【信息科学与控制工程】　　doi：10.11809/bqzbgcxb2019.01.031

基于百科知识的军事装备知识图谱构建与应用

车金立，唐力伟，邓士杰，苏续军

（陆军工程大学石家庄校区火炮工程系，石家庄　050003）

摘要：为解决信息时代网络中的军事装备数据分布较为稀疏，数据间缺乏良好的关联与组织，导致知识难以被高效利用的问题，提出了一种军事装备知识图谱的构建方法，该方法通过网络爬虫不断获取原始百科数据，利用高质量的百科知识对知识图谱构建过程中的知识抽取、知识融合、知识图谱的储存与更新等关键技术进行研究，并在已构建的知识图谱基础上实现了军事装备领域的知识问答。该方法有效利用了网页中的零散军事装备数据，实现了军事装备知识图谱的构建。

关键词：军事装备；知识图谱；网络爬虫；知识问答

本文引用格式：车金立，唐力伟，邓士杰，等. 基于百科知识的军事装备知识图谱构建与应用[J]. 兵器装备工程学报，2019，40(1)：148－153.

Citation format：CHE Jinli，TANG Liwei，DENG Shijie，et al. Construction and Application of Military Equipment Knowledge Graph Based on Encyclopedia Knowledge[J]. Journal of Ordnance Equipment Engineering，2019，40(1)：148－153.

中图分类号：TJ01；TP391　　**文献标识码**：A　　**文章编号**：2096－2304(2019)01－0148－06

Construction and Application of Military Equipment Knowledge Graph Based on Encyclopedia Knowledge

CHE Jinli，TANG Liwei，DENG Shijie，SU Xujun

（Department of Artillery Engineering，Army Engineering University，Shijiazhuang 050003，China）

Abstract：In the information age，the distribution of military equipment data is sparse in the network. And there is a lack of good association and organization for military equipment data. These all lead to the difficulty of efficient use of knowledge. To solve the above problems，a method of constructing military equipment knowledge graph was proposed. This method obtains the original encyclopedia data through the web crawler，and uses these high quality encyclopedia knowledge to research the key technologies such as knowledge extraction，knowledge fusion，storage and update of knowledge graph in the process of constructing knowledge graph. On the basis of the constructed knowledge graph，knowledge questions and answers in the field of military equipment were also realized. The method effectively utilizes the scattered military equipment data in the webpage，and realizes the construction of knowledge graph of military equipment.

Key words：military equipment；knowledge graph；web crawler；knowledge question and answer

伴随信息时代的发展，各个行业中的业务数据呈爆炸式增长，已逐渐成为一种重要资源，并发挥着不可或缺的作用。与此同时，在军事装备领域中，也拥有大量装备类型、装备型号及装备参数等数据。这些数据蕴含着丰富的价值信息，然而，相关人员在面对这些海量的军事装备数据时，却无法高效地从中获取关键知识，从而也无法依据数据指导军事装备

收稿日期：2018－08－05；**修回日期**：2018－09－29

作者简介：车金立(1994—)，男，硕士研究生，主要从事自然语言处理与知识图谱构建研究。

工作[1]。通过调研发现,导致数据中价值信息继续沉睡的首要问题是,这些数据缺乏紧凑有效的组织结构与直观形象的可视化查询方式,没有形成相应的军事装备知识体系,难以进行深层次的数据挖掘与应用。

因此为解决上述问题,本文将知识图谱(knowledge graph,KG)技术引入到军事装备领域。知识图谱是借助信息处理技术从海量文本数据中抽取领域实体、属性及其相互关系,从而构成结构化知识的一种知识表示形式。知识图谱可以将具有复杂关系的知识通过图形化的方式准确形象的表达出来,这种形成体系的知识更符合人类的学习习惯,有助于相关人员快速地从海量数据中学习关键知识及其相互关系,更好地把握知识结构。知识图谱目前在通用领域、金融领域、医疗领域及情报领域等均有广泛应用,且发挥着前所未有的作用[2]。

