基于知识图谱的电力行业与外部数据融合研究

刘锦隆[1],马 进[2],邹 双[3],宋立华[4],王秋琳[4]
(1. 国网四川省电力公司, 四川 成都 610041;
2. 四川凯普顿信息技术股份有限公司,四川 成都 610046;
3. 四川公众项目咨询管理有限公司,四川 成都 610041;
4. 福建亿榕信息技术有限公司, 福建 福州 350001)

摘 要:针对电力行业与外部多源异构数据融合利用的难题,提出了一种基于知识图谱的内外部数据融合应用思路。通过对现有文本信息提取以及知识图谱构建主流技术路线的分析,提出优化的文本信息提取方案及知识图谱构建方案,实现对多源异构数据融合建模工作的支撑。技术研究成果在某省电力公司开展了工程项目过程管理领域风险识别试点应用,验证了所提技术方案的可行性。

关键词:多源异构数据融合;文本信息提取;知识图谱构建;企业管理风险识别
中图分类号:TP391 **文献标志码**:A **文章编号**:1003-6954(2020)06-0026-05

Research on Fusion of Electric Power Industry Data and External Data Based on Knowledge Graph

Liu Jinlong[1], Ma Jin[2], Zou Shuang[3], Song Lihua[4], Wang Qiulin[4]
(1. State Grid Sichuan Electric Power Company, Chengdu 610041, Sichuan, China;
2. Sichuan Captain Information Cooperation, Chengdu 610046, Sichuan, China;
3. Sichuan Gongzhong Project Consultation Management Co., Ltd., Chengdu 610041, Sichuan, China;
4. Fujian Yirong Information Technology Co., Ltd., Fuzhou 350001, Fujian, China)

Abstract: Aiming at the problem of fusion and utilization of electric power industry data and external multi-source heterogeneous data, an application idea of internal and external data fusion based on knowledge graph is proposed. Through the analysis of the existing text information extraction and the mainstream technology route of knowledge map construction, an optimized text information extraction scheme and knowledge map construction scheme are proposed to support the modeling work of multi-source heterogeneous data fusion. A pilot application of risk identification in the field of engineering project process management are carried out in a provincial electric power company, which verifies the feasibility of the proposed technical scheme.

Key words: multi-source heterogeneous data fusion; text information extraction; knowledge graph construction; enterprise management risk identification

0 引 言

近年来,随着数据中台的建设及电力大数据技术的充分发展,电力行业逐步实现了行业数据的逻辑集中,对各类电力信息系统产生的海量数据的采集、存储、加工、处理和全价值链的分析处理能力也得到逐步完善。基于大数据的分析已在电力企业获得广泛应用并将更加深入,为管理提升、优化整合及服务转型提供技术支撑[1]。

然而,互联网及政务大数据的高速发展同样导致行业外部数据的爆发式增长,来自互联网的公开数据资源逐渐成为电网企业经营风险监测、电网运行维护、供应商优选及各类决策分析中不可或缺的重要组成部分。有必要开展内外部数据融合,将电力行业内部数据与外部的互联网公开数据及社会数据中的关键元数据及信息字段提取出来,融合形成

统一的数据结构,开展数据分析及挖掘利用。

要实现内外部多源异构数据的融合,首先要解决的是融合数据的统一表示。学术界先后提供网络本体语言(web ontology language, OWL)、资源描述框架(resource description framework, RDF)等数据表示方案。2012 年,谷歌公司提出知识图谱技术,由于其兼具严谨且务实的数据表示能力以及包括图数据库、图嵌入、图挖掘等成熟的技术配套,逐步成为多源数据融合表示的主流方法,是当前将多源数据融合在产业中落地的首选。

下面提出一种基于知识图谱实现企业内外部数据融合利用的技术方法,依托文本信息提取等人工智能技术的应用,将来自互联网的、难以分析的非结构化数据转换为结构化数据,而后进一步利用知识图谱技术,实现内外部数据的融合,为各电力应用需求提供支撑。基于所述技术开展了电力营销客服领域的应用研发并在某电力公司进行了部署使用,验证了该技术的可行性。

1 相关技术现状

所提出的基于知识图谱的电力行业与外部数据融合,主要涉及以下几个方面的关键技术。

1.1 文本信息提取技术

文本信息提取,也称为命名实体抽取,是指从文本中提取出特定的实体[2]。在实际项目中最常用的是专有命名实体提取。不同于通用命名实体,专有命名实体通常带有更多的限定,比如从电力营销退补工单中提取出“故障开始时间”,而不是所有的“时间”,所以不能采用通用命名实体的预训练模型。因此,专有命名实体通常也成为“关键信息抽取”,下面以“关键信息抽取”作为简称。

