文章编号: 1006-2475(2019) 09-0035-06

水利信息知识图谱的构建与应用

冯 钧,徐 新,陆佳民

(河海大学计算机与信息学院,江苏 南京 211100)

摘要: 近年来知识图谱技术作为一种用于描述客观世界中概念、实例及其关系的新方法,得到了人们的广泛关注,利用知识图谱可以有效拓展搜索结果的广度。目前水利行业采用的基于关键字的搜索技术难以利用对象间关系进行信息检索。为此,本文首先提出一种面向水利对象数据的知识图谱构建方法,用以实现水利信息知识图谱的构建。然后,提出基于推理规则的知识推理方法,利用隐藏在水利信息知识图谱中的知识实现智能数据检索。最后,将上述技术应用于水利领域,实现水利信息知识图谱构建与检索系统。通过该系统可以有效利用水利对象之间的关系,充分发挥水利信息资源的价值。

关键词: 水利信息资源; 知识图谱; 知识推理

中图分类号: TP391. 3 文献标识码: A doi: 10. 3969 / j. issn. 1006-2475. 2019. 09. 007

Construction and Application of Water Conservancy Information Knowledge Graph

FENG Jun, XU Xin, LU Jia-min

(College of Computer and Information, Hohai University, Nanjing 211100, China)

Abstract: In recent years, knowledge graph technology has been widely used as a new method for describing concepts, entities and their relationships in the objective world. The use of knowledge graph can effectively expand the breadth of search results. At present, the keyword-based search technology adopted by the water conservancy industry is difficult to use the relationship between objects for information retrieval. To this end, this paper first proposes a knowledge graph construction method for water conservancy object data, which is used to realize the construction of water conservancy information knowledge graph. Then, a knowledge reasoning method based on inference rules is proposed, and the knowledge hidden in the knowledge graph of water conservancy information is used to realize intelligent data retrieval. Finally, the above technology is applied in the field of water conservancy, and the water conservancy information knowledge graph construction and retrieval system is realized. Through this system, the relationship between water conservancy objects can be effectively utilized, and the value of water conservancy information resources can be fully utilized.

Key words: water conservancy information resources; knowledge graph; knowledge inference

0 引 言

水利信息资源是指水利部门或为水利部门采集、加工、处理的信息资源[1]。随着我国水利事业的不断发展,水利部门积累了大量的水利信息资源。但这些信息资源存在管理分散、服务目标单一、利用效率低下的问题,制约了水利信息化发展[2]。为此,水利部门通过对象建模将水利信息资源转换为水利对象数据,用以支持数据共享。这些水利对象数据对象种类众多,对象属性多样,对象间关系丰富。目前使用的基于关键字的搜索技术难以利用对象间关联关系(例如三峡枢纽水库与长江、太湖与苕溪之间的关联关系)进行信息检索与推荐。如何表达和利用水利对象及其关系,为用户提供全面准确的信息,成为亟需解决的问题。

2012 年,Google 首次提出了知识图谱的概念[3],并成功应用于语义搜索上,提高了检索结果的质量。知识图谱本质上是一种叫做语义网络的知识库,即具

收稿日期: 2019-03-15; 修回日期: 2019-03-27

基金项目: 国家重点研发计划项目(2018YFC0407901,2017YFC0405806) ; 国家自然科学基金资助面上项目(61602151)

作者简介: 冯钧(1969-) ,女,江苏常州人,教授,博士生导师,博士,研究方向: 时空数据管理,智能数据处理与数据挖掘,水利信息化,E-mail: fengjun@ hhu. edu. cn; 徐新(1995-) ,男,江苏盐城人,硕士研究生,研究方向: 知识图谱,自然语言处理; 陆佳民(1983-) ,男,江苏南通人,讲师,博士,研究方向: 分布式数据处理,知识图谱,水利信息化。

有有向图结构的一个知识库，其中图的节点代表实例或者概念，而图的边代表实例/概念之间的各种语义关系[4]。知识图谱的出现，为人们提供了一种更好地组织、管理和理解海量信息的方法，同时，也成为知识检索、智能问答[5]、个性化推荐[6]等应用的基础。

