“科学知识图谱”与“Google知识图谱”比较分析*
——基于知识管理理论视角

冯新翎[1,3] 何 胜[1,3] 熊太纯[2] 武群辉[2] 柳益君[1,3]

(1. 江苏理工学院计算机工程学院 常州 213001;2. 江苏理工学院图书馆 常州 213001;
3. 常州市云计算与智能信息处理重点实验室 常州 213001)

摘 要 [**目的/意义**]随着大数据技术发展,“Google知识图谱”(knowledge graph)引起广泛关注,由于在中文译名上与图书情报领域中的“科学知识图谱”(Mapping knowledge domain)相似,容易引起误解。[**方法/过程**]基于知识管理理论,从知识获取、知识组织、知识存储、知识共享和知识创新的角度比较分析,并讨论大数据情景下两类知识图谱在相关领域的应用前景。[**结果/结论**]分析结果表明,两者之间既有本质的区别又有紧密的联系,在大数据时代,两者在知识创新方面的融合和发展将会带来知识管理领域科学范式的变革。

关键词 科学知识图谱 谷歌知识图谱 语义网 大数据

中图分类号 G350 **文献标识码** A **文章编号** 1002-1965(2017)01-0149-05

引用格式 冯新翎,何 胜,熊太纯,等.“科学知识图谱”与“Google知识图谱”比较分析[J].情报杂志,2017,36(1):149-153.

DOI 10.3969/j.issn.1002-1965.2017.01.026

Comparison and Analysis of Mapping Knowledge Domain and Google Knowledge Graph*
——Based on the Theory of Knowledge Management

Feng Xinling[1,3] He Sheng[1,3] Xiong Taichun[2] Wu Qunhui[2] Liu Yijun[1,3]

(1. School of Computer Engineering, Jiangsu University of Technology, Changzhou 213001;
2. Library, Jiangsu University of Technology, Changzhou 213001;
3. Cloud Computing Intelligent Information Processing, Changzhou 213001)

Abstract [**Purpose/Significance**] With the development of big data technology, "Google knowledge graph" has caused widespread concern. Since the Chinese translation for "knowledge graph" is similar to the "Mapping knowledge domain" in the Library and Information field, so it's easy to cause misunderstanding. [**Method/Process**] Based on the knowledge management theory, the study compares and analyzes "Google knowledge graph" and "Mapping knowledge domain" from the perspectives of the knowledge acquisition, knowledge organization, knowledge storage, knowledge sharing and knowledge innovation, and discusses their application prospect in related fields under the big data background. [**Result/Conclusion**] Analysis results show that "Google knowledge graph" and "Mapping knowledge domain" have essential differences in nature but they also have a tight connection. In the era of big data, the integration and development of them in the knowledge innovation aspect will revolutionize the field of knowledge management science paradigm.

Key words mapping knowledge domain Google knowledge graph semantic web big data

收稿日期:2016-08-07 修回日期:2016-09-20

基金项目:国家社会科学基金一般项目“基于大规模网络分析方法和内存计算技术的高校图书馆大数据应用模式与实证研究”(编号:15BTQ016)研究成果之一。

作者简介:**冯新翎**(ORCID:0000-0001-5873-6019),女,1981年生,硕士,讲师,研究方向:大数据挖掘与应用;**何 胜**(ORCID:0000-0001-6762-8271),男,1971年生,博士,副教授,研究方向:大数据挖掘与应用;**熊太纯**(ORCID:0000-0002-8623-4793),男,1971年生,研究馆员,研究方向:图书馆学;**武群辉**(ORCID:0000-0002-3720-9006),男,1973年生,副研究馆员,研究方向:图书馆学;**柳益君**(0000-0002-5381-6084),女,1978年生,副教授,研究方向:大数据挖掘与应用。

0 引 言

自从 2002 年由 Rasmussen 等学者在 65 届美国信息科学与技术学会会议上发表题为“Visualizing knowledge domains”的文献[1],将可视化方法及工具应用于图书情报领域知识管理的研究在国外学界逐步兴起。国内方面,陈悦、刘则渊等提出将这一研究方法的中文译为“科学知识图谱绘制”[2]。随后,“科学知识图谱”或“知识图谱”概念在国内图情领域得到广泛应用,成为知识管理的重要方法[3,4]。