文本中关键信息的抽取,比较主流的有以下两类方案[3-4]:

1)基于规则的关键信息提取

基于规则的关键信息提取,是将人工观察到的待提取信息在上下文中出现的规律,固化为事先定义好的程序规则,通常使用正则表达式或基于正则表达式的扩展文本模式匹配技术。以正则表达式为例,其提供了一系列面向文本匹配的特殊语法,通过组合应用,实现对特定模式文本内容的准确提取。例如,提取电子邮件的正则表达式为/^(\w)+(\.\w+)*@(\w)+((\.\w{2,3}){1,3})$/,其中的\w代表任意字符,{2,3}代表出现 2 次或 3 次,该正则表达式可以识别出 xxxx@xxxx.xxx 格式的电子邮件地址。正则表达式表达灵活,可以匹配几乎任何模式的文字。应用正则表达式的前提是对拟提取信息的“模式”或“规则”要非常明确。

2)基于序列标注的方法

主流算法是序列标注算法中的条件随机场(conditional random field, CRF)算法。根据特征提取方式的不同,又可以分为人工特征工程 + CRF 以及深度学习特征提取 + CRF,后者最主流的技术路线是 BI - LSTM(双向循环长短文本记忆神经网络) + CRF。具体过程如图 1 所示。

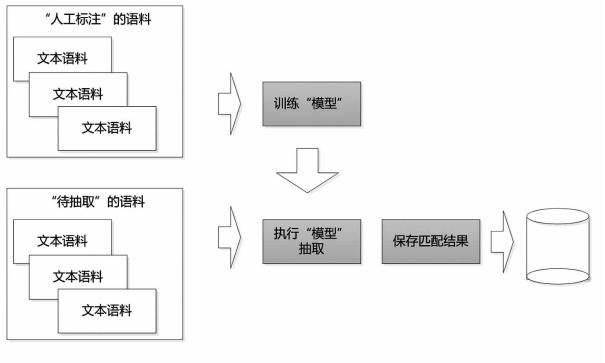

图 1 基于序列标注技术的信息抽取过程

基于“序列标注”的方法具有适应性强的优点,其不需要由专家编写规则,所以对于没有明显模式(人工很难观察到特定规则)的关键信息的抽取具有较强的抽取能力;其缺点是需要一定数量的、已经标注好的语料作为导入,这部分语料的标注工作需要人工编写。待抽取的关键信息越缺乏模式、要求抽的结果越精确,需要导入的语料就越多。另外,序列标注算法的缺点是不够稳定,其执行过程是“黑盒”(不像规则判断能够回溯),准确率不由算法决定,而主要取决于用以训练的标注语料是否和目标测试语料比较一致,所以该方法构建出的“抽取模型”常常难以判断是否能够满足业务对于抽取准确性的要求。

综上所述,现有两种文本关键信息抽取的技术路线各有以下优缺点和适用范围:

1)基于规则的方法抽取效果稳定,且不需要实现进行人工标注,但是限定性太强,匹配的范围比较小,对于没有固定模版的文本抽取不适用;

2)序列标注方法匹配范围较大,但需要事先准

备较多的人工标注语料,且抽取效果不稳定,准确性难以预估,对抽取准确性有比较严格要求的场景不太适用。

1.2 知识图谱技术

知识图谱是一种由节点和边组成的图数据结构,本质上是结构化的语义知识库。通过把专业业务领域中多源异构信息连接在一起,得到实体关系网络,知识图谱能够提供从“关系”的角度去分析问题的强大能力。最初,知识图谱的提出主要用于解决与实体相关的智能问答问题,目前已经广泛地应用到信息检索、数据挖掘领域。在电力行业,基于知识图谱之上的关系计算、图挖掘等能力与异常分析、静态分析、动态分析等数据挖掘方法相结合,可用于企业经营风险分析中的反欺诈、不一致性验证以及电网设备故障分析、灾害防御预警、主数据质量优化等领域。

