

一种适用于电力标准文献的内容揭示方法^①

王宏¹, 周育忠¹, 王昕¹, 甘克勤², 周洁²

¹(南方电网科学研究院, 广州 510080)

²(中国标准化研究院, 北京 100191)

摘要: 传统的电力标准文献信息揭示方法主要分为外部描述的信息揭示和内容特征的信息揭示, 但都难以对电力标准文献的具体内容进行全面描述, 更难以适应以知识单元为基础的内容指标比对需求. 为破解上述难题, 本文将语义网理论、传统信息管理技术和标准化理论进行集成, 提出了电力标准文献内容揭示, 设计了电力行业标准文献内容揭示系统, 并在电力行业进行了实践与推广应用, 取得了比传统检索更优的效果.

关键词: 电力行业; 标准文献; 内容揭示; 揭示系统

A Content Revelation Method for Electric Power Standard Literature

WANG Hong¹, ZHOU Yu-Zhong¹, WANG Xin¹, GAN Ke-Qin², ZHOU Jie²

¹(Electric Power Research Institute, China Southern Power Grid, Guangzhou 510080, China)

²(China National Institute of Standardization, Beijing 100191, China)

Abstract: Traditional information revelation methods of electric power standard literature are mainly divided into external description and internal feature. But both ways are difficult to conduct a comprehensive description of electric power standard literature's details and adapt to the demands of content index comparison based on knowledge unit. In order to solve the above-mentioned problems, semantic web theory, traditional information management technology and standardization theory are integrated creatively in this paper. An electric power standard literature content revelation method is proposed and a system of electric power standard literature content revelation is designed, which is in application and dissemination in the power industry, and got better effect than traditional retrieval.

Key words: electric power industry; standard literature; content revelation; revelation system

1 现状与目标

电力标准资源作为一种产品, 其价值的实现依赖于对其利用的程度, 利用越充分, 其价值实现就越大. 本文研究并实践一种针对电力行业标准文献内容揭示的方法, 目的是实现电力标准文献内容与用户需求最大化的匹配, 这主要取决于两个方面: ①资源中蕴含的内容能否被准确的揭示; ②这种揭示与用户需求的描述是否一致^[1].

电力标准文献内容中的知识是内容揭示的对象, 鉴于知识结构和演化过程的复杂性, 内容揭示的有效方法一直是本体学、目录学和现代知识信息处理理论研究的中心, 是学术界和电网企业标准应用单位长期探讨的课题^[2]. 本文综合应用信息管理学、语义互联网

理论和信息处理技术, 并结合了标准化原理、标准化体系与体例结构分析, 在电力行业标准领域构建了以体例元数据为基础的标准内容揭示体系, 研发了以本体类与属性描述为检索手段的标准内容揭示检索系统.

1.1 传统电力标准文献信息揭示能力的不足

传统的电力标准文献信息揭示方法可归纳为外部描述的信息揭示和内容特征的信息揭示两类. “外部描述的信息揭示”是通过标准文献外部特征的揭示实现文献外部信息有序化并存储于关系型数据库中(如 Oracle 数据库), 以达到控制文献的目的, 以电力行业标准文献为例, 其外部描述内容包括: 标准号、标准题名、分类号、主题词等; “内部特征的信息揭示”主要通

① 收稿时间:2016-07-25;收到修改稿时间:2016-09-20 [doi: 10.15888/j.cnki.csa.005740]

过文献内容特征实现文献内部信息的有序化,并存储于 Key-Value 数据库中(如 Redis 数据库),以达到控制文献内容的目的,以电力行业标准文献为例,其内部描述内容包括:篇章段落信息(包括前言、适用范围、附录等)、表格信息、图片信息、指标信息等。

随着信息化技术的发展,电网企业对标准文献资源的揭示质量提出越来越高的要求,标准文献资源内容揭示作为核心环节,其传统的揭示方式正受到来自各方面的挑战,传统的分类、主题词、文摘对内容知识揭示的局限性日益凸显,主要体现在:

① 难以对电力标准文献的内容进行全面描述。作为一种技术类规范性文件,电力标准文献一般从多个角度对标准化对象进行描述,如设备产品类标准一般会规定该产品的原产地范围、术语、定义、原材料、产品分类、等级、规格等要求。而传统的分类、主题词、关键词等元数据描述,一方面限于标引词有限不能完全覆盖资源的内容;另一方面,限于标引者或加工人员精力和能力所限、认知和理解能力不同,难以有效、准确、全面的揭示一些隐含内容,难以满足电网用户需求。

