Research and Implementation of Intelligent Report Search Technology*

ZHANG Fan, WANG Honggang, YANG Zhi*
(*Big data center of State Grid Corporation of China, Beijing* 100033, *China*)

Abstract: In view of the weak links and key areas of report management, combined with the needs of report application, big data processing and artificial intelligence technology were introduced to improve the report processing efficiency, data quality and analysis ability. Based on the construction technology of ontology knowledge base, combined with the characteristics of power grid business, the ontology knowledge base of power system was established. According to the inference rules of power grid ontology, the storage and association relationship of the retrieved information in the ontology knowledge base were explored. Based on the operation data, equipment management, grid model and parameters of large power grid, this paper analyzed the storage, analysis and retrieval technology of data from structured and unstructured aspects. The results show that the intelligent report search technology has the ability of global, intelligent and transparent service, which effectively improves the efficiency of data asset discovery and utilization. It is suitable for the business areas with more report requirements and supports the application requirements of data management in various fields of the company.

Key words: report management; knowledge base; power system; inference rules; retrieval
EEACC: 8110B; 0290Z **doi**: 10.3969/j.issn.1005-9490.2020.05.021

智能化报表搜索技术的研究与实现*

张 帆*, 王宏刚, 杨 志
(国家电网有限公司大数据中心, 北京 100033)

摘 要: 针对报表管理的薄弱环节和重点关注领域, 结合报表应用的需求, 引入大数据处理和人工智能技术, 以改善报表的处理效率、数据质量及分析能力。基于本体知识库构建技术, 结合电网业务特性, 建立了电力系统的本体知识库, 根据电网本体的推理规则探究被检索信息在本体知识库中的存储、关联关系; 基于大电网运行数据、设备管理以及电网模型、参数等, 从结构化和非结构化两方面分析数据的存储、解析及检索技术。结果表明, 智能化报表搜索技术具有全局化、智能化以及透明化的服务能力, 有效提高了数据资产的发掘和利用效率, 适用于报表需求较多的业务领域中, 支撑公司各领域数据管理应用需求。

关键词: 报表管理; 知识库; 电力系统; 推理规则; 检索
中图分类号: V242.3 **文献标识码**: A **文章编号**: 1005-9490(2020)05-1066-05

电力信息系统数据生产来源多样化、数据体量大、数据分布广[1-2]。从物理结构上看, 数据之间呈现离散状态, 但是实际情况是电力系统数据在深层逻辑上存在着紧密的关联关系。而报表的作用是对分散在不同位置的数据, 按照特定的分析目标进行多源聚合, 并加工计算, 形成具有业务参考价值的数据集, 反馈电力系统运营状况, 辅助业务决策[3-5]。报表智能化报表检索技术能够辅助人工发现数据之间的关联关系, 快速生成具有丰富业务内涵的价值信息[6-8]。

国内外已有大量关于报表检索的相关研究。文献[9]指出大量复杂的手工报表无法满足信息技术发展, 传统意义上的报表受固定模式的约束, 不能满足医院的客户化需求, 而提出了应用交叉报表解决集成报表的缺陷, 并证明其能够为医院提供便捷有效且高经济效益的解决方案; 文献[10]应用微软 Excel 统

项目来源: 国家电网有限公司科技项目(5500/2019-83002B)
收稿日期: 2020-02-20 **修改日期**: 2020-05-11

计表自带的函数运算功能,提出了一种嵌套函数公式,能够自动查找特征值并按照用户需求输出相应的函数;文献[11]从传统报表结构出发,以用友插件为报表载体,以报表的元数据为基础,提出了模糊判断和精确判断两种检索机制。实践证明,该方法大大提高了报表全文检索的易用性以及可扩展性;文献[12]描述了基于 Windows 操作环境下的 Foxpro 软件设计开发的查新与文献检索数据库管理系统,介绍了该系统的特点,查询、浏览、统计、报表等数据库及数据库管理的方法;文献[13]提出了一种具有语义查询与检索、报表个性化设计和生成、权限管理等多功能的报表系统;文献[14]说明了信息检索和报表生成模块性能的高低和功能的完善与否对评价网络安全审计系统的影响,证明文章所提的信息检索和报表生成模块能够对一定时期内的系统所记录的审计和网络会话记录进行条件检索和信息筛选并生成统计报表;文献[15]借鉴开源报表设计软件 BIRT,开发报表模板库系统,使用户较快地完成制表工作,给出报表模板库存储与检索系统的系统结构,并详细阐述存储结构的设计,以及检索系统的算法与编码实现过程。