知识图谱的构建大致可以划分为两个主要步骤。首先,结合相关领域的业务知识,将业务领域的关键概念实体及其之间的关系,描述为知识图谱本体;而后,利用实体抽取[5]、实体消歧及链接[6]、实体关系抽取[7]、知识推理[8]等技术,从实际业务数据中抽取出实体、关系相关信息字段,进行消歧融合,按照知识图谱的本体进行“填充”,获得知识图谱数据实例,进行知识图谱存储。知识图谱技术应用的主要挑战包括知识图谱构建过程自动化水平不高以及数据本身存在错误、冗余而导致的数据噪声等问题。

下面将应用知识图谱,开展电力行业数据与外部数据的知识融合,为相关业务的趋势洞察及辅助决策分析等提供数据支撑。

2 基于知识图谱的数据融合

所提出的电力行业数据与外部数据融合方案,主要依托两个步骤。首先是将外部的非结构化文本数据进行关键信息提取,将难以分析的非结构化数据转换为结构化数据;而后通过对相关领域业务实体关系的分析,设计知识图谱结构,将外部数据与电力行业结构化数据融合到知识图谱中,为下一步各类高级应用提供数据基础。下面重点介绍其中的关键环节。

2.1 文本预处理

为了开展基于非结构化文本的分析及知识图谱构建,需要对数据进行一定的预处理步骤,包括:

1)中文分词。待分析中文文本通常是连续的文字序列,不能直接进行分析,需要通过中文分词,将其切分为有意义的词。中文分词技术在信息检索等文本分析挖掘领域都有广泛应用。所提方法主要采用的是基于统计语言模型的序列标注方法,其基本原理是:首先准备标注数据,并基于人工标注的词性和统计特征,对待分词的文本进行建模与参数训练,该模型即可一定程度上描述词元素相对于上下文的分布;而后,利用模型对待提取文本中分词出现的概率进行预测,将概率最大的词作为分词分析结果。这类序列标注算法的代表是 CRF 算法。

2)构建词向量模型。计算机不能直接处理文本、图像、声音等内容,需要将其转化为数字特征后才能处理,词向量就是文本中的单词转后的计算机能够处理的数字化特征。所采用的是词嵌入(word embedding)技术。其基本过程是将文本嵌入到一个数学空间里,从而使得文本中语义相似的词用相似的向量表示。采用的具体模型是 word2vec。

2.2 文本信息提取

在第 1.1 节中,介绍了目前文本信息提取的主流技术路线。针对现有文本关键信息抽取的缺点,提出一种创新的方法,能够充分利用现有技术方法的优点,同时很大程度上规避其缺点,因而具有广泛适用性,其主要原理和步骤如下:

1)利用“规则抽取”准确性高、匹配范围小的特点,编写少量的规则,实现从大量的语料中匹配出少量但准确的抽取对象,并作为后续过程的导入。

2)将步骤 1 中获得的抽取结果,切割出一定比例(如 80%),作为导入到“自动序列标注”方法的训练语料,替代“人工标注”过程。

3)利用步骤 2 得到的训练语料,结合开源的“自动序列标注”类算法,构建“抽取模型”。

4)利用步骤 3 的结果,对步骤 2 切割出的、剩余的语料(如 20%)进行自动化抽取,并对抽取结果进行自动判断;如果模型自动判断的准确性尚未达到业务要求,则前往步骤 1,编写更多的正则表达式,形成更多的“标注语料”,作为模型训练导入;如果模型自动判断的准确性已经达到业务要求,则停止该过程,并将该模型作为文本抽取最终模型部署应用。

完整过程如图 2 所示。

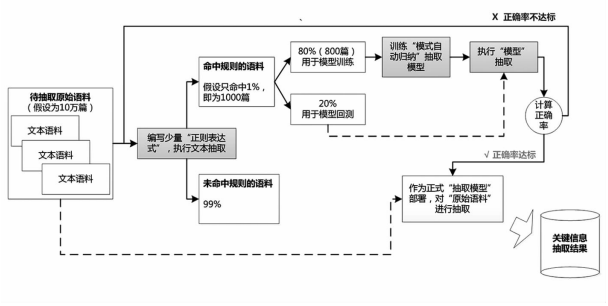

图 2　文本信息提取技术过程

相对于现有技术方案，所提方案在文本关键信息抽取问题上取得以下提升：用基于少量规则的“规则判断”替代“人工标注”获得初始标准语料，大大降低了初期人工投入；对“序列标注”模型的抽取结果进行自动化回测，确保模型的准确性符合业务需求；整个过程是可增量迭代的。如果“序列标注”训练得到的模型的抽取效果不理想，仅需要增加编写少量的规则（正则表达式），执行同样的过程循环，即可有效提升模型抽取效果，前期编写的规则不会被废弃。

