计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2021,30(2):97−102 [doi: 10.15888/j.cnki.csa.007796]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于改进决策树的配电网多源数据快速检索①

柯　强[1], 陈志华[1], 胡经伟[1], 陈焕军[2], 邳志旺[2], 张　晗[2], 周雪松[3]

[1](国网黄冈供电公司 经济技术研究所, 黄冈 438701)
[2](天津楚能电力技术有限公司, 天津 300392)
[3](天津理工大学 电气电子工程学院, 天津 300384)
通讯作者: 陈志华, E-mail: chenzhihua_hg@163.com

摘　要: 当前, 电网中含有海量的多源信息数据, 但是由于数据体量大、种类多、维度高, 难以实现高效有效的数据检索. 因此本文根据实际电力运行系统的数据结构及多源数据库样本分析, 提出了一种基于互信息的改进决策树算法作为数据挖掘内核, 并提出适用于电力系统的并行处理架构, 可实现多源数据的快速、有效信息检索, 并有效处理实时数据. 在搜索时根据代表性特征子集直接从多源信息原始数据提取信息, 判断索引信息量并排序形成决策树模型, 通过 Spark MapReduce Python 数据分解并行检索实现多源数据同时提取, 缩短检索时间. 本文以某区域电网数据库为算例进行模拟验证, 结果表明: 该方法可以实现配电网的多源异构信息提取, 有效避免重复数据, 满足在线工程决策要求.

关键词: 决策树; 并行计算; 信息检索; 多源异构

引用格式: 柯强,陈志华,胡经伟,陈焕军,邳志旺,张晗,周雪松.基于改进决策树的配电网多源数据快速检索.计算机系统应用,2021,30(2):97–102. http://www.c-s-a.org.cn/1003-3254/7796.html

Fast Multi-Source Data Retrieval Method for Distribution Network Based on Improved Decision Tree

KE Qiang[1], CHEN Zhi-Hua[1], HU Jing-Wei[1], CHEN Huan-Jun[2], PI Zhi-Wang[2], ZHANG Han[2], ZHOU Xue-Song[3]

[1](Economic and Technical Research Institute, State Grid Huanggang Power Supply Company, Huanggang 438701, China)
[2](Tianjin Chuneng Electric Power Technology Company, Tianjin 300392, China)
[3](School of Electrical and Electronic Engineering, Tianjing University of Technology, Tianjin 300384, China)

Abstract: At present, the power grid contains a large number of multi-source information data, but due to the large size of the data types and high multi-dimensions, it is difficult to achieve effective data retrieval.According to the data structure of actual power operation system and multi-source database sample analysis, an improved decision tree algorithm based on mutual information is proposed as the kernel of data mining, and a parallel processing architecture suitable for power system is put forward, which can retrieve multi-source data fast and efficiently. The information is directly extracted from the original data of multi-source information according to the representative feature subset during searching. The index information is judged and sorted to form the decision tree model, and multi-source data is extracted simultaneously through Spark MapReduce Python data decomposition and parallel retrieval, so as to shorten the retrieval time. Taking a regional power grid database as an example to simulate and verify, the results show that the method can realize multi-source heterogeneous information extraction of power distribution network, effectively avoid duplicate data, and meet the requirements of online engineering decision.

Key words: decision-making tree; parallel computing; information retrieval; multi-source heterogeneous

① 基金项目: 国家自然科学基金 (51877152)
Foundation item: National Natural Science Foundation of China (51877152)
收稿时间: 2020-06-25; 修改时间: 2020-07-27; 采用时间: 2020-08-10; csa 在线出版时间: 2021-01-27

随着智能电网的不断建设, 电网中运行和维护所产生的数据量呈指数形式增长, 电网数据不断增加, 电力行业开始进入大数据时代[1]. 电力大数据除了包含大数据的广义 4V 特征, 即数据量庞大 (Volume)、数据类型多 (Variety)、数据变化速度快 (Velcoity)、数据价值密度不高 (Value) 的性质, 还携带了能源行业的特有印记, 包含大量多维时空数据、关联关系数据以及实时响应数据等. 此外, 这些电力数据分别集成在不同的信息管理系统上, 如贯通调度管理系统 (Outage Management System, OMS)、生产管理系统 (Production Management System, PMS)、地理信息系统 (Geographic Information System, GIS)、用电信息采集系统等. 而不同管理平台之间数据不能相互兼容, 并且它们之间还含有大量的重叠数据. 另一方面, 这些多源数据也存在互补关系, 其中蕴含着丰富的电力运行信息, 如文献 [2] 提出基于多源数据的线路保护通道及故障定位方法等, 采用多源数据获得更多电网有效信息. 因此如何利用现有信息管理系统的信息, 快速准确的检索所需信息是配电网多维数据融合管理系统建设的基础和关键.