为提供海量数据的智能检索服务,Google 公司率先构建了统一结构化的 Knowledge Map[5],即以语义网[6](Semantic Web)和领域本体[7](Ontology)为其关键技术的大规模语义网络知识库。Knowledge Map 按照中英文字面的含义也译为“知识图谱”,近年来主要出现在国内计算机领域相关文献中[8-10]。另外,中国中文信息学会(http://www.cipsc.org.cn/)已经连续召开三届“中文知识图谱研讨会”,讨论语义网知识库的开发和应用,“知识图谱”概念在中文信息处理领域已经被普遍使用。随着语义网技术在图情领域中的广泛应用,以语义网知识库为特征的“Google 知识图谱”势必引入到图情领域,从而在名称上与作为知识管理重要方法的“科学知识图谱”相冲突,因此有必要厘清各自概念的内涵,并予以比较和分析。

两类知识图谱都属于知识管理范畴,在知识管理过程中不同阶段扮演不同角色,完成各自功能。以下首先分析各自术语一般涵义,追溯相关理论渊源,重点以知识管理过程为主线进行比较分析,并总结各自适用领域。另外,因为“Google 知识图谱”是大数据时代的产物,所以还讨论了两类知识图谱在大数据环境下的应用前景。

1 两类知识图谱的涵义

“科学知识图谱”(Mapping knowledge domain)将科研活动的主体(如研究人员,机构,团队)或具有某个共同特征的学科领域群体(如学科知识单元、知识群体)作为研究对象,是知识管理过程中的一种分析方法,广泛应用于科学计量、引文分析、知识创新预测等,包含“科学图”(Science mapping)、“文献计量图”(Bibliometric mapping)、“文献图”(Literature mapping)等内容[11]。陈超美认为知识图谱是揭示知识演化进程和机制的信息可视化方法[12]。陈悦等将知识图谱定义为能绘制、挖掘、分析和显示知识之间的相互联系,并能提供知识共享以及促进科研合作的可视化工具[2]。综合来看,“科学知识图谱”的共同特点是基于图形分析的可视化手段,显示知识演化进程和知识结构,辅助知识管理的方法和工具[13]。

“Google 知识图谱”(Google Knowledge Graph)是在知识管理过程中,为应对海量知识检索挑战,由 Google 公司提出并构建的基于语义网的大规模知识库。基于本体和语义网技术,“Google 知识图谱”通过描述现实世界中的各种实体(概念)及其复杂关系,将多种异构的知识库关联起来,并构建基于图(Graph)的统一的结构化语义网络知识库,在此基础上实现智能检索和知识推理。为方便表述,本文的“Google 知识图谱”是建立在语义网技术基础上所有类似 Google 知识库架构的知识库统称,如搜狗知立方(www.sogou.com)和百度知心(www.baidu.com)在架构上类似 Google 知识库,就属于“Google 知识图谱”类型。

2 两类知识图谱的区别

2.1 相关理论渊源 以科学主体和学科知识为研究对象的“科学知识图谱”,用图形方式直观呈现科学主体(或学科知识)网络结构、知识单元互动和知识群体演化等隐含的复杂关系,其产生有深刻的理论渊源。相关支撑理论有揭示网络结构和演化关系的“社会网络分析”理论,强调知识创新的“知识单元离散和重组”理论,尤其是科学史和科学哲学领域中,库恩提出的“科学发展模式”理论[14,16]。库恩认为,科学发展进程实质是通过新旧“范式”交替更迭的模式,不断推动科学创新和科学革命。“科学知识图谱”是“跟踪科技前沿、选择科研方向、开展知识管理并辅助科技决策”的重要方法和工具[15],以助益科技活动、强化知识管理等方式有力地促进了旧范式突破和新范式诞生,从而积极推动科学发展的进程。

作为大数据时代产物的“Google 知识图谱”,紧密依存大数据理论,以及关注数据规范性和关联性的本体和语义网理论。由于信息技术飞速发展引起了数据生成,传播与存储方式的巨大变革,为更全面、精准和高效获取知识以及发现创新知识,“Google 知识图谱”以本体建模为手段,通过领域概念术语的规范化,推动知识全面共享,借助于语义网络分析理论挖掘并发现新知识,应用语义网知识库关联方法实现海量知识的分布式存储。

