医学知识库领域研究可视化分析

罗爱静 梁朝聪

（1 中南大学湘雅三医院 长沙 410013 2 中南大学信息安全与大数据研究院 长沙 410013 3 医学信息研究湖南省普通高等学校重点实验室（中南大学） 长沙 410013）

〔摘要〕 以 Web of Science 数据库为数据源，利用可视化软件 CiteSpace、SATI、Ucinet 对医学知识库领域研究文献进行可视化分析，揭示该领域研究力量、核心期刊、核心作者、文献共被引情况、热点等内容。

〔关键词〕 知识库；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.1673－6036.2017.04.013

Visualized Analysis of Medical Knowledge Base Study *LUO Ai－jing, LIANG Chao－cong, 1The Third Xiangya Hospital of Central South University, Changsha 410013, 2Information Security and Big Data Research Institute of Central South University, Changsha 410013, 3Key Laboratory of Medical Information Research (Central South University), College of Hunan Province, Changsha 410013, China*

〔**Abstract**〕 Taking Web of Science database as the data source, the paper visually analyzes the literatures about medical knowledge data using the visualization software, such as Citespace, SATI and Ucinet, reveals the research power, core journals, core authors, situation of the literature referred and hot topics in this field.

〔**Keywords**〕 Knowledge base; Medicine; Visualized analysis

1 引言

知识库是知识工程中结构化、易操作、易利用、全面有组织的知识集群，是针对某一（或某些）领域问题求解的需要，采用某种（或若干）知识表示方式在计算机存储器中存储、组织、管理和使用的互相联系的知识集合[1－2]。20 世纪 90 年代以来，知识库在各行业和学科领域取得了快速发展，尤其在医药卫生行业，知识库已成为医学信息学重要的研究方向。医学信息种类繁多，数量巨大，包括疾病、药品、辅助检查、手术信息等显性知识，也包括存在于医务人员脑海中的诊疗疾病的能力、经验和技能等隐形知识，且这些知识相互关联，密不可分[3－4]。医学知识库的应用为医学知识组织和存储提供了一个良好的思路。

随着医学知识库的快速发展及文献资料数量的不断增加，目前对医学知识库研究领域还没有较详细的相关综述出现，而利用可视化软件，采用一般计量、共现、聚类、中心性、多维尺度、社会网络分析等数据分析方法，挖掘和呈现出节点链接图、引文网络图谱等可视化数据结果，能够直观地反映学科领域的发展轨迹、知识基础、研究热点等内容[5－6]。因此，本研究利用可视化工具，结合科学计量方法，对已有研究进行整理，绘制该领域知识图谱，分析医学知识库的研究现状，以期为医学知识库领域的研究提供参考。

〔收稿日期〕 2016－12－15

〔作者简介〕 罗爱静，博士，教授，博士生导师，发表论文 60 余篇；通讯作者：梁朝聪，硕士研究生。

2 资料与方法

以 ISI 出版的 Web of Science 数据库中的文献为数据源，构建检索式：主题 = "knowledge base" OR "knowledge database" OR "knowledge repository" + 主题 = "medical" OR "medicine" OR "medical science" OR "clinical" OR "health care"，索引 = SCI - EXPANDED、SSCI、A&HCI、CPCI - S、ESCI、CCR - EXPANDED、IC，文献类型限定为 Article 和 Proceedings Paper，检索时限为 1999 年至今，检索日期为 2016 年 6 月 30 日。排除不相关文献后，得到 1 658篇相关文献。采用 CiteSpace、SATI、Ucinet 等可视化软件和工具对医学知识库领域的文献时间分布、核心力量、核心期刊、核心作者、文献被引、研究热点等方面进行分析。

3 载文分布

对每年的文献刊载量分析可在一定程度上揭示医学知识库研究的热度。从 1999 年至今医学知识库领域的每年发文量统计，见图 1，1999 年起该领域每年的文献数量总体上呈递增趋势，其中 1999 - 2005 年呈稳步增长状态，2005 - 2007 年呈快速增长状态，2007 - 2011 年发文量虽有所起伏，但总体上呈上升趋势，且高于上一阶段，研究热度较稳定。2011 - 2013 年发文量又一次激增，由 2011 年的 105 篇增长到 2013 年的 169 篇。到目前为止，医学知识库领域研究仍保持着较高的研究热度。