2.3　知识图谱构建

完成文本信息提取后，将外部的非结构化数据转换为结构化的关键字段信息；而后，结合相关业务知识，即可开展知识图谱本体结构的设计。采用文献[9]所述的知识图谱表示方法对知识图谱进行建模。首先，根据业务域知识，参照电力公共数据模型（common information model，CIM），设计知识图谱的结构，并完成图数据库中对应图结构的创建；而后，将文本信息提取的结果按照图谱结构组织，调用图数据库接口导入实体关系数据，即可将电力行业结构化数据及外部非结构化文本的关键字段整合到知识图谱中。

知识图谱本质上是一种图数据结构，通常采用图数据库进行存储。这里采用开源图形数据库 Neo4J 存储知识图谱。

如图 3 所示，Neo4J 为知识图谱的创建、应用提供了全过程支撑。在数据入库阶段，可采用批量导入的方式将抽取好的命名实体与实体关系导入至图形数据库中；在知识探索阶段，采用 Cypher 语言查询所有节点及关系即能获取整个知识图谱的全貌，也可搜索所需的节点及关系信息，可以为使用者提供个性化的知识服务；在应用集成阶段，采用编程的方式可以调用 Neo4J 的 RREST API 接口进一步开发知识图谱界面。

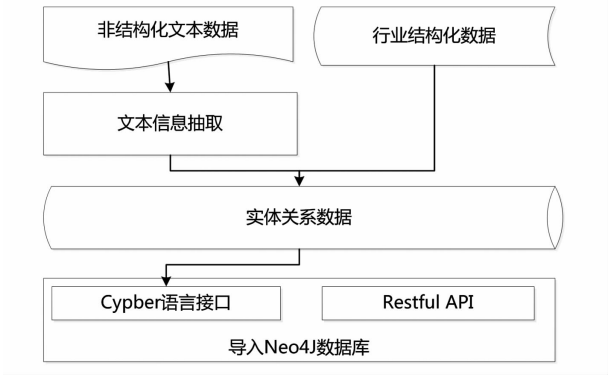

图 3　基于 Neo4J 的知识图谱存储及利用

2.4　基于知识图谱的内外部数据融合分析

通过上述步骤构建的知识图谱，能较好地实现内外部数据的融合，并为数据分析提供丰富的关系查询、计算手段。在实际应用中，可基于知识图谱这一良好的数据结构，结合业务需求与规则，开展更为丰富的多维风险防控及辅助决策。主要包括以下两种方法。

1）基于图规则。当前，知识图谱主要以图数据库为载体。以所选用的 Neo4J 为例，其提供的 Cypher 查询语言支持丰富的实体、属性及关系计算查询能力，可以高效地执行多维条件关系查询，筛选出符合特定关系条件的实体关系集合。在执行风险分析、辅助决策时，可以将相关的业务规则转换为图规则 Cypher 语句进行查询。基于图规则的知识图谱挖掘也是目前工业界使用的主流方案。

2）基于图神经网络。图神经网络的关键思想是将知识图谱的实体和关系转化为连续的向量空间，从而能够与深度学习算法结合，基于项目风险等样本的知识图谱表示进行建模，基于图结构的相似性识别出潜在的同类风险。图神经网络目前是知识图谱挖掘应用的研究热点，有很大的应用潜力，但由于相关的理论、算法还在持续发展完善过程中，目前还未在工业界得到广泛应用。

3　应用实例

所提出的技术方案在电网工程项目管理风险预警领域进行了实践，基于电网工程项目管理过程中涉及的电力行业及外部异构数据构建风险知识图谱，开展基于知识图谱的项目管理风险预测。

