DOI:10.20033/j.1003-7241.(2025)12-0125-05

知识图谱抽取算法的设计

方志宁[1]，王严博[2]，王　勇[3]，战科宇[2]

（1.国电电力宁夏新能源开发有限公司，宁夏 宁川 750000；2.中国搜索信息科技股份有限公司 北京 100013；
3.中国测绘科学研究院地理空间大数据中心 北京 100142）

摘　要：为了提高知识图谱抽取的准确率和速度，提出一种基于 Transformer 的知识图谱抽取算法。该算法利用 Bert 和 FastText 对文本数据进行预处理，通过 Transformer 模型有效提取特征，并采用 SCR 算法精确完成实体命名任务。进一步地，算法创新地将命名后的实体再次处理并输入模型，深入挖掘实体、关系与尾实体间的关联性，从而高效完成知识抽取。实验结果显示，该算法在实体抽取任务中表现出色，准确率高达 95%，F_1 值也达到了 95.26%，充分验证了算法各组件的有效性。此外，该算法成功应用于构建电力系统知识图谱，其准确全面、易用性强的特点以及卓越的可视化效果，为用户提供了直观且易于理解的知识展示方式，具有广阔的应用前景。

关键词：Transformer；知识图谱；抽取算法；词嵌入；标签与实体；关联性

中图分类号：TP311.13；TH-39　　**文献标志码**：A　　**文章编号**：1003-7241(2025)12-0125-05

Design of knowledge graph extraction algorithm

FANG Zhining[1], WANG Yanbo[2], WANG Yong[3], ZHAN Keyu[2]

(1. Guodian Electric Ningxia New Energy Development Co., Ltd., Ningchuan 750000, China;
2. China Search Information Technology Co., Ltd., Sanyuan Street, Dongcheng District, Beijing 100013, China;
3. Geospatial Big Data Center, Chinese Academy of Surveying and Mapping, Haidian District, Beijing 100142, China)

Abstract: In order to improve the accuracy and speed of knowledge graph extraction, a Transformer based knowledge graph extraction algorithm is proposed. This algorithm preprocesses text data using Bert and FastText, effectively extracts features through the Transformer model, and accurately completes entity naming tasks using the SCR algorithm. Furthermore, the algorithm innovatively reprocesses the named entities and inputs them into the model, delving deeper into the correlations between entities, relationships, and tail entities, thus efficiently completing knowledge extraction. The experimental results show that the algorithm performs well in entity extraction tasks, with an accuracy of up to 95% and an $F1$ value of 95.26%, fully verifying the effectiveness of each component of the algorithm. In addition, the algorithm is successfully applied to construct a knowledge graph of the power system, with its accurate and comprehensive characteristics, strong usability, and excellent visualization effect, providing users with an intuitive and easy to understand way to display knowledge, and has broad application prospects.

Keywords: Transformer; knowledge graph; extraction algorithm; word embedding; tags and entities; relevance

0　引言

在大数据的挖掘和分析过程中，可视化分析逐渐崭露头角，成为研究热点。在可视化分析中，知识图谱中的实体、关系、属性一般被称为三元组，在知识图谱这一有向图中利用拓扑结构进行串联和可视化呈现。在形成知识图谱时[1-3]，可以将不同的数据进行知识抽取，该知识可被应用在多种场合中[4,5]。

对于知识图谱的抽取，有许多学者进行了研究，例如：王书鸿等提出的电力设施的知识图谱抽取算法[6]，使用远程监督学习完成实体关系抽取，该算法存在无法体现词的位置，对于陌生词汇或使用率低的词汇无法处理等问题，因此会导致提取时准确率下降。周俊等提出的农业相关的知识图谱抽取算法[7]，提供构建知识图谱的基础信息，由于改进 DASREL 算法对于数据的依赖性很强，若数据量相对较少则无法抽取的准确率。吕东东等[8]提出了农业产品相关的知识图谱抽取算法[9]，在现有的农产品标准之上结合正则包装器，使该算法的实用性降低。

Transformer 具有强大的并行计算、表达能力和可视化属性[10-11]，并且 Transformer 能够进行长距离依赖建模，具有较好的灵活性和易理解性[12-14]，因此本文提出基于 Transformer 的知识图谱抽取算法。使用该算法能够高效地抽取知识图谱中实体和关系。