将知识图谱技术应用于水利领域，使用水利对象数据构建的水利信息知识图谱，可以利用水利对象间的关联关系，扩展检索结果的广度。本文针对水利对象数据，提出一种水利信息知识图谱的构建方法；并在此基础上实现了基于推理规则的知识推理，用以进一步挖掘隐藏在水利信息知识图谱中的知识；最后将上述技术应用于水利信息知识图谱构建与检索系统，实现了水利信息的智能检索与推荐。

1 相关研究

1.1 通用知识图谱

通用知识图谱面向全领域，包含了大量的常识性知识，强调知识图谱的广度。目前国外在通用知识图谱领域已经有了许多成果，例如利用从维基百科中抽取的知识构建的 DBpedia[7]；整合维基百科、WordNet 和 GeoNames 的知识所形成的 Yago[8]；利用从互联网中挖掘出的知识构建的 NELL[9] 以及 Concept Graph[10]、Freebase[11]、Wikidata[12] 等。近年来，国内也出现了许多通用知识图谱，例如复旦大学发布的融合了百科数据及部分领域知识的 CN_DBPedia[13]；上海交通大学发布的包含中文维基百科、互动百科、百度百科三大百科数据的 zhishi. me[14]；基于中英文百科的 XLore[15] 以及百度的知心、搜狗的狗立方等。这些覆盖面广泛的通用知识图谱，可以为普通用户在智能问答、个性化推荐等方面提供更好的服务。

1.2 垂直领域知识图谱

垂直领域知识图谱通常面向某一具体领域，更注重知识的深度和完备性，知识的粒度更细。例如通过从 EMR 信息中提取医疗事实，构建的乳腺肿瘤知识图谱[16]；利用文物信息对构建的文物本体进行实例化操作，形成的文物知识图谱[17]；从不同的软件资源中提取信息构建软件知识实例，形成的软件知识图谱[18]；从碳交易领域数据中抽取三元组，再将其转换为关联数据，构建的碳交易领域知识图谱[19] 以及地理领域的 GeoNames[20]、影视领域的 IMDB[21]、音乐领域的 MusicBrainz[22] 等。这些面向某一特定领域的知识图谱有助于充分发挥领域数据的价值，同时为后续的智能化应用研究奠定基础。尽管目前出现了一些领域知识图谱，但在水利领域知识图谱技术还没有得到广泛关注，同时，这些领域知识图谱多停留在构建方面，对于知识图谱的应用还缺少考虑。

2 水利信息知识图谱构建

水利对象数据中蕴含着大量水利信息知识，是构建水利信息知识图谱的重要数据来源。本文主要利用水利对象数据构建水利信息知识图谱，下面首先给出水利信息知识图谱相关概念的定义，再给出利用水利对象数据构建水利信息知识图谱的流程。

2.1 水利信息知识图谱的相关概念

定义 1 （概念层 G_C）概念层是知识图谱的核心，描述了知识图谱的数据模式，规范了实例层中的事实。在水利信息知识图谱中，概念层 $G_C=(C, P_C, N_C, R_C)$，其中 C 表示图谱中的概念节点，如水库概念、河流概念；P_C 表示概念节点的属性边，如水库概念的属性边有工程等别、水库类型；N_C 表示属性值类型节点，如工程等别的属性值类型节点为整型、水库类型的属性值类型节点为字符型；R_C 表示概念节点和概念节点之间的关系，如水库概念与河流概念之间存在流入关系。

定义 2 （实例层 G_E）实例层由一系列事实组成。在水利信息知识图谱中，实例层 $G_E=(E, P_E, N_E, R_E)$，其中 E 表示图谱中的实例节点，如三峡枢纽水库、长江；P_E 表示实例节点的属性边，如工程等别、水库类型；N_E 表示属性值节点，如工程等别的属性值节点为 1、水库类型的属性值节点为山丘水库；R_E 表示实例节点与实例节点之间的关系，如三峡枢纽水库与长江之间存在流入关系。

定义 3 （水利信息知识图谱 G）水利信息知识图谱 $G=(G_C, G_E, R)$，其中 G_C 表示水利信息知识图谱的概念层；G_E 表示水利信息知识图谱的实例层；R 表示 G_C 中的概念节点与 G_E 中的实例节点之间的关系，如水库概念与三峡枢纽水库之间存在实例概念间关系。同时每一个实例节点 E 只与一个概念节点 C 存在关系 R。