谷歌在2012年率先提出知识图谱的概念,并将其用于增强搜索引擎的性能,提高知识搜索的质量[3]。此后,知识图谱就在国内外掀起了研发热潮,通用领域知识图谱的构建技术逐渐走向成熟。例如基于维基百科数据构建的多语言知识图谱DBpedia[4],其规模庞大并覆盖了125种语言,在全世界都有广泛的应用。以及具有较高置信度的YAGO[5],其拥有的大部分实体都经过了高质量的评估,具有较高的准确率,但所包含实体的丰富程度低于DBpedia。另外由微软公司所构建的Probase[6]也极具代表性,是拥有概念最多的知识图谱。除此之外,国内许多机构也展开了对通用知识图谱的探索,例如上海交通大学构建的首个中文通用领域知识图谱Zhishi.me[7]、复旦大学构建的知识图谱、百度构建的"知心"和搜狗构建的"知立方"等。这些通用知识图谱的数据大多都来自于百科知识及已有的结构化数据,覆盖范围较为广泛。此后,一部分学者开始对垂直知识图谱进行研究,垂直知识图谱就是针对特定领域构建的知识图谱,其构建过程依赖该领域的行业数据,覆盖范围较窄。如面向影视领域的IMDB、面向音乐领域的MusicBrainz、面向中医药领域的知识图谱[8]及面向植物领域的知识图谱[9]等,这些垂直知识图谱在构建过程中也有相当一部分数据来源于互联网中的网页。通过总结以上知识图谱的研究可以发现,领域中现有的结构化数据以及百科网页中的半结构化数据是构建知识图谱的关键数据源。与此同时,在面向军事装备领域时,由于该领域具有一定的保密性,相关数据难以获取,且实体间关系较为复杂,面向非结构化数据抽取知识时准确性有限,构建军事装备知识图谱面临一系列困难。

为此,本文主要基于数据质量相对较高的半结构化百科知识构建军事装备知识图谱,对构建流程中的知识抽取、知识融合、知识图谱的构建与储存以及知识图谱的更新维护等重要技术进行研究,为装备工作者提供全面准确的军事装备知识体系。并在此基础上,实现了基于军事装备知识图谱的知识问答,辅助装备工作者快速准确地获取所需知识。

1 知识图谱技术架构

1)知识图谱的定义

知识图谱实质上是一种以图形的方式形象表示现实概念及其相关关系的结构化知识库,其更加符合人类对现实世界的认知,是一种更为合理的管理海量知识的方式,其基本的网状结构如图1所示。另外,也正是这种储存结构使得知识图谱可扩充性较好,便于修改知识及添加新的知识。

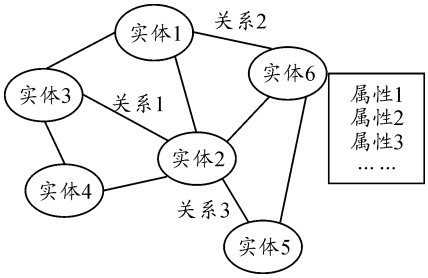

图1 知识图谱的网状结构

知识图谱的基本单元为<实体1—关系—实体2>及<实体—属性名称—属性值>这种三元组,图1中的节点代表实体,实体是知识图谱中最基本的元素,描述现实世界中某一概念分类中的个体,如步枪分类中的具体型号步枪,并且每个实体应拥有唯一的ID来与其他实体进行区分。图1中的边代表实体间的相互关系,实体通过彼此关系相互连接,进而构成完整清晰的结构化知识网络。另外,实体可包含多个属性,描述实体可能具有的属性、特性及参数等内在特征,例如射程、杀伤力等。

2)知识图谱构建的技术流程

构建知识图谱的主要目标就是综合利用多种信息处理技术将蕴藏在非结构化文本数据、半结构化网页及表格数据及结构化数据中具有价值的知识抽取出来,以三元组的形式储存在计算机中形成知识体系,便于人与计算机的进一步学习与分析利用,

