● 何 胜[1,2]，熊太纯[3]，叶飞跃[1,2]，李仁璞[1,2]，冯新翎[1,2]
(1. 江苏理工学院计算机工程学院，江苏 常州 213001；2. 常州市云计算与智能信息处理重点实验室，江苏 常州 213001；3. 江苏理工学院图书馆，江苏 常州 213001)

基于语义网的高校图书馆学科知识服务方案研究*

摘 要：开展学科知识服务是高校图书馆学科馆员的主要任务之一。文章首先分析了大数据环境下高校图书馆学科知识服务面临的挑战，针对学科知识库构建、智能检索与知识问答以及学科服务平台建设等问题，基于语义网设计了一种面向高校图书馆学科知识服务问题的解决方案并给出应用案例，帮助学科馆员应对大数据环境下学科知识服务的挑战，促进该领域进一步发展。

关键词：高校图书馆；学科知识服务；语义网；本体

Abstract: It is one of the main tasks for subject librarians to provide subject knowledge services in university libraries. Firstly, the challenges are analyzed for subject knowledge service of university libraries in big data environment. To solve the problems including the construction of subject knowledge database, intelligent retrieval and question answering as well as subject service platform construction, the paper provides a solution of subject knowledge service problems in university libraries based on semantic web and applies the solution in a practical case in order to help subject librarians deal with the challenges of subject knowledge service in big data environment and promote the field for further development.

Keywords: university libraries; subject knowledge service; semantic web; ontology

数据开放和知识共享时代，开展知识服务是图书馆服务的主要任务，即以知识的搜寻、组织、分析为基础，根据用户的问题和所处的环境，提供有效的支持知识创新和知识应用的服务[1]。学科知识服务一般是指学科馆员基于图书馆资源（图书，数据库等）提供的知识服务。针对包括高校教师和学生在内的用户在学科知识获取、知识应用和知识创新方面的需求，利用图书馆相关资源，由专业学科馆员提供的知识组织、知识检索、知识挖掘和知识可视化等专业知识服务[1-2]。近年来，由于高校图书馆资源数字化迅猛发展，以各类数据库为基础的图书馆学科数据具有种类多，总量大，数据增加迅速，价值密度低等大数据特征，对这些数据库的恰当融合，智能检索以及有效挖掘的需求急速增长[3-5]。如何基于现有的数字资源，合理构建学科知识库，实现智能检索和知识问答，并建立学科知识服务平台以充分发挥学科馆员的主导作用是大数据时代高校图书馆学科知识服务面临的重要问题[1-4]。

1 大数据背景下高校图书馆学科知识服务

对比传统的学科知识服务，大数据环境下学科知识库的构建，知识检索与问答以及学科馆员主导作用的发挥面临新的挑战。

1.1 学科知识融合与知识挖掘

学科知识库是学科馆员或用户用于检索和查询的数据库，大数据环境下，数据来源众多，不同格式数据纷繁复杂，有效融合异构数据库，建立统一的结构化知识库是学科服务的基础。

万维网联盟 W3C 提出的语义网技术为数据库融合和应用提供了技术支撑。语义网（Semantic Web）[7] 结合本体技术（Ontology）[8]，以图（Graph）模型来描述现实世界中的各种实体及其复杂关系，其中的节点表示实体，而节点之间的边用来刻画实体的属性或实体之间的关系。按照资源描述框架（Resource Description Framework，RDF）[9] 和属性图（Property Graph）[10] 组建规则，实体、属性和属性值组建成为语义三元组（语义网的基本单位），大量的语义三元组构成语义网数据库。将各种异构的语义网数据库有机链接起来，即数据关联技术（Linked Open Data，LOD）[11]。当前，针对一定规模的多样性数据库的关联和融合取得了较好的进展[12-14]。然而在大数据背景下，由于不同的数据库之间，数据的重复度高，冗余数据多，如何应用数据关联技术融合海量的学科知识库需要深入研究。

价值密度低是大数据的另一个特征，如何从海量的低

* 本文为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”的成果之一，项目编号：15BTQ016。

价值密度的各类数据库挖掘出高价值知识以支持知识创新，是学科知识服务的重要任务[4]。基于云计算平台提高计算性能，并采用数据挖掘和分析工具从海量数据中获取知识是知识创新的重要手段[3]，近来，以图数据库为基础，应用大规模网络分析和网络挖掘的知识发现方法成为大数据分析的热点方法[15-16]。

