DOI:10.20033/j.1003-7241.(2026)08-0166-04

一种基于电力海量分布式数据智能检索方法

宋长青, 任帅辉, 谢剑翔, 梁籍云, 罗 新

(广东电网有限责任公司广州供电局, 广州 510000)

摘 要:针对常规电力海量分布式数据检索采用语义近似并行方法,忽略数据点实时值对属性值影响而导致召回率低的问题,本文提出一种电力海量分布式数据智能检索方法。首先,分析电力系统基本架构,明确分布式数据的节点拓扑与流向。其次,采用三元法对分布式数据进行"数据源标识—属性字段—实时测值"的结构化表述。在此基础上,通过一致性哈希法对数据点实时值进行预处理,将动态实时数值映射至稳定哈希空间以消除分布偏差,进而提取数值区间与变化趋势等属性值特征,建立关键词倒排索引实现智能化检索。实验结果表明,所提方法的平均召回率达0.88,较传统方法提升约22%,检索结果与目标数据一致性显著提高。该方法通过融合实时值动态影响与哈希映射机制,有效满足了电力海量分布式数据检索对高召回率与高准确性的现实工程需求。

关键词:智能化检索;分布式数据;关键词索引;智能化检索;数据检索;电力海量数据

中图分类号: TP391.3;TP311.13;TM73 **文献标志码**: A **文章编号**: 1003-7241(2026)08-0166-04

A intelligent retrieval methods for massive distributed data of electric power

Song Changqing, Ren Shuaihui, Xie Jianxiang, Liang Jiyun, Luo Xin

(Guangdong Power Grid Co., Ltd. Guangzhou Power Supply Bureau, Guangzhou 510000, China)

Abstract: To address the low recall rate issue in conventional semantic approximate parallel methods for massive distributed power data retrieval, which arises from ignoring the impact of real-time values of data points on attribute values, this paper proposes an intelligent retrieval method for massive distributed power data. Firstly, the basic architecture of the power system is analyzed to clarify the node topology and data flow of distributed data. Secondly, a ternary method is used to provide a structured representation of distributed data in the form of 'data source identification—attribute field—real-time measurement.' On this basis, real-time values of data points are preprocessed using a consistent hashing method, mapping dynamic real-time values into a stable hash space to eliminate distribution bias, followed by extracting attribute value features such as numerical ranges and variation trends to establish a keyword inverted index for intelligent retrieval. Experimental results show that the proposed method achieves an average recall rate of 0.88, an improvement of approximately 22% over traditional methods, with significantly enhanced consistency between retrieval results and target data. By integrating the dynamic impact of real-time values with the hash mapping mechanism, this method effectively meets the practical engineering requirements of high recall and high accuracy in the retrieval of massive distributed power data.

Keywords: intelligent retrieval; distributed data; keyword index; intelligent retrieval; data retrieval; massive power data

在电力系统运行的过程中,将会产生海量的电力方面的分布式数据。在海量分布式数据存储库中,实现数据的智能化检索,能够实现对电力数据的高效取用,进一步提高电力系统的服务质量。对此,很多研究人员提出了数据检索方法。

文献[1]针对海量数据的特点,对数据的来源路径进行追踪,建立数据传输过程的特征模型,对面向的数据对象进行表达,分析数据组织结构。搭建三维数据空间,在该空间中分析数据的时空属性,得到数据的分片策略。在该空间中建立R树空间索引,将数据索引与数据查询关键词进行对应,实现分布式存储数据的并行检索。实验结果表明该方法检索结果不够全面。文献[2]基于异构数据的多分辨率特点,对数据进行预处理,并进行关联分析。建立面向动态的时空数据组织模型,在原有数据检索模式的基础上兼容异构数据的多检索模式。对数据检索的应用需求进行演化,采用漂移壳的方法对数据进行多层级的时空剖分,并进行编码,实现数据的时空框架建模,得到带粒子的通量模型数据。对模型数据进行重组处理,利用HBase协处理器技术构架数据的物理存储模型,编写查询算法,代入到该模型中进行训练验证。实验结果表明该方法检索结果精准度不高。文献[3]对数据源以及数据模态进行分析,对数据对应的语义进行区分。引入哈希与字典学习阐述数据语义一致性,对数据检索目标进行评价。将数据特征代入低维空间中,计算数据与语义的近似性。采用矩阵分解的方法将目标评价结果进行转化,结合中间变量方法对转化结果进行求解,解决非凸问题。利用网络

