面向智能搜索应用的水利知识图谱构建

高凤宁[1]，高祥涛[2]，曹 帅[2]，朱向荣[1]，司存友[2]，胡 伟[1]

（1. 南京大学 计算机科学与技术系，江苏 南京 210023；2. 江苏省水文水资源勘测局，江苏 南京 210029）

摘要：在水利知识图谱的基础上，结合字符串相似度以及 word2vec 生成的词嵌入的余弦相似度，设计了关联属性的语义查询算法，实现了异构水利知识的融合。在网页端搭建了一个智能搜索应用，通过用户实验，验证了基于水利知识图谱的智能搜索应用可以降低水利领域从业人员的数据获取难度，增强对水利领域相关知识的理解。

关键词：水利知识图谱；知识融合；词嵌入；语义搜索

中图分类号：TM734 **文献标识码**：B **文章编号**：1007-7839(2021)10-0059-06

Construction of water conservancy knowledge graph for intelligent search application

GAO Fengning[1], GAO Xiangtao[2], CAO Shuai[2], ZHU Xiangrong[1], SI Cunyou[2], HU Wei[1]

(1. *Department of Computer Science and Technology, Nanjing University, Nanjing* 210023, *China*; 2. *Jiangsu Hydrology and Water Resources Survey Bureau, Nanjing* 210029, *China*)

Abstract: On the basis of water conservancy knowledge graph, combined with the string similarity and the cosine similarity of word embedding generated by word2vec, semantic query algorithm of associated attributes was designed to achieve the integration of heterogeneous water conservancy knowledge. After building an intelligent search application on the web, user experiments verified that the intelligent search application based on the water conservancy knowledge graph could reduce the difficulty of data acquisition for practitioners in the water conservancy field and enhance the understanding of relevant knowledge in the water conservancy field.

Key words: water conservancy knowledge graph; knowledge fusion; word embedding; semantic search

随着科学技术的发展，大数据在各个领域都引起人们的高度重视并得到了广泛应用，为人们获得更为深刻、全面的洞察能力提供了前所未有的空间与潜力。伴随大数据时代的到来，大数据成为推动数字经济发展的关键生产要素，成为建设数字中国的关键创新动力，同时也成为重塑国家竞争优势的重大发展机遇。2017 年我国水利部正式印发《关于推进水利大数据发展的指导意见》，该意见是水利部深入贯彻党中央提出的国家大数据战略、国务院《促进大数据发展行动纲要》等系列决策部署的重要举措，旨在水利行业推进数据资源共享开放，促进水利大数据发展与创新应用。

现有的水利信息化工作仍存在标准化和规范化相对滞后、普及程度较低、发展水平较低等问题[1]。知识图谱是一种新型的知识表示方法和数据管理模式。国务院《新一代人工智能发展规划》

收稿日期：2021-05-24

基金项目：江苏省水利科技项目(2019046)

作者简介：高凤宁(1997—)，男，硕士研究生，研究方向为知识融合。E-mail：ngao. nju@gmail. com

通信作者：胡伟(1982—)，男，副教授，博士，主要从事知识图谱研究。E-mail：whu@nju. edu. cn

中明确将知识图谱列为新一代人工智能关键共性技术。知识图谱将领域中的异构知识结构化,构建起知识间的关联,结合大数据与深度学习,已成为推动互联网和人工智能发展的核心驱动力之一,对于水利信息的组织管理和智能应用也具有重要价值。

为了更好地发挥知识图谱在信息组织与管理方面的作用[2],加强对水利资源的整合与利用,本文采用关系数据库转 RDF(Databases to RDF,D2R)技术构建了面向水利领域的知识图谱,并设计实现了基于字符串相似度结合词嵌入(word embedding)的余弦相似度的属性相似度计算算法,实现了水利知识图谱的融合,并在此基础上搭建了基于水利知识图谱的网页端智能搜索应用。

1 水利知识图谱的构建

伴随着知识图谱的不断演变与发展,面向特定领域的知识图谱在现阶段获得了广泛应用。针对结构化的关系型数据库,采用关系数据库转 RDF(Databases to RDF,D2R)技术构建了面向水利领域的知识图谱,并进行可视化展示。