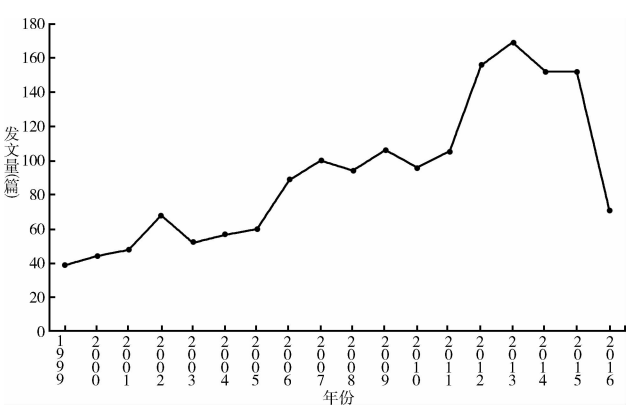

图 1 医学知识库领域研究文献发文量时间分布

4 知识图谱可视化分析

4.1 研究力量

对地区和机构进行可视化分析，可揭示医学知识库领域研究的核心力量分布，发现国家和研究机构之间的社会关系，为评估国家或机构的学术影响力提供一个新的视角。将数据导入到 CiteSpace 中，Node Types 选为 Country 和 Institution，阈值设置为 T30，其他设置为默认值，得出医学知识库领域研究的核心力量分布图谱，见图 2。中心度测量的是某节点对经过该节点并彼此相连接的另外两个节点的控制能力，一定程度上表征了某节点与其他节点所具有的广泛和密切的联系以及在整个网络中的重要地位和作用[7]。因此可从发文量和中心度两个角度对医学知识库领域的核心力量进行分析。

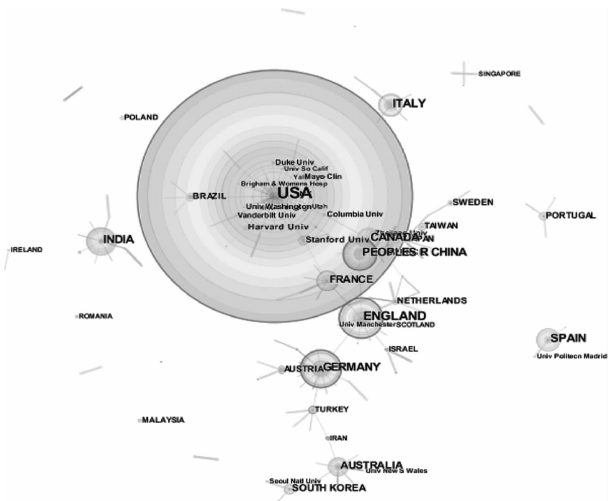

图 2 医学知识库领域研究国家和机构分布图谱

由图 2 可知，在对医学知识库的研究中，不同国家（地区）的研究实力不尽相同：美国的发文量最多，共 619 篇，占总发文量的 33.7%；英国其次，共发文 108 篇，占总发文量的 5.9%；随后是德国、中国等。美国的发文量远高于其他国家，表明美国在该领域研究中处于遥遥领先的地位。从中心度来看，最高的是英国，中心度为 0.40；其次为美国、德国和中国，中心度依次为 0.38、0.28 和 0.22，表明英国、美国、德国和中国在该领域研究中具有较高的学术影响力。从机构分布来看，发文量排名前 10 的机构分别是斯坦福大学（29 篇）、哈

佛大学（21 篇）、浙江大学（19 篇）、哥伦比亚大学（18 篇）、梅奥医学中心（18 篇）、范德比尔特大学（17 篇）、杜克大学（16 篇）、华盛顿大学（16 篇）、犹他州大学（14 篇）和南加州大学（13 篇）。主要分布在高等院校和医疗机构中，发文量排名前 10 的机构中除浙江大学外，其余都是美国的科研机构。表 1 列出了中心度排名前 10 的机构，可见美国的研究机构无论是在发文量还是中心度上都占据优势，其他研究机构研究力量相对较薄弱。