1.2 学科知识检索与知识问答

学科知识检索是图书馆学科服务中获取知识的基本方法。传统的知识检索方法，是以检索的关键词匹配数据库字段的方式，在海量数据环境下，这种匹配方式容易产生大量似是而非的检索结果，准确度不高。学科知识问答是知识服务的新功能，用户一般提供查询语句而非单个关键词，且要求精准快速回答相关问题。知识问答需要先准确分析查询语句并按语义拆分该语句为单个分词，再通过分词查询知识库，从而给出高相关度的答案。

语义网查询技术以结构化的实体和链接实体的网络关系为基础，能够查询到较为深入、广泛和完整的结构化知识[17]，可以应用于图书馆学科知识检索与知识问答。

1.3 学科馆员的主导作用

因缺乏方便的知识服务平台和有力的分析工具，学科馆员所提供的知识服务难以适应大数据背景下的需求。学科馆员所熟悉的传统知识检索手段和分析工具，高校教师和研究人员也能方便获取并使用；学科馆员虽有一定的学科知识背景，但是就某一确定的研究领域，所涉及的知识深度，还难以比肩高校中长期从事相关领域研究的教师，导致学科馆员难以及时提供高校教师急需的创新型知识，在学科知识服务中的主导作用难以有效发挥。因此，需要利用信息技术，构建大数据下的图书馆学科知识平台作为学科服务的支撑，使得学科馆员能够高效地开展海量数据下知识检索、知识问答和知识可视化等服务，从而发挥主导作用。

从上述分析可以看出，语义网是学科知识服务领域的重要技术，但在大数据环境下面临挑战。近年来，为应对海量数据检索问题，以 Google 公司为代表的企业纷纷构建大规模知识库，以语义网和本体为其关键技术，在海量数据的分析和挖掘的基础上，提供智能检索和知识问答服务，引起企业界和学界关注。现有基于语义网的知识库包括维基百科知识库 DBPedia[18]，YAGO[19]，中文知名知识库有搜狗知立方[20]和百度知心[21]等，不少知识库的存储规模达到亿万级。将知识库构建、知识检索和问答以及知识挖掘等一系列技术引入高校图书馆学科知识服务，将会促进该领域大数据应用的落地，从而提高服务质量。

2 基于语义网的学科知识服务方案

针对大数据环境下高校图书馆学科知识服务面临的问题，提出一种新的学科知识服务方案，见图 1。其中，“学科知识采集与清洗”为知识服务提供数据基础，“学科知识库构建与知识存储”给出学科知识库组建方法，包括“知识检索、知识问答和可视化”三种功能的“学科知识服务平台”是学科馆员开展知识服务的软件平台。

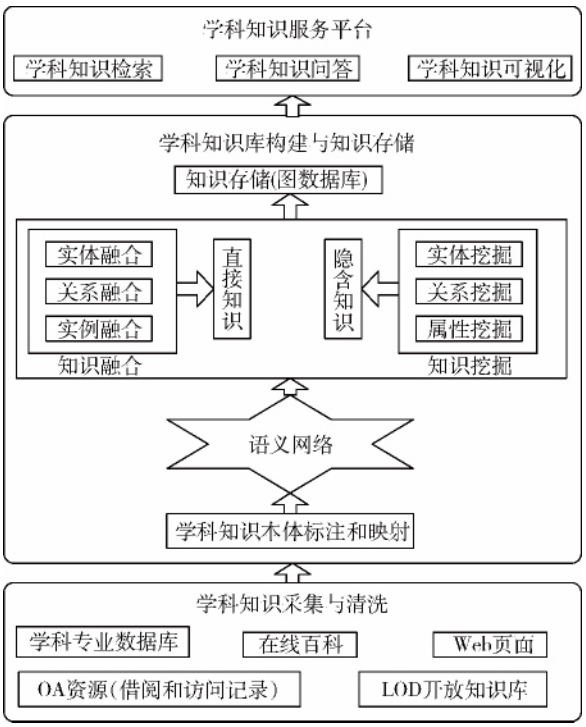

图 1 大数据下基于语义网的高校图书馆学科知识服务方案

2.1 学科知识采集与清洗

学科知识数据库来源，根据其产生方式的不同分为以下几类，学科专业数据库是指万方，维普等通过签约等方式获取使用权的知名数据库，或者图书馆自建的有本校特色的专业数据库或文献库；在线百科是由用户、专家等互联网络使用者共同参与编辑并反复完善而构建起来的知识库，库中包含大量结构化的学科知识，如百度百科、维基百科等；Web 页面数据来源于互联网网页，包括文本、图片及音视频等海量信息，需要通过信息抽取和挖掘技术，产生新的学科数据；OA 资源来自于图书馆日常服务过程中产生的各种搜索、浏览和借阅的日志数据，其特点是数据历史性强，数据增加迅速；LOD 是以 RDF 格式在 Web 上发布各种开放的关联知识库。