1.1 知识图谱简介

知识图谱是知识工程在大数据环境中的成功应用,知识工程作为人工智能领域的一个重要分支,经历了很长时间的演变和发展历史。

万维网之父 Tim Berners - Lee 于 1998 年提出语义网(Semantic Web)的概念,于 2001 年正式发表相关论文[3],由此揭开了世界范围内语义网研究的序幕。从 2006 年开始,大规模结构知识资源的出现和网络规模信息提取方法的进步,使得大规模知识获取实现了自动化,并且在网络规模下运行。大规模的知识图谱不断涌现,并逐渐在大型行业和领域中正得到广泛的运用。例如 2012 年由谷歌推出的知识图谱、Facebook 图谱搜索,以及微软 Satori 等,已成为驱动语义搜索、机器问答、智能推荐的强大动力引擎。

现阶段知识图谱的发展和应用,除了通用的大规模知识图谱,例如 DBpedia[4]、YAGO[5]和 Wikidata[6]等,各行业领域如商业、金融、生命科学等也在建立领域相关的知识图谱,并且广泛应用,在智能客服、商业智能等真实场景体现出巨大的应用价值,而更多知识图谱的创新应用仍有待开发。

1.2 水利知识图谱的构建

知识图谱,可以理解为一张由知识点相互连接而成的语义网络,具有很强的描述能力,可以用来更好的查询复杂关联信息,从语义层面理解用户意图,改进搜索质量。知识图谱以三元组 <h, r, t> 的三元组形式存储实体的属性和关系,h 代表头实体,r 代表关系,t 代表尾实体。

本文构建的水利知识图谱,数据来源于江苏省水利云平台,目前采用关系型数据库存储管理。如表 1 所示,字段代表不同属性,后面六列为属性的描述,包括一级分类、二级分类、三级分类、单位、数据库(实例名)和表中文名(表英文名)。不难看出,数据的结构化程度很高,是典型的关系型数据库。

要将关系型数据库数据转为 RDF 模型[7],相关工作考虑将关系数据库模式与本体进行映射[8],本文采用更轻量级的 D2R(Database - to - RDF)技术[9],将关系型数据库发布为知识图谱。D2R 主要包括 D2R 服务器,D2R 查询引擎以及 D2R 查询映射语言。D2R 的主要框架如图 1 所示。

D2R 查询映射的主要功能是定义将关系型数据转换成 RDF 模型的映射规则。首先利用 D2R 提供的查询映射语言,根据表格型的水利数据生成预定义的映射文件。然后针对数据特点,对映射文件进行修改,将数据映射到本体上去。主要有以下 2 种映射规则:数据库的层级信息描述作为本体中不同的类;数据库中的字段作为属性。

得到本体之后,可以对本体中的数据进行查询。D2R 服务器提供了对 RDF 数据进行查询访问的接口,以供上层的 RDF 浏览器、SPARQL 查询[10]客户端以及传统的 HTML 浏览器等调用。

通过构建 SPARQL 查询语句,D2R 查询引擎将 RDF 数据的查询语言 SPARQL 转换为关系型数据库数据的查询语言 SQL,并将 SQL 查询结果转换为 RDF 三元组或者 SPARQL 查询结果,以此实现整个查询的流程。

本文将建成的水利知识图谱进行了处理,组织成了图的表示形式,并可视化展示。首先对关系数据中的字段名和描述进行了预处理,删除了不相关的序列号以及日期,同时对双引号、空格、制表符等字符进行了处理。然后将关系数据中的各个属性字段与其对应的属性值,以(key, value)键值对的形式存储,再转化成有向图的表示形式。有向图含有 17 329 个节点与 34 713 条边,包含了水利数据中的关联路径信息。为了便于直观展示,图 2 中仅展示了前 5 级关联路径所构成的有向图。

表 1　来源数据样例

字段	一级分类	二级分类	三级分类	单位	数据库(实例名)	表中文名(表英文名)
降水量(P)	专题数据	水文	水文	水文局	基础水文数据库(NEWDB)	日降水量表(HY_DP_C)
险情代码(DNGCD)	专题数据	防汛防旱	防汛防旱	防办	实时工情(SSGQNEW)	抢险动态信息表(REI_FLFTTR)
县(TOP_XIAN)	基础数据	水利普查	水利工程	水文局	水利普查数据库(SLPCSYS)	堤防工程(P206_0)
村(TOP_CUN)	基础数据	水利普查	灌区	水文局	水利普查数据库(SLPCSYS)	灌溉面积(P701_0)