2.2 知识管理视角 已有的相关研究文献对知识管理的过程划分并不完全一致,但一般包括知识获取(采集)、知识组织、知识存储(检索)、知识共享和知识创新等阶段[17-19]。两类知识图谱的共性在于二者都是服务于知识管理过程,区别在于二者分别参与不同的过程,完成不同的功能,如图 1 所示,“科学知识图谱”本质是知识管理的方法,一般与知识获取、知识组织、知识共享和知识创新密切相关,“Google 知识图

谱”本质是知识库,参与了知识获取、知识组织、知识存储和知识创新过程。

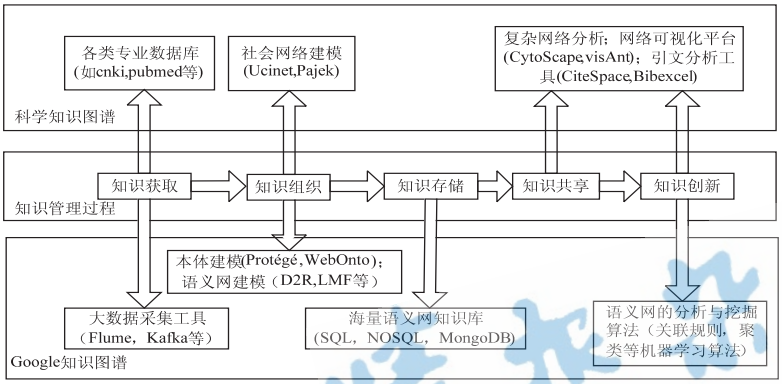

图 1　基于知识管理的两类知识图谱比较

2.2.1　知识获取　以知识收集和整理为主要功能的知识获取是知识管理的首要环节。“科学知识图谱”一般利用已构建的专业数据库,这些数据大多来自于科学引文索引（SCI)、社会科学引文索引（SSCI)、艺术与人文引文索引（A&HCI)、中文社会科学引文索引（CSSCI)等数据库或其收录的核心期刊文献,如美国医学文献数据库(pubmed)、中国知网数据库(cnki)等,数据类型有期刊论文、会议论文、专利、基金、出版物等,这些专业的数据资源具有客观、准确的特点;另外,也将社会媒体数据、网站日志、人物履历数据等不属于文献的数据作为其知识获取的来源。

“Google 知识图谱”是从包含各种结构化的数据库(如各类专业数据库)和非结构化的来自于互联网、物联网、云计算平台的海量数据(如图片、视频、日志等)获取知识。应用信息领域的专业工具,如 Flume(flume. apache. org),Kafka（kafka. apache. org)等,将结构化和非结构化数据导入和整合,并通过抽取、转换和装载工具(Extraction,Transformation,Loading,ETL)形成结构化的知识。

2.2.2　知识组织　知识组织是指对获取的知识进行表示、分类、编码使其有序化,以利于知识应用和管理,应用信息技术对知识建模是知识组织的核心环节[19]。“科学知识图谱”一般使用社会网络建模方法:基于各类专业数据库中的知识,依据相关需求,如科学家合作研究、引文分析、生物模块预测等,将知识抽象成节点,而节点之间的关系抽象成边,从而构建成网络模型,各类模型因节点关系的不同而具有不同的网络结构。如科学家合作网络可以将科学家作为抽象节点,以是否共同发表论文确定节点间是否存在边连接,构建社会网络分析模型。相关的建模工具有 Ucinet(www. analytictech. com/ucinet/),Pajek(vlado. fmf. uni-lj. si/pub/networks/pajek/)等。

在知识组织过程中,“Google 知识图谱”一般首先分析实体(即现实世界的各种概念)的元数据(即实体属性,用于表述实体的特征),依据元数据构建本体模型,再依据实体之间语义关联构建语义网。按照语义网的构建规则,每个实体有唯一标识符(identifier),实体之间存在关联,也称作关系(relation)。“Google 知识图谱”一般以图(graph)模型来描述语义关系:其中的节点表示实体,而节点之间的边用来刻画属性或关系。实体、属性和属性值以 W3C 提出的资源描述框架 RDF[20]或属性图(property graph[21])为构建规则,构成语义三元组,是语义网基本单位。在大数据背景下,大量的语义三元组的相互链接即构成大规模的语义网络知识库,其中本体建模工具有 Protégé(protege. stanford. edu),WebOnto(kmi. open. ac. uk/technologies/name/webonto);语义网建模工具有 D2R(d2rq. org/d2r-server),LMF(code. google. com/p/lmf/)等。

