绕大数据的技术方法展开，包括云计算、社交媒体、网络、数据库、算法、演化分析/数据挖掘等。

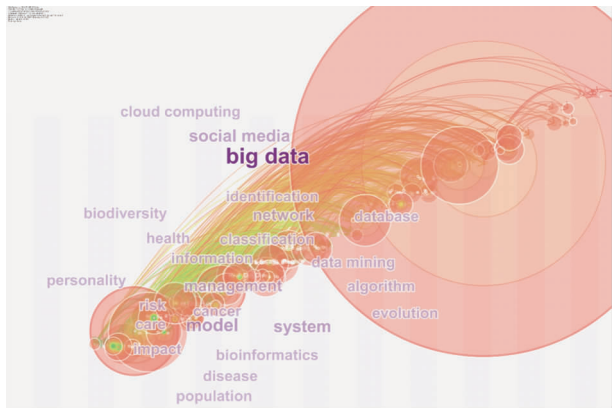

图 5 生物医学大数据研究领域的研究前沿知识图谱（CiteSpace）

5 讨论

5.1 挑战

通过 VOSViewer 分析和 CiteSpace 分析可知，生物医学大数据处理的关键功能由大数据的应用支持[8]，而生物医学大数据处理的根本在于数据挖掘，数据挖掘是一个决策支持过程[9]，通过数据挖掘能够很好地辅助临床，促进生物医学的发展，如疾病的早期检测，预防各种致命疾病与个性化疾病管理和监测以及医学影像信息的数据利用等[10-13]。在大数据应用过程中也面临着各种挑战。大数据应用程序应该是用户友好、透明的[14]，而卫生保健中的大多数数据是非结构化的（如大部分来自于自然语言处理）[15]，零散、分散、很少标准化，使得数据采集和清理复杂化[16]。此外，医学大数据的存储和传输成本高，一旦生成数据，与保护和存储数据相关的成本很高[17]。

5.2 机遇

虽然存在的挑战很多，但是挑战之下拥有很大潜力，通过改进与完善大数据技术与方法可以将挑战化为机遇。Hadoop 是能够对结构化或半结构化的海量数据进行分布式处理的软件框架[18]，为提高医学影像的检索效率，Yao Q A 等研发了基于 Hadoop 的医学影像检索系统[19]，陆婷娟等利用 Hadoop 技术、结合分布式文件系统和集中存储二者的优点和

医学影像的特点设计了一套二者相结合的医学影像“在线一归档”二级存储架构，解决了影像存储与传输系统的扩展性和可用性问题[20]。此外，大数据研究需要与云计算结合，充分利用云计算的分布式并行计算能力对海量、复杂的数据进行处理。李萍介绍了云计算、大数据时代医院信息化的 3 个转变：一是基础架构平台向云计算的转变；二是信息管理向数据集成平台的转变；三是终端多样化的转变[21]。大数据与云计算结合是生物医学大数据的建设方向，具有发展前景。

6 结语

随着大数据、云计算等新技术的迅速发展和广泛应用，基于大数据的数据分析、基于云计算的数据共享正在逐步成为现实，大数据研究广泛应用于生物医学。科学知识图谱可显示学科的发展进程与结构关系，通过 CiteSpace 分析和 VOSViewer 分析得到的科学知识图谱有效展示了国际生物医学大数据研究领域的研究力量、热点和前沿。就研究力量而言，我国在生物医学大数据研究领域发文量虽较高，但是中心性不强，需要加强与各国的合作，提高我国在生物医学大数据研究领域的权威性。中国科学院作为我国科研中坚力量，在国际生物医学大数据研究领域具有重要地位，要继续发展中国科学院的战略地位，形成以中国科学院为牵头单位的生物医学大数据研究机构合作线。我国作者也要加强在生物医学大数据领域的研究，努力提高文章的质量。研究热点和研究前沿主要围绕生物医学大数据研究的应用与挑战，在未来发展过程中需要加强大数据技术的应用，将挑战化为机遇。生物医学大数据的应用还处在萌芽阶段，也是我国实现快速发展的契机，国家需予以重视，生物医学大数据的应用研究必将成为未来大数据研究的热点。

参考文献

1 刘倩丽，关健. 中国电子健康档案的应用现况与展望[J]. 中国健康教育，2015，31（10）：969－970，979.

2 张春丽，成彧. 大数据分析技术及其在医药领域中的应用[J]. 标记免疫分析与临床，2016，23（3）：327－333.

3 Langer SG. Challenges for Data Storage in Medical Imaging Research [J]. J Digit Imaging，2011，24（2）：203－207.

4 Dean J，Ghemawat S. MapReduce：simplified data processing on large clusters [J]. Communicaations of the ACM，2008，51（1）：107－113.

5 Boyd D，Crawford K. Critical Questions for Big Data Provocations for a Cultural，Technological，and Scholarly Phenomenon [J]. Information Communication &Society，2012，15（5）：662－679.

6 Schadt E E，Linderman MD，Sorenson J，et al. Computational Solutions to Large－scale Data Management and Analysis [J]. Natrure Reviews Genetics，2010，11（9），647－657.

7 Manyika J，Chui M，Brown B，et al. Big Data：the next frontier for innovation，competition，and productivity [R]. McKinsey Global Institute，2011.