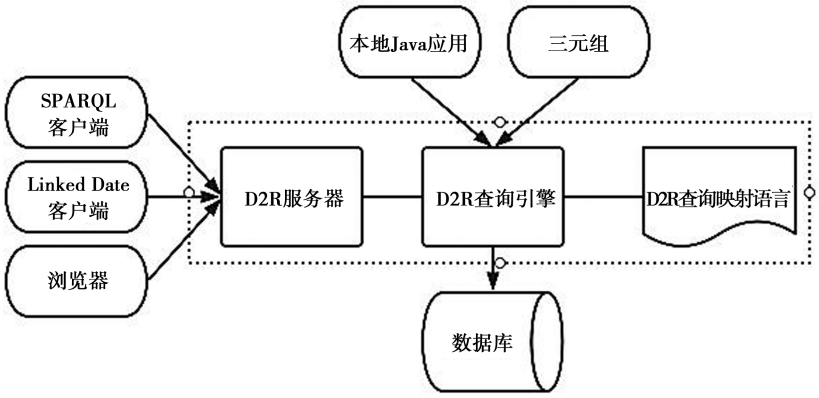

图 1　D2R 主要框架

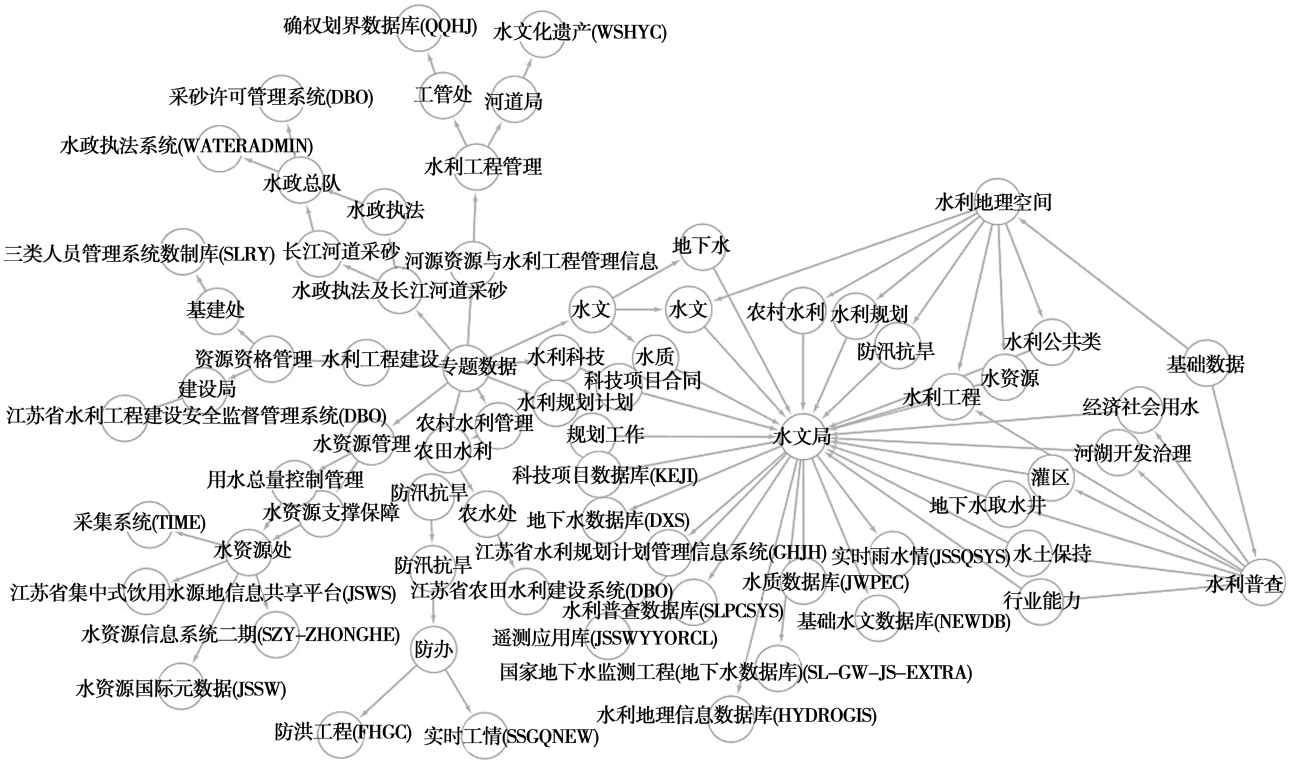

图 2　水利知识图谱部分展示

2　水利知识融合

知识图谱中通常有多种不同来源的数据，每种来源的数据的管理组织方式也各不相同，这形成了知识的多源异构性。因此，知识图谱的构建，通常伴随知识融合[11]。知识融合，就是将来自多个来源的关于同一个实体或概念的描述信息融合起来。知识融合一般包括实体层面(数据层面)和本体层

面(概念层面)的融合。不同的知识图谱,收集知识的侧重点不同,对于同一个实体,有的知识图谱可能侧重于其本身某个方面的描述,有的知识图谱可能侧重于描述实体与其他实体的关系。实体层面的融合,主要是将不同知识图谱中的实体进行对齐,找出等价实体。本体层面的融合,主要是找出等价或为包含关系的概念或者属性。知识融合过程主要采用相似度计算、聚类、表示学习等技术实现。

针对所用数据结构化程度高、规范化程度高的特点,本文设计了一种查找水利知识图谱本体层面的相似属性的算法,针对知识图谱本体层面中的每一个属性,分别求该属性和其他属性的中文字段的字符串相似度$score_1$和英文字段的字符串相似度$score_2$,并赋予相应的权重α和β,用以计算最终的相似度得分。属性中英文字段的字符串相似度,使用 Python 标准库 difflib 提供的 SequenceMatcher 进行计算。

