基于 Solr 的科技成果查新系统的构建研究

温慧明[1],宫晓辉[2]
(1. 煤炭科学研究总院 经济与信息研究分院,北京 100013;
2. 国网电力科学研究院 中电普华信息技术有限公司,北京 100096)

摘 要:随着各大国企,甚至是私营企业的快速发展,其科研项目和科技成果的数量呈现指数级增长,然而,企业的内部查新工作还是人工来完成,难度较大,因此文中从提高企业内部科技成果的查新效率出发,基于 Solr 搜索应用服务器这一核心平台,研究基于 Solr 的科技成果查新系统的设计和实现。首先简要介绍了 Solr 的概念、特性和系统架构,然后介绍了基于 Solr 引擎的科技成果检索查新系统的功能结构和系统架构,最后阐述了系统的界面和功能的具体实现,特别是检索查新和对比查看功能的设计和实现。

关键词:科技查新;Solr;科技成果
中图分类号:TP39 **文献标识码**:A **文章编号**:1673-629X(2014)06-0067-04
doi:10.3969/j.issn.1673-629X.2014.06.017

Research of Sci-tech Achievement Novelty Search System Based on Solr

WEN Hui-ming[1], GONG Xiao-hui[2]
(1. Economic and Information Research Branch of China Coal Research Institute, Beijing 100013, China;
2. Zhongdian Puhua Information Technology Co., Ltd., China Electric Power Research Institute, Beijing 100096, China)

Abstract: With the rapid development of state-owned enterprise, the number of scientific and technological achievements and the scientific research projects present increase exponentially, however, internal sci-tech novelty search of the enterprises is still by hand to complete, the difficulty is large, therefore, in order to enhance the enterprise internal sci-tech novelty search efficiency, based on the solr search application server core platform, research design and implementation of sci-tech novelty search system based on solr. Briefly describe the solr concepts, system architecture and features, then introduce function structure and system architecture of sci-tech novelty search system based on solr, finally describe the concrete realization of system's interface and the function, especially the novelty search and retrieval contrast to check the function design and implementation.

Key words: sci-tech novelty search; Solr; sci-tech achievements

0 引 言

近年来,开放源码软件《open-source software》作为一个新的理念被迅速推广,它的使用、修改和分发也不受许可证的限制[1]。这样企业的信息化建设如果基于开源软件进行功能定制和增值开发,则可以将有限的技术力量和资金投入建设适应自身业务需要的信息化系统。而 Solr[2-4] 就是开源软件中的一个典型案例,Solr 采用 Java5 开发,是一个高性能基于 Lucene[5-7] 的全文搜索服务器。同时对其进行了扩展,提供了比 Lucene 更为丰富的查询语言,实现了可配置、可扩展并对查询性能进行了优化,是一款非常优秀的全文搜索引擎[8-11]。Solr 的出现为科技成果检索查新系统的实现提供了强有力的保障。

另一方面,经过许多年的发展,各大企业的科研项目和科技成果的数量呈现指数级增长,伴随出现了申报过程繁琐、科技查新难度大和管理过程复杂等一系列问题,而目前各大企业已经有自己的科研管理系统,但是不包括检索查新的关键功能[12-14],所以为了避免科研项目低水平重复和成果鉴定、评奖失准等问题,在

收稿日期:2013-07-30 **修回日期**:2013-11-05 **网络出版时间**:2014-02-24
基金项目:国家"863"高技术发展计划项目(SS2012AA061303)
作者简介:温慧明(1983-),男,硕士,研究方向为图形图像处理、数据库、数字化矿山、煤炭信息化。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20140224.0858.012.html

企业内部搭建科技成果检索查新系统来实现科研项目和成果的在线申报、在线查新和统一管理是迫在眉睫。

文中在简要介绍开源的全文检索引擎 Solr 的基础上,设计并实现了基于 Solr 引擎的科技成果检索查新系统,实现了对企业内所有的科研项目和成果的申报管理和检索查新[15]。

1 Solr 简介

1.1 Solr 概述

Solr 是一个高性能,采用 Java5 开发,基于 Lucene[14]的全文搜索服务器。同时对其进行了扩展,提供了比 Lucene 更为丰富的查询语言,实现了可配置、可扩展并对查询性能进行了优化,并且提供了一个完善的功能管理界面,是一款非常优秀的全文搜索引擎;由于它具有灵活的 XML 格式配置和支持多种客户端语言,所以易于加入到 Web 应用程序中,能为多种数据格式提供索引、检索、分布式搜索及层面搜索、命中醒目显示、强大的查询缓冲并支持多种输出格式(包括 XML/XSL 和 JSON 格式)的功能。2006 年,Apache Software Foundation 在 Lucene 顶级项目的支持下得到了 Solr,先后共发布了很多版本。如今 Solr 已经广为人知,并且许多公司都已经使用 Solr 去构建自己的搜索引擎:AOL、Disney、Apple、阿里巴巴和安居客等[2]。