目前正处于泛在电力物联网建设的关键时期[3], 这需要更加精确、高效和个性化的电网多源数据检索. 因此急需建设适合电网数据特点的信息检索方法. 文献 [4] 以电网事故信息为基础形成数据仓库, 并利用分类与回归算法进行检索. 文献 [5] 针对电力数据多维度的特点, 提出了基于流形排序的电网截面数据检索方法. 由于电网数据类型复杂, 文献 [6] 设计一种基于 B+树及倒排索引的双层混合索引结构可同时对字符型及数值型数据进行检索. 文献 [7] 分析了海量电网状态监测数据管理平台结构与功能, 提出基于 MapReduce 的海量数据检索方法. 文献 [8] 首先采用模糊特征分组聚类方法对电力数据进行分组并提取特征向量, 然后使用云计算技术实现分布式检索. 目前电力系统数据检索技术大多直接从大量数据中检索出满足用户查询需求的记录, 消耗时间长且精确度不高, 并且上述大多数方法仅适合文本数据和 Web 数据检索, 不适用于含多源数据的电力系统.

决策树方法是一种适用于数据分类、检索的方法, 能保证检索精度的同时, 提高信息检索速度. 目前常用的决策树算法主要是 Quinlan 在 1986 年提出的 ID3 算法[9], 它采用信息熵作为判断分类的依据, 通过衡量系统的有序程度来进行区分. ID3 算法选择信息最大的属性来对样本进行分割, 可以提高算法的速度和精度, 但是它以信息增益作为判断标准, 更倾向于选择具有更多值的属性. 除此之外, 还有 C4.5[10], SPRINT[11], PUBLIC[12] 等改进算法, 它们在一定程度上弥补了 ID3 算法的不足, 可以处理更多的实际问题.

基于上述分析, 本文根据实际电力运行系统的数据结构及多源数据库样本分析, 提出了采用基于互信息的改进决策树算法进行快速信息检索. 通过该算法根据代表性特征子集对数据进行分类, 直接从多源信息原始数据提取信息, 并通过并行多任务处理的方式多源数据同时提取, 可以有效处理实时数据, 实现配电网的多源异构信息提取. 最终, 基于该算法提出了电力系统应用的电网辅助决策平系统模型构建, 并在仿真数据库中进行了验证.

1 改进的决策树信息检索算法

决策树算法是经典的数据挖掘算法, 其算法时间复杂度较低, 分类速度快, 可以适用于海量数据的快速检索分类. 对于数据分类检索而言, 输出结果的准确性和完整性至关重要. 而决策树方法是树形结构形式的分类器, 通过一系列的分类规则, 实现数据分类, 对多元数据分类具有较好的效果.

决策树算法从信息量最大的根节点开始, 按每个样本的属性作为不同的分类节点 (子节点), 将不同属性值作为不同分支, 直到当前节点属于同一类或相同属性值为止. 决策树算法中的属性排序将大大影响决策树的分类效果及速度, 因此需要按照某种度量将属性进行排序, 进而保证决策树算法的效果. 目前, 决策树算法在图像辨识、故障诊断等多个方面获得了广泛的应用. 本文结合配电网数据信息特征, 对决策树算法进行修改并构建一套检索系统.

在构建决策树分类模型时, 最重要的问题是建立一个高效的属性评价系统. 对于多源异构的电力系统数据库来说, 存在了大量的冗余数据和互补数据, 需要更加有效的分割方式, 仅通过信息熵作为分类标准对于一些情况不够鲁棒, 会识别出大量无效数据, 难以达到应用要求. 本文在此提出一种基于互信息适用于电力系统多源异构的改进决策树算法, 可以有效解决重叠数据分割.