表 1 医学知识库领域研究中心度排名前 10 的机构

序号	机构名称	中心度	国家
1	斯坦福大学	0.15	美国
2	北卡罗来纳州立大学	0.13	美国
3	芝加哥大学	0.04	美国
4	哥伦比亚大学	0.03	美国
5	庆熙大学	0.03	韩国
6	塞萨洛尼基亚里士多德大学	0.02	希腊
7	京都大学	0.02	日本
8	哈佛大学	0.01	美国
9	范德比尔特大学	0.01	美国
10	犹他州大学	0.01	美国

4.2 核心期刊

分析某领域的文献来源期刊可揭示该领域的核心期刊分布，对核心期刊文献共引频次的分析则能够反映出这一期刊所刊登文献的利用率及其含金量[8]。将 Node Types 类型选为 Cited Journal，阈值设置为 T30，其他设置为默认值，得出期刊共被引知识图谱，见图 3。

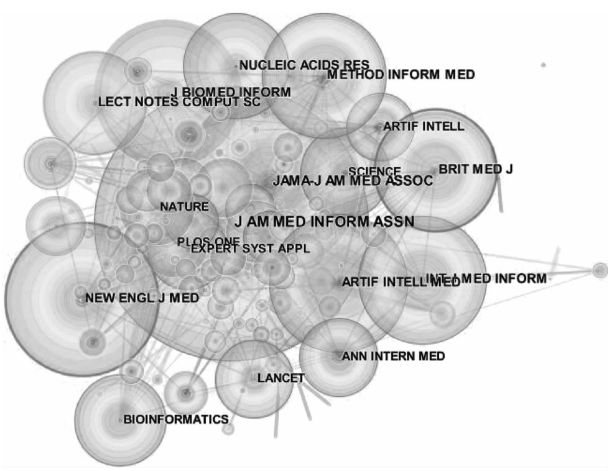

图 3 医学知识库领域研究期刊共被引知识图谱

由图 3 可知，医学知识库研究共被引期刊主要为《美国医学信息学会杂志》（*J AM MED INFORM ASSN*）、《美国医学协会杂志》（*JAMA – J AM MED ASSOC*）、《新英格兰医学杂志》（*NEW ENGL J MED*）、《生物医学信息杂志》（*J BIOMED INFORM*）、《人工智能在医学领域的应用》（*ARTIF INTELL MED*）、《国际医学信息学杂志》（*INT J MED INFORM*）等。这些期刊涉及医学信息、医学、计算机科学等领域，刊登了大量有关医学知识库的研究文献，是该领域的重要文献来源，对该领域研究起到支持作用。从中心度来看，《新英格兰医学杂志》（*NEW ENGL J MED*）、《英国医学杂志》（*BRIT MED J*）、《柳叶刀》（*LANCET*）、《生物信息学杂志》（*BIOINFORMATICS*）等中心度较高，均超过 0.18，表明这些期刊刊载的医学知识库研究文献质量高，对该领域的研究影响较大，起重要的支撑作用。其中 *BRIT MED J*、*NEW ENGL J MED*、*BIOINFORMATICS* 既是高被引期刊，又是高中心度期刊，表明这些期刊在该领域研究中占有较强的核心地位。

4.3 核心作者

通过作者的发文量和被引频次分析可识别出医学知识库领域研究的核心作者群。表 2 列出了医学知识库领域研究发文量前 15 的作者，他们是医学知识库研究的高产作者，对该领域的研究贡献较大。

表 2 医学知识库领域研究发文量前 15 的作者

排名	作者	发文量（篇）
1	David W. Bates	9
2	Christopher G. Chute	8
3	U Hahn	8
4	Hongfang Liu	8
5	MA Musen	7
6	Wajahat Ali Khan	7
7	Muhammad Afzal	7
8	Adam Wright	7
9	ZH Wu	7
10	Maqbool Hussain	7
11	Sungyoung Lee	7
12	Shiri Gordon	6
13	Jacques Bouaud	6
14	KlausPeter Adlassnig	6
15	Blackford Middleton	6