1.2 Solr 体系架构

Solr 是基于 Lucene 的全文搜索服务器[5-7],专注于企业应用,重点开发了支持搜索服务相关的管理模块和接口。Solr 的体系架构[2-4]如图 1 所示。

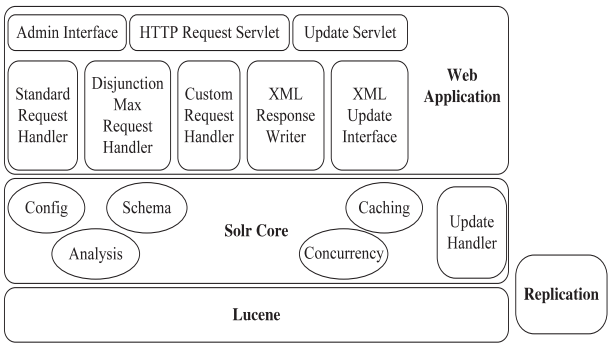

图 1 Solr 体系架构图

(1) Web 应用层提供 Web 查询接口与界面,还提供了索引更新接口。其中管理员 Web 接口界面可以查看搜索引擎的基本信息,自定义查询语句进行查询;提交的查询请求通过 HTTP 请求处理器调用 Solr Core 进行处理,把最后的处理结果以 XML、JSON 等格式返给前台。

(2) 中间层由六大独立核心模块组成,是 Solr 引擎的核心,其中配置模块(Config)主要用于搜索服务器配置参数文档并进行加载和解析;同样索引模式定义模块(Schema)主要用于索引模式参数文档的加载与解析;分析模块(Analysis)顾名思义就是对查新语句进行分析处理;并发控制模块(Concurrency)提供建立索引和读取索引的并发控制处理机制;缓存机制模块(Caching)用于实现对文档和分面数据的缓冲提高查询效率;更新处理器(Update Handler)主要用于各种数据资源的索引处理。

(3) 最底层为全文索引工具 Lucene,负责底层的文本索引和文档查询。索引复制模块(Replication)是一个相对独立的模块,其功能主要用于支持分布式索引和检索。

2 基于 Solr 引擎的科技成果检索查新系统设计

2.1 系统总体功能

基于高性能的 Solr 的科技成果检索查新系统必须具有的功能结构图如图 2 所示。

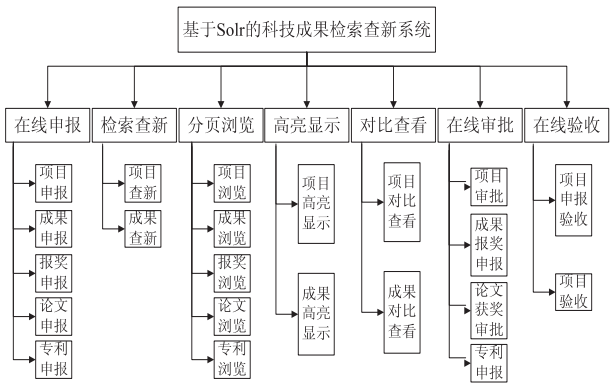

图 2 系统功能结构图

该系统整体包含七个部分,在线申报是对科研项目、科研成果、成果报奖、优秀论文和专利的网上申报功能,检索查新是对科研项目的科技成果的查新处理[12-14],检索有多少个项目或成果是与其重复的,重复率多少的功能,为了提供友好的界面浏览效果,分类浏览和高亮显示的功能,对比查看的功能是用于查看要查新的文档和库里的文档的重复内容,并且重复内容要高亮显示,同时也提供了对项目的网上验收功能。

2.2 系统体系架构

基于 Solr 的科技成果检索查新系统的体系架构图如图 3 所示。

(1)最底层为数据层,为上层服务提供数据源,一种是数据库数据源:为上层的在线申报、在线审批以及在线验收的功能服务,另一种是索引文件数据源:为检索查新、对比查看、分页浏览和高亮显示功能服务;