上述研究文献在不同领域对报表搜索相关便捷方法展开研究,具有一定的借鉴意义。本文对电力系统领域报表数据的检索技术展开研究。首先,基于本体知识库构建技术,搭建电力系统的本体知识库;接着,根据电网本体的推理规则对被检索信息在本体知识库中的存储、关联关系进行探究;最后,基于大电网运行数据、设备管理以及电网模型、参数等,从结构化和非结构化两方面分析数据的存储、解析及检索技术。

1　本体知识库的提出

知识作为实践认知和经验的一种表现形式,知识库由知识和知识处理方式两部分组成,多种知识库又构成了知识域,涵盖了基本规则、定义概念、不同的知识处理方法以及逻辑查询语言等[16-17]。本体作为一个能够共享的概念化范围,描述了某个领域的知识结构以及内在关系,如电网的知识阐述了不同电气设备数据间的潜在关系,依据大电网知识所建立的大电网本体知识库是描述电力系统领域数据含义及内在联系的基础。一般来说,在电力系统领域中,特定的电气设备、报表文件或是抽象的知识理论都可以被当做本体对象。

建立电力系统本体知识库,为电网运行数据查询做推理以及知识积累,基于电力系统本体知识库对电力系统本体做出分析,探究电力系统本体及其内在联系。电力系统本体知识库由知识本体定义以及本体间的关系两部分构成。其中,知识本体定义作为电力系统本体知识库的基础,是电力系统领域分析方向和研究问题的抽象概述;而本体间的关系,则是阐述电力系统中各本体的相互关系。电力系统本体知识库的作用主要表现为以下两点:

(1)能够对电力系统本体做出精准定义,谨防出现来自不同数据库信息的异构现象;

(2)能够进行知识本体的推理,本体具有一定的逻辑思维特性,基于本体对用户输入的检索语句进行联想扩句,能够进一步推断出其未能完整表达的需求。

2　电力系统本体知识库的构建

本文对当前电网运行中存在的不同数据进行分析,将这些数据划分为结构化以及非结构化两种。结构化数据包括电网设备、电网运行数据以及业务流程等;非结构化数据包括电网运行期间的采集图像、视频等。如图 1 所示,为电力系统本体知识库结构图。

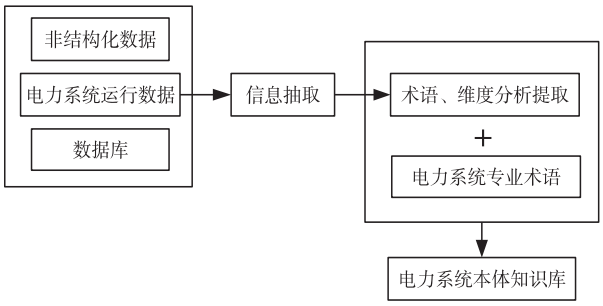

图 1　电力系统本体知识库结构图

在图 1 中,电力系统本体采用 Web Ontology Language(OWL)作为描述语言,其具有能够明确定义各术语的含义以及阐明不同术语间相关关系的功能[18]。如图 2(a)所示,应用 OWL 本体进行电力系统领域专业数据、设备及两者间关系的阐述,这段本体语句表达了“换流器”是“换流站”的子类这一概念;除此之外,OWL 还能够描述本体的基本属性,如图 2(b)所示,表示“额定电压”是“换流器”的基本属性;图 2(c)中 OWL 列举了一个“整流站”的本体实例“整流器”。

图 2(d)中对整流器的基本属性进行描述,指出“整流器”是归属于换流站的设备,还描述了整流器具有“额定容量”这一基本属性,此外还指明了换流站本体,如此一来,基于换流站能够找到整流器属性,也能够查询到其所属换流站,基于一个本体能够检索出与其具有关联性的其他本体及其属性,此外,还能具有迭代查询功能。

```
    < OWL：Class rdf：ID=“换流站” / >
    < OWL：Class rdf：ID=“换流器” / >
    < rdfs：subClassOf rdf：resource = “#换流站” / >
< / OWL：Class >
```

(a)

```
< OWL：ObjectProperty rdf：ID = “额定电压” >
    < rdfs：domain rdf：resource =“换流器”/ >
< OWL：ObjectProperty >
```