1　基于 Transformer 的知识图谱抽取

1.1　知识图谱构建过程

知识图谱的构建由以下步骤组成，分别为实体命名、知识抽取、特征融合以及存储，知识图谱构建过程如图

*基金项目：省部级自然科技基金项目(CEZB210008182)

收稿日期：2024-05-08

1 所示。

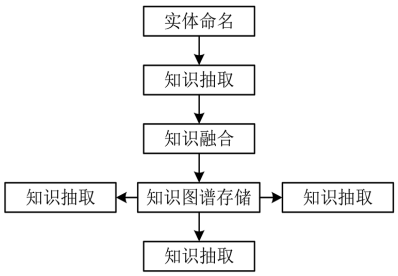

图 1　知识图谱构建过程

在知识图谱的构建过程中,最为重要的便是前两步,实体命名和知识抽取,二者直接影响了知识图谱构建的结果,若知识抽取时,实体、关系以及尾实体三者的关联产生问题,则知识图谱的全面性和准确性会大打折扣。

1.2　知识图谱抽取

利用基于 Transfoemer 的深度学习模型完成知识图谱的抽取结构构建,知识图谱抽取结构如图 2 所示。

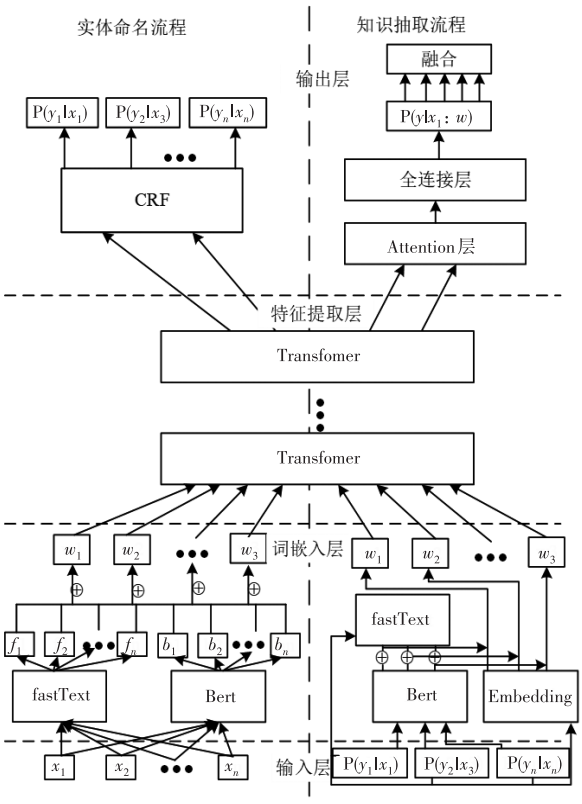

图 2　知识图谱抽取结构

在进行知识图谱抽取时共分为三步:第一步为实体命名,第二步则是知识抽取,第三步是将抽取出的知识进行融合便可以得到最终的知识图谱。

在该深度学习网络中,分别通过输入层、词嵌入层、特征提取层、输出层完成知识图谱抽取的实体命名(左侧)与知识抽取(右侧),实体命名与知识抽取过程均采用 Transformer 模型完成特征提取,且知识抽取过程中的输入层的输入数据为实体命名过程的输出结果。

1.2.1　实体命名

1)输入层。将文本内信息作为实体命名过程的输入数据进行实体命名。

2)词嵌入层。在该层中,共使用了两种嵌入技术分别是 fastText 以及 Bert。根据输入文本信息词的向量给出未来生成的知识图谱中索引的词向量序列是该层的主要任务。两种嵌入技术的拼接词向量是该层的输出结果。

① fastText 虽然对比其他的嵌入技术本身可能没有太大优势,但是该技术除了嵌入外,还能够构造出词的 n-grams 特征,该特征可以实现不同词间的结构信息共享,这样使用率低的文本或不使用的文本也能够存在向量表示结果,可以从已知文本获取未知文本的信息。

② Bert 是谷歌提出的一种具有迁移学习能力,并且使用了大量语料进行训练的模型,该模型具有极强的泛化能力,通过微调参数便可以在不同语言中发挥强大的效果。在本文中使用 Bert 模型只需要进行微调,既能够节省大量的训练时间,又可以提升仅使用 fastText 模型带来的一些缺点。