由于学科数据的多源、异构以及收集过程中难以避免的各种错误，使得这些数据一定程度上含有噪音，且部分数据出现冗余，甚至缺失的特点。为获得具有一致、正确

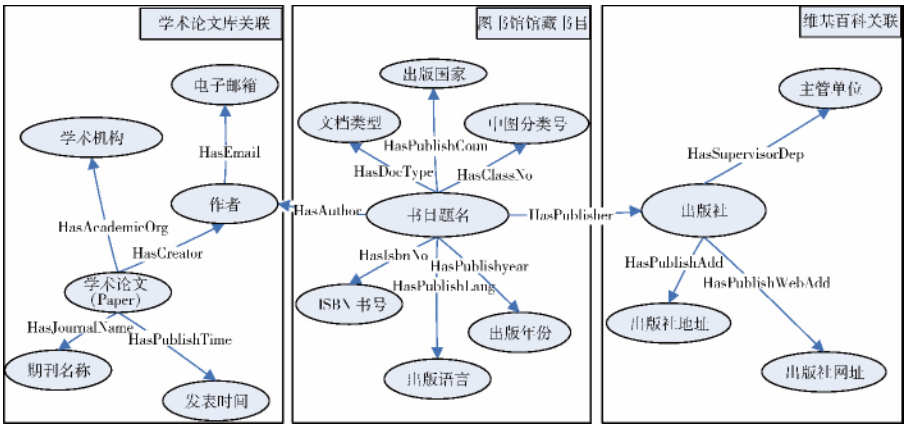

图 2　以高校图书馆馆藏书目库为基础库的学科知识库结构

和完备[22]的高质量学科数据，需要对数据进行清洗，即使用相关工具（如 Extraction Transformation and Loading，抽取、转化和装载工具）检查数据并除去数据中所有明显的重复、错误和不一致[15]。

2.2　学科知识库构建与存储

经过数据清洗的不同类型的学科数据需要先对照相应本体进行标注和映射，并链接形成具有本体语义的语义网络。在此基础上，进行知识融合和知识挖掘，将产生的直接知识和隐含知识存储于图数据库（Graph DataBase，GDB，如 Neo4j 图数据库），形成具有 RDF 结构的统一学科知识库。

知识融合包括实体融合、关系融合和实例融合[23]，在语义网基础上，针对不同的数据库，合并现有的知识，对其中语义元素进行去重，消歧并归一化；知识挖掘包括实体挖掘，关系挖掘和属性挖掘，应用数据挖掘等信息技术（如“图聚类算法”“关联规则挖掘”等机器学习算法）从学科知识库中产生新的知识，用于发现隐含知识，是知识创新的重要手段[24]。

以下是将图书馆馆藏书目库作为基础库，通过关联“维基百科”和“学术论文库”构建学科知识库的案例，包括知识库结构设计，本体库选择和知识库的语义网模型构建三个部分。知识库结构见图 2，依据中国机读目录（China Machine Readable Catalogue，CNMARC）的图书著录规则，图书馆馆藏元数据集合对应“书目题名”“作者”“出版社”“ISBN 书号”“出版语言”“出版年份”“中图分类号”等属性。通过“作者”属性关联学术论文库，通过“书目题目”“作者”和“出版社”等属性关联维基百科，限于篇幅，案例中给出“出版社”属性的关联示意图。学术论文库的元数据集合对应“作者”“学术论文”“期刊名称”“学术机构”“电子邮箱”“发表时间”等属性。

依据图 2 学科知识库结构，经过细致调研后，选择 W3C 组织推荐的知名本体库，确定关联关系和对应的属性类别，见表 1。

依据表 1 的本体构建出高校图书馆学科知识实体元素之间的关联图，见图 3，形成 16 组具有 RDF 结构语义网三元组数据模型。通过对“学术论文库”及“维基百科”等海量数据库的实时关联检索，以及针对检索结果的数据记录中重复的实体，关系和实例的去重、消歧后存入数据库，完成知识库的构建。