与信息熵相似, 互信息也是由信息论的概念衍生而来, 它可以表示两个变量之间相互依赖性的度量[13]. 信息熵可以从原始数据中选择一个有代表性的特征子集, 直接从原始数据中提取出需要的信息, 但需要满足

一些条件[14]. 然而, 互信息利用互信息判断不同属性之间的相互包含关系, 选择低冗余特征子集, 在数据挖掘领域的特征选择方面具有更加突出优势. 同时, 基于互信息的决策树的构造过程更加直观. 但是, 互信息也倾向于选择多值属性, 为此本文增加权因子, 平衡各不同类别.

当样本集的一个属性均匀分布在所有类别中, 则与类别的互信息为 0, 说明该属性与类别的关系较弱. 如果一个属性在不同类别的分布上有显著的差异, 那么它们之间就会有大量的相互信息, 说明属性和类别之间存在显著的关系. 通过计算类别和不同属性之间的相互信息, 可以实现最优属性分割. 对于一个样本属性, 其与类别的相关性可以表示为互信息:

$$I(x,y)=\log\frac{p(x_i,y_i)}{p(x_i)p(y_i)'} \tag{1}$$

其中, $p(x_i)$表示属性x的值为i的概率; y为样本类别, $p(y_j)$表示类别为j的概率; $p(x_i,y_j)$为属性为i类别为j时, 属性x与类别y的联合概率. 当互信息越大, 属性与类别的相关性越大. 在计算时需要考虑两个相关变量的分布概率, 因此采用平均互信息, 并增加权因子$1/C$, 其由各类数决定:

$$C_i^{-1}=\exp\left(-\frac{n_i}{\sum n_i}\right)m_0 \tag{2}$$

其中, n_i为属性i的数据量; m_0为人工常量, 根据问题对权值进行微调.

最终本文的互信息定义为:

$$MI(x)=\frac{1}{C}\sum_{i,j}p(x_i,y_i)\log\frac{p(x_i,y_i)}{p(x_i)p(y_i)'} \tag{3}$$

其中, $1/C$为权因子, 由各类数决定; $MI(x)$表示属性x与类别之间的互信息.

通过基于互信息的决策树模型可以实现数据分类和数据筛选, 数据筛选算法流程具体如下:

输入: 候选数据集 (D 个数据)、索引关键词 (S 个).

步骤 1. 根据索引结构构建决策树模型 T.

步骤 2. 由式 (2) 更新权值矩阵 C.

步骤 3. 计算$MI(x)$, 降序排序索引.

步骤 4. 筛除不相关数据, 得到子分类数据D_n.

步骤 5. 精简不相关分支, 得到精简决策树T_n.

步骤 6. 迭代计算, 重复 2~4 次, 得到最终数据集 *Output*.

在检索过程中, 数据互信息越大与可能筛除更多无用数据的概率成正比, 互信息排序越靠前表明是查询的可能性越大, 可以大大减少不相关信息, 提高检索速度和准确率. 此外, 每次迭代过程中都不断精简决策树, 可以进一步提升计算准确度, 确保在当前数据集下得到最佳排序.

本文提出的算法的计算效果与数据自身属性也有很大的关系. 当数据具有确定的分类属性时, 比如本文中的电力系统数据, 根据每个属性的分类结果进行筛除时不会将有效信息进行错筛, 因此可以保证检索信息的准确性, 进而不会错误的筛除相关数据, 可以保证最终输出结果的完整性. 但是当数据属性分类不确定时, 每次的分类结果难以达到百分之百的准确, 而且输出结果对准确率要求较高, 所以如果仍采用步骤 4 将可能会将部分数据错误筛除. 此时可以省略步骤 4, 以保证所有的数据完整性, 但是相应会增加计算负担.