为了结合属性本身的语义信息,考虑利用 word2vec[12] 词嵌入(word embedding)模型。为了将文本表示的数据转化为计算机可理解和计算的形式,一般采用 one-hot 编码的方法将文本转为词嵌入。但是 one-hot 编码一般较为稀疏,占用较大的存储空间,而且词与词之间的向量是正交关系,没有任何语义关联。为了克服这一缺点,充分结合属性之间的语义信息,word2vec 使用一层神经网络将 one-hot 编码映射到分布式形式的词嵌入。

Word2vec 有 2 种训练词向量的方式:CBOW 和 Skip-Gram。CBOW 模型是通过上下文的内容预测中间的目标词,而 Skip-Gram 则相反,通过目标词预测其上下文的词,两者互为镜像。通过最大化词出现的概率,训练模型可得到各个层之间的权重矩阵,词嵌入就是从这个权重矩阵里面得来的。本文实验中用到的是 CBOW 模型,其模型结构如图 3 所示。

CBOW 模型的输入层是由 one-hot 编码的输入上下文$\{x_1,\cdots,x_C\}$组成,其中窗口大小为C,词汇表大小为V,隐藏层是N维的向量,最后输出层是也被 one-hot 编码的输出单词y。被 one-hot 编码的输入向量通过一个$V\times N$维的权重矩阵W连接到隐藏层,隐藏层通过一个$N\times V$维的权重矩阵W'连接到输出层。

假设知道输入与输出权重矩阵的大小,第一步就是计算隐藏层 h 的输出,该输出就是输入向量的加权平均:

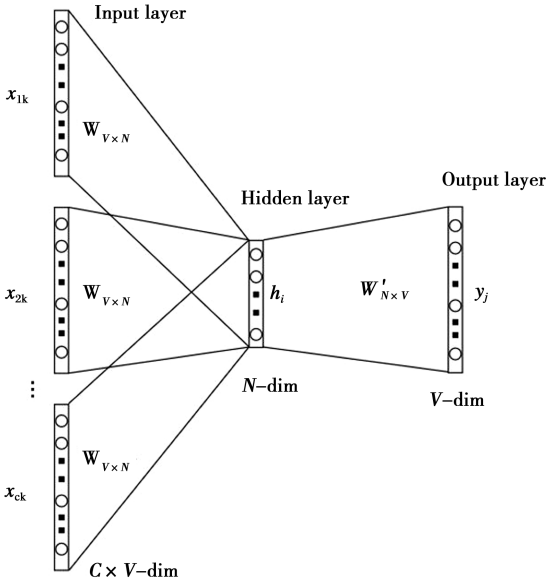

图 3 CBOW 模型

$$h=\frac{1}{C}W\cdot\left(\sum_{i=1}^{C}x_i\right) \tag{1}$$

第二步就是计算在输出层每个结点的输入:

$$u_j=v'^{T}_{w_j}\cdot h \tag{2}$$

式中,$v'^{T}_{w_j}$是输出矩阵W'的第j列。

最后计算输出层的输出y_j:

$$y_{c,j}=p(w_{y,j}\mid w_1,\cdots,w_c)=\frac{\exp(u_j)}{\sum_{j'=1}^{V}\exp(u_{j'})} \tag{3}$$

权重矩阵W与W'可以通过反向传播算法以及随机梯度下降来学习。首先给这些权重赋一个值进行初始化,然后按序训练样本,逐个观察输出与真实值之间的误差,并计算这些误差的梯度,在梯度方向纠正权重矩阵。

得到属性字段的词嵌入向量表示以后,用余弦相似度计算向量间的距离,代表两个属性之间的语义相似程度,并赋予权重$1-\alpha-\beta$,结合上文的中文字段的字符串相似度$score_1$和英文字段的字符串相似度$score_2$,得到最终的相似度得分,公式如下所示。

$$score=\alpha\cdot score_1+\beta\cdot score_2+(1-\alpha-\beta)\cdot score_3 \tag{4}$$

式中,$score_1$代表属性中文字段的相似度,$score_2$代表属性英文字段的相似度,$score_3$代表词嵌入的距离。

算法 1 查找属性p的 top-k 个相似属性

输入:水利知识图谱$\mathbb{KG}$,属性p,权重α,β,参数 k,阈值 threshold

输出:属性p的 top-k 个相似度大于阈值 threshold 的相似属性

1　对𝕂𝔾中的每一个其他的属性 p'：

2　计算属性中文字段的相似度$score_1$；

3　计算属性英文字段的相似度$score_2$；

4　计算属性的词嵌入之间的余弦相似度$score_3$；

5　属性之间的相似度 $score = \alpha \cdot score_1 + \beta \cdot score_2 + (1-\alpha-\beta) \cdot score_3$；

6　$score\_list.append(score)$；

7　$result = score\_list.sort(k, threshold)$；/ * 找出 top - k 个相似度大于阈值的相似属性 * /

8　$returnresult$；

算法 1 给出了计算属性相似度的算法。低维向量之间的距离，采用余弦相似度进行计算。几何中夹角余弦可用来衡量两个向量方向的差异，机器学习中借用这一概念来衡量样本向量之间的相似程度。假定 A 和 B 是两个 n 维向量，A 为 $A(A_1, A_2, \cdots, A_n)$，B 为 $B(B_1, B_2, \cdots, B_n)$，则 A 与 B 的余弦相似度为

$$\cos(\theta) = \frac{A \cdot B}{|A| \times |B|} = \frac{\sum_{i=1}^{n}(A_i \times B_i)}{\sqrt{\sum_{i=1}^{n}(A_i)^2} \times \sqrt{\sum_{i=1}^{n}(B_i)^2}} \tag{5}$$

余弦值的范围在[-1,1]之间，值越趋近于 1，代表两个向量的方向越接近，语义相似度也就越高；越趋近于 -1，它们的方向越相反，语义相似度也就越低；接近于 0，表示 2 个向量近乎于正交。