收稿日期:2025-06-05;**录用日期**:2025-06-30

作者简介:宋长青(1972—),男,高级工程师,研究方向为工程管理。

引用本文:宋长青,任帅辉,谢剑翔,等.一种基于电力海量分布式数据智能检索方法[J].自动化技术与应用,2026,45(8):166-169.(Song Changqing, Ren Shuaihui, Xie Jianxiang, et al. A intelligent retrieval methods for massive distributed data of electric power[J]. Techniques of Automation and Applications, 2026, 45(8): 166-169.)

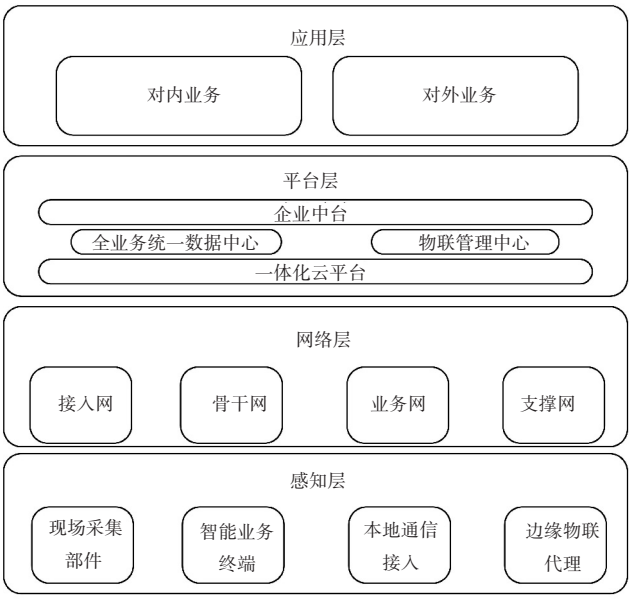
学习的方法获取数据之间的模态关系,进而分析数据层结构,形成节点邻接矩阵形式,根据数据语义相似度引入处理框架,分割为轻量化的检索子任务,得出数据检索结果。实验结果表明该方法检索效率不高。

考虑到上述文献所提出的数据检索方法无法满足电力数据的智能化检索需求,本研究以电力海量分布式数据为研究对象,提出了一种电力海量分布式数据智能检索方法。该方法通过分析电力系统架构,计算数据属性值,来对应检索关键词,实现精准度较高的海量分布式电力数据智能化检索过程。

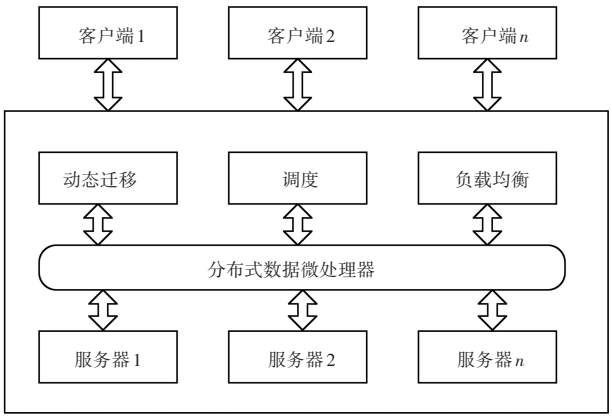
1 电力海量分布式数据智能检索

1.1 预处理海量分布式电力数据

针对引入电力系统架构进行分析,得到以四层基础架构为核心的基本架构,如图 1(a)所示[4]。在该架构中,通过计算机集群实现物联网节点之间的数据共享,将海量数据进行分布式实时存储,形成分布式实时数据库,其基本架构如图 1(b)所示[5-6]。



(a) 物联电力系统结构图



(b) 分布式数据库结构图

图 1 电力海量分布式数据基础框架

Fig. 1 Basic framework of massive distributed power data

在如图 2 所示的分布式数据库架构中,使用三元组的方法对数据进行表述,分别为数据点的实时值,数据点的当前状态以及数据点实时值的产生时间。以数据点的实时值对数据进行分析[7-10]。采用一致性哈希的方法对数据进行映射处理,数据映射处理是指将数据从一种形式或表示转换为另一种形式的过程。首先,将数据实时值的关键值 K 进行首尾相接,得到一个圆环,作为物联网节点数据分配临时的令牌,对多个节点的 K 值区间进行管理[11]。本研究将 K 值的映射值设置为 32 位。通过映射,得到源数据点在圆环上的相邻节点,完成对海量分布式电力数据的预处理。