2 并行计算

电力系统所产生的数据可以区分为静态数据和动态数据, 其中在较大时间尺度中数据不发生变化时, 可以认为是静态数据 (例如, PMS、GIS 等数据库), 除此之外, 在小时间尺度中不断更新或者累积的数据称为动态数据. 电力系统中数据庞大, 尤其是对于动态数据上千节点的数据采集会造成巨大的数据累积. 动态数据的处理对于能否有效挖掘关键数据至关重要, 但许多算法直接在一定时间内忽略最新的动态数据, 而采用历史数据, 这对于实时变化的电力系统来说可能会影响巨大. 在此, 我们通过技术处理实时更新的数据以实现动态数据的挖掘. 此外, 不同的关键词检索, 也将对静态数据的数据处理产生不同的要求. 这些对于计算机的要求将大大增加.

决策树算法的时间及空间复杂度均为$O(2^n)$, 计算时间和内存量随数据和索引量增长而急剧增长. 尤其处理大数据时, 采用单进程对数据进行处理会速度缓慢, 浪费大量计算机资源. 因此, 使用并行处理方法将算法并行化十分必要, 对挖掘进行加速, 对计算资源实现充分的利用. 同时处理多个任务主要有进程分支和线程派生的实现方式, 在此我们采用线程派生, 相比于进程分支可以提升计算效率, 线程同步易于控制.

本文提出的基于 Spark MapReduce 的并行决策树算法是由多个 map (映射) 以及 reduce (归约) 函数组组成, 并支持转换 (transformations) 和行动 (actions), 它们的实现基于 RDD. 多个数据行组成一个 RDD, 数据行

的内容可以是数字、数组或者是混合类型的数据. Spark MapReduce 将计算资源分为一个 master 节点和多个 worker 节点, master 向空闲的 worker 分配工作.

多源数据可以通过各数据库实现物理分割, 但是对于大型数据库中还需要进行分块. 在一个完整的任务中, master 先将数据分块 (split 0–6),然后将生成的决策树分别赋值给不同 worker. 被分配了 map 任务的 worker 获取并提取数据行, 接受输入数据, 通过对中间变量键值对的操作, 并暂存到内存中; master 向分配了 reduce 任务的 worker 通知缓存信息, 将键相同的键值对发到不同 reduce 函数中, 进而归纳合并生成简化键值对, 并生成最终计算结果.

如图 1 所示, 即为整个数据挖掘引擎并行工作的流程图, 通过对各个数据库分类并行的进行数据获取和解析并通过决策树进行数据挖掘, 可以大大的提高数据处理效率并降低计算时间.