2.2.3　知识存储　知识存储是将知识存储到各种载体,如书籍、期刊或数据库中。“科学知识图谱”本质是知识管理的分析方法,一般较少涉及知识存储过程。

“Google 知识图谱”本质是以语义三元组为基础的结构化的海量知识库。依据知识应用目的可以分为通用知识图谱和行业知识图谱,如表 1 所示。通用知识图谱一般指常识性知识,如维基百科(Wikipedia)、百度知心等百科类知识库,其中“Google 知识图谱”已经包含超过 5 亿个实体,35 亿个属性和相互关系;行业知识图谱则是指具有行业领域知识特征的结构化知识库,如 Geonames 知识库是存储基因组本体的知识库,Linked Movie Database 是存储影视本体的知识库,阿里巴巴知识库则是存储商品本体的知识库等。

表 1　“Google 知识图谱”举例

	名称	来源
通用知识图谱	Google 知识图谱	www. google. com
	Wikipedia(维基百科)	en. wikipedia. org
	百度知心	www. baidu. com
	搜狗知立方	www. sogou. com
行业知识图谱	Linked Movie Database	datahub. io/dataset/linkedmdb
	Geonames	www. geonames. org
	阿里巴巴知识库	www. etao. com

2.2.4　知识共享和知识创新　知识共享和创新阶段主要涉及个体或组织(或群体)的知识学习以及知识传播,关注创新型知识的产生环境、机制和方法。“科学知识图谱”侧重于知识共享,兼具知识创新功能,而“Google 知识图谱”则只偏重于知识创新。基于社会网络分析方法,“科学知识图谱”依据社会网络模

型和聚类分析工具,能准确构建成员关系密切的社区及发现社区关键人物,在此基础上构建知识共享的网络路径,如通过社区中的关键人物共享和传播知识;基于 cytoSpace(www.cytoscape.org/)和 visAnt(visant.bu.edu/)等网络可视化平台,"科学知识图谱"能应用聚类等算法从纷繁复杂的知识网络中发现创新型知识,借助可视化工具清晰展示知识结构和脉络,绘制知识地图,以导航方式显示知识之间的重要动态联系,方便用户把握知识来源、知识流动和知识汇聚过程的来龙去脉。

"Google 知识图谱"的长处是应用机器学习算法发现创新型知识。通过关联规则、图聚类等算法,分析所构建的语义网知识库,形成创新型知识,在此基础上基于实体检索方法提供智能检索和个性化推荐功能,为用户提供高质量的知识服务。

2.3 **适用研究领域** 除了本文重点讨论的图书情报领域之外,"科学知识图谱"的应用主要还集中在科学学、管理学和教育学等诸多领域。用于展示各领域的学科结构,可视化学科研究内容,揭示学科间的关系,以及识别和分析学科发展新趋势和预测学科前沿等。尤其对于科学学领域,在疏理科学发展历史,描述以科学家(团体)为代表的科学主体之间的科研合作情况,以及科技政策辅助分析和决策咨询等方面发挥日益重要的作用[22-23]。

"Google 知识图谱"的应用重点集中在信息科学领域,依照万维网联盟 W3C 制定的领域本体规范,主要由大型互联网企业构建实施,以推进知识创新和提供高水平知识服务为目标,目前涉及的行业和部门有证券、医疗、商业、娱乐、图书馆和情报行业等。

3 两类知识图谱的联系

两类知识图谱都是以图(Graph)为基础构建网络模型,在网络分析的基础上服务于知识管理,所有网络分析的现存的理论和方法都可以应用于两类知识图谱的分析,在这些方法中,具有代表性的是网络聚类分析和可视化分析方法。

3.1 **网络聚类分析** 聚类分析是将分析对象根据彼此之间的亲疏关系或相似程度分成不同的类群,密切关联或相似程度高的对象归到同一类群。对于"科学知识图谱",如共词分析过程中,将学科或主题中的关键词作为分析对象,利用网络模型中词与词之间的亲疏关系,应用聚类分析,发现隐藏的密切关联的类群,从而揭示学科或主题的结构与演化规律。对于"Google 知识图谱",为构建统一的结构化知识库,需要对含义相同但表述不同的实体归一化,即实体消歧或实体对齐过程,具体过程如下:以实体对象为聚类中心,利用空间向量模型等方法定义实体对象之间的相似度,应用聚类方法,分析不同表述的实体的相似程度,将相似程度高的那些实体归并为同一实体对象,并分配全局唯一标识,完成实体消歧和对齐。