1.2 提取分布式数据属性值特征

通过数据的维度对映射处理后的数据进行筛选,本研究以 770 维作为分界线,将满足 770 维及以上的数据采集到统一的数据空间中,将未满足 770 维的数据进行处理,进行补零填充操作,如公式(1)所示[12]。

$$W_i = \left(\frac{L \oplus 770}{l}\right)^2 \tag{1}$$

式中,i 表示数据点,W_i 表示数据维度,L 表示损失度,l 表示汉明距离度量。

完成数据补零后,将两种数据对应的矩阵进行乘法求和,如公式(2)所示。

$$f_i = 1*1 + 0*0 + 0*1 + 0*0 = 0*0 + 1*1 + 0*0 + 0*0 = 2 \tag{2}$$

式中,f 表示数据的特征属性。

将其与卷积核继续进行乘法求和,得到映射后的 3 * 1 矩阵,利用该矩阵,对数据的特征进行提取。对数据的稀疏特征量进行分析,如公式(3)所示。

$$Q = \frac{\omega(f_i - 1)^{W_i}}{\theta} \tag{3}$$

式中,Q 表示稀疏特征量,ω 表示分布间隔,θ 表示均值系数。

利用计算所得的数据稀疏特征量,得到数据的特征属性值,如公式(4)所示。

$$r = (wQ - v)^2, v \in [0,1] \tag{4}$$

式中,r 表示数据特征属性值,w 表示高斯混合参数,v 表示混合角点参数。

通过上述过程,完成数据属性值特征的提取过程。

1.3 智能化检索海量分布式数据

基于提取所得的海量分布式数据特征属性值的不同,建立检索关键词与数据属性值之间的索引关系[13-15]。通过该索引关系实现分布式数据的智能化检索过程。

计算数据检索关键词与数据属性值之间的轮廓系数,如公式(5)所示[16]。

$$k = \frac{y - r}{\max(r, y)} \tag{5}$$

式中,k 表示轮廓系数,y 表示轮廓均值。

根据公式(5)的计算结果,建立数据检索的对应索

引,如公式(6)所示。

$$g = \frac{\sum_{i=1}^{k} r \cdot d}{T_i} \tag{6}$$

式中,g 表示数据检索索引关系,d 表示智能化检索数据源,T_i 表示检索关键词词频。

根据建立的关键词索引,将数据检索的关键词与数据属性值进行关联,实现电力海量分布式数据的智能化检索过程。

2 实验

2.1 实验准备

对本文所提方法进行验证。本次实验采用模拟仿真的方法进行。从实验结果分析本文所提电力海量分布式数据智能化检索方法的有效性。

本次实验数据模拟来源于某电网系统数据信息。选用该数据集中公开的数据信息部分,已作过预处理,不会对电力系统用户的隐私造成侵犯或泄露。对该数据集中的数据信息进行随机筛选,得到以10个物联网节点为基站的电力物联数据,共计40万余条。每个物联网节点的数据信息数量并不完全一致,但总体上相差不大。每条数据信息的长度基本一致,平均为63个字节,主要具有7个属性字段。

完成模拟数据样本的筛选后,采用数据模拟软件对数据样本进行模拟仿真,本研究采用ART模拟软件进行实验数据模拟。使用模拟所得的实验数据进行数据导入测试。首先对数据导入的实时效率进行测试,对插入数据的耗时进行测定,采用多线程导入方法进行测试,得到测试结果如图2所示。

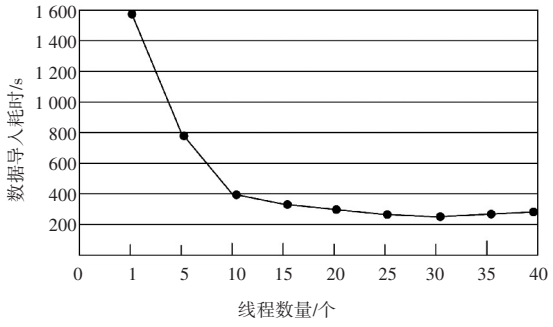


图2 数据导入测试

Fig. 2 Data import test

根据图2可知,本次实验所使用的数据导入方法所产生的数据导入耗时结果满足本次实验的需求,并且以数量为30个线程时,耗时达到最短,基于此结果,本次实验采用30个线程的方法进行数据导入。利用线路连接将实验数据传输至本次实验所使用的计算机中,如图3所示。