(b)

```
< OWL：Class rdf：ID=“整流站”/ >
< OWL：oneOf rdf：parseTypeID=“Collection”/ >
    < OWL：Thing rdf：about =“整流器”/ >
< OWL：oneOf >
< OWL：Class >
```

(c)

```
< Measurement>
< observedSubject rdf：resource = “整流器” / >
< observedphenomenon rdf：resource = “换流站” / >
< observed Value >
    < Quantity >
< quantityUnit rdf:resource = “额定容量” / >
    < quantityValue rdf:datatype = “&xsd;float” >65
< / quantityValue>
    < / Quantity >
< / observedValue >
< / Measurement >
```

(d)

图 2 OWL 语言描述

3 电力系统本体知识库的推理规则

在搭建好电力系统本体知识库后,需要基于知识库来检索本体,找出不同本体间的特性关系,以获取检索结果实现客户需求。为此,本文结合电网业务对 OWL 本体的描述语句以及推理规则的实现方式进行分析,基于搜索关键词在本体知识库中进行本体查询及推理,以探究满足深层次检索需求的方法。

已知电力系统知识库能够提供关系推理和属性查询推理两种推理服务[19]。将本体推理结果用 Q 表示,本体属性描述为 P,本体描述为 O,内在联系描述为 R,推理条件描述为 C,不同属性的权重占比为 W。

(1)本体间关系推理,如推理结果由本体、内在联系以及推理条件三部分组成,进行推理后能够获得这个本体中涉及特定联系的所有本体及其描述信息。假设推理条件为“额定容量 = 65”,则能够获得“换流站”本体以及包括该本体的属性信息。若推理条件为“额定容量 = 80”,则无法找到满足该推理条件的本体,则无返回结果。

(2)本体属性查询推理,如 $P=((O_1,\cdots,O_n),W)$,包括本体定位查询 $Q=(O_1,\cdots,O_n)$。$(O_1,\cdots,O_n)$ 为具有承接特性的本体,如:根据继承关系对换流站、整流站、整流器进行从大到小顺序排列,得到的检索路径有利于快速找到本体及其描述。这样能够快速找到本体及其描述。基于 P 可得到本体 O 按照权重 W 过滤后的属性,如可以获得整流站中整流器额定电压、额定电流等全部属性值。

4 算例应用

本文基于 Lucene 搜索引擎对智能化报表搜索技术进行研究,选择 Lucene 分词器 IKAnalyzer 对检索条件以及非结构化文档的信息切词,结合电力系统本体知识库,以电力系统本体及属性作为词库,为 IKAnalyzer 切词提供支持。

4.1 结构化数据

结构化数据所包含的电网设备、电网运行以及业务流程数据都是基于数据库进行存储的,为此,需要基于 SQL 语句对电力系统本体与其基本属性进行搜索。

在生成索引文件时,需提供各本体的数据类型属性,结构化数据应提供的属性主要包括:数据源,表/视图,字段名,检索条件等。基于所设定的检索关键词在索引文件中进行搜索,根据关键词出现频率作为权重,对满足要求的全部搜索结果进行依次展示。例如:输入检索条目“换流站整流站整流器的额定电压”,IKAnalyzer 则将检索条目分成“换流站”、“整流站”、“整流器”、“额定电压”这些电气专业词汇知识本体,接着基于索引文件对有关元数据的扩展属性进行检索。

如图 3 所示,对该结构化数据查询过程的解析为:首先,查找“整流站”对应的站点 ID,语句描述为:SELECT ID AS 站点 ID FROM 换流站运行管理表 WHERE 厂站命名 = ‘整流站’;接着,查找“额定电压”,因为额定电压是“换流器”的基本属性,因此将二者合并生成 SQL 语句:SELECT 额定电压 FROM 整流站参数表 WHERE 厂站命名 = ‘换流器’ AND 站点 ID = 站点 ID,分别执行这两个检索语句就能得到检索结果。

```
整流站：[ip, datasource, psidp, idpps, 整流
站管理表，厂站命名，整流站，ID，站点ID]
整流器：[ip, datasource, psidp, idpps,整流
器参数表，厂站命名，整流器，，]
额定电压：[ip, datasource, psidp, idpps,整流
站参数表，额定电压，，，]
```