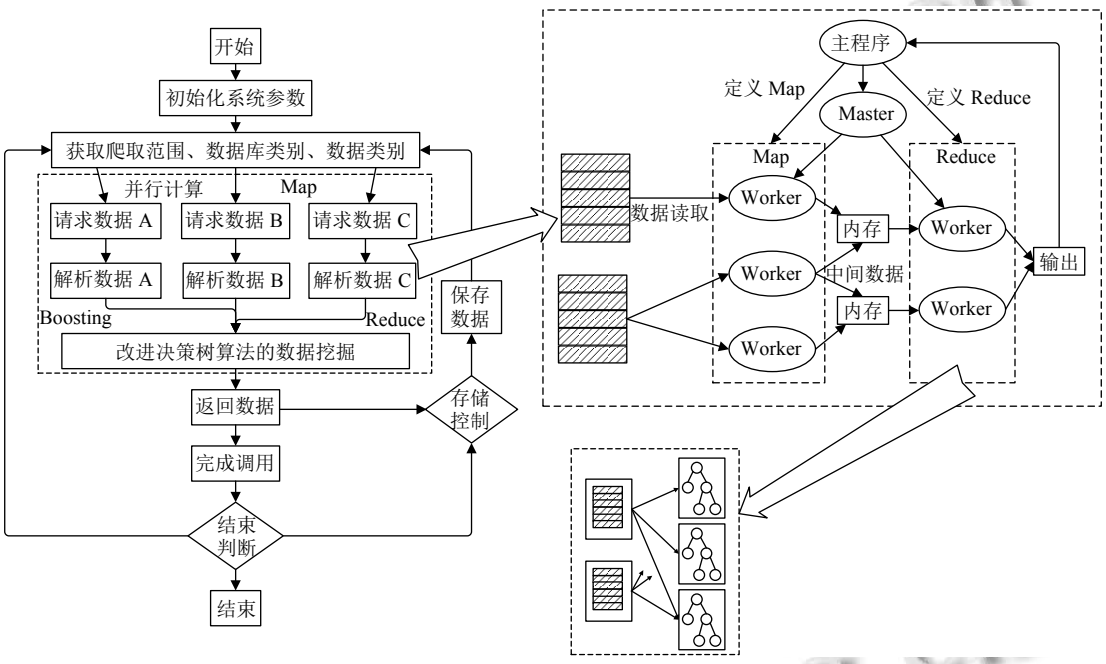

图 1 数据挖掘并行计算流程图

3 基于决策树算法的电力系统信息检索应用模型

基于互信息的决策树算法能有效地度量属性与不同数据库数据之间的关联, 区分不同属性对分类的重要性. 通过增加权因子, 可以对在实际复用情况极好的数据库, 在分类策略中适当考虑增加其权重, 提高检索效率及准确率. 根据电力系统的实际需求, 提出了如图 2 所示基于决策树算法的电网辅助决策系统应用模型.

在此我们考虑了电力系统中的多个异构数据库, 包括调度管理 (OMS)、生产管理 (PMS)、地理信息 (GIS)、用电信息采集等系统, 还有监控和数据采集系统 (Supervisory Control And Data Acquisition, SCADA)、同步相量测量装置 (Phasor Measurement Unit, PMU) 等实时数据.

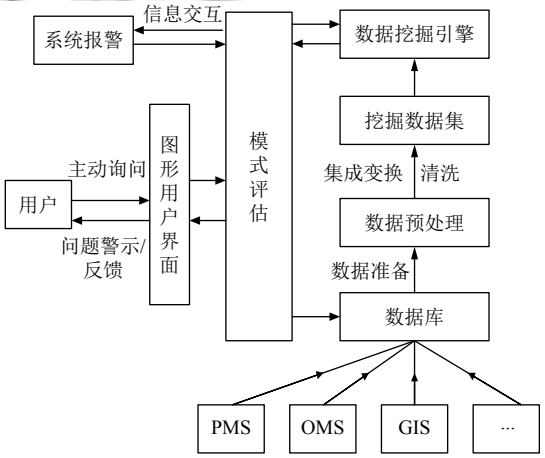

图 2 基于决策树算法的电网辅助决策系统应用模型

该模型构架可以实现对用户主动问询的关键词进行快速检索和数据挖掘, 结果经过校验后反馈给用户,

如果用户不满意可以追加搜索或进行调整, 实现对电网运行的辅助决策; 此外, 电网运行出现警告时, 可以通过该系统获得除了实时测量系统和故障预测系统之外的数据, 实现多源数据共用. 下面是每个部分的运行功能:

(1) 数据处理

由于数据库中往往存在空缺数据和不一致数据等, 因此需要对数据进行预处理, 提高数据质量. 数据清理可以处理掉遗漏数据和异常数据, 数据集成将多源数据构成统一数据集, 数据变换可以规范化数据成一个适合的描述形式. 通过数据处理得到挖掘数据集以便开展进一步的数据挖掘工作.

(2) 数据挖掘

数据挖掘部分是整个系统的核心部分, 其由一组功能模块构成, 包括数据挖掘、数据查询、数据存储等. 本挖掘任务具有明显的实时性、准确性要求, 为此我们将互信息决策树与 Spark MapReduce 相结合, 计算速度更快、树的规模更小. 对于实时数据我们采用 Boosting 的技术, 可以对新加入数据产生新决策树并形成决策树集合, 增强数据挖掘的深度.

(3) 模式评估

对于最终数据有效性的验证与评估必不可少. 首先, 通过数据库数据反复比对, 保证结果的准确性. 此外, 在一些模糊条件下, 难以得到满意的结果, 为此数据也交由用户进行修改和评价, 并不断返回修正, 直到得出用户满意的结果.

(4) 用户交互

用户可以主动发起数据检索, 设置数据要求, 同时也可以对得到的结果进行修正并进一步完善检索要求, 最终在交互界面中得到最终结果.

(5) 系统主动问询

除了用户主动咨询数据辅助决策外, 我们还留有系统接口, 该系统可以有效的筛除冗余数据, 提高数据集成度和利用率, 为此, 可以对现有故障预测系统、调度系统进行数据支撑.