3)特征提取层。特征提取层的主要作用便是抽取文本中词向量的特征,通过结合词在文本中的语境,得出文本中的词向量。本文中特征抽取层运用了 Transformer 模型,与 RNN 网络模型相比较,本文的 Transformer 模型不仅可以避免梯度爆炸和梯度消失的问题,还有更强大的并行计算能力。Transformer 模型运用了注意力机制,对全局的文本进行编码,使特征提取能力进一步提升,在知识图谱的实体命名中 Transformer 模型的特征抽取能力远超于 CNN 与 RNN 类网络,并且 Transformer 模型可以通过增加内部模块的方式提高模型的拟合能力。

4)输出层

条件随机场(conditional random filed, CRF)利用特征函数 $F(y \mid x)$ 构造标签与标签间关系,输出层的公式表达为

$$y^* = \arg\max F(y \mid x) \tag{1}$$

1.2.2　知识抽取

使用基于 Transformer 的深度学习模型进行知识抽取的结构与实体命名的结构十分相似。

1)输入层。输入数据为实体命名过程的输出结果。

2)嵌入层。该层中使用的工具与实体命名时的工具相同,均为 fastText 与 bert。

3)特征提取层。与实体命名过程一致,均利用 Transformer 模型完成。

4)输出层。使用 self-attention(自注意力机制), self-attention 的引入使基于 Transformer 的深度学习模型能够得到文本中整个句子的表征。在知识抽取时,全连接层的激活函数为 softmax,获取句子中实体对应的关系概率。在建立指数图谱中,除了最初定义的若干个关系类型外,还需要增加 other 类型,该类型用于表示不属于最初定义

的关系类型。

1.3 基于 Transformer 模型的特征提取

利用基于 Transformer 的深度学习模型进行知识图谱抽取时,在特征提取层利用 Transformer 模型完成知识图谱抽取时实体命名与知识抽取的特征提取。下文以知识抽取过程的特征提取为例对 Transformer 模型进行详细分析。

1.3.1 Transformer 模型结构

在 Transformer 模型中内部有多个编码器和解码器进行堆叠组成,原始实体输入编码器,正确实体输入解码器,利用监督学习完成模型的训练。Transformer 模型的编码器和解码器在训练和测试时的工作内容不同,训练时原始实体输入编码器,正确实体输入解码器;测试时原始实体输入编码器,解码器负责校对输出信息。Transformer 结构如图 3 所示。

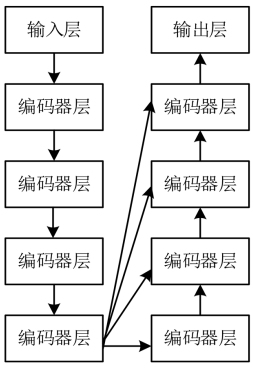

图 3 Transformer 结构

在 Transformer 中编码器层包含了自注意力机制层和全连接层。在一个编码器层输出时运用 ReLU 激活函数,公式为

$$\mathrm{FFN}(x)=\max(0,xW_1+b_1)W_2+b_2 \tag{2}$$

式中,W_1、W 为权值;b_1、b_2 为偏置系数;x 为计算的实体值。

1.3.2 Transformer 编码与解码

在 Transformer 中编码器与解码器内部结构如图 4 所示。

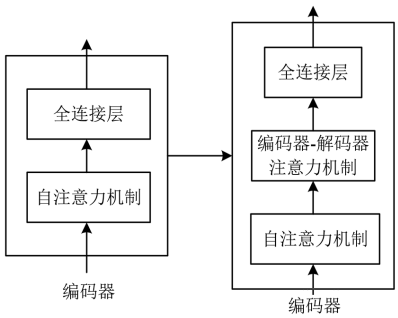

图 4 编码器与解码器内部结构

若要建立知识图谱则需要数量足够大的嵌入面模型进行训练,可是在现实情况很难找到如此海量的数据进行训练,难以有效表示小样本关系。基于平移假设可知

$$h+r=t \tag{3}$$

则

$$r=t-h \tag{4}$$

式中,h、r、t 为知识图谱中的三元组集合 $\tau=\{(h,r,t)\}\subseteq E\times R\times E$,三者分别代表了实体、关系以及尾实体。