表 1　学科知识库的本体库选择及关联关系

本体库（命名空间）	属性	关联关系
rdf（http://www.w3.org/1999/02/22-rdf-syntax-ns#）	rdf：type	书目题名—文档类型
	rdf：country	书目题名—出版国家
dc（http://purl.org/dc/elements/1.1/）	dc：Creator	学术论文—作者
	dc：publisher	书目题名—出版社
	dc：language	书目题名—出版语言
	dc：identifier	书目题名—中图分类号
	dc：date	书目题名—出版日期
dcterms（http://purl.org/dc/terms/）	dcterms：isPartOf	学术论文—期刊名称
	dcterms：issued	学术论文—发表时间
	dcterms：rightsHolder	出版社—主管单位
vCard（http://www.w3.org/2001/vcard-rdf/3.0#）	vCard：ORG	学术论文—学术机构
	vCard：Email	作者—电子邮箱
	vCard：ARD	出版社—出版社地址
foaf（http://xmlns.com/foaf/0.1/）	foaf：name	书目题名—作者
	foaf：homepage	出版社—出版社网址
bibo（http://purl.org/ontology/bibo/）	bibo：isbn	书目题目—ISBN 书号

2.3　学科知识服务平台

学科知识服务平台主要面向学科馆员，包括学科知识检索、学科知识问答和学科知识可视化三个部分。

1）学科知识检索。用户输入关键词或查询词，依据学科知识库中结构化的三元组，搜索引擎识别其中涉及的学科实体或属性，按重要性高低展现与查询实体相关的知识卡片；通过实时计算显示并推荐高相关度的一类实体或属性，在搜索界面显示“其他人还搜了”效果，供用户参考。

2）学科知识问答。通过语义框架分析准确理解用户所提出的语句含义，给出精确的答案，而非大量相关性不

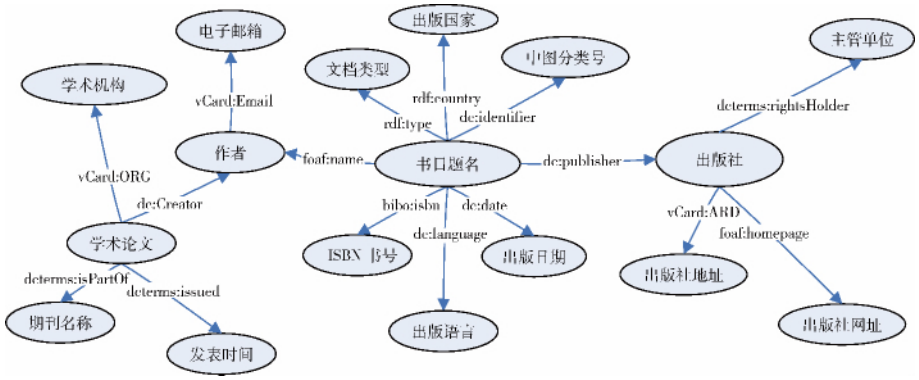

图 3　高校图书馆学科知识库的语义网模型

高的网页链接列表。在上述已构建的学科知识库案例的基础上，设计的问答流程如图 4 所示，共包括以下步骤：①用户以语句的形式提出“问题”；②“问题理解”——为准确理解用户问题，应用语义分词技术，分割并抽取语句中的关键词；③“语义框架分析”——利用机器学习算法分析结果，确定问题的焦点，问题类型，确定检索知识库所需的关键参数；④“SPARQL 查询”——自动产生 SPARQL 查询语句检索学科知识库；⑤给出精确“答案”。

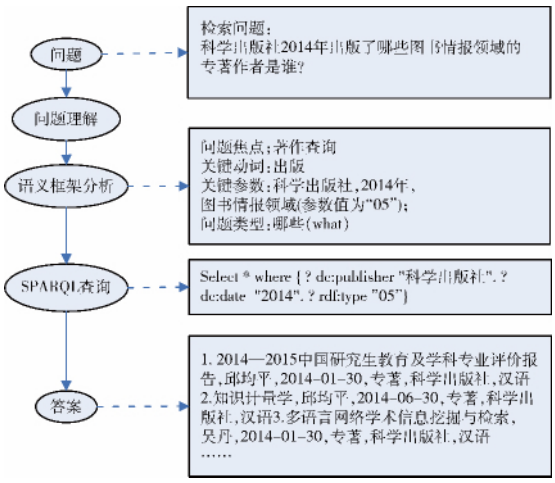

图 4　基于学科知识库的知识问答

3）学科知识可视化。通过知识地图和网络可视化方法给出知识脉络和相互关系。建立于学科知识库基础上的知识地图是学科知识服务的重要方式，通过知识导航显示知识实体之间的动态联系，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。基于语义网的网络可视化能够通过清晰独特的网络方式动态展现知识结构，揭示知识关系进而预测学科知识前沿。