若 2 个实例节点 E_1 和 E_2 之间存在实例间关系 R_E，且 E_1、E_2 分别与概念节点 C_1、C_2 存在实例概念间关系 R_1、R_2，则 C_1、C_2 之间也必然存在 R_C，且关系 R_C 与关系 R_E 有相同的名称。如长江与三峡枢纽水库之间存在流入关系，三峡枢纽水库与水库概念、长江与河流概念之间存在实例概念间关系，则水库概念与河流概念之间也存在流入关系。图 1 为部分水利信息知识图谱的结构。

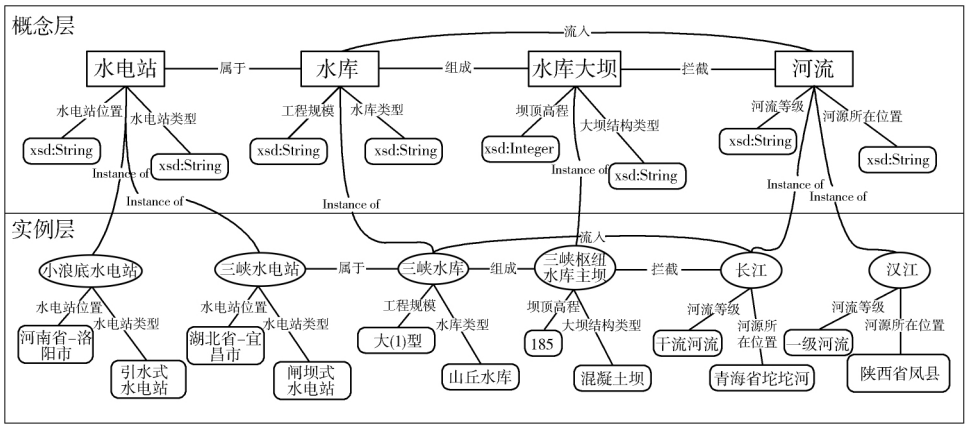

图 1　水利信息知识图谱结构

现有的水利对象数据存储在关系数据库中,主要由对象名录表、对象基础信息表和对象关系表这 3 类表组成,其中对象名录表保存了水利对象的对象名称、对象代码等信息; 对象基础信息表保存了不同水利对象的特征信息; 对象关系表记录了 2 个水利对象之间的关系。对象基础信息表与对象名录表相关联,对象名录表与对象关系表相关联。图 2 为部分水利对象数据。

水库　　水库基础信息表

ID	水库代码	水库名称	工程等别	水库类型	…
21026	42050600030	三峡枢纽水库	1	山丘水库	…
22002	330182000385	新安江水库	1	山丘水库	…

水库名录表

ID	水库代码	水库名称	对象建立时间	…
2010	420506000030	三峡枢纽水库	2012/12/31	…
2011	330182000385	新安江水库	2012/12/31	…

水库与河流关系表

ID	水库代码	河流代码	…
1101	42050600030	F00 0000 0000S	…
1132	42038100006	FF00 0000 000L	…

河流　　河流名录表

ID	河流代码	河流名称	对象建立时间	…
1000	F00 0000 0000S	长江	2012/11/31	…
1001	FF00 0000 000L	汉江	2012/11/31	…

河流基础信息表

ID	河流代码	河流名称	河源所在位置	河流级别	…
10030	F00 0000 0000S	长江	青海省-格尔木市	干流河流	…
11006	FF00 0000 000L	汉江	陕西省-凤县	一级河流	…

图 2　水利对象数据数据库片段

2.2　水利信息知识图谱的构建方法

本文的大多知识图谱都采用自顶向下的构建方法,依次构建知识图谱的概念层与实例层。下面将给出水利信息知识图谱概念层与实例层的具体构建方法。

2.2.1　概念层构建

1) 构建概念节点 C。根据在编的《水利对象分类与编码总则》,结合水利对象数据,在领域专家的帮助下,确定概念节点 C,例如水库概念节点。每一个概念节点对应一张对象名录表和一张对象基础信息表,如水库概念节点对应水库对象名录表和水库基础信息表。