图 3 结构化数据查询

4.2　非结构化数据查询

如图 4 所示,为非结构化数据索引文件的生成过程,要搜索电力系统运行期间生成的非结构化数据,智能化报表搜索技术首先基于非结构化数据的通用格式以及数据类型选取合适的 Parser,不同的 Parser 会根据文件类型的不同解析出相应的文本信息,如:HTML 解析器、PDF 解析器等;接着,通过 IKAnalyze 从解析出的文本信息中提取出与检索关键词有关的内容信息,如:检索关键词的频率、在文件中的位置等,需要说明的是,Lucene 搜索的核心是对被检索文件进行预处理,建立文件检索机制的过程,在处理信息量较大且文档信息稳定的条件下,建立索引能够有效提升检索效率;最后,IKAnalyzer 将获取到的有关信息写入 Index 文件中。

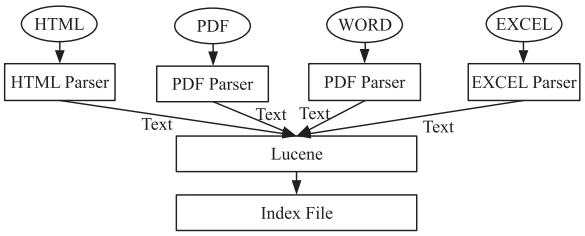

图 4　非结构化数据索引文件的生成

如图 5 所示,为基于电力系统本体知识库的索引生成方式。基于智能化报表检索技术的核心检索机制,找出电力系统及其有关知识本体,进而生成本体及其属性的描述信息。在索引文件中,描述文件的不同本体有且仅有唯一的元数据,每个本体均具备基本属性、辅助属性与 ID 描述。基于 Lucene 将本体描述文件转化为倒排索引以便于检索使用,生成的索引表中的每一项都包括一个属性值和具有该属性值的各记录的地址,Lucene 检索机制能够根据属性值找到相应的记录文件。

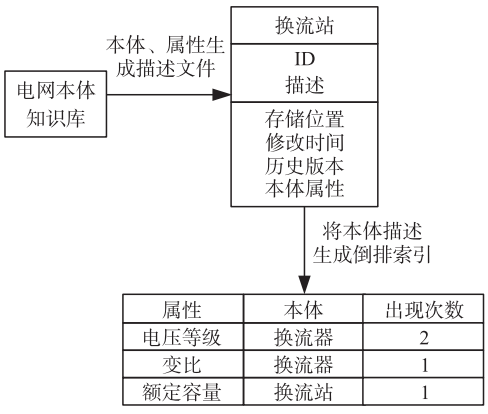

图 5　基于电力系统本体知识库的索引生成方式

4.3　应用算例

(1)智能化报表检索技术在电力系统运行数据检索的应用

电力系统运行过程所产生的数据量庞大,采用智能化报表检索技术能够有效提高电力系统运行管理的效率。如图 6 所示,为运行信息智能检索结果。通过电力系统本体知识库搜索关键词“2017 年 12 月 11 日”、“直流电压值”、“交流电压值”进行推理获取对应的本体以及与其具有关联性的本体,并结合结构化与非结构化检索技术对数据做出检索与展示。

图 6　运行信息智能检索结果

(2)智能化报表检索技术在电力系统事故分析中的应用

掌握电力系统常见故障以及对应的处理措施,提前做好事故预案,有利于降低电网故障率。但随着电力系统规模的不断扩大,电网运行情况日渐复杂,工作人员想要及时地获取相关信息,需要应用智能化报表检索技术快速获取所有电网事故的相关资料,如:事故原因,造成的结果以及故障处理措施等。

以图 7 所示的故障检索处理流程为案例进行说明,输入“2018 年台风山竹侵害电网情况”进行搜索,智能化报表检索技术通过对“台风山竹”与“侵害电网”两个本体进行推理,从而获取与这两个本体相关的一些知识本体,如:台风线路跳闸、短路故障以及断线故障等。把这些相关的知识本体当做深度搜索条件,在结构化与非结构化数据中进行深度