因此通过实体嵌入便可计算关系的原始嵌入 R_0 为

$$R_0=S_t-S_h \tag{5}$$

式中,S_t、S_h 为头尾的嵌入实体。若要得到实体和关系的位置信息,需要将对应的三元组位置嵌入与训练嵌入加和,具体公式为

$$h_t=h_0+h_{\mathrm{pos}} \tag{6}$$

$$t_i=t_0+t_{\mathrm{pos}} \tag{7}$$

式中,h_0、t_0 为原始的头尾实体嵌入;h_{pos}、t_{pos} 为头尾实体位置嵌入。

根据原始嵌入与嵌入位置信息以及上一模块的关系嵌入组合为三元嵌入,完成参考集和查询集的结合,输入至 L 层 Transformer 中

$$E_i^1=\mathrm{Transformer}(h_i,R_0,t_i) \tag{8}$$

$$E_i^l=\mathrm{Transformer}(E_i^{l-1}),\ l=(2,3,\cdots,L) \tag{9}$$

式中,E_i^l 为经过了 l 层编码层后的实体或关系特征,该特征对实体的语义角色进行编码。能够识别不同实体关系的细粒度含义。

2 实验分析

2.1 实验对象

为保证本文算法构建的知识图谱效果,以南方电网的文本数据为基础,构建电力系统知识图谱。使用的文本数据包括了历史故障信息、电网分布情况、故障修理步骤、电网规章制度等多种信息。该电力公司的共管理输配电线路约 8 万公里,总装机量约为 1.73 亿千瓦供电面积超过 100 万平方公里,售电量超过 5 300 亿千瓦时。

为保证实验的顺利进行,对 Transformer 进行超参数设置,超参数设置如表 1 所示。

表 1 超参数设置

超参数	参数设置
cahr_embedding	60
bigram_embedding	60
Transformer_Layers	4
Transformer_HiddenSize	256
Transformer_Head	4
BiLSTM_Layers	1
BiLSTM_HiddenSize	256
batch_size	16
优化器	SGD
学习率	0.000 8
epochs	50
dropout	0.15

2.2 实验数据

使用本文算法进行知识图谱的三元组的数据抽取,三元组数据抽取结果及时间如表 2 所示。

表 2　三元组数据抽取结果及时间

例句	抽取结果	抽取时间/ms
输电线路采用双回线运行	运行方式:双回线运行	1
为确保输电线路通道畅通应定期巡检线路	定期巡检线路,保证道路畅通	3
遇到恶劣天气情况,应停止输电保证安全	恶劣天气停止输电	2
发生线路故障时切断电源并执行备用应急预案	发生紧急情况,切断电源、启动应急预案	3

通过表 2 可以清晰观察到,基于 Transformer 模型的深度学习网络在知识抽取任务中展现了出色的性能。它能够准确快速地从不同文本中抽取出实体和关系,进而在构建知识图谱过程中实现时间的高效利用。这一优势为知识图谱的迅速形成提供了重要支持。

本文算法进行训练时,准确率及 F_1 值变化结果如图 5 所示。

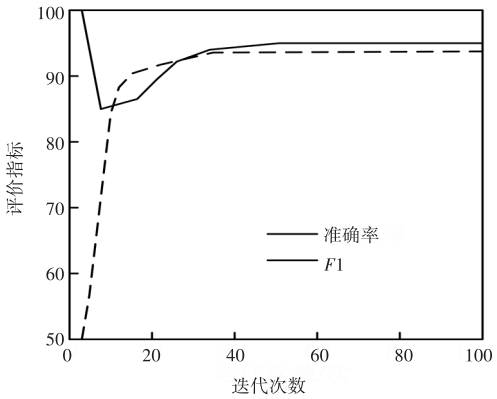

图 5　准确率及 F_1 值变化结果

通过图 5 可以观察到,本文提出的算法在训练过程中的准确性变化。在迭代开始时,准确率出现明显的下降,达到了最低值 85%。然而,随着迭代次数的增加,准确率逐渐上升,显示出模型在训练过程中的优化和学习。在迭代次数达到 60 次后,准确率更是达到了最高值 95%,这表明模型已经充分学习了数据中的特征和模式。这一趋势验证了本文算法的有效性和模型的学习能力。

为验证本文算法中各个部分的有效性,进行消融实验,其中-LSTM 为使用 LSTM 代替 Transformer、-CNN 为使用 CNN 网络代替 Transformer、-Attention 为去掉 Attention 机制,消融实验结果如表 3 所示。

表 3　消融实验结果

模型	准确率/%	F_1
Transformer	95	95.26
-Attention	90.27	90.55
-CNN	82.68	83.51
-LSTM	87.94	85.26