2) 构建属性边 P_C 和属性值类型节点 N_C。抽取概念节点对应的对象基础信息表中的字段及字段类型作为概念节点 C 的属性边 P_C 和属性值类型节点 N_C。如抽取水库基础信息表中的工程等别字段作为水库概念的属性边,其对应的属性值类型节点则为工程等别字段的类型,即整型。

3) 构建概念间关系 R_C。构建概念关系二分图,通过二分图映射确定概念节点 C 之间的关系 R_C。其中概念关系二分图的定义如下:

定义 4　(概念关系二分图 G_B) $G_B = (V_C, V_T, E_B)$,其中 V_C 代表概念节点,V_T 代表关系表节点,E_B 代表概念节点 V_C 与关系表节点 V_T 之间的边。

首先构建概念关系二分图,利用从数据库中抽取的关系模式,得到不同对象名录表和对象关系表之间的关系。再将对象名录表和对象关系表映射为图中的节点,将概念节点也视为图中的节点,用概念节点替换相应的对象名录表节点,得到概念关系二分图。

再利用概念关系二分图映射算法,将二分图映射到只有概念节点集合 V_C 的图中,即可得到概念节点间关系 R_C,从而完成水利信息知识图谱概念层的构建。

算法 1　概念关系二分图映射算法。

输入: $G_B = (V_C, V_T, E_B)$

输出: 水利信息知识图谱概念层

while(V_T != null)

从 V_T 中取出一个关系表节点 A;
if(A 与 2 个概念节点相连)
获得与 A 相连的 2 个概念节点 B 和 C;
在 B 和 C 之间添加关系 R_C;
删除 A 与 B 和 C 之间的边;
else
获得与 A 相连概念节点 B;
添加从 B 指向 B 自身的关系 R_C;
删除 A 与 B 之间的边;
end if
指定关系触发词描述 R_C;
end while

图 3 为关系概念二分图的映射示意图,圆形节点为部分概念节点,矩形节点为部分对象关系表节点,将二分图映射到只含有概念节点集合的图中,即可得到概念节点之间的关系。

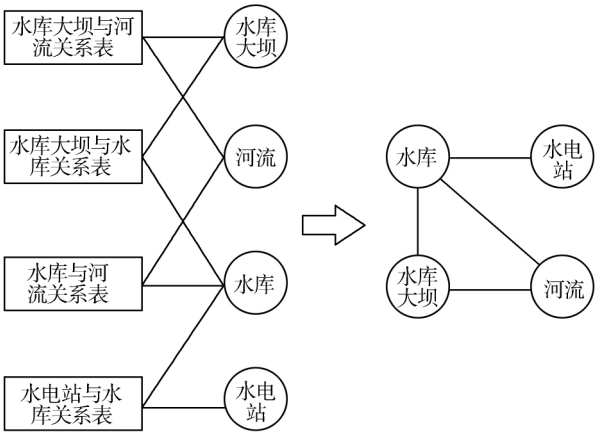

图 3 二分图映射示意图

2.2.2 实例层构建

1) 构建实例节点 E 和概念与实例间关系 R。

根据 G_C 中概念节点 C 对应的对象名录表,抽取水利对象名称,构建相应的实例节点 E,如三峡枢纽水库、长江。同时在 G_C 中的概念节点 C 与 G_E 中相应的实例节点 E 之间添加实例概念间关系 R。如在三峡枢纽水库节点与水库概念节点之间添加关系 R。

2) 构建属性边 P_E 和属性值节点 N_E。

根据 G_C 中 C 的属性边 P_C 和属性值类型节点 N_C,从对象基础信息表中抽取与 P_C 相同的字段作为实例节点 E 的属性边 P_E,抽取对应的字段值作为其属性边 P_E 的属性值节点 N_E。例如,对于三峡枢纽水库,根据先前构建的水库概念节点的属性边,从水库基础信息表中抽取相同名称的字段,如水库类型、工程等别等作为三峡枢纽水库节点的属性边;抽取对应的字段值,如山丘水库、1 等,作为属性边的属性值节点。