知识图谱构建的技术流程如图2所示,主要就是通过信息抽取、知识融合、图谱的构建与储存等核心技术从海量的异构数据源出发,抽取其中的实体、属性等关键知识及其相互关系,再完成实体对齐与实体消歧等数据融合工作,然后采用符合质量要求的知识构建图谱,并利用图数据库对构建好的知识图谱进行储存。另外,由于信息时代中数据随时在高速增长,人们对于知识的认知需求也在持续提升,因此为了保证知识图谱的时效性,还需按照整个技术流程将新的知识不断地加入到已构建完成的知识图谱当中,实现对知识图谱的更新与维护。

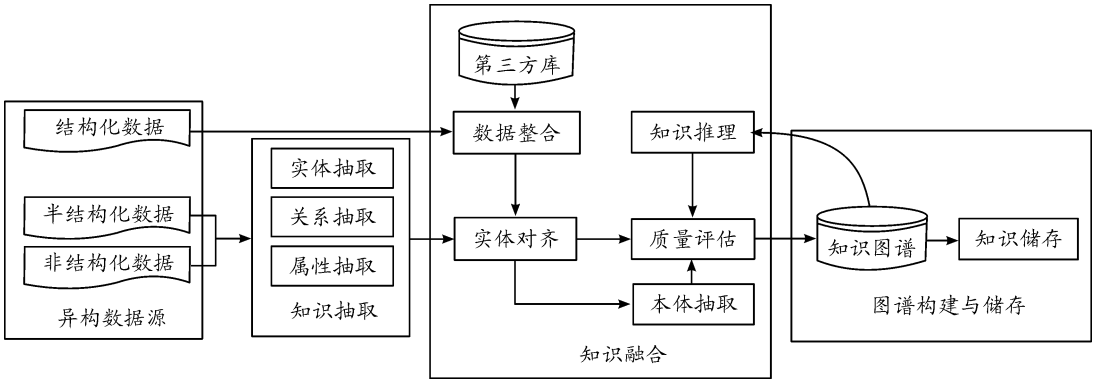

图 2 知识图谱构建的技术流程

目前,按照知识图谱的构建方式,主要分为自顶向下与自底向上两种方式[10]。其中,自顶向下构建知识图谱的方式相对简单,但这种方式十分依赖已构建好的高质量结构化知识库,需要首先从结构化数据源中抽取出相应的本体模式,定义好概念之间的层次关系,然后再将从多种数据源中学习到的实体及属性添加到定义好的概念体系当中。在知识图谱研发的初期,许多公司都采用这种高效的方式,例如,谷歌构建的知识图谱中有很大一部分数据都来自 Freebase 数据库,DBpedia 是在维基百科数据基础上构建完成的。

然而,随着自然语言处理等知识抽取技术的快速发展,自底向上的构建方式逐渐成为了主流。该方式的主要思想是利用知识抽取技术先从异构数据源中抽取实体,然后再对抽取到的实体进行抽象,归纳出概念间的体系结构,进行本体的构建。这种方式有利于抽取出新的本体模式,可以构建出更为丰富全面的概念体系,且自动化程度更高,Knowledge Vault[11] 则是通过这种方式构建的典型代表,实现了对现有知识库的补充与完善。

2 基于百科知识的军事装备知识图谱构建

2.1 军事装备知识图谱构建流程

依据知识的覆盖范围进行划分,面向特定领域构建的军事装备知识图谱应属于垂直知识图谱,它相比于强调知识广度的通用知识图谱而言,更强调知识的深度、知识的准确性及知识的层次结构[12]。因此为使军事装备工作者及相关人员快速准确地把握关键知识及知识体系,军事装备知识图谱应具有很强的领域针对性与专业性,需要根据领域相关度更高且丰富准确的领域数据进行构建。

在分析军事装备领域数据时可以发现,由于保密性等原因,现有结构化数据十分稀少,收集较为困难,而非结构化数据中知识权威性较低,且抽取准确率有限,难以达到可用性标准。因此,可信度较高的半结构化百科知识将作为构建军事装备知识图谱的重要数据源,发挥关键作用。并且,为满足垂直知识图谱的知识准确性及正确层次结构的要求,本文将采用自顶向下和自底向上相结合的方式来构建军事装备知识图谱,其构建流程如图 3 所示。