② 难以适应以知识单元为基础的内容指标比对。在实际电力生产活动中,电力标准的工具书属性,使得电力标准内容指标比对成为电网企业标准文献的重要应用场景。而标准内容与指标比对的基础就是基于本体理论和体例归纳对标准文献内容的有效组织;传统的揭示组织方式,包括“外部描述信息揭示”和“内部特征信息揭示”都难以支撑电力标准内容与指标的比对。

③ 难以支撑开展知识演化与知识推理的研究与实践。电力标准文献中的指标是电力行业应用的重要基础数据之一,针对兴起于工业化大生产的标准开展知识演化与知识推理的研究,必能为即将到来的工业 4.0 提供有效支撑。而传统的揭示方式显然难以支撑这种研究的开展。

1.2 电力标准文献内容揭示的目标

针对传统电力标准文献信息揭示能力的不足,电力标准文献内容深度揭示旨在实现以下目标:①全面描述电力标准文献内容,实现围绕设备产品或标准化对象的技术指标的精确查询、体系查询。②支撑电力标准指标多维度比对分析业务,实现针对产品或标准化对象的标准体系比对、体例结构比对、技术指标比

对等。③支撑开展电力标准知识演化与知识推理研究,探索有效的电力行业应用。

2 理论技术实用

“电力标准文献内容揭示”属于集成创新的技术,将语义网理论、传统信息管理技术和标准化理论进行集成,形成可指导电力标准实践的“标准文献内容揭示”理论方法与实用。

2.1 语义互联网理论与实用

随着互联网技术的发展,语义互联网(semantic web)正逐渐将 Internet 变成一个巨大的全球化知识库。语义环境下,信息资源的内容描述包括两点:①对本体概念的规范化处理;②显示概念本体之间的关系^[3]。主要技术方法包括:可扩展标记语言(XML)、资源描述模型(RDF/RDF Schema)和知识本体(Ontology)等,这些适合语义 web 技术的知识描述和揭示方法提供了信息内容表示的标准化框架^[4]。

资源描述框架 RDF 是 Resource Description Framework 的缩写,RDF 来将元数据描述成为数据模型:一个 RDF 文件包含多个资源描述,而一个资源描述是由多个语句构成,一个语句是由资源、属性类型、属性值构成的三元组,分别对应自然语言中的主语、谓语和宾语。电力行业标准文献内容揭示方法借鉴 RDF 三元组的资源描述方法,建立“产品-体例-揭示内容”的资源描述框架,其中,产品主要归纳本标准描述的标准化对象,体例则归纳同类标准(产品、方法、安全、环保、卫生、基础、管理等)的体例结构,揭示内容则为描述该产品对应体例的内容描述,分别对应标准文献内容的主语、谓语、宾语。

2.2 传统信息管理技术与实用

在电力标准文献管理中,传统信息管理技术通过分类号^[5]、主题词、关键词等元数据对资源的描述来实现“外部描述信息揭示”,电力标准文献内容揭示的实操过程中,应用标引和分类技术,开展基于设备产品和揭示内容的标引和分类工作,不仅解决了传统技术的局限性(如前文所述),更解决了“电力标准文献内容揭示”过程中电网设备产品和揭示内容抽取通用性和准确性的问题。

2.3 标准化理论与实用

作为技术类规范性文件,电力行业的标准化工作者不仅遵循 GB/T 1.1-2009《标准化工作导则 第1部

分: 标准的结构和编写》, 按照一定的体例结构编写标准, 更需遵循 GB/T 13016-2009《标准体系表编制原则和要求》, 按照一定的原则和要求, 构建标准体系. 标准文献书写规范、成体系的特性, 是电力标准文献内容揭示方法的重要理论基础和方法依据.

本方法在确定揭示的领域范畴之后, 首先将电力标准按其描述特性划分为设备产品、方法、安全、环保、卫生、基础、管理等各种类型, 然后针对不同类别的标准, 分别分析其“产品-体例-揭示内容”的标引和分类原则.

3 系统研发与实现

通过将本文提出的电力标准文献内容揭示方法应用于南方电网企业实践, 完成了 100 余项变压器类国家标准、行业标准、企业标准文献内容的揭示加工, 并研发了电力行业标准内容揭示系统. 本章从功能设计、系统架构和实现效果角度, 阐述电力行业标准内容揭示系统.

3.1 功能设计

电力标准文献内容揭示系统的数据检索功能包括: 简单检索、高级检索和检索结果展示.