4 仿真分析

4.1 改进决策树算法性能分析

为了验证改进决策树算法的效果和一般适用性, 我们针对几种经典数据集与经典的 ID3 算法进行对比, 结果如表 1, 其中 m_0 为本文算法的人工常量.

从结果中可以看出, 与 ID3 算法相比, 本文提出的基于互信息的方法更能提高分类的精度, 其具有更低的冗余度. 对于 ID3 算法无论是精度较低的数据集 (如 Car) 还是精度较高的数据集 (如 Krvs), 改进方法都可以有效提升准确率, 说明了本文提出方法具有普遍性和精确性.

表 1 ID3 算法与本文算法在不同数据集中的精度对比 (%)

数据集	ID3算法	本文算法	
		$m_0 = 1$	m_0灵活调整
Krvs	99.5982	99.6721	99.7154
Car	90.0226	93.9170	94.2674
Solar-flare1	94.6145	95.8461	96.1193

需要特别说明的是, 本文所提算法中的权因子中的人工常量需要根据数据集本身进行微调, 通过表格对比可以看出, 权因子的高效选取会对精度提升具有较大帮助. 该部分的目的是针对不同的数据集特点, 将资深行业从业人员的多年经验充分融合到算法中, 例如在本文第 3 部分中信息检索模型中加入了用户交互检验, 可以有效自行调整人工常量. 但是当人工保持恒定时, 仍然对比 ID3 算法有一定精度提升, 对于一般行业从业者也可以有较好的效果.

4.2 电力系统信息检索应用模型分析

为验证本文提出模型的实用性和鲁棒性, 在此模拟某北方区域电网的实际情况, 使用本文所提方法进行仿真分析. 在实验室环境下搭建平台, 模拟电网中包含 3 个数据库 (GIS、PMS、SCADA) 的历史数据, 并模拟 1 路实时的数据输入, 用户检索某变电站 A 主变 #2 异常信息, 时间限值为 3 个月. 如图 3 所示, 为在该检索关键词下对 3 个数据库的决策树的信息检索过程.

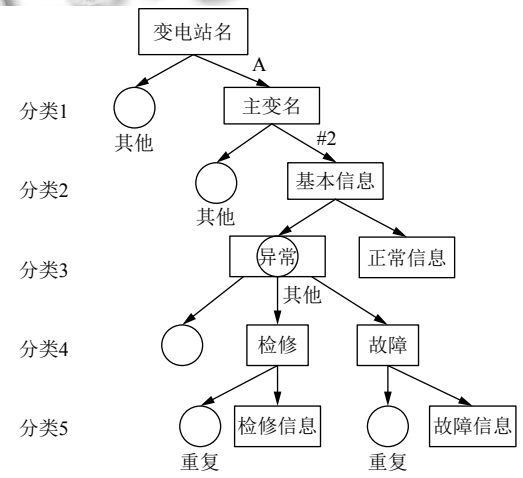

图 3 决策树检索实现流程

从图 3 中可以看出, 基于互信息决策树的检索方法对于分类结果是一种从下而上的搜寻方法. 先将大

量的数据筛除, 可用提高数据处理效率, 最终的重复分类可以筛除冗余数据. 表 2 和表 3 是检索 PMS 数据库中 2160 条数据的检索结果, 通过互信息决策树的分类和对各个分类器中数据的验证分析, 最后获得可用数据 6 个. 对 PMS 数据库的检索分析在 Inter CORE i7 & RAM 8 GB 的台式机上验证, 花费时间 0.0125 s.

表 2 PMS 数据关键词互信息结果

属性	变电站名	变压器名	基本信息	异常/故障
变电站名	0.883	0.654	0.549	0.301

表 3 PMS 数据各层分类结果

	分类1	分类2	分类3	分类4	分类5
分类器中数据量	20	15	12	10	3 & 5
可用数据量	1	3	5	5	2 & 4

针对 3 个数据库的数据进行检索, 并不断提高数据量级别, 检索花费时间如表 4 中所示. 在低量级时, 并行计算效果不明显, 随着数据量不断提升, 基于 Spark MapReduce 的并行计算优势不断显现, 平均计算时间不断降低. 仿真实验验证了该模型更适合海量电网多源异构数据的检索, 充分说明了基于互信息决策树的分类方法的适用性, 并行计算的可行性和交互计算时间的迅速性.