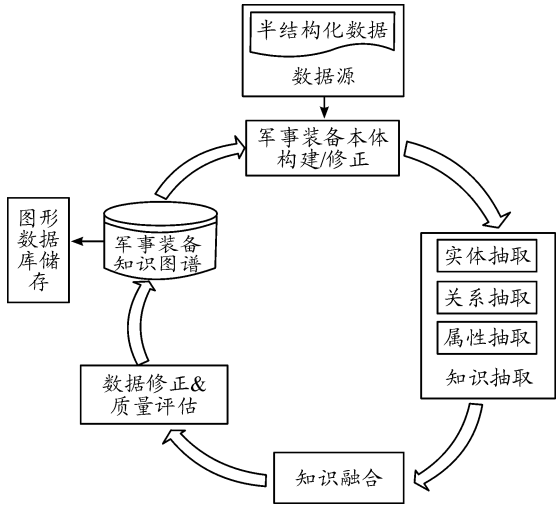

图 3 军事装备知识图谱构建流程

该流程首先通过定义知识图谱的本体模式来保证知识层次结构的正确性,然后再针对定义好的本体模式抽取领域数据中的相关知识,再经过知识融合、质量评估等阶段筛选出满足准确率要求的三元组进行军事装备知识图谱的构建,并用图数据库进行储存。整个流程是一个循环更新的过程,需要根据新的数据来更新或修正已构建好的知识图谱。

由于军事装备知识图谱的特殊性,应首先构建本体。本体是对领域内不同实体概念、属性及其相互关系进行抽象和约束而得到的一种规范,是知识图谱中最核心的部分,主要表现为树形结构[13]。构建本体的主要目的是为知识图谱定义明确的数据模式,正确的数据模式可以保证知识图谱中的实体具有较强的体系关系,是构建知识图谱的基础。由于军事装备知识图谱主要为领域相关人员提供服务,对本体准确率的要求较高,因此本文在综合考虑军事装备领域专业知识及数据源的基础上,通过对军事装备分类及其之间关系进行剖析,确定军事装备核心概念,得到具有树形结构的数据模式图,总共分为 8 大类、100 小类,由于数量较多,图 4 中有所省略。每个装备实体还包含各自的属性,如国家、射程、战斗全重、最大速度等。

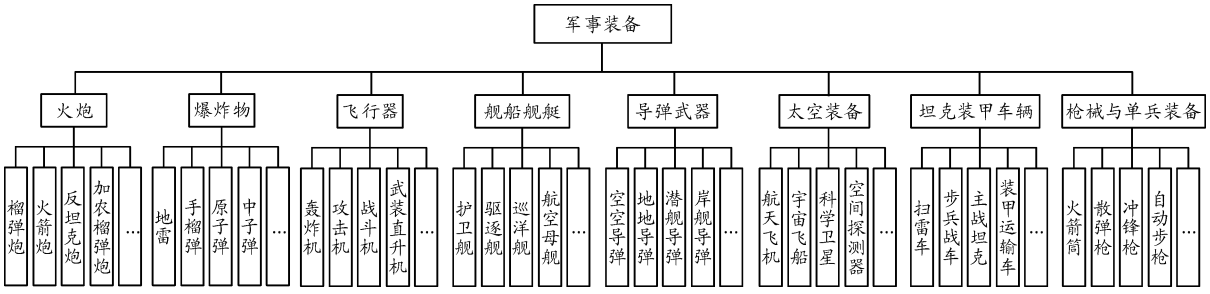

图 4　军事装备本体结构示意图

2.2　百科数据抽取

现实世界中数据是零散的,如何获取数据是构建知识图谱所面临的又一大难题。在信息时代,互联网中拥有着世界上最为庞大的数据,互联网中的网页可以为构建知识图谱提供海量的数据。百科网站是这些海量数据中的一种高质量半结构化知识,其中的词条大部分由人工进行编辑,包含了文章标题、分类标签、信息框等大量规范化知识,可信度较高,利用这些数据构建知识图谱非常具有价值。