3) 构建实例间关系 R_E。

根据对象关系表构建实例间关系 R_E,并根据概念间关系 R_C 的名称,确定实例间关系 R_E 名称,完成水利信息知识图谱的构建。如根据水库与河流关系表,三峡枢纽水库与长江存在关系,则在这 2 个实例之间添加关系。同时,由于三峡枢纽水库、长江分别与水库概念、河流概念存在实例概念间关系,而水库概念和河流概念之间存在流入关系,则在这 2 个实例间也存在流入关系。

3 基于水利信息知识图谱的知识推理

基于知识图谱的知识推理旨在基于已有的知识图谱,推理得到新的事实[23]。本文根据推理规则进行知识推理,挖掘隐藏在水利信息知识图谱中的水利知识。首先在已有的水利信息知识图谱的基础上,结合水利领域知识,定义推理规则。表 1 为部分推理规则及其解释。例如,在图谱的概念层中,河流概念与水库概念之间存在流入关系、水库概念和水电站概念之间存在属于关系,而在河流概念与水电站概念之间并无关系。但结合领域知识可知水电站与河流之间存在位于关系,因此,可以定义推理规则 1,通过水库得到水电站所在的河流。

表 1 水利信息知识图谱部分推理规则及解释

编号	推理规则	解释
1	(河流,流入,水库),(水电站,属于,水库)→(水电站,位于,河流)	水电站位于流入该水电站所属水库的河流上
2	(泵站,拥有,取水口),(泵站口,位于,湖泊),(湖泊,属于,流域分区)→(取水口,属于,流域分区)	取水口属于拥有该取水口的泵站所在湖泊的流域分区
3	(桥梁,位于,河段),(河段,属于,河流)→(桥梁,横跨,河流)	桥梁横跨该桥梁所在河段所属的河流

再通过推理规则的实例化,将抽象的概念替换为具体的实例,通过推理即可得到隐藏在水利信息知识图谱中的知识。例如,在应用推理规则 1 获得三峡水电站所在河流时,先将推理规则中的水电站概念替换为三峡水电站,利用水利信息知识图谱对规则体中的事实进行匹配,得到具体的水库实例,即三峡枢纽水库。再将水库概念替换为具体的水库实例,重复上述步骤,即可得到具体的河流实例,也即三峡水电站所在的河流。利用如表 1 所示的推理规则,通过知识推理可以实现对水利信息知识图谱的进一步挖掘。

4 系统实现与展示

4.1 系统总体架构

本文利用水利对象数据,通过搭建水利信息知识图谱构建与检索系统,实现水利信息知识图谱的构建和水利信息的智能检索与推荐。水利信息知识图谱

构建与检索系统按照本文提出的知识图谱构建方法构建水利信息知识图谱,并应用了基于推理规则的知识推理方法,实现了图谱检索和推理检索 2 种不同的知识检索方式。该服务系统主要分为 3 层: 数据存储层、业务逻辑层和应用层。其中应用层主要由概念管理模块、知识构建模块、推理规则管理模块、知识检索模块、系统管理模块等组成。业务逻辑层利用水利对象数据,通过 Jena 构建水利信息知识图谱,并使用 Fuseki 实现对水利信息知识图谱的查询。数据存储层使用 Jena TDB 作为知识图谱的持久化工具。其系统架构如图 4 所示。