对作者进行共被引分析，将 Node Types 类型选为 Cited Author，阈值设置为 T25，其他设置为默认值，得到作者共被引知识图谱，见图 4，图中节点代表被引作者，节点的大小代表被引频次。从图中可知，除匿名作者群体外，被引频次最高的作者是 BODENREIDER O，被引频次为 55 次，BODENREIDER 博士的研究主要集中在生物医学领域术语和本体论，对一体化医学语言系统进行了深入探讨，为医学知识库的研究奠定了坚实的基础。其次是 BATES DW，被引频次为 53，BATES DW 是生物医学信息领域的著名教授，致力于研究医学信息技术，尤其是医学决策支持研究。排名第 3 的是 CIMINO JJ，被引频次为 43，CIMINO JJ 是著名的生物信息学教授，对医学受控词汇进行了深入研究，推动了一体化医学语言系统项目的启动，创建了基于鉴别诊断知识库的诊断决策支持系统 Dxplain[9]。另外 RECTOR AL、HRIPCSAK G、MUSEN MA、SHAHAR Y、ZADEH LA、FRIEDMAN C、PELEG M 等人的被引频次也较高。

从中心度来看，中心度排名前 5 的作者分别是 BATES DW(0.1)，BODENREIDER O(0.09)，RECTOR AL(0.09)，CIMINO JJ(0.07)，FRIEDMAN C(0.05)。这 5 位作者中心度与被引频次均排名靠前，对医学知识库领域研究具有较高影响力；但结合发文量、被引频次、中心度可知，少有兼顾产量与高影响力的作者出现，研究作者群体较分散。

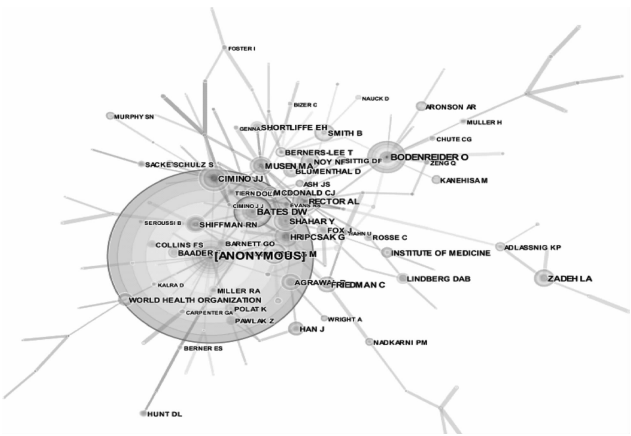

图 4 医学知识库领域研究作者共被引知识图谱

4.4 文献共被引

将 Node Types 选择为 Cited Reference，阈值选择为 T20，Pruning 设置为 Pathfinder，其他设置为默认值，得到医学知识库领域研究共被引文献知识图谱，见图 5。图 5 中共有 290 个节点、449 条连线，节点代表被引文献，节点大小表示被引频次，节点内圈中的颜色及厚薄度表示文献不同时间段被引频次，节点含有紫色光圈表明其具有中心性。

图 5 医学知识库领域研究文献共被引知识图谱

由图 5 可知，对医学知识库领域研究贡献度最大的是 BODENREIDER O 在 2004 年发表的"The Unified Medical Language System (UMLS): integrating biomedical terminology" 一文，在该领域中被引频次为 23，中心度为 0.3。该文提到的 UMLS 是由美国国立医学图书馆开发的生物医学词汇库，集成在 UMLS 的超级叙词表词汇包括 NCBI 分类目、基因本体论、医学主题词 (MeSH)、OMIM 数据库和数字解剖标志知识库等。该文对 UMLS 术语资源集成、属于一体化原则、外部链接引用、UMLS 数据访问等进行了说明[10]。贡献度其次的是 DE CLERCQ PA 在 2004 年发表的"Approaches for Creating Computer - interpretabte Guidetines that Facititate Decision Support" 一文，在该领域中被引频次为 23，中心度为 0.17。该文回顾了制定和实施有利于决策支持的计算机可理解的临床指南通用方法，涉及指南表示、采集、验证和执行等方面，介绍了 5 种方法用于建立基于计算机的临床指南[11]。对医学知识库领域研