(2)核心业务层主要是依赖功能强大的 Solr 搜索引擎,该层首先负责索引参数文档的配置,引入 IK 中文分词器对申报的项目和成果进行分词,并建立索引。

同时使用 IK 中文分词器配合 Solr 引擎解析查新请求,并进行检索,最后将检索后的结果以标准的格式返回前台,例如:XML 格式等。

(3)用户应用层是系统的重点,主要是提供用户界面,实现用户与系统交互的 7 大功能,支持分页浏览、高亮显示和对比查看等。

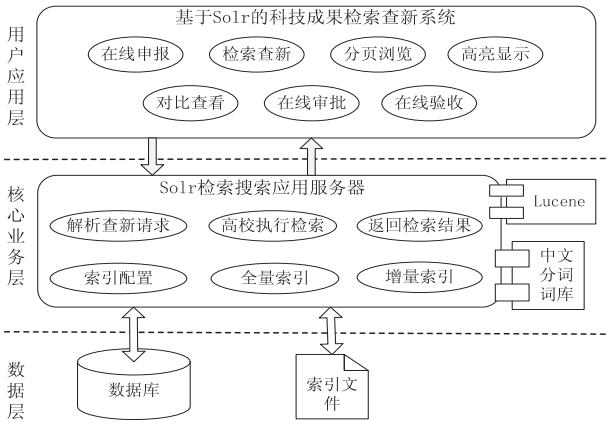

图 3　基于 Solr 的科技成果检索查新系统的体系架构图

3　基于 Solr 引擎的科技成果检索查新系统实现

3.1　集成 Solr 和 IKAnalyzer 中文分词器

(1)集成 Solr3.5。

首先从 Apache Solr 的官网下载 apache-solr-3.5.0 软件包,将其解压后有两个目录:dist 和 example,其中 dist 目录下有一个 War 包 apache-solr-3.5.0.war,example 目录下包含了一些样例数据和一些 Solr 主目录。然后使用 Eclipse 开发工具建立自己的工程 ns_solr,再将 apache-solr-3.5.0.war 解压,把解压后的两个文件 favicon.ico、index.jsp 和 Admin 文件夹拷贝到工程的 jsp 文件夹下,将解压后的 WEB-INFO/lib 文件夹下的 jar 包拷贝到工程中,并将 WEB-INFO/Web.xml 文件与工程里的 Web.xml 文件做整合,接下来要在工程 ns_solr 的配置文件 Web.xml 里指定 Solr 主目录。为简单起见,可以将前面提到的 example 下的 solr 文件夹复制到 D 盘,然后修改工程 ns_solr 里的 Web.xml,将标签“<env-entry-value>”的内容改为 Solr 主目录的物理路径,即“d:\solr”。为验证 Solr 是否正确安装,先将该工程部署到 tomcat6.0 里,在浏览器中访问该地址:http://localhost:8080/solr/admin/,如果出现 Solr 引擎的管理主界面则表示配置成功。

(2)集成 IKAnalyzer 中文分词器。

该研究下载的是 IK Analyzer 3.2.8[8],解压后将 IKAnalyzer3.2.8.jar 拷贝到工程 ns_solr 的 lib 目录下,然后将 ext_stopword.dic 和 IKAnalyzer.cfg.xml 拷贝到工程 ns_solr 的 src 根目录下;接下来就是配置 solr 的 Schema.xml 文件,替换创建索引和查询检索的分词器为 org.wltea.analyzer.solr.IKTokenizerFactory 即可完成集成。

3.2　索引文件的创建和更新

(1)索引参数和索引结构配置。

为了使 Solr 能够对科研计划项目和成果创建索引,需要对 Solr 进行设置,具体步骤如下:首先配置索引性能参数,即修改 solrconfig.xml 文件,该研究采用 Solr 的默认的性能参数配置;其次配置索引结构,即修改 Schema.xml 文件,该文件通过定义 fieldType、fields、copyField 等几个主要标签来配置索引的主体结构。首先定义三种字段类型,即 String、Text_general 和 Date,分别与 solr.StrField、solr.TextField 和 solr.TrieDateField 类对应,其中 String 和 Text_general 字段类型需要配置自定义的分词器,该系统采用 IKAnalyzer 中文分词器来进行中文分词[8],然后在 fields 节点内定义了 id(唯一标识符)、name(项目名称)、subject(项目主要内容)、keywords(项目关键词)和 text(复制字段)五个具体的字段,id 属于 string 字段类型,别的都属于 text_general 字段类型。对可能存在多值的字段(如作者和关键词)设置 multiValued 为 true;其中 text 字段是为满足对所有字段进行统一检索而建立的一个复制字段,通过<copeField>的标签将所有的字段内容复制到该字段中即可完成索引结构的配置。

(2)基于 Solrj 的索引创建。

首先基于 Solrj 创建索引的方式有两种,一种是通过 CommonsHttpSolrServer 来创建,一种是通过 EmbeddedSolrServer 来创建。其中 CommonsHttpSolrServer 是通过 URL 访问 Solr 搜索应用服务器来创建索引,意味着 Solr 搜索应用服务器与现有工程是独立分开的,而 EmbeddedSolrServer 则是应用在嵌入 Solr 搜索应用服务器的工程里,所以该研究使用 EmbeddedSolrServer 来创建索引,代码片段如图 4 所示。

