信息检索技术在文档管理中的应用*

蒋春茂[1],宁 芊[1],傅贺平[2]
(1.四川大学 电子信息学院,四川 成都 610065;
2.中国石油工程设计西南分公司,四川 成都 610041)

摘 要:为方便信息管理,在已有标准电子文档基础上,运用信息检索原理及信息检索实现技术,研究以关键字查询方法为重点的标准平台索引服务,建立起一个标准共享平台。标准共享平台运行结果表明,信息检索技术能方便地运用于文档的管理中。

关键词:标准共享平台;信息检索;关键字
中图分类号:TP39 **文献标识码**:A **文章编号**:1674-7720(2013)18-0080-02

The application of information retrieval technology in document management

Jiang Chunmao[1], Ning Qian[1], Fu Heping[2]
(1.School of Electronic Information, Sichuan University, Chengdu 610065, China;
2.China Petroleum Engineering Co.Ltd.(Southwest Company), Chengdu 610041, China)

Abstract: In order to manage the information. It established a standard sharing platform with the existing standard what is made of electronic document. Study the principles of information retrieval technology and the method of information retrieval technology what focus on the keyword query methods, discuss the indexing service provided by the standard platform. The results show that the information retrieval technology can be easily applied to the management of the document.

Key words: standard sharing platform; information retrieval; keyword

对石油的开采、运输等工程设计过程中需要用到大量文档信息(国外标准、国家标准、企业标准等)。为方便标准的管理和使用,实现技术有形化及知识共享和积累,将标准的管理与信息检索技术相结合,建立了与数字图书馆[1]类似的能通过关键词查询或任何经过定义的方式获得所需信息的系统。通过该系统,用户可以随时随地、方便而快捷地查找并获得统一、准确的标准信息。

本文以中国石油公司的标准电子词典开发项目为背景,整个项目是通过已有的标准电子文档建立一个标准共享平台[1-2]。讨论以关键字查询为重点的信息检索技术的基本思想。

1 信息检索技术

信息检索技术的基本原理[3]是通过对大量的、分散无序的文献信息进行搜集、加工、组织、存储,建立检索系统,并通过一定的方法和手段使存储与检索这两个过程所采用的特征标识达到一致,以便有效地获得和利用信息源。其核心思想是用户信息需求与文献信息集合的比较和选择,是两者匹配的过程。

信息检索的一般过程是检索系统将文档集合中的文献对象进行标引,用户将需要查找的信息(即信息需求)表达成查询,以信息提问的方式提交给检索系统,则检索系统运用预先设定的匹配算法[4]进行计算,检索出查找对象,并最终输出满足用户需要的结果。信息检索主要过程如图1所示。

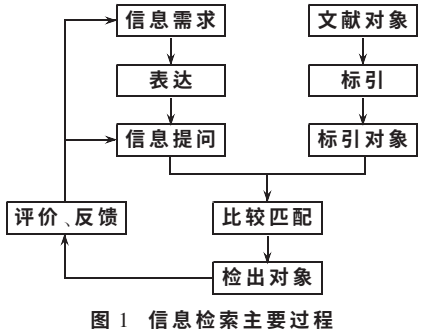

图1 信息检索主要过程

*基金项目:四川省应用基础科技计划项目(2011JY0014)

2 信息检索技术的实现

2.1 标准平台提供的检索功能

对用户的信息需求，标准共享平台提供分类目录查询和输入关键字查询两种方式。输入关键字查询需要用户输入所需查询信息的标题或标准编号中的字、词或者发行部门等信息，计算机通过事先设置的算法返回用户查找的信息。

2.2 关键字检索技术

由于标准平台的开发面向已有的电子文档，因此平台的关键字检索设计主要在于解决关键字索引及如何查询索引问题。

按照输入关键字查询要求，将标准名称及对应的标准编号与标准内容通过序号建立一一对应关系[5]，可将每篇文档的标题看成是全文信息。利用倒排索引[6-7]思想，一方面将所有标准名称及编号中的每个字按照其首字母先后顺序建立一个索引，称为词表，它包含一个记录表项，记录表项记录了出现这个字的标题所在地址信息及其编号情况。另一方面将出现的各个词项的文档标题或编号构成一个文件，即记录文件。例如，表 1 展示了部分标准信息，针对这些标准中的词条建立倒排索引，部分内容如表 2 所示。

表 1 部分标准信息

序号	标准名称	标准编号
1	保温层下的腐蚀控制	NACE SP0198
2	油气田及管道站场外腐蚀控制技术规范	Q/SY 1186-2009
3	埋地钢质弯管聚乙烯复合带防腐层技术标准	Q/SY GJX 113-2010

表 2 倒排索引部分内容

关键字	文档 ID	出现位置
腐蚀	1，2，3	5
埋地	3	8
…	…	…

由表 2 可知，包含关键字“腐蚀”的文档序号为 1、2、3，其标题构成一个文件。标准平台设置的每条索引结构包括文档发布年份、类型、名称等，采用倒排索引表示对应的文档，并通过主键来唯一确定一条索引数据。其数据结构如表 3 所示。

表 3 标准平台中一条索引的数据结构

主键	属性						全文分词
key	标准名称	标准代码	地址	大小	发布年份	类型	关键字

在查找索引词表问题上，由于词表是按其首字母顺序进行位置排序，当输入一个字时，利用二分法[3]找到其首字母所属段词表，然后再对此段词表进行顺序搜索，直到查找到此字在词表中的位置，同时找到此字的记录表项，根据记录表项中所记录的标准名称所在位置及序号找到相应的文档标题。当输入不止一个字时通过 AND 操作，找到同时包含输入字的文档标题，通过之前与全文信息建立的对应关系即可找到文档信息。

3 信息检索技术在系统中的实现

在搜索框内输入关键字，便可显示含有关键字的标准或文档信息，如果没有符合的标准，系统则输入“无此项信息”。图 2 为输入关键字“腐蚀”后的检索示意图。

图 2 检索示意图

从对标准电子词典的测试情况发现，运用倒排检索及二分法对词表进行查询的方式所建立的标准共享平台有非常高的正确率，满足用户要求。

标准电子词典的成功开发，是将信息检索技术与企业文档管理相结合思想应用到实际工作中的一个例子，它使得用户通过登录系统输入关键字便可以对所需标准进行搜索等操作。将工作人员从繁重的、重复的手工劳动中解放出来，提高了管理部门的管理水平。

参考文献

[1] 黄如花，王梅，黄晓斌，等.数字图书馆原理与技术[M].湖北：武汉大学出版社，2005.

[2] 席生长，胡宏涛.信息检索技术在中石油勘探与生产分公司门户内的应用研究[J].福建电脑，2008(1)：102-103.

[3] SHAFFER C A，张铭，刘晓丹，等.数据结构与算法分析(C++版)[M].北京：电子工业出版社，2002.

[4] 闻玉彪，贾时银，邓世坤，等.一种改进的最大匹配中文分词算法[J].计算机技术与发展，2011，10(21)：92-98.

[5] 王斌.从信息检索到搜索引擎[J].术语标准化与信息技术，2009(4)：38-43.

[6] 刘兴宇.基于倒排索引的全文检索技术研究[D].武汉：华中科技大学，2004.

[7] 王泽胤.全文信息检索的快速索引文件结构及系统的设计与实现[D].吉林：吉林大学，2009.

(收稿日期：2013-04-23)

作者简介：

蒋春茂，女，1989 年生，硕士研究生，主要研究方向：智能控制。

宁芊，女，1969 年生，博士，副教授，主要研究方向：智能控制。