究贡献度排名第 3 的是 SMITH B 在 2007 发表的 "The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration" 一文，在该领域中被引频次为 23，中心度为 0.17。该研究提到开放生物医学本体（OBO）是一系列的本体集合，"OBO Foundry"作为核心，其目标是将生物医学研究中产生的数据形成单一的、一致的、累计扩大并在算法上易于操作的整体[12]。对高被引文献进行综合分析，其研究主要涉及本体研究、统一医学语言、数据交换格式和准则规范、术语描述逻辑等方面，体现了医学知识库领域的基础性研究，为该领域的研究和发展奠定了坚实的基础。

4.5 研究热点

研究热点是某一阶段、某一领域的研究学者集中关注的研究主题。关键词是对文章主题的高度凝练与概括，对高频的关键词进行分析可以确定医学知识库领域的研究热点。利用 SATI 软件对数据进行分析，得出该领域研究的高频关键词，见表 3。另外，利用 Unicat 软件中的可视化工具 Netdraw 进行中心度分析并绘制医学知识库领域的关键词共现知识图谱，见图 6，图中节点代表关键词，节点大小代表关键词的中心度，节点之间连线表示关键词的共现关系。

表 3 医学知识库领域研究的高频关键词

序号	关键词	频次
1	Ontology	62
2	Knowledge base	51
3	Knowledge	39
4	Database	32
5	Data mining	31
6	Data	25
7	Medical informatics	24
8	Decision support	23
9	Expert system	23
10	Diagnosis	23
11	Knowledge representation	22
12	Decision support systems	21
13	Repository	20
14	Fuzzy logic	20
15	Semantic Web	19
16	Classification	19
17	Clinical trials	18
18	Knowledge management	18
19	Medical	18
20	Natural language processing	17

结合表 3 和图 6 进行分析，出现频次最高的关键词是本体（Ontology），并且中心度也最高，表明本体在医学知识库领域研究中占有重要地位，基于本体的医学知识库研究具有较高研究热度；Knowledge base，Repository 等高频词作为表示知识库的主题概念，因此出现频次也较高；从数据挖掘（Data mining）、知识表示（Knowledge representation）、语义网络（Semantic Web）、模糊逻辑（Fuzzy logic）等高频词以及规则（DICOM、Rule）、信息抽取（Information Extraction）、支持向量机（Support vector machine）等关键词聚集发现，医学知识库中的知识组织、知识表示、知识挖掘是研究热点；从诊断（Diagnosis）、决策支持（Decision support）、专家系统（Expert system）、决策支持系统（Decision Support Systems）等关键词聚集发现临床决策支持功能与应用是该领域的一个研究热点；从糖尿病（Diabetes）、癌症（Cancer）、电子病历（Electronic Health Records）、流行病学（Epidemiology）等关键词聚合发现专科疾病知识库和电子病历知识库研究等是该领域的应用研究热点；另外，影像存储与传输系统（PACS）、医学图像（Medical imaging）、质量提升（Quality improvement）、云计算（Cloud computing）等关键词聚合，但关键词节点聚合较少，共现关系不紧密，说明医学图像处理、存储及医学图像知识库构建虽然作为医学知识库研究的一个方向，但是研究热度不高。