8 Raghupathi W，Raghupathi V. Big Data Analytics in Healthcare：promise and potential [J]. Health Inf Sci Syst，2014，（2）：3.

9 Deepa V K，Geetha J RR. Rapid Development of Applications in Data Mining [C]. IEEE International Conferenceon Green High Performance Computing，2013.

10 Fernandes L，O'Connor M，Weaver V. Big Data，Bigger Outcomes [J]. J AHIMA，2012，83（10）：38－43.

11 Hsieh JC，Li AH，Yang CC. Mobile，Cloud，and Big Data Computing：contributions，challenges，and new directions in telecardiology [J]. Int J Environ Res Public Health，2013，10（11）：6131－6153.

12 宁康，陈挺. 生物医学大数据的现状与展望 [J]. 科学通报，2015，60（Z1）：534－546.

13 罗志辉，吴民，赵逸青. 大数据在生物医学信息学中的应用 [J]. 医学信息学杂志，2015，36（5）：2－9.

14 Song TM，Song J，An JY，et al. Psychological and Social Factors Affecting Internet Searches on Suicide in Korea：a big data analysis of Google search trends [J]. Yonsei Med J 2014，55（1）：254－263.

15 Jee K，Kim GH. Potentiality of Big Data in the Medical Sector：focus on how to reshape the healthcare system [J]. HealthcInform Res，2013，19（2）：79－85.

16 Mancini M. Exploiting Big Data for Improving Healthcare Services [J]. J e－Learning Knowledge Soc，2014，10（2）：1－11.

（下转第 17 页）

且无危重记录的普外科、眼科、耳鼻喉科患者，可以不必过多依赖“大医院”的诊疗；就诊距离在 200～300 公里内及 300 公里以上的患者，急、重症患者居多，大部分就诊于肿瘤科、心内科、ICU 等重症科室，对高质量医疗服务的需求增大，当地的医疗水平满足不了患者的医疗需求，故选择距离更远的“大医院”就诊。

5.2 相关对策及建议

为正确引导患者的就医路径，合理分配医疗资源，国家大力推进分级诊疗政策，鼓励远程医疗服务发展。通过推进分级诊疗政策，引导优质医疗资源下沉，提升基层医生诊断水平，以常见病、多发病、慢性病分级诊疗为突破口，培养科学合理的就医秩序，逐渐形成基层首诊、双向转诊、急慢分治、上下联动的分级诊疗模式。发展远程医疗服务，打破患者的地域和时间壁垒，提高医疗服务可及性，缩短偏远地区患者的就诊时间，同时也节省路途花费。基于远程医疗技术建立医疗联合体，提高医疗网络中专家医师、普通医生及医疗设备的利用率[11]，为我国医疗资源的合理配置提供新思路。

参考文献

1 高阔，甘筱青. 患者选择就医单位分布及其影响因素分析［J］. 中国卫生事业管理，2014（7）：516－517，545.

2 张容瑜，尹爱田，Shi Lisheng，等. 就医行为及政策影响因素研究进展［J］. 中国公共卫生，2012，28（6）：861－862.

3 李淑玲，乔钰涵，张伟. 新农合农民就医行为与认知影响因素的实证研究［J］. 工业工程与管理，2014，19（4）：98－103.

4 钱东福，尹爱田，孟庆跃，等. 甘肃省农村居民选择住院医疗机构的影响因素研究［J］. 中国卫生经济，2008，27（1）：40－43.

5 刘晓莉，段占祺，陈文，等. 四川省农村居民就医行为现状调查及对策分析［J］. 卫生软科学，2016，30（1）：30－33.

6 黄建军，曾玉和. 病人选择就诊医院影响因素的 logistic 回归分析［J］. 中国医院统计，2006，13（2）：119－121.

7 余辉，张力新，刘文耀. 计算机辅助医学知识发现系统研究——糖尿病并发症流行病学数据挖掘［J］. 生物医学工程，2008，25（2）：295－299.

8 刘尚辉，王露，郑德禄. Apriori 关联规则在甲状腺结节病案分析中的应用［J］. 中国卫生统计，2011，28（2）：178－179.

9 Montella A. Identifying Crash Contributory Factors at Urban Roundabouts and Using Association Rules to Explore Their Relationships to Different Crash Types［J］. Accident Analysis & Prevention，2011，（4）：1451－1463.

10 左篙，张雄，刘礼德. 基于数据挖掘的门诊信息资源分析［J］. 现代生物医学进展，2013，13（23）：4568－4592，4594.

11 赵杰，崔震宇，蔡雁岭，等. 基于远程医疗的资源配置效率优化研究［J］. 中国卫生经济，2014，33（10）：5－7.

（上接第 11 页）

17 Mohr DC，Burns MN，Schueller SM，et al. Behavioral Intervention Technologies：evidence review and recommendations for future research in mental health［J］. Gen Hosp Psychiatry，2013，35（4）：332－338.

18 Taylor RC. An Overview of the Hadoop/MapReduce/HBase Framework and Its Current Applications in bioinformatics［J］. BMC Bioinformatics，2010，11（Suppl 12）：S1.

19 Yao QA，Zheng H，Xu ZY，et al. Massive Medical Images Retrieval System Based on Hadoop［J］. Journal of Multimedia，2014，9（2）：216－222.

20 陆婷娟，戚小平. 基于 Hadoop 的医学影像数据平台应用研究［J］. 世界复合医学，2015，（3）：223－226.

21 李萍. 云计算与大数据时代医院信息化的三个转变［J］. 中国医院管理，2013，33（12）：80－81.