① 简单检索

针对电网设备产品(类)和揭示内容(技术指标)的简单检索要求达到提问式检索的效果, 选项包括: 含下层标准、含上层标准、全库. (上、下层标准是通过本体类表实现了向上和向下兼容)

② 高级检索

提供逐层引导式检索功能, 包括:

查类, 逐级展开电网设备产品分类, 引导用户逐步精准定位关注的产品;

查特性, 展开体例分类, 引导用户逐步精确定位到关注的体例;

查指标, 展开揭示内容分类, 引导用户逐步精确定位到关注的揭示内容.

选项包括: 含下层标准、含上层标准、全库.

③ 检索结果

检索结果包括以下项目: 标准种类、检索对象、属性类型、技术指标、内容、内容注释、来源、相关标准、标准体系(取值: 体系名称或者其他).

3.2 系统架构

系统架构包括硬件与网络拓扑、软件架构.

① 软件与网络拓扑

系统配置如图 1 所示, 可根据实际硬件、网络情况进行调整.

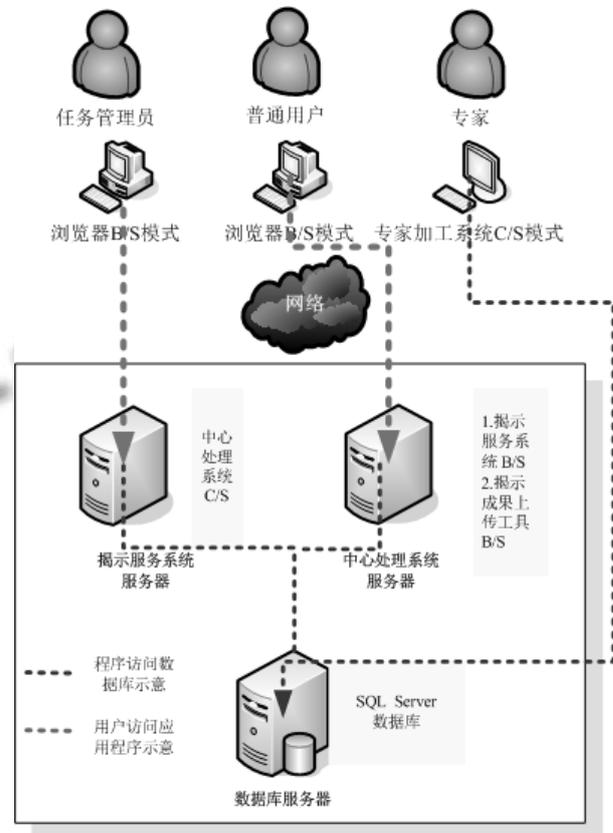


图 1 电力标准内容揭示系统硬件与网络拓扑

网络: 可以是内部局域网络, 也可以是公网, 根据需要进行配置.

服务器: 需要 2 个服务器, 分别是数据库服务器, 用于安装 SQL Server 2005; 揭示服务系统服务器, 用于安装电力标准内容揭示服务系统, 安装.NET2.0、IIS 6.0, 操作系统使用 Windows 2008、Windows 2003.

客户端: 普通用户与任务管理员应用客户端系统的浏览器, 比如 IE、FireFox 来访问程序, 现在主流配置 PC, Windows XP、Windows Vista、Windows 7 操作系统; 专家使用在客户端系统的专家加工系统进行加工.

② 软件架构

电力标准内容揭示系统采用分层的体系结构, 详细架构如图 2 所示.

从下向上分别是:

数据层: 包含电力标准内容揭示数据库、电力标

准文献数据库、电力标准文献全文,其中,内容揭示数据库通过关系型数据库设计表达了“三元组的资源描述方法”,核心数据关系如图3所示。

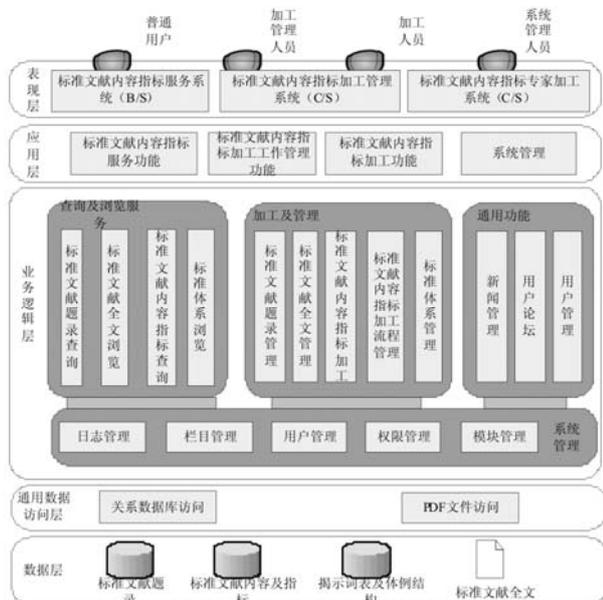


图2 电力标准内容揭示系统软件架构

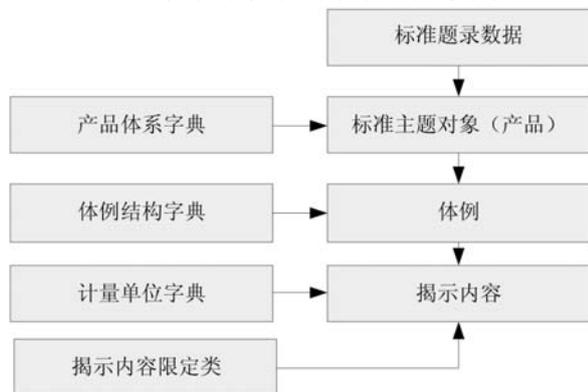


图3 内容揭示核心数据“三元组”关系图

通用数据访问层,包含访问 SQL Server 数据库的通用算法;

业务逻辑层:根据业务逻辑,实现业务流程,为展现层提供算法调用,由于核心数据“三元组”逻辑关系相对复杂,本系统采用 Lucene 的索引技术,提升核心数据表的检索效率。

应用层:实现标准文献内容指标服务、加工及其工作管理以及系统管理的功能。

展现层:应用 Windows 桌面程序与 IE 等浏览器展现给用户,包含一系列 html、aspx 文件。由于引导检索逐层展开的特点(详见 3.3),展现层大量采用 AJAX 技术,结合业务逻辑层的索引技术,提升了引导式检索的用户体验。

3.3 实现效果

简单检索与高级检索实现效果如图4所示。高级检索中,主题词的检索输入词为“变压器”时,点击“查类”按钮,则展示所有“变压器”相关的上位产品和下位产品概念,同时,展示了所有变压器相关的指标及其基本分类。



图4 电力标准内容揭示系统简单检索与高级检索示意

标准内容揭示系统与标准题录信息检索和标准全文检索效果对比,如表1所示。

表1 标准内容揭示检索效果比对

查询效果	标准内容揭示检索	标准题录信息检索	标准全文检索
查全率	基于产品、体例、揭示内容及其上下位、同义词等字典关系,确保查全率。如检索“电力设备”可涵盖所有“变压器”标准	仅针对标题、适用范围、主题词检索,无法深入标准具体指标内容,难以确保查全率。	尽管针对全文内容进行检索,但缺少上下位和同义词的字典支撑,难以确保查全率,如检索“电力设备”难以涵盖“变压器”标准
查准率	基于产品、体例、揭示内容细粒度的加工结果,确保查准率。如检索“变压器 结构”,可根据检索人的预期,将变压器中与结构有关的安全条款等内容进行展示。	针对标题、适用范围、主题词的简单匹配,匹配结果是没有同时命中“变压器”和“结构”的标准,进而切词后分别检索“变压器”和“结构”出现大量噪音。	针对标准全文的检索,不可避免的产生大量噪音。如检索“变压器 结构”仅国标命中百余条,不乏描述变压器标准的文档结构的内容。

4 总结

本文描述了一种适用于电力标准文献的内容揭示方法,从现状与目标、理论技术实用以及系统研发与实现的角度,描述了该方法的理论基础与实际应用,最后展现其应用结果,并与传统检索方式进行了对比。

本文的研究成果已经在电力行业内部上线使用,取得了良好的效果,极大提升了标准的使用查询效率,实证了电力标准文献的内容揭示方法的有效性。

参考文献

- 1 常唯.标签在数字学术资源内容揭示中的作用研究.图书馆杂志,2007,26(1):46-52.
- 2 张敏,邓胜利.基于内容揭示的信息资源控制的演进.图书情报工作,2009,53(2):117-120.
- 3 张帆.信息组织学.北京:科学出版社,2005.
- 4 卢巧云.XML:数字图书馆信息组织的基础技术.情报科学,2003,21(9):960-962.
- 5 胡永明,周洁.国际标准分类法及其在我国的应用研究.世界标准信息,1998,(11):26-28.