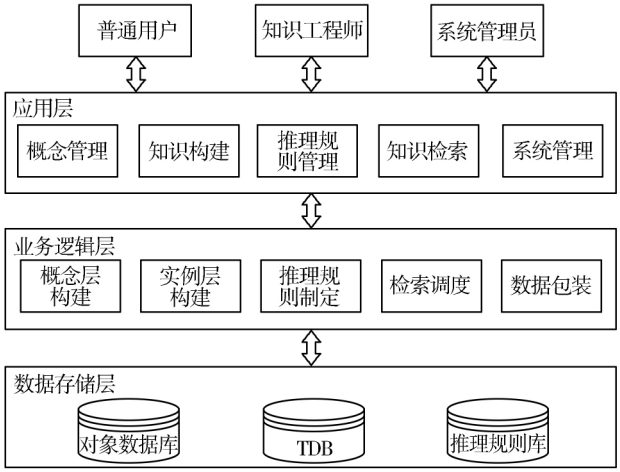

图 4　水利信息知识图谱构建与检索系统架构

4.2　系统主要功能模块

1) 概念管理模块。

该模块利用水利对象数据,按照本文提出的概念层构建方法,构建并维护水利信息知识图谱的概念层。

2) 知识构建模块。

该模块利用水利对象数据,结合概念管理模块构建的水利信息知识图谱的概念层,按照本文提出的实例层构建方法,构建并维护水利信息知识图谱的实例层。

3) 推理规则管理模块。

该模块根据水利信息知识图谱的概念层,结合水利领域知识,对推理规则进行管理,实现推理规则的定义、修改、删除等操作。

4) 知识检索模块。

该模块实现了对水利信息知识图谱的 2 种不同知识检索方式,并实现检索结果的图形可视化。知识检索可分为图谱检索和推理检索 2 种方法。图谱检索可以查询显性存储在水利信息知识图谱中的知识,而推理检索通过知识推理,利用规则库中的推理规则,可以发掘隐式存储在水利信息知识图谱中的知识。

5) 系统管理模块。

该模块主要负责系统用户角色和权限的分配。该系统将用户划分为普通用户、知识工程师、系统管理员。普通用户可以使用 2 种不同的知识检索方式查询水利信息知识; 知识工程师是指具有一定水利领域知识的用户,可以进行水利信息知识图谱的构建,并且可以定义推理规则; 系统管理员负责系统的日常维护以及用户角色和权限的分配。

4.3　系统实现展示

水利信息知识图谱构建与检索系统通过 ECharts 实现对知识检索结果的图形可视化。使用图谱检索方式时,利用对象间的关联关系,实现了信息检索与推荐。例如查询“太湖”时,得到太湖与苕溪、梁溪河等存在直接关系,与汾湖、甘玲水库等存在间接关系。图 5 为使用图谱检索方式查询“太湖”时的图形可视化结果。

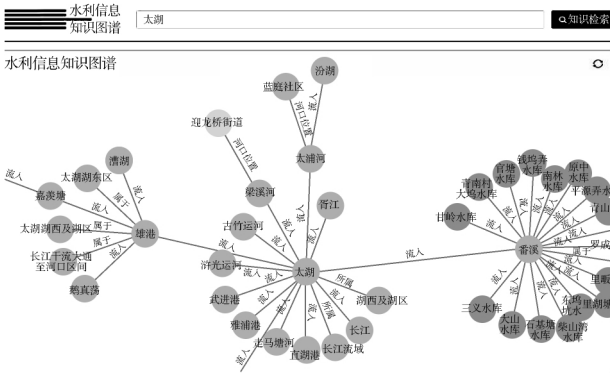

图 5　图谱检索结果

使用推理检索方式时,不仅利用了对象间的关联关系,同时还使用知识推理方法,利用隐藏在水利信息知识图谱中的知识,实现了智能数据检索。例如查询“三峡水电站所在河流”时,不仅可以利用对象间的关联关系得到与三峡水电站相关的实例,同时也可以利用表 1 中的推理规则 1 推理得到三峡水电站所在河流为长江。图 6 为使用推理检索方式查询“三峡水电站所在河流”的结果。

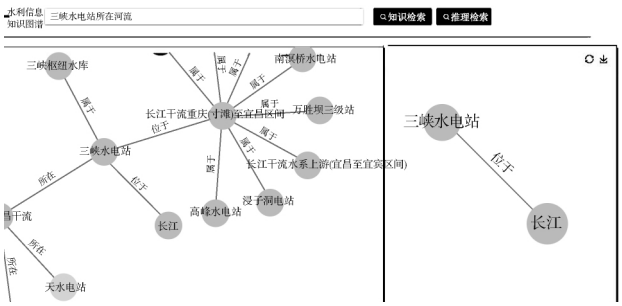

图 6　推理检索结果

图 7 为使用水利领域传统的检索方式检索“三峡水电站所在河流”的结果。从图 6 和图 7 中可以发现,传统的检索方式无法检索出三峡水电站所在河流,而基于知识图谱的检索方式可以有效地进行查