通过表 3 的详细数据展示,可以清晰观察到在电力系统文本数据的实体和关系抽取任务中,对模型各部分进行替换或取消后的效果变化。值得注意的是,在取消了 Attention 之后,抽取效果发生了一定程度的变化,准确率由原本的 95%下降至 90.27%,同时 F_1 值也呈现下降趋势。而对于其他部分的调整,如使用 LSTM 或 CNN 替换 Transformer,准确率和 F_1 值会明显下降,这表明这些部分对模型的抽取能力有一定的贡献但是特征提取的效果较差。综上所述,这些实验结果表明本文提出的算法在电力系统文本数据的实体和关系抽取任务中具有优越性。

本文算法形成的电力系统部分知识图谱如图 6 所示。

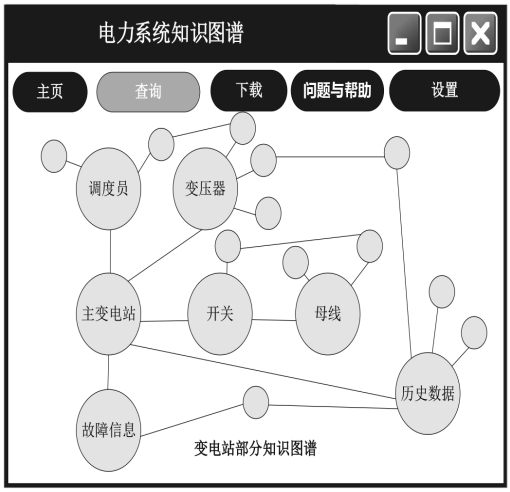

图 6　电力系统部分知识图谱

通过图 6 的展示,可以清晰地看出本文算法生成的电力系统部分知识图谱的优越性。该知识图谱信息全面、精准,能够迅速而简便地展现不同实体之间的相互关联情况。其简洁明了的布局使得用户能够快速获取所需的知识,并深入了解实体之间的关系。此外,本文算法生成的知识图谱使用简单,不仅易于上手,还能够保持高频率的更新,确保图谱内容与最新知识保持一致。综上所述,本文算法具备多种优点,是一个高效、实用的知识图谱抽取算法,为不同领域的知识管理和应用提供了有力支持。

利用本文算法生成的知识图谱结合其他算法,应用于变电站故障检测之后,准确率分析结果如表 4 所示。

表 4　准确率分析结果

故障名称	故障原因	准确率/%
接地故障	设备接地不良或地线断裂	97
断路器故障	断路器无法正常开关	95
变压器故障	变压器内部或外部原因损坏	91
电缆故障	电缆老化、机械损伤等	93
自动装置故障	保护装置自动装置等故障	93

通过观察表 4 的数据,可以发现接地故障和断路器故障的诊断率非常高,超过了 95%。这可能是因为这两种故障发生的频率较高,导致相关数据量较为充足,从而使得模型能够更好地学习和识别这些故障。相对而言,变压器、电缆和自动装置等设备的故障由于发生频率较低,导致模型对于这些故障的诊断准确率略低,但也达到了 91%以上的水平。这些数据表明,本文算法所抽取的知识图谱在实际应用中具有很好的实用效果,能够有效地提高设备故障诊断的准确率和效率。

3 结束语

设计了一种基于 Transformer 的知识图谱抽取算法。在电力系统知识图谱生成方面展现出了卓越性能,其迅速而精确地抽取实体与关系的能力,为知识图谱的构建提供了坚实基础,进一步印证了本文算法在抽取能力上的突出优势,彰显了其独特价值,能够显著提高故障诊断的准确度,为电力系统的稳定运行提供有力支持,有助于及时发现并排除故障,确保电力系统的正常运行,更对于提升电力系统的整体性能与可靠性具有深远意义。

参考文献

[1] 袁满, 刘梦琪, 牟梦宁. 基于 MDR 的知识图谱语义关系及表示标准化模型研究[J]. 情报学报, 2023, 42 (7):832-841.

[2] 张宁豫, 谢辛, 陈想, 等. 基于知识协同微调的低资源知识图谱补全方法[J]. 软件学报, 2022, 33(10):3531 -3545.

[3] 阮利, 温莎莎, 牛易明, 等. 基于可解释基拆解和知识图谱的深度神经网络可视化[J]. 计算机学报, 2021, 44(9):1786-1805.