3　结束语

以语义网和本体为关键技术的学科知识服务方案，具有以下 3 个方面的优点。

学科知识库的深度融合和有效挖掘。以图为数据结构组建语义网络来描述知识之间的联系，更易于融合相关学科的不同来源的异构数据源，有利于形成有统一语义网数据表达格式的知识库；另外，在构建学科知识库的过程中，能方便地应用大规模网络分析和挖掘算法分析语义网络，产生大量的创新型知识，从而构建出高质量的学科大数据知识库。

学科知识的精准检索和智能问答。基于语义网的大规模学科知识库由结构化的实体和链接实体的网络关系构成，在检索时，采用的关键词匹配实体的技术，循着网络链接关系能更加快速而精准地找到所需的信息；建立于学科知识库上的智能问答能通过对查询语句的语义解析，准确理解用户问题和用户查询意图，从而提供智能知识问答，使得学科馆员的知识服务快捷高效。

学科知识的清晰导航和直观可视化。依据基于语义网的结构化学科知识库，方便构建学科知识地图，提供知识导航服务。基于图（Graph）的语义网络可视化技术，能够通过显示节点和边的网络拓扑结构直观理解和把握学科知识的演化脉络和关联。□

参考文献

[1] 徐恺英，刘佳，班孝林. 高校图书馆学科化知识服务模式研究 [J]. 图书情报工作，2007，51（3）：53-55.
[2] 高俊芳. 云计算下的高校图书馆学科知识服务研究 [J]. 现代情报，2013，33（9）：54-58.
[3] 韩翠峰. 大数据时代图书馆的服务创新与发展 [J]. 图书馆，2013（1）：121-122.
[4] 苏新宁. 大数据时代数字图书馆面临的机遇和挑战 [J]. 中国图书馆学报，2015（6）：4-12.
[5] 白如江，冷伏海. “大数据”时代科学数据整合研究 [J]. 情报理论与实践，2014，37（1）：94-99.
[6] 何胜，熊太纯，周冰，等. 高校图书馆大数据服务现实困境与应用模式分析 [J]. 图书情报工作，2015，59（22）：50-54.
[7] Semantic web architecture [EB/OL]. [2016-06-11]. http://www.w3.org/2000/Talks/1206-xml2k-tbl/.
[8] Gruber T R. A translation approach to portable ontology specifications [J]. Knowledge Acquisition，1993，5（2）：199-220.
[9] RDF 1.1 Concepts and Abstract Syntax [EB/OL]. [2016-06-11]. http://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/.
[10] Property Graph Model [EB/OL]. [2016-06-11]. https://github.com/tinkerpop/blueprints/wiki/Property-Graph-Model.

（下转第 106 页）

机制，分析师徒契合度对隐性知识的影响。在企业型组织中，选择合适的隐性知识转移的主体和受体，不仅会提升知识转移的绩效，还会促进徒弟对所接受的隐性知识更好地融合吸收和创新，进而实现企业知识资产的增加。考察师徒之间的契合度对隐性知识转移是十分有必要的，是隐性知识转移不可忽视的一个因素。同时，本研究仍然存在很多不足，有关契合度对师徒制隐性知识转移绩效的影响等方面，有待进一步的探索和分析。□

参考文献

[1] 王晓蓉，李南. 企业师徒制中隐性知识转移路径及其微观过程研究 [J]. 情报理论与实践，2012，35（6）：26-30.

[2] 李南，王晓蓉. 企业师徒制隐性知识转移的影响因素研究 [J]. 软科学，2013，27（2）：113.

[3] 万涛. 隐性知识转化为显性知识的评价判断规则研究 [J]. 管理评论，2015，27（7）：66-75.

[4] 秦亚欧，刘岩. 虚拟社区知识转移的情境研究 [J]. 情报科学，2015，33（3）：41-44.

[5] 单汨源，冯彦，张人龙. 基于 Multi-Agent 的创新型组织隐性知识传播模型研究 [J]. 科技管理研究，2015（17）：92-95.