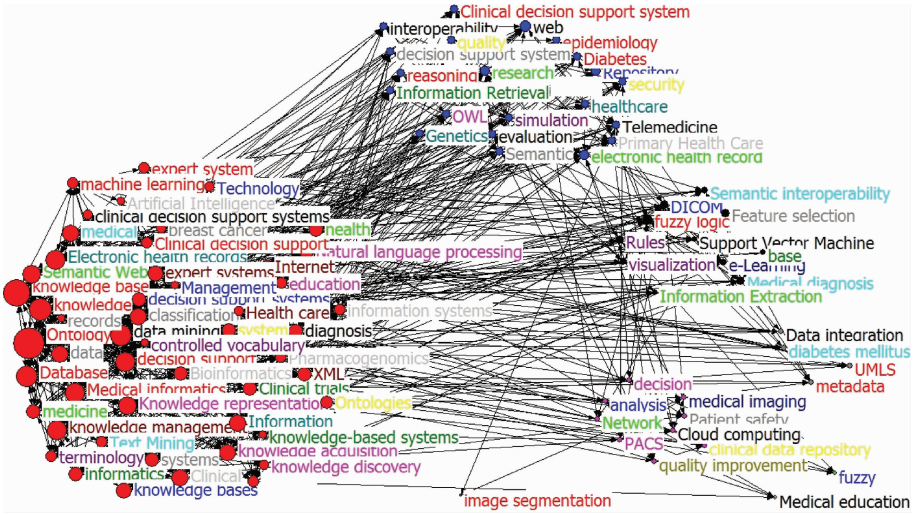

图 6 医学知识库领域关键词共现知识图谱

5 结语

医学知识库作为知识库研究的一个分支，其研究热度总体上呈上升趋势。该领域的研究机构主要分布在高等院校和医疗机构中，美国的发文量遥遥领先于其他国家，英国、美国、德国和中国在该领域有较高学术影响力；该领域文献主要发表在医学信息、医学、计算机科学等领域的期刊；少有兼顾高产和高影响力的作者，核心作者较分散；对医学知识库领域研究产生重要影响的文献主要涉及本体研究、统一医学语言、数据交换格式和准则规范、术语描述逻辑等方面；另外医学知识库领域的研究热点主要分布在本体医学知识库研究，医学知识库中的知识组织、知识表示、知识挖掘，医学知识库临床决策支持功能与应用，专科疾病知识库，电子病历知识库构建应用研究等方面。

参考文献

1 鄢珞青. 知识库的知识表达方式探讨 [J]. 情报杂志，2003，(4)：63 - 64.

2 吴顺祥，吉国力. 数据库系统与知识库系统的对比分析 [J]. 计算机工程与应用，1999，(9)：83 - 85.

3 蒋立辉，王伟. 医学知识库与医学知识的获取 [J]. 医学信息，2006，19 (9)：1500 - 1502.

4 张文举，李娜. 基于知识服务的医学知识服务系统研究 [J]. 中华医学图书情报杂志，2007，16 (5)：1 - 5.

5 Chen C M. CiteSpace II：detecting and visualizing emerging trends and transient patterns in scientific literature [J]. Journal of the American Society for Information Science and Technology，2006，57 (3)：359 - 377.

6 邓君，马晓君，毕强. 社会网络分析工具 Ucinet 和 Gephi 的比较研究 [J]. 情报理论与实践，2014，37 (8)：133 - 138.

7 杨利军，魏晓峰. 基于知识图谱的国外社会网络分析领域可视化研究 [J]. 情报科学，2011，29 (7)：1041 - 1048.

8 胡德华，种乐熹，邱均平. 信息行为研究的可视化分析 [J]. 图书馆杂志，2015，(9)：49 - 54.

9 Barnett G O，Cimino J J，Hupp J A，et al. DXplain. An Evolving Diagnostic Decision - support System [J]. JAMA，1987，258 (1)：67 - 74.

10 Bodenreider O. The Unified Medical Language System (UMLS)：integrating biomedical terminology [J]. Nucleic Acids Res，2004，32 (Database issue)：D267 - D270.

11 de Clercq P A，Blom J A，Korsten H，et al. Approaches for Creating Computer - interpretabte Guidetines that Facititate Decision Support [J]. Artificial Intelligence in Medicine，2004，31 (1)：1 - 27.

12 Smith B，Ashburner M，Rosse C，et al. The OBO Foundry：coordinated evolution of ontologies to support biomedical data integration [J]. Nature Biotechnology，2007，25 (11)：1251 - 1255.