询,从而为用户提供了全面准确的信息。

图 7 传统检索结果

5 结束语

本文以水利对象数据为基础,提出了水利信息知识图谱构建方法;实现了基于推理规则的知识推理方法;并将上述技术应用于水利信息知识图谱构建与检索系统,构建了水利信息知识图谱,实现了智能数据检索与推荐。本文对水利领域知识图谱构建与应用进行了探索,但由于主要考虑使用水利对象数据构建水利信息知识图谱,对于互联网中非结构化文本的利用不充分,在未来的工作中,将进一步挖掘蕴含在互联网文本中的水利知识,添加到水利信息知识图谱中,用以丰富水利信息知识图谱。

参考文献:

[1] 朱跃龙,许峰,冯钧,等. 水利信息资源目录体系构建研究[J]. 水利信息化,2010(2):4-8.

[2] 成建国,冯钧,杨鹏,等. 水利数据资源目录服务关键技术研究[J]. 水利信息化,2014(6):18-21.

[3] AMIT S. Introducing the Knowledge Graph [EB/OL]. (2012-05-01) [2019-03-15]. http://googleblog.blogspot.pt/2012/05/introducing-knowledge-graph-things-not.html.

[4] 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.

[5] HIXON B, CLARK P, HAJISHIRZI H. Learning knowledge graphs for question answering through conversational dialog [C]// Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2015:851-861.

[6] LU C, LAUBLET P, STANKOVIC M. Travel attractions recommendation with knowledge graphs [C]// European Knowledge Acquisition Workshop. Springer, 2016:416-431.

[7] BIZER C, LEHMANN J, KOBILAROV G, et al. DBpedia: A crystallization point for the Web of data [J]. Journal of Web Semantics, 2009,7(3):154-165.

[8] SUCHANEK F M, KASNECI G, WEIKUM G. YAGO: A large ontology from Wikipedia and WordNet [J]. Journal of Web Semantics, 2008,6(3):203-217.

[9] CARLSON A, BETTERIDGE J, KISIEL B, et al. Toward an architecture for never-ending language learning [C]// AAAI the 24th Conference on Artificial Intelligence. 2010:1306-1313.

[10] WANG Z Y, WANG H X, WEN J R, et al. An inference approach to basic level of categorization [C]// Proceedings of the 24th ACM International Conference on Information and Knowledge Management. 2015:653-662.

[11] YUE B, GUI M, GUO J H, et al. An effective framework for question answering over freebase via reconstructing natural sequences [C]// Proceedings of International Conference on World Wide Web Companion. ACM, 2017:865-866.

[12] WMF. Wikidata [EB/OL]. [2019-03-15]. http://www.wikidata.org/wiki/Wikidata:Main_Page.

[13] XU B, XU Y, LIANG J Q, et al. CN-DBpedia: A never-ending Chinese knowledge extraction system [C]// International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems. Springer, 2017:428-438.

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

[15] WANG Z G, LI J Z, WANG Z C, et al. XLore: A large-scale English-Chinese bilingual knowledge graph [C]// International Semantic Web Conference (Posters & Demos). 2013,1035:121-124.

[16] 崔洁,陈德华,乐嘉锦. 基于 EMR 的乳腺肿瘤知识图谱构建研究[J]. 计算机应用与软件,2017(12):128-132.

[17] 林炀平. 文物知识图谱构建与检索关键技术研究与实现[D]. 杭州:浙江大学,2017.

[18] 李文鹏,王建彬,林泽琦,等. 面向开源软件项目的软件知识图谱构建方法[J]. 计算机科学与探索,2017,11(6):851-862.

[19] 王良萸. 面向碳交易领域的知识图谱构建方法[J]. 计算机与现代化,2018(8):114-119.

[20] ROWLINGSON B. Geonames: Interface to www.geonames.org Web Service [EB/OL]. (2014-12-19) [2019-03-15]. http://ftp.iitm.ac.in/cran/web/packages/geonames/2014.

[21] IMDB. Home Page of IMDB [EB/OL]. (2017-06-28) [2019-03-15]. http://www.imdb.com.

[22] MetaBrainz Foundation. Musicbrainz Home Page [EB/OL]. (2016-06-06) [2019-03-15]. http://musicbrainz.org/.

[23] 官赛萍,靳小龙,贾岩涛,等. 面向知识图谱的知识推理研究进展[J]. 软件学报,2018,29(10):74-102.