如图3所示,本次实验所使用的计算机包括两台不同型号的计算机。为应对海量的电力分布式数据,确保数据处理和检索过程的效率和准确性,实验选取操作系统Centos7.4、内存64 GB、CPU为Intel Xeon E5-2620V4的主控计算机,5.7.1MySQL、硬盘为4TB 7200 r/min、内存为128 GB的主控计算机作为实验计算机设备。完成实验准备后,在上述环境中开展本次电力海量分布式数据智能检索模拟仿真测试。



图3 实验环境照片

Fig. 3 Experimental environment photo

2.2 电力海量分布式数据智能检索

对本文所提出的电力海量分布式数据智能检索进行可行性测试。对本文所提出的检索方法进行可视化实现。采用[0,0,0,1 000,1 000,1 000]物联网节点的实验模拟数据进行测试实验。在该物联网节点的实验数据中随机筛选出部分数据对本文所提出的数据智能化检索方法进行测试,检索语句包括用电量、营销数据、窃电行为数据等多个检索关键词。经过测试,得到测试结果如表1所示。

表1 数据值智能化检索

Tab. 1 Intelligent retrieval of data values

检索语句量	总用时/ms	条用时/ms
1条	4 623	0.054 9
3条	7 356	0.097 5
5条	16 596	0.341 8
10条	34 658	0.954 3
15条	89 667	1.678 2
20条	156 989	2.684 4

由表1可知,应用本文提出的数据智能化检索方法,所得出的数据检索用时随着检索语句关键词条数的增加而逐渐增加,说明本文所提方法对于数据检索的查询较为全面,检索得出的结果可信度较高,表明了本文所提方法的可行性。从这一检索结果可以初步判断,本文所提出的电力海量分布式数据智能化检索方法具备较好的应用效果,表明了方法的可行性。

2.3 结果评价指标

为了更加直观地体现出本研究所提方法的有效性,将本次模拟仿真实验的结果评价指标设置为检索结果的召回率。利用公式对检索实验结果的描述,进行召回率的计算如下:

$$R = \frac{\vartheta}{\vartheta + \psi} \tag{7}$$

式中,R 表示检索结果的召回率;ϑ 表示检索结果的真正例;ψ 表示检索结果的假正例。

召回率的计算结果直接反映了电力海量分布式数据智能化检索结果,在所有的数据样本中检索正确的数据量,表明了检索结果的正确度。计算所得的召回率结果数值越高,则表明该方法得出的数据智能化检索结果可信度越高,具备更强的实践应用性。

2.4 结果分析与讨论

为了体现本研究所提方法实验结果的有效性,分别应用文献[1]和文献[2]所提出的数据检索方法作为本次模拟仿真检索实验的对比方法,与本文所提出的数据智能化检索方法进行召回率结果对比,开展本次电力海量分布式数据智能化检索实验。

在本次检索实验中,为使不同检索方法所导入的物联网节点数据来源及数据量均保持一致,数据导入线程始终采用 30 个线程,保证本次检索实验结果不被其他实验变量所干扰,维持变量唯一,确保实验结果的科学性。同时,为了减少实验误差,本次实验共进行 5 个轮次,最终的召回率结果取 5 个轮次实验结果的平均值。