1）数据梳理与采集

包括电力行业数据和外部数据两个部分。其

中,电力行业数据主要包括项目合同主数据、项目执行过程数据、供应商评价数据等;外部数据包括招标公告、中标公告、企业工商数据、企业失信数据等。

2)信息提取及知识图谱构建

利用第 2 章提出的文本清洗及关键信息提取方法,将内外部数据中关键的实体字段提取出来,并汇集到对应的业务实体及其关系,主要数据结构如图 4 所示。

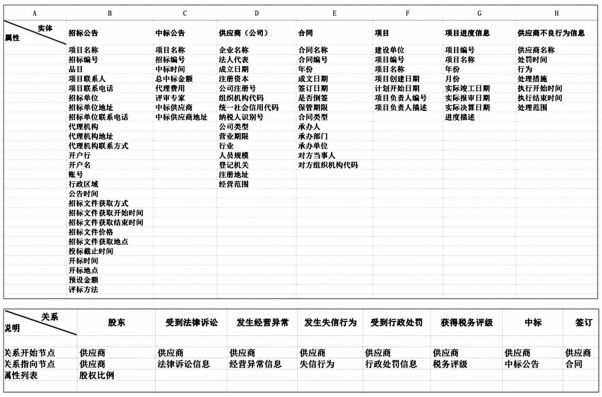

图 4　电网工程管理风险知识图谱实体关系结构

编制脚本,将提取出的实体、属性及关系信息按照图数据库 Neo4J 的结构导入到知识图谱中,形成的最终知识图谱。

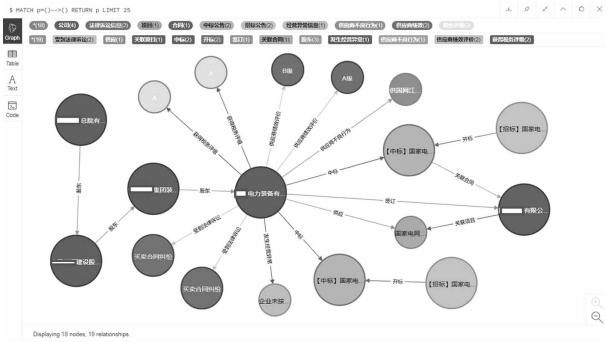

图 5　工程项目过程管理领域风险知识图谱

图 5 展示了所构建的知识图谱的局部内容,通过执行图查询“MATCH P = =( ) - - >( ) RETURN p LIMIT 25”获得,所以限定展示最多 25 个节点。图 5 中展示了公司、法律诉讼、项目、合同、招标公告、中标公告、供应商不良行为、供应商绩效等节点以及节点之间的 10 类关系。项目过程环节的主要信息,如采购、招标、中标、执行等主要环节的实体、属性及关系都在风险知识图谱中进行了清晰的展示。

基于知识图谱中完整、规范的数据,结合图查询、图计算等技术,即可进行如下探查和分析:

1)关联交易合规性风险分析。从工商数据中获取电力公司与相关供应商的股权关系,维护到知识图谱中,形成“公司 - 公司股权关系”,与“公司 - 项目中标关系”“公司 - 项目采购关系”结合,通过图规则查询,即可筛查出同一甲方单位采购的项目中,被具有股权关系的乙方单位中标的个数、金额与比例,与设定的阈值比较,即可识别出关联交易的规模,识别是否存在关联交易风险。

2)项目执行过程管控风险。从项目执行过程中产生的不同电子文件中抽取项目关键属性(包括项目编号、项目名称、项目合同签订时间、项目开工时间、项目约定工期、项目实际完成时间等),整合到知识图谱。通过图规则查询,即可筛查出哪些项目签订时间晚于开工时间,即“倒签核查”风险;还有项目已开展的时间大于合同约定工期,即“工程延期”风险。还可以进一步将风险项目所对应的征信数据进行对比,如果在存在上述风险的同时,供应商在征信方面也存在已知风险,则相应增加其风险等级。此场景体现了知识图谱对多源数据融合的价值。

3)项目单位履约及承载力不足风险分析。提取历史中标公告、合同数据中的关键属性,如甲方、乙方、项目开始时间、项目工期、项目金额等,整合到知识图谱中,即可通过图规则汇总项目单位在一段时间内承担的项目个数、金额,将当前数据与历史同期数据比较,即可筛查出相对于历史同期承担的项目个数、金额超过一定阈值的情况,识别可能存在的“乙方承载力不足“风险,进行提前预警。

4　结　语

行业数据与外部数据的融合分析,是大数据技术向纵深领域发展的主要趋势之一。以电力行业现实需求为背景,针对外部非结构化数据与行业内部数据融合分析问题,提供基于自然语言处理的非结构化文本关系信息提取方法以及融合知识图谱的构建方法,实现支撑业务所需的多源异构数据的融合,为上层高级业务应用提供坚实的知识数据基础。在某省电力公司开展的工程项目过程管理领域风险识别试点应用,能够较好地将内外部异构数据整合到

(下转第 38 页)

Technical Framework of Smart Grid Data Management from Consortium Blockchain Perspective [J/OL]. Proceedings of the CSEE. https://doi.org/10.13334/j.0258-8013.pcsee.181971.