[6] 温丹丹，杨岚，张建华. 知识传播中基于参与者环境和行为的信任评价机制研究 [J]. 情报科学，2015，33（3）：85-89.

[7] 王欣，孙冰. 企业内知识转移的系统动力学建模与仿真 [J]. 情报科学，2012，30（2）：173-178.

[8] 杨波. 系统动力学建模的知识转移演化模型与仿真 [J]. 图书情报工作，2010，54（18）：89-94.

[9] 王进. 领导部属契合度伦理认同与知识共享关系的实证研究 [J]. 科技管理研究，2013（2）：146-149.

[10] 朱卫未，于娱，施琴芬. 隐性知识转移势差效应机理研究及主体需要层次分析 [J]. 科技进步与对策，2011，28（3）：122-125.

[11] 周密，赵文红，宋红媛. 基于知识特性的知识距离对知识转移影响研究 [J]. 科学学研究，2015，33（7）：1059-1068.

[12] 张喜征，聂振. 企业间知识距离测度模型及其应用研究 [J]. 科技进步与对策，2009，26（22）：160-163.

[13] 陈伟，潘伟，杨早立. 知识势差对知识治理绩效的影响机理研究 [J]. 科学学研究，2013，32（12）：1864-1871.

[14] 张宝生，张庆普. 基于扎根理论的隐性知识流转网成员合作意愿影响因素研究 [J]. 管理学报，2015，12（8）：1224-1229.

[15] 蒋天颖，孙伟. 关系嵌入强度、知识吸收能力与集群企业技术创新扩散 [J]. 情报杂志，2012，31（10）：201-206.

[16] 汤中彬，张扬，乔长蛟. 人际情报网络隐性知识共享影响因素分析及网络模式构建 [J]. 情报科学，2015，33（9）：100-104.

[17] 蔡小筱，张敏. 虚拟社区中基于熟人关系的知识共享研究综述 [J]. 图书馆学研究，2015（2）：2-11.

[18] 段钊，张文静，卢新元. 组织成员间知识转移的博弈仿真分析 [J]. 情报科学，2011，28（11）：1724-1732.

作者简介：李伟，男，1993 年生，硕士生。

郭东强，男，1957 年生，博士，教授。

录用日期：2016－08－08

（上接第 110 页）

[11] 关联开放数据 [EB/OL]. [2016-06-11]. http://linkeddata.org/.

[12] 夏翠娟，刘炜，赵亮，等. 关联数据发布技术及其实现——以 Drupal 为例 [J]. 中国图书馆学报，2012，38（1）：49-57.

[13] 盛东方，孙建军. 基于语义搜索引擎的学科知识服务研究——以 GoPubMed 为例 [J]. 图书情报工作，2015（4）：113-119.

[14] 庄倩，常颖聪，何琳，等. 基于关联数据的科学数据组织研究 [J]. 情报理论与实践，2016，39（5）：22-26.

[15] 王元卓，靳小龙，程学旗. 网络大数据：现状与展望 [J]. 计算机学报，2013，36（6）：1125-1138.

[16] 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.

[17] 孙建军，徐芳. 基于关联数据的学科网络信息深度聚合框架构建 [J]. 图书馆，2015（7）：50-54.

[18] 维基知识图谱 [EB/OL]. [2016-06-11]. http://wiki.dbpedia.org/.

[19] YAGO 知识图谱 [EB/OL]. [2016-06-11]. http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/.

[20] 搜狗知立方 [EB/OL]. [2016-06-11]. http://baike.sogou.com/v66616234.htm.

[21] 百度知心 [EB/OL]. [2016-06-11]. http://baike.baidu.com.

[22] WANG R Y，KON H B，MADNICK S E. Data quality requirements analysis and modeling [C]. IEEE Computer Society. Proc. 9th ICDE. Vienna：Austria，1993：670-677.

[23] 知识图谱——机器大脑中的知识库 [EB/OL]. [2016-06-11]. http://www.kmcenter.org/html/s3/201508/21-16619.html.

[24] 知识图谱技术原理介绍 [EB/OL]. [2016-06-11]. http://www.kmcenter.org/html/s78/201601/11-16906_2.html.

作者简介：何胜，男，1971 年生，博士，副教授。

熊太纯，男，1971 年生，硕士，研究馆员。

叶飞跃，男，1960 年生，博士，教授。

李仁璞，男，1976 年生，博士，教授。

冯新翎，女，1981 年生，硕士，实验师。

录用日期：2016－07－11