本文主要采用网络爬虫技术从百科网页中获取知识,并在爬虫过程中使用 Partial PageRank 策略来优先爬取重要程度较高的网页,其主要流程如图 5 所示。

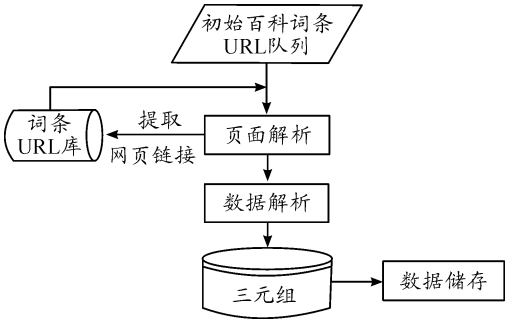

图 5　数据爬取流程

该流程首先将“军事装备”的百科词条链接作为初始的 URL,然后对该词条 URL 进行页面解析,提取出页面中所包含的其他词条链接添加到词条 URL 库中用于爬取军事装备相关百科词条上的数据。而数据解析是对百科网页中所包含的知识进行定位,进行所需知识的爬取,最后将爬取到的实体、属性及相互关系等知识以三元组的形式进行储存。

本文在该爬虫框架的基础上辅助以人工筛选的方式总共获取了 5858 条军事装备词条页面,每条军事装备词条页面如图 6 所示。每个词条对应一个实体,词条标题即为实体名称。每个实体拥有分类标签,列举其所属类别,可从中挖掘实体间存在的关系,以 < 实体 1,关系,实体 2 > 形式的三元组进行储存。除此之外,每个词条中的特定位置上还有描述其对应实体属性的信息框,通常以表格形式出现,可从中抽取出实体的属性信息,以 < 实体,属性名称,属性值 > 形式的三元组进行储存。本文通过数据解析模块对这些关键知识进行抽取而构建三元组,并通过分类关系将抽取到的各个实体对应到构建好的军事装备本体中,以实现知识图谱的构建。

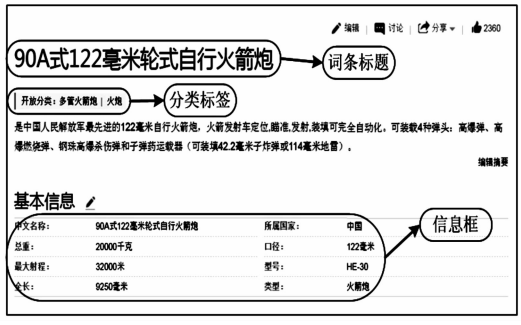

图 6　军事装备百科词条页面

百科知识抽取将百科页面中的半结构化数据转换成了三元组形式的关键知识,但这些关键知识中仍存在部分冗余的信息,主要存在具有不同实体名称但实际对应现实世界中同一实体的现象,需要对这些冗余的知识进行知识融合,保证所构建知识图谱的准确性。由于本文所构建的军事装备知识图谱对准确性的要求较高,因此本文将采用聚类算法对这些冗余的知识进行筛选,然后再使用人工进行评估。该方法首先使用词袋模型[14]对实体名称进行向量化表示,并计算各实体名称的相似度,然后设定阈值筛选出相似度较高的实体名称,最后再通过人工决定舍取,实现实体对齐及属性值的整合。由于百科知识中只是部分存在这种冗余的信息,使用该方法并不会造成较高的人工成本。

2.3　知识图谱的储存更新与维护

在经过复杂的信息处理之后,零散的知识已被整理成有序的结构良好的知识,接下来则需要对这些知识进行储存。军事装备知识图谱中的节点及关系数量众多,且随着知识图谱的更新与维护,其数量不断增长,节点及关系也需要不断地修正以保证知识的准确性,给知识图谱的储存带来一系列困难,因此知识图谱的储存方式对进一步管理及应用知识具有较大影响。目前,图形数据库由于其包含的节点、关系及属性三种元素可以与知识图谱中的关键知识很好的映射,在储存知识图谱方面获得了很好的效果,成为了主流方式。Neo4j 作为图形数据库的典型代表,可轻松稳定地储存及管