3.2 **网络可视化分析** 网络可视化将复杂网络数据以清晰的网络视图展现出来,帮助研究者洞察其中隐藏的知识和规律。"科学知识图谱"能够利用相关可视化工具,如 citeSpace[12] 显示节点之间的关系,找出具有重要地位的文献、作者、学科和群体,绘出网络视图,构建显示知识关系的知识地图等。"Google 知识图谱"基于知识库中的语义网络模型,构建出基于图的大规模网络,应用网络可视化分析工具发现海量实体中蕴含的创新型知识并绘图展示。

4 大数据环境下两类知识图谱的应用分析

随着资源数字化进程的急速推进,众多领域的数字资源具有数据增加迅速,总数据量大,种类繁多且价值密度低等大数据特征,这将会给两类知识图谱在知识管理各阶段的相互关系及未来发展产生明显影响。

4.1 **知识获取和组织阶段的相互借鉴** "科学知识图谱"的数据一般依赖于现成的数据库获取知识,并在此基础上构建网络模型组织知识。海量数据下,特别是关联数据技术(Linked Open Data,LOD)已成为数据库技术发展的潮流,借助多种数据库关联,能更加全面地融合各种知识和产生创新型知识。因此借鉴"Google 知识图谱"的理念,从互联网和云计算系统中收集数据,以及关联多种异构数据库来构建知识库,是大数据时代"科学知识图谱"获取知识的重要手段;另一方面,在社会网络建模过程中,融入语义网的构建方法,在不同的节点间嵌入强语义关联,能够使得社会网络具有推理能力,实现网络分析的智能化。

"Google 知识图谱"可以借鉴"科学知识图谱"中的社会网络分析方法,如中心性、凝聚子群和核心-边缘结构等方法,从上述多个角度分析语义网实体之间的结构和关系,从而有利于全面解析语义网络的特征。

4.2 **知识存储和共享阶段的各自发展** "科学知识图谱"区别于"Google 知识图谱"重要功能在于能通过网络分析发现社团和社团中的关键人物,基于网络路径分析方法实现社团中成员的知识共享。大数据环境下,社团规模急剧扩大,可以达到百万以上的级别。大规模社会网络分析对计算机硬件以及相关算法的性能将提出更高的要求,可以预见,基于分布式计算机集群的云计算技术将会成为"科学知识图谱"大规模网络分析的主要手段。

"Google 知识图谱"则需要建立知识库,以存储海量的结构化语义网知识。基于分布式存储技术以取得

更大存储容量,另外优化分布式数据库的增、删、改、查以获取更优的管理性能是当前需要迫切解决的问题,代表技术如 Hadoop 平台上的分布数据库 NoSQL 技术等。

4.3　**知识创新阶段的深度融合**　应用数据挖掘算法从网络中发现知识是知识创新的重要手段,由于两类知识图谱在分析方法上同属于网络分析范畴,有关网络分析算法和工具能够相互通用并深度融合。针对海量数据挖掘的聚类和关联挖掘等属于“Google 知识图谱”的机器学习算法,可以集成到“科学知识图谱”相关的软件工具中,以提高算法和工具分析性能;另一方面,“Google 知识图谱”可以利用“科学知识图谱”中的可视化算法和工具展现大规模语义网络,清晰显示海量知识实体之间的复杂关系。

5　结　语

作为知识管理领域的重要分析方法,“科学知识图谱”以社会网络分析和可视化为核心方法,广泛应用于科学学、管理学和图书情报学等诸多领域,已经有了近 15 年的发展历程,其支撑理论的研究,体系方法的完善和应用成果方面都取得长足的发展。“Google 知识图谱”则是为顺应大数据发展的潮流而提出的基于语义网的海量知识库,从 2012 年至今不过 4 年时间,但是发展起点较高,伴随关联数据和机器学习研究的兴起,近年来在企业界发展势头迅猛。