[7] 崔志伟. 区块链金融:创新、风险及其法律规制[J]. 东方法学,2019(3):87-98.

[8] 华劼. 区块链技术与智能合约在知识产权确权和交易中的运用及其法律规制[J]. 知识产权,2018(2):13-19.

[9] Franklin M, Halevy A, Maier D. From Database to Data-spaces[J]. ACM SIGMOD Record, 2005, 34(4):27-33.

[10] 薛腾飞,傅群超,王枞,等. 基于区块链的医疗数据共享模型研究[J]. 自动化学报,2017,43(9):1555-1562.

[11] 吴红瑶. 基于关联数据的科学数据共享模型研究[D]. 沈阳:辽宁师范大学,2018.

[12] 熊维祥. 基于区块链技术的学分认证系统研究[D]. 北京:北京邮电大学,2018.

[13] 瞿海妮,庞徐玮,尤鸣宇,等. 电力大数据的应用价值及其共享平台分析与设计[J]. 经营与管理,2017(7):104-108.

[14] 蒋雷雷,代作松,秦宾. 基于 Hadoop 架构的电力企业数据共享模型研究[J]. 通信电源技术,2018,35(1):97-100.

[15] 龚钢军,魏沛芳,孙跃,等. 区块链下电力数据的统一监管与共享交易模型[J]. 信息技术与网络安全,2019,38(3):57-62.

[16] 李旭,王合建. 基于区块链的电力数据共享机制[C]// 中国电机工程学会电力信息化专业委员会. 2019 电力行业信息化年会论文集,北京:人民邮电出版社,2019:188-191.

[17] 李杰,柴焰明,杨燕,等. 区块链智能合约技术的原理与应用[J]. 云南电力技术,2018,46(6):12-18.

作者简介:

佘佐超(1990),男,助理工程师,从事信息系统安全防护工作;

李 喆(1989),男,硕士,工程师,从事 5G、电力北斗技术试点应用等工作;

刘浩宇(1991),男,硕士,工程师,从事数据质量管理、电力大数据分析等工作。

(收稿日期:2020-10-15)

(上接第 30 页)

知识图谱,利用图规则挖掘,实现典型风险的验证,验证了所提方案的有效性和可行性。

前面只重点阐述了基于行业数据与外部数据构建融合知识图谱的过程,对图神经网络、图嵌入等基于图的挖掘分析未开展深入讨论,这也是后续进一步研究的方向。

参考文献

[1] 李志,费晓璐,郭振. 基于数据中台的电力企业数据资产管理方法研究[J]. 电力信息与通信技术,2020,18(7):76-81.

[2] 林广和,张绍武,林鸿飞. 基于细粒度词表示的命名实体识别研究[J]. 中文信息学报,2018,32(11):62-71.

[3] 冯蕴天,张宏军,郝文宁. 面向军事文本的命名实体识别[J]. 计算机科学,2015,42(7):15-18.

[4] 陈锋,翟羽佳,王芳. 基于条件随机场的学术期刊中理论的自动识别方法[J]. 图书情报工作,2016,60(2):122-128.

[5] 翟社平,段宏宇,李兆兆. 基于 BILSTM_CRF 的知识图谱实体抽取方法[J]. 计算机应用与软件,2019,36(5):269-274.

[6] 曾宇涛,林谢雄,靳小龙,等. 基于多维信息融合的知识库问答实体链接[J]. 模式识别与人工智能,2019,32(7):642-651.

[7] 鄂海红,张文静,肖思琪,等. 深度学习实体关系抽取研究综述[J]. 软件学报,2019,30(6):1793-1818.

[8] 张仲伟,曹雷,陈希亮,等. 基于神经网络的知识推理研究综述[J]. 计算机工程与应用,2019,55(12):8-19.

[9] 李涛,王次臣,李华康. 知识图谱的发展与构建[J]. 南京理工大学学报,2017,41(1):22-34.

作者简介:

刘锦隆(1987),男,工程师,主要从事电力信息化建设工作;

马 进(1977),男,硕士,高级工程师,主要从事人工智能、信息化建设相关工作;

邹 双(1983),男,硕士,高级工程师,主要从事信息化研究建设相关工作;

宋立华(1982),男,硕士,高级工程师,研究方向为人工智能自然语言处理、知识图谱技术应用;

王秋琳(1980),男,高级工程师,研究方向为电力人工智能与电力信息化工作。

(收稿日期:2020-10-16)