经过实验,得到不同数据检索方法的召回率结果如图 4 所示。

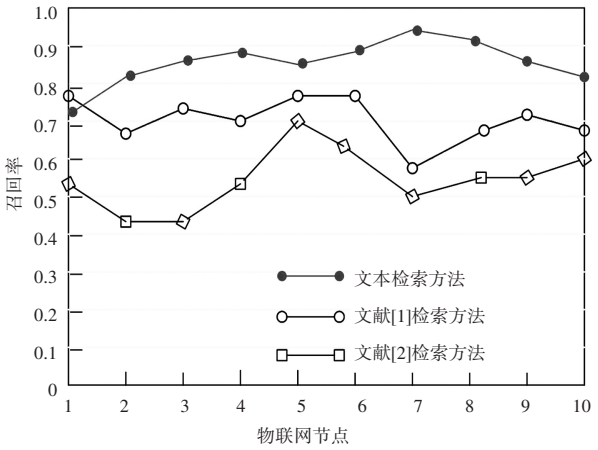


图 4 不同检索方法召回率结果

Fig. 4 Recall results of different retrieval methods

由图 4 可知,在本次检索实验中,本文所提方法的检索结果召回率较高,最高能够达到 0.93,最低为 0.72,而文献[1]和文献[2]所提方法的最高召回率分别为 0.78 和 0.70,最低召回率分别为 0.59 和 0.43。在多个物联网节点的数据智能化检索的召回率结果中,本文所提方法的平均召回率为 0.88,其他两种方法的平均召回率结果分别为 0.67 和 0.56,由此可见,本文所提方法能够大幅度提高检索结果的召回率。

从对比实验结果可以看出,本文所提出的电力海量分布式数据智能化检索方法能够有效应对电网系统中,分布式数据的检索实践。数据智能化检索结果的平均召回率较高,说明该方法的检索结果与检索目标的一致性较高。在电力数据检索实践工作中,将运维人员需要的数据信息精准呈现,为工作人员的日常工作提供高效的数据支持,具备较强的应用性。

3 结论

对海量的分布式数据进行智能化检索对于电力系统的日常工作起着十分重要的作用。对此,本研究提出一种电力海量分布式数据智能检索。经过实验可知,该方法检索结果的召回率较高,表明本研究内容能够有效提高数据检索的精准度,助力电力系统的日常运维工作实践,推动电力技术的应用与发展。

参考文献

[1]谭凯中,秦勃,何亚文. 面向过程的海洋时空数据分布式存储与并行检索[J]. 中国海洋大学学报(自然科学版),2021,51(11):94-101,134.
[2]王慈枫,邹自明,胡晓彦,等. 一种面向动态地球磁层的数据组织模型和高效检索[J]. 地球物理学进展,2022,37(4):1492-1499.
[3]魏秀卓,赵慧南. 基于 Mapreduce 的多源多模态大数据检索方法研究[J]. 计算机仿真,2021,38(4):422-426.
[4]吴义三,肖红,陈希球. 船舶岸基通信网络动态数据自动检索方法[J]. 舰船科学技术,2021,43(24):133-135.
[5]李盼盼,赵浩,林慧恩. 基于相似树查询的隐私大数据定向检索算法[J]. 计算机仿真,2021,38(11):429-432,437.
[6]王宏志,燕钰. 深度学习驱动的跨模态数据检索[J]. 哈尔滨理工大学学报,2021,26(1):9-16.
[7]宁海涛. 基于分块主色法的图像无序激增数据检索仿真[J]. 计算机仿真,2021,38(1):375-379.
[8]曾飞,杨雄,苏伟,等. 基于区块链与数据湖的电力数据存储与共享方法[J]. 电力工程技术,2022,41(3):48-54.
[9]招景明,张捷,宋鹏,等. 一种高效的基于云边端协同的电力数据采集系统[J]. 电网与清洁能源,2022,38(5):49-55.
[10]崔阳阳,赵洪山,曲岳晗,等. 基于残差 U 型网络的低压台区电力缺失数据补全方法[J]. 电力系统自动化,2022,46(9):83-90.
[11]谢清玉,张耀坤,李经纬. 面向智能电网的电力大数据关键技术应用[J]. 电网与清洁能源,2021,37(12):39-46.
[12]严康,陆艺丹,覃芳璐,等. 配电网用户侧异构电力物联设备网络风险量化评估[J]. 电力系统保护与控制,2023,51(11):64-76.
[13]杨维永,刘苇,崔恒志,等. SG-Edge:电力物联网可信边缘计算框架关键技术[J]. 软件学报,2022,33(2):641-663.
[14]宋亭. 基于决策树的图书馆文献资源智能检索方法[J]. 自动化技术与应用,2024,43(4):63-66.
[15]朱彦名,徐潇源,严正,等. 面向电力物联网的含可再生能源配电网运行展望[J]. 电力系统保护与控制,2022,50(2):176-187.
[16]赵鹏,蒲天骄,王新迎,等. 面向能源互联网数字孪生的电力物联网关键技术及展望[J]. 中国电机工程学报,2022,42(2):447-458.