正如刘则渊和陈超美等在相关文献中将“科学知识图谱”的“图”和“谱”分别释义为“可视化的知识图形”和“序列化的知识谱系”[15,24],非常形象地将知识网络的各种复杂的互动、交叉和演化关系勾画出来。在大数据时代,“科学知识图谱”将面临的是大规模网络单元的互动、交叉和演化的挑战,需要基于海量数据进行组织、疏理和挖掘,并在此基础形成创新型知识,而这正是“Google 知识图谱”的优势所在,因此二者在通过方法和工具上的进一步融合,从而促进知识创新方面将有极其丰富的发展空间。另外,“Google 知识图谱”中,以语义网模式绘制的“图”和领域本体规范下的“谱”,将给“科学知识图谱”理论研究和实践应用增添新的活力,也必将推动知识管理领域的新旧范式变革与更迭。

参 考 文 献

[1] Rasmussen E, Atkins H B, Borner K, et al. Visualizing knowledge domains[J]. Proceedings of the American Society for Information Science & Technology, 2002, 39(39): 476-477.

[2] 陈　悦, 刘则渊. 悄然兴起的科学知识图谱[J]. 科学学研究, 2005, 23(2): 149-154.

[3] 曹树金, 吴育冰, 韦景竹, 等. 知识图谱研究的脉络、流派与趋势—基于 SSCI 与 CSSCI 期刊论文的计量与可视化[J]. 中国图书馆学报, 2015, 41(219): 19-31.

[4] 李明鑫, 王　松. 近十年国内知识图谱研究脉络及主题分析[J]. 图书情报工作, 2016(4): 93-101.

[5] 谷歌知识图谱[EB/OL]. [2016-08-06]. https://www.google.com/intl/es419/insidesearch/features/search/knowledge.html.

[6] Semantic web architecture[EB/OL]. [2016-08-06]. http://www.w3.org/2000/Talks/1206-xml2k-tbl/.

[7] Gruber T R. A translation approach to portable ontology specifications[J]. Knowledge Acquisition, 1993, 5(2): 199-220.

[8] 王元卓, 贾岩涛, 刘大伟, 等. 基于开放网络知识的信息检索与数据挖掘[J]. 计算机研究与发展, 2015, 52(2): 456-471.

[9] 苏永浩, 张　驰, 程文亮, 等. CLEQS —基于知识图谱构建的跨语言实体查询系统[J]. 计算机应用, 2016, 36(S1): 204-206.

[10] 杜亚军, 吴　越. 微博知识图谱构建方法研究[J]. 西华大学学报(自然科学版), 2015, 34(1): 27-35.

[11] Cobo M J, López-Herrera A G, Herrera-Viedma E, et al. Science mapping software tools: Review, analysis, and cooperative study among tools[J]. Journal of the American Society for Information Science and Technology, 2011, 62(7): 1382-1402.

[12] 陈超美. CiteSpace II: 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.

[13] 秦长江, 侯汉清. 知识图谱——信息管理与知识管理的新领域[J]. 大学图书馆学报, 2009, 27(1): 30-37.

[14] 库恩 T S. 科学革命的结构[M]. 上海: 上海科学技术出版社, 1980.

[15] 刘则渊, 陈　悦, 侯海燕. 科学知识图谱: 方法与应用[M]. 北京: 人民出版社, 2008.

[16] 陈　悦, 陈超美, 刘则渊, 等. CiteSpace 知识图谱的方法论功能[J]. 科学学研究, 2015, 33(2): 242-253.

[17] Staab S, Studer R, Schnurr H P, et al. Knowledge processes and ontologies[J]. IEEE Intelligent Systems, 2001, 16(1): 26-34.

[18] Armistead C. Knowledge management and process performance[J]. Journal of Knowledge Management, 1999, 3(2): 143-154.

[19] 王　昊, 谷　俊, 苏新宁. 本体驱动的知识管理系统模型及其应用研究[J]. 中国图书馆学报, 2013, 39(204): 98-110.

[20] RDF 1.1 concepts and abstract syntax[EB/OL]. [2016-08-06]. http://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/.

[21] Property graph model[EB/OL]. [2016-08-06]. https://github.com/tinkerpop/blueprints/wiki/Property-Graph-Model.

[22] 杨思洛, 韩瑞珍. 国外知识图谱的应用研究现状分析[J]. 情报资料工作, 2013(6): 16-18.

[23] 梁永霞, 李正风. 基于 CSSCI 的中国科技政策研究的知识图谱[J]. 中国科技论坛, 2010(10): 86-93.

[24] 刘则渊, 陈超美, 侯海燕. 迈向科学学大变革的时代[J]. 科学学与科学技术管理, 2009, 30(7): 5-12.

(责编: 贺小利; 校对: 王平军)