设置不同权重 α 和 β 时，本文充分结合数据本身特定，观察到数据中的属性字段，其中文字段的相似度在较大程度上会决定属性的相似程度，英文字段的决定作用次之，语义层面的作用再次，因此，将 α 初始化为一个较大的权重，使得不同权重之间满足：

$$\alpha > \beta > 1-\alpha-\beta \tag{6}$$

然后通过实验，对上述权重进行微调，得到合适的权重取值。最终设置 $\alpha = 0.6$，$\beta = 0.3$，该权重能在数据集上取得较好的实验效果。

3　基于水利知识图谱的智能搜索应用

基于知识融合的水利知识图谱，在网页端设计实现智能搜索应用，更好地利用属性之间的关联关系，达到搜索人员的搜索意图。通过用户打分实验，验证了智能搜索应用具有良好的可用性。

3.1　智能搜索应用可视化

为降低水利专业从业人员的检索难度，加深用户对特定水利本体的了解程度，充分利用属性之间的关联关系，本文设计并实现了基于水利知识图谱的智能搜索应用[13]，以网页的形式访问。智能搜索应用的网页采用 Java 语言开发，采用的是 Spring Boot 框架。图 4 展示了智能搜索应用的结构框架。

用户可登录网页，打开智能搜索引擎，在搜索框输入意向词条，根据智能词条提示选择知识图谱中真实存在的相应属性，查询其相似属性。查询结果以列表形式返回，根据属性相似度从高到低排列，供用户参考。另外，本文还将相似属性的词向量在低维空间中进行了展示，可以更直接地反映相似属性之间的语义相似情况，给用户提供更直观的感受。

例如，用户可以在搜索框输入词条“降水量”，智能词条提示就会显示知识图谱中与该属性相对应的真实存在的属性词条，用户可以选择对应的词条。假设选择词条“降水量(P)”，然后点击查询按钮，即可查询该属性的相似属性。点击图示按钮，还可以查看词嵌入的词向量在三维空间中的具体分布情况。

图 5 展示了三维空间中词嵌入的分布情况。可以看出，与属性词条的语义更接近的其他属性，在方向上会与代表属性词条的词向量更接近，同时计算出的余弦相似度也会更近。

3.2　用户打分实验

为了对开发的智能搜索应用软件的可用性进行评估，本文设计了用户打分实验，以便深入了解开发的智能搜索应用软件对具有不同专业背景的用户而言其可用性的情况。

系统可用性量表(System Usability Scale，SUS)[14]是一种用于可用性检测的建议问卷调查量表，能很好地评估产品在特定使用环境下为特定用户用于特定用途时的有效性、效率和用户主观满意度。本文实验中邀请了具有不同专业背景的 30 名志愿者参与测评，其中有 5 名水利领域的从业人员，10 名知识图谱相关背景的学生(包括 4 名本科生和 6 名研究生)和 15 名不具备知识图谱相关背景的其他专业的学生，以保证用户的背景多样性和实验结果的公平性。

打分实验采用 30 位志愿者的平均分作为最终的实验得分，SUS 平均分为 87.88，中位数为 88，方差为 8.28，这表明本文开发的智能搜索应用软件具有较好的可用性。