[4] 李永卉, 周树斌, 周宇婷, 等. 基于图数据库 Neo4j 的宋代镇江诗词知识图谱构建研究[J]. 大学图书馆学报, 2021, 39(2):52-61.

[5] 钟卓, 唐烨伟, 钟绍春, 等. 人工智能支持下教育知识图谱模型构建研究[J]. 电化教育研究, 2020, 41(4): 62-70.

[6] 王书鸿, 郑少明, 刘中硕, 等. 面向某地区电网继电保护装置缺陷知识图谱构建的实体关系抽取[J]. 电网技术, 2023, 47(5):1874-1887.

[7] 周俊, 郑彭元, 袁立存, 等. 基于改进 CASREL 的水稻施肥知识图谱信息抽取研究[J]. 农业机械学报, 2022, 53(11):314-322.

[8] 吕东东, 陈俊华, 毛典辉, 等. 农产品标准领域知识图谱实体关系抽取及关联性分析[J]. 农业工程学报, 2022, 38(9):315-323.

[9] 张宇, 于合龙, 郭文忠, 等. 基于知识图谱的番茄种植管理可视化查询[J]. 农机化研究, 2024, 46(3):8-13.

[10] 梁礼明, 何安军, 李仁杰, 等. 跨尺度跨维度的自适应 Transformer 网络应用于结直肠息肉分割[J]. 光学精密工程, 2023, 31(18):2700-2712.

[11] 徐伟. 一种基于知识图谱的电力设备信息自适应检索方法[J]. 自动化技术与应用, 2025, 44(10):110-114.

[12] 陈孟元, 韩朋朋, 刘金辉, 等. 动态遮挡场景下基于改进 Transformer 实例分割的 VSLAM 算法[J]. 电子学报, 2023, 51(7): 1812-1825.

[13] 张锋, 张朔严, 乔利红, 等. 基于知识图谱的变电站配置文件智能校核技术研究[J]. 电测与仪表, 2024, 61(4):64-72.

[14] 黄思蓓, 张磊. 工业互联网安全知识图谱设计研究[J]. 自动化仪表, 2021, 42(12):90-99.

作者简介:方志宁(1967—),男,本科,高工,研究方向:信息化。

(上接第 50 页)

[11] MATHIEU J L, KOCH S, CALLAWAY D S. State estimation and control of electric loads to manage real-time energy imbalance[J]. IEEE Transactions on power systems, 2012, 28(1):430-440.

[12] ZHOU X, SHI J, TANG Y, et al. Aggr-egate control strategy for thermostatically controlled loads with demand response[J]. Energies, 2019, 12(4):683.

[13] 姚垚, 张沛超. 大规模变频空调参与电力系统辅助服务的协调控制方法[J]. 电力系统自动化, 2018, 42(22):127-134.

[14] JU P, JIANG T, LI H, et al. Hiera- rchical control of air-conditioning loads for flexible demand response in the short term[J]. IEEE Access, 2019(7): 184611-184621.

[15] CHEN G, LIU D. Adaptive robust eco- nomic dispatch and real-time control of distribution system considering controllable inverter air-conditioner clusters[J]. Frontiers in Energy Research, 2023(10):1017892.

[16] 郭旭歆, 高赐威, 王朝亮, 等. 基于中央空调虚拟储能模型的调峰策略研究[J]. 电力需求侧管理, 2022, 24(4):42-46.

[17] 王蓓蓓, 胡晓青, 顾伟扬, 等. 分层控制架构下大规模空调负荷参与调峰的分散式协同控制策略[J]. 中国电机工程学报, 2019, 39(12):3514-3528.

[18] HU X, NUTARO J. A priority-based control strategy and performance bound for aggregated HVAC-based load shaping [J]. IEEE Transactions on Smart Grid, 2020, 11(5):4133-4143.

[19] WANG H, CHEN H, LI Y, et al. A Review of air conditioning load aggregation in Distribution networks[J]. Frontiers in Energy Research, 2022(10):890899.

[20] VITTAL V, MCCALLEY J D, ANDERSON P M, et al. Power system control and stability[M]. Newark, NJ:John Wiley&Sons, 2019.

[21] 王锡凡, 方万良, 杜正春, 等. 现代电力系统分析[M]. 北京:科学出版社, 2003.

[22] 陈虹. 模型预测控制[M]. 北京:科学出版社, 2013.

作者简介:俞 乾(1976—),男,博士,教授级高级工程师,研究方向:电力系统。