表 4 多源海量数据仿真时间对比

序号	检索条数(万条)	花费时间(s)	序号	检索条数(万条)	花费时间(s)
1	0.8	0.0338	4	32.1	2.1147
2	1.58	0.0717	5	63.4	4.1662
3	11.73	0.6094	6	130.6	7.8231

5 结论

如何有效利用电力系统中的海量多源异构数据来帮助运行人员辅助分析决策是一个难题, 目前只能通过单一数据库进行判断, 可能造成信息不完整. 针对不同数据库之间存在信息冗余、数据结构不同、数据处理量大、难以迅速提取等问题. 本文得到以下结论:

(1) 提出了一种基于改进决策树的信息检索算法, 通过互信息判断索引信息量, 对多源数据关键词快速分类检索.

(2) 通过并行计算处理实现了多源数据库并行处理, 可以有效接纳处理实时数据, 大大提升计算效率.

(3) 基于本文算法, 提出了面向电力系统应用的电网辅助决策平系统模型, 可以实现对当前电网快速信息检索以辅助用户决策.

(4) 本文所提方法在百万级数据量的仿真系统中得到了验证, 在大数据量的系统仍可保持较高计算速度.

参考文献

1 薛禹胜, 赖业宁. 大能源思维与大数据思维的融合 (一) 大数据与电力大数据. 电力系统自动化, 2016, 40(1): 1–8. [doi: 10.7500/AEPS20151208005]

2 高鹏翔. 基于多源数据融合的配电网运行故障特征信息提取技术研究 [硕士学位论文]. 北京: 华北电力大学 (北京), 2019.

3 杨挺, 翟峰, 赵英杰, 等. 泛在电力物联网释义与研究展望. 电力系统自动化, 2019, 43(13): 9–20, 53. [doi: 10.7500/AEPS20190418015]

4 任锦标. 基于数据仓库及决策树算法的电网事故事件信息智能检索方法研究. 集成电路应用, 2019, 36(12): 86–87.

5 曲朝阳, 孙立擎, 潘峰, 等. 基于流形排序的电网截面数据检索. 科学技术与工程, 2016, 16(15): 239–244. [doi: 10.3969/j.issn.1671-1815.2016.15.043]

6 龙禹, 吴尚远, 高骞, 等. 基于 B+树的电力大数据混合索引设计与实现. 自动化与仪器仪表, 2018, (9): 67–69.

7 黄华林, 庞欣婷. 基于 Hadoop 的数据资源管理平台设计. 计算机应用与软件, 2018, 35(7): 329–333. [doi: 10.3969/j.issn.1000-386x.2018.07.059]

8 杜红军, 李巍, 张文杰, 等. 基于云计算技术的电力大数据分布式检索系统. 电网与清洁能源, 2018, 34(9): 19–24. [doi: 10.3969/j.issn.1674-3814.2018.09.004]

9 Quinlan RJ. Induction of decision trees. Machine Learning, 1986, 1(1): 81–106.

10 Quinlan RJ. C4.5: Programs for machine learning. San Mateo: Morgan Kaufmann Publish, 1993.

11 Shafer JC, Agrawal R, Mehta M. SPRINT: A scalable parallel classifier for data mining. Proceedings of the 22th International Conference on Very Large Data Bases. Bombay, India. 1996. 544–555.

12 Rastogi R, Shim K. PUBLIC: A decision tree classifier that integrates building and pruning. Data Mining and Knowledge Discovery, 2000, 4(4): 315–344. [doi: 10.1023/A:1009887311454]

13 Tang PS, Tang XL, Tao ZY, *et al*. Research on feature selection algorithm based on mutual information and genetic algorithm. Proceedings of the 2014 11th International Computer Conference on Wavelet Active Media Technology and Information Processing. Chengdu, China. 2014. 403–406.

14 Ding C, Peng HC. Minimum redundancy feature selection from microarray gene expression data. Journal of Bioinformatics and Computational Biology, 2003, 3(2): 185–205. [doi: 10.1142/S0219720005001004]