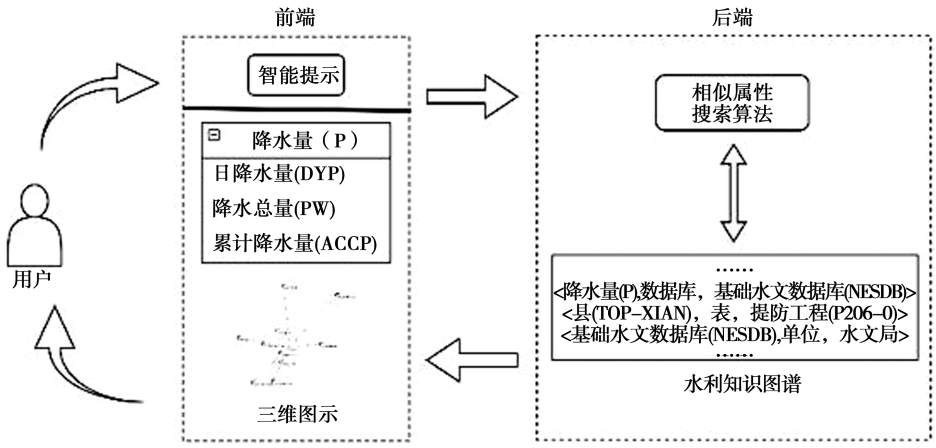

图 4 智能搜索应用结构框架

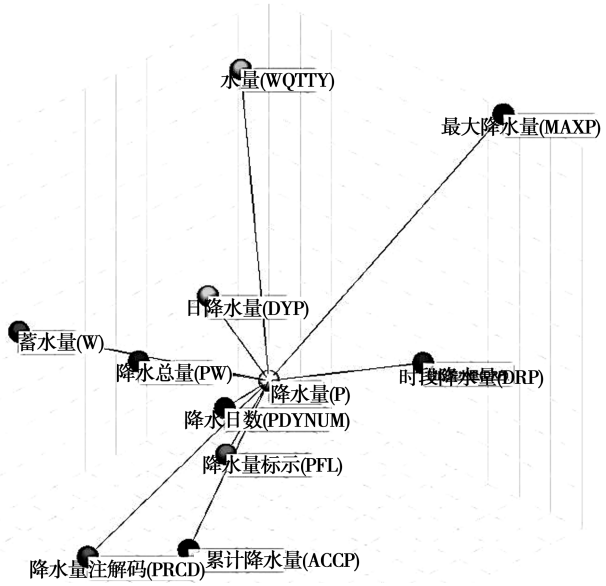

图 5 三维空间中的词嵌入

4 结 语

本文通过构建水利知识图谱,设计相似属性查找算法,充分结合数据特点以及数据潜在的语义信息,实现了本体层面的知识融合。基于融合后的水利知识图谱,本文开发了网页端的智能搜索应用,可以搜索相似属性,降低了水利领域从业人员的检索难度,并进行了三维空间上词嵌入的可视化展示,更加直观地增强对水利领域相关本体的理解。

未来计划尝试对水利知识图谱中的实体进行匹配,进一步提升水利知识图谱的融合程度,更好地挖掘相关水利知识的潜在语义信息,对水利领域的数据存储管理具有良好的应用价值。

参考文献:

[1] 葛召华, 张中坤, 李博. 基于知识图谱的水利数据垂直搜索应用[J]. 山东水利, 2018(5):1-2.

[2] 王鑫, 邹磊, 王朝坤. 知识图谱数据管理研究综述[J]. 软件学报, 2019(7):2139-2174.

[3] BERNERSLEE T, HENDLER J, LASSILA O. The Semantic Web, Scientific American[J]. Scientific American, 2001, 284(5):34-43.

[4] LEHMANN J, ISELE R, JAKOB M, et al. DBpedia – a large – scale, multilingual knowledge base extracted from wikipedia[J]. Semantic web, 2015, 6(2):167-195.

[5] MAHDISOLTANI F, BIEGA J, SUCHANEK F. YAGO3: a knowledge base from multilingual wikipedias[C]. Seventh Biennial Conference on Innovative Data Systems Research, 2015.

[6] VRANDECIC D, KRTOETZSCH M. Wikidata: A free collaborative knowledgebase[J]. Communications of the ACM, 2014, 57(10):78-85.

[7] DA N B, GUHA R V. RDF Vocabulary Description Language 1.0: RDF Schema[R/OL]. (2003-12-15)[2014-02-10]. http://www.w3.org/TR/rdf – schema/.

[8] 贾存鑫, 胡伟, 柏文阳, 等. SMap:基于语义的关系数据库模式与 OWL 本体间映射方法[J]. 计算机研究与发展, 2012, 49(10):2241-2250.

[9] BIZER C. D2R MAP-a database to RDF mapping language[C]//ACM, The 12th Internanonal World Wide Web. Budapest, 2003.

[10] Prud Hommeaux E, Seaborne A. SPARQL Query Language for RDF[S]. 2008.

[11] HUANG J, HU W, BAO Z, et al. Crowdsourced Collective Entity Resolution with Relational Match Propagation [C]. ICDE, 2020.

[12] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[J]. arXiv, 2013.

[13] CHENG G, LIU D, QU Y. Fast algorithms for semantic association search and pattern mining[J]. IEEE Transactions on Knowledge and Data Engineering,2019(99):1-1.

[14] JOHN BROOKE. SUS:a retrospective[J]. Journal of Usability Studies, 2013, 8(2):29-40.