理上亿的节点及关系，具有较高的可靠性及扩展性，并支持完整的 ACID 事务，可高效地完成对知识图谱的储存、更新、管理与检索。

因此，本文将采用 Neo4j 图形数据库对军事装备知识图谱进行储存，并对其进行可视化，使军事装备领域知识具有更强的可读性，方便使用者快速获取及理解军事装备领域知识及知识结构，图 7 是军事装备知识图谱的部分实例。图 7 中不同颜色的节点代表不同类型的实体，节点之间的边描述了实体间具有的关系。

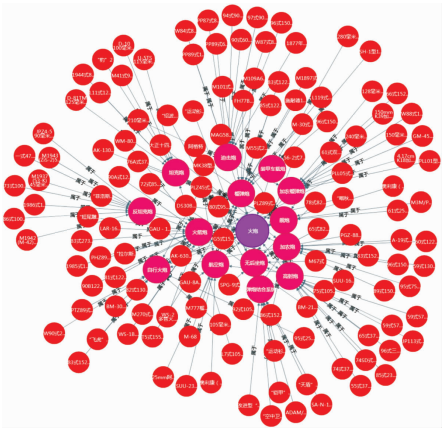

图 7 军事装备知识图谱实例

构建知识图谱并不是一蹴而就的，需要不断地根据新的知识来更新与丰富知识图谱，以保证其所包含领域知识的时效性。本文将采用增量更新的方式对军事装备知识图谱进行更新，即首先利用爬虫程序持续获取百科页面中的数据并以三元组的形式进行储存，选取其中置信度较高且经过去重操作的三元组作为新增知识，再依据这些新增知识对已构建的军事装备知识图谱进行更新或修正，并借助部分人工进行干预。

知识图谱的更新主要包括本体更新和数据更新两方面。本体更新即为数据模式的更新，主要是通过对新增知识进行抽象以发现新的实体类型，实现对本体元素的增加或修正。本体更新所涉及的元素相对较少，但都会直接对知识图谱中的实体及相互关系造成影响，因此需要较强的人工干预。数据更新则相对简单，对整个知识图谱所造成的影响也较小，其主要目标是继续发现新的实体、关系或属性，不断丰富和扩展军事装备知识图谱。

3 基于军事装备知识图谱的知识问答

基于百科知识构建的军事装备知识图谱实现了对网络中零散信息的良好组织，将其整合为了更加符合人类认知习惯的结构化知识，是一种高效管理与利用知识的方式。然而为了更好地利用和展示知识，发挥军事装备知识图谱的潜在使用价值，本文基于军事装备知识图谱实现了的知识问答，辅助相关使用人员快速准确地获取所需知识。

知识问答是知识图谱最主要的应用之一，其目的是回答使用者通过自然语言所提出的问题[15]。由于知识图谱中数据的精度、关联度及结构化程度较高，基于知识图谱的知识问答具有较好的性能。本文在已构建的军事装备知识图谱的基础上，基于模板匹配的方式[16]实现了知识问答，其主要流程如图 8 所示。该流程的输入为使用者所提出的自然语言问题，在经过分词，模板匹配和查询三个主要步骤之后，返回用户的查询结果。

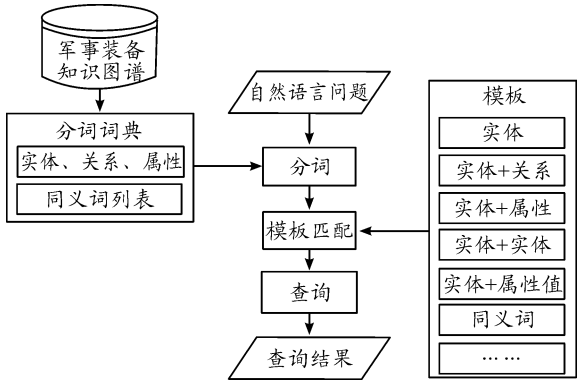

图 8 知识问答流程

分词时所采用的方法为借助分词词典的正向匹配法，分词词典则基于军事装备知识图谱进行构建，其包含了知识图谱中的实体名、关系名、属性名及属性值，并加入了上述名词的同义词列表。分词的目的是划分出输入问题中所包含的知识图谱中的关键词，并确定这些关键词在知识图谱中的类型，加入同义词列表是为了保证在使用者输入较随意时的分词准确性与鲁棒性。然后将分词后得到的结果与预先定义的模板进行匹配，找到匹配率最高的模式。例如当匹配到“实体 + 属性”的模式时，用户最有可能查询的为该实体所包含属性的属性值。最后再依据匹配到的模式将自然语言问题转换为 Neo4j 图形数据库的 Cypher 查询语言，在军事装备知识图谱中完成知识的查询，并将可视化查询结果返回给使用者。