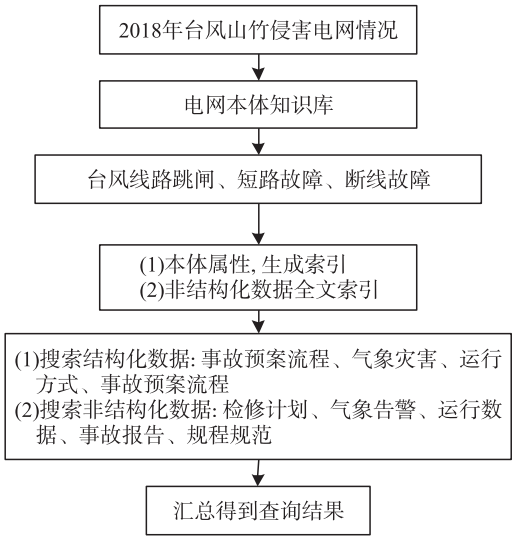

图 7　故障检索处理流程

检索,获取相应的报表检索结果。基于检索结果,用户能够掌握在历史电网事故的详细信息,通过分析检索结果提供的报表数据,能够为今后防范同类电网事故的发生提供经验指导,以增强电网的稳定运行能力,提升工作人员的业务水平和管理经验。

5 结语

应用报表搜索技术对分散在不同位置的数据,按照特定的分析目标进行多源聚合,形成具有业务参考价值的数据集,对辅助人工发现数据之间的关联关系,快速生成具有丰富业务内涵的价值信息具有重要意义。文章基于本体知识库构建技术,构建了电力系统本体知识库。按照电网本体的推理规则对被检索信息在本体知识库中的存储、关联关系进行探究,基于大电网运行数据、设备管理以及电网模型、参数等,从结构化和非结构化两方面分析数据的存储、解析及检索技术,应用结果充分证明本文研究成果能够为增强运行数据智能分析以及科学决策管理能力提供技术支撑,对提高调度数据检索准确性、智能性具有重要意义。

参考文献:

[1] 李航. 复杂数据源报表模型的研究与实现[D]. 开封:河南大学,2008.

[2] 王理勤. 查新与文献检索数据库管理系统的开发[J]. 大学图书馆学报,1998,16(5):13-14.

[3] 杨争林,宋燕敏,曹荣章,等. 基于 Excel 的电力市场报表管理系统[J]. 电力系统自动化,2003,27(1):27-30.

[4] 张金辉. 基于局域网工业监控生产报表系统的软件设计与实现[J]. 自动化与仪器仪表,2015(7):18-20.

[5] 金风. 报表软件生成档案管理台账的方法[J]. 档案与建设,2015(2):88-90.

[6] 高军,李晓明. 基于三层结构的电力信息系统智能报表的研究[J]. 电力科学与工程,2003(03):25-28,55.

[7] 王越,孙骏,陆叶,等. 药品不良反应报表综合管理系统的研究与应用[J]. 药物流行病学杂志,2006(2):83-84.

[8] 段培娟. 支持多维数据分析的 Web 报表研究与实现[D]. 昆明:昆明理工大学,2012.

[9] 朱文娟. 医院交叉报表的设计与维护[J]. 计算机光盘软件与应用,2013(19):285-285.

[10] 格艳,王玉冰. 利用 Microsoft Excel 嵌套函数自动检索引水月报特征值研究与应用[J]. 科技信息,2013(5):475-475.

[11] 王瑞军. 基于元数据模型的电子政务检索系统的设计与实现[J]. 软件,2014,35(3):33-34.

[12] 曹宇,叶健辉,于永超. 基于电网本体知识库的智能搜索研究与实现[J]. 电力与能源,2016,37(1):1-6.

[13] 程孝侠. 支持语义查询的报表系统设计与实现[D]. 武汉:华中科技大学,2017.

[14] 孙磊. 网络安全审计系统之信息检索及报表生成的设计与实现[D]. 济南:山东大学,2005.

[15] 赵雷,王鹏恺. 报表模板库存储与检索系统开发[J]. 硅谷,2010(22):63-65.

[16] 寇波云,许振纲,邓卫. 家蚕品种资源数据库管理系统之二——查询检索报表输出功能[J]. 四川蚕业,1987(2):40-43.

[17] 张浩,蔡晖,刘孟觉,等. 类 Excel 跨平台报表系统[J]. 自动化技术与应用,2010(4):19-22.

[18] 王小玲. 图像检索技术识别虚假原始凭证探析[J]. 财会通讯,2010(7):123.

[19] 姜寒. 报表指标模型设计及灵活定制开发[J]. 科技视界,2016(26):165-165.

张 帆(1990—),男,硕士,工程师,从事电力大数据分析,493584158@qq.com。