4 结论

本文在充分阐述知识图谱的发展、定义及其一般构建流程的基础上，分析了军事装备知识图谱构建的特殊性，提出了针对军事装备领域的知识图谱构建方法。该方法预先构建了军事装备本体，通过网络爬虫技术不断获取百科网页中高质量的百科知识，再通过知识融合去除冗余的知识，将最后得到的关键知识以三元组的形式储存在 Neo4j 图形数据库中形成网状结构的知识体系，并对构建完成的军事装备知识图谱持续更新。另外，本文利用已构建的军事装备知识图谱实现了军事装备领域的知识问答，辅助装备工作者高效获取知识。

下一步工作将对非结构数据中的知识抽取技术进行研

究,提高抽取准确率,并利用抽取得到的知识逐步扩增本体的层级结构。同时还需进一步开发知识问答过程中灵活多样的人机交互界面。

参考文献:

[1] 王家其,贾红丽,尹承督,等. 基于大数据的部队装备信息管理应用[J]. 兵器装备工程学报,2017(11):99-102.

[2] 漆桂林,高桓,吴天星. 知识图谱研究进展[J]. 情报工程,2017,3(1):4-25.

[3] PUJARA J,MIAO H,GETOOR L,et al. Knowledge graph identification[C]//International Semantic Web Conference. Springer-Verlag New York,Inc. 2013:542-557.

[4] BIZER C,LEHMANN J,KOBILAROV G,et al. DBpedia-a crystallization point for the Web of data[J]. Web Semantics Science Services & Agents on the World Wide Web,2009,7(3):154-165.

[5] SUCHANEK F M,KASNECI G,WEIKUM G. YAGO:a large ontology from wikipedia and wordnet[J]. Web Semantics Science Services & Agents on the World Wide Web,2007,6(3):203-217.

[6] WU W T,LI H S,WANG H X. Probase:a probabilistic taxonomy for text understanding[C]//Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data,ACM New York,USA,2012:481-492.

[7] NIU X,SUN X R,WANG H F,et al. Zhishi. me-weaving Chinese linking open data[C]//International Conference on the Semantic Web. Springer-Verlag,2011:205-220.

[8] 阮彤,孙程琳,王昊奋,等. 中医药知识图谱构建与应用[J]. 医学信息学杂志,2016,37(4):8-13.

[9] 赵明,杜亚茹,杜会芳,等. 植物领域知识图谱构建中本体非分类关系提取方法[J]. 农业机械学报,2016,47(9):278-284.

[10] 徐增林,盛泳潘,贺丽荣,等. 知识图谱技术综述[J]. 电子科技大学学报,2016,45(4):589-606.

[11] DONG X,GABRILOVICH E,HEITZ G,et al. Knowledge vault:a web-scale approach to probabilistic knowledge fusion[C]//Proc of the 20th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. New York:ACM,2014:601-610.

[12] 胡芳槐. 基于多种数据源的中文知识图谱构建方法研究[D]. 上海:华东理工大学,2015.

[13] 刘峤,李杨,段宏,等. 知识图谱构建技术综述[J]. 计算机研究与发展,2016,53(3):582-600.

[14] MIKOLOV T,CORRADO G,CHEN K,et al. Efficient estimation of word representations in vector space[C]//International Conference on Learning Representations. 2013:1-12.

[15] 毛先领,李晓明. 问答系统研究综述[J]. 计算机科学与探索,2012,6(3):193-207.

[16] ZHENG W,ZOU L,LIAN X,et al. How to build templates for RDF question/answering:an uncertain graph similarity join approach[C]//ACM SIGMOD International Conference on Management of Data. ACM,2015:1809-1824.

(责任编辑 杨继森)