图 4　创建索引代码片段

3.3　用户界面和功能的设计和实现

由 2.1 节知道该系统将实现 7 大功能模块,其中检索查新、高亮显示和对比查看是系统核心和难点,而在线申报、在线审批、在线验收和分页浏览是辅助功

能,实现起来相对比较容易,所以本节重点讲解检索查新和对比查看二大功能模块的设计和实现,其中高亮显示暗含在这两个模块中实现。

(1)检索查新功能和界面的设计和实现。

在该系统中检索查新包括检索、计算查重率和按查重率排序三步骤,是在线申报后进行查重处理的功能,以项目为例,基于 Solrj 的 EmbeddedSolrServer 构建检索查新语句,而查询条件中必须有高亮显示的设置,如图 5 所示。

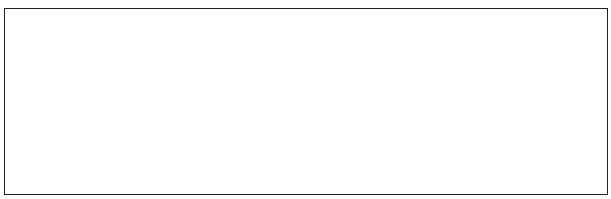

图 5 设置高亮显示查询条件的代码片段

接下来计算查重率(重复率),首先处理目标文档中高亮显示的内容,将文档内容按句分开,如果该句中高亮显示的内容字数超过 70% 则将该句内容整体高亮显示,否则整体不高亮显示,并记录高亮显示的文本;反过来根据目标文档构建检索语句查询当前文档的内容,用同样的方式处理该文档的高亮显示内容,并记录高亮显示文本,将该文本中高亮显示的内容字数与该文本的总字数的比值作为该文档与目标文档的查重率;以此类推,处理所有的目标文档和当前文档的高亮显示内容,并计算查重率;最后根据当前文档与各个文档的重复内容的累计和计算当前文档的总查重率(重复率),并根据目标文档与当前文档的查重率进行排序展示。

(2)对比查看功能和界面的设计和实现。

通过上面的检索查新的功能实现之后,对比查看功能的实现就相对简单多了,就是将检索查新功能实现过程中记录的每一对当前文档和目标文档的高亮显示内容从 Session 中获取处理显示即可,上面是显示该文档与目标文档的查重率,然后将界面左右分割,左边显示当前文档,右边显示目标文档,两个文档的内容都可以通过自己的滚动条上下拖动,方便与两个文档重复内容的对比;对比查看的界面设计图太大,所以此处忽略显示。

从查新系统的实验运行情况看,系统能够较好地满足用户的科技成果申报和查新的要求,并能较为准确地计算查新文档的查重率,可以通过对比查看的功能来实现对文档重复内容的详细对比,为查新报告的出台提供了非常客观和直接的依据。

4 结束语

文中根据企业的科技查新工作实际情况,将优秀的全文检索引擎 Solr 应用于科技成果检索查新系统,满足了企业现阶段的成果查新使用,为企业建立内部资源库并研制全文检索查新系统提供了初步的范例。由于该系统目前只集中式部署于一些中型企业,下一步将要研究开展系统的分布式检索查新、可视化检索查新等方面的工作,以提高系统的整体性能和实用性,充分发挥 Solr 在建设科技成果检索查新服务系统中的优越性能。

参考文献:

[1] 黄琳喆. 论开源软件的知识产权保护问题[D]. 兰州:兰州大学,2009.

[2] 鲜国建,赵瑞雪. 基于 Solr 的中文农业期刊文摘检索系统的构建研究[J]. 现代图书情报技术,2011,27(6):51-58.

[3] 霍 庆,刘培植. 使用 Solr 为大数据库搭建搜索引擎[J]. 软件工程,2011,32(6):11-14.

[4] 王小森. 基于 Solr 的搜索引擎的设计与实现[D]. 北京:北京邮电大学,2011.

[5] 劳志佳. 基于 Lucene 3.5 搜索技术的研究与实现[J]. 现代计算机:下半月版,2012(6):70-73.

[6] 罗 刚. 解密搜索引擎技术实战 Lucene&Java 精华版[M]. 北京:电子工业出版社,2011.

[7] 林碧英,赵 锐,陈良臣. 基于 Lucene 的全文检索引擎研究与应用[J]. 计算机技术与发展,2007,17(5):184-186.

[8] Feng Xia,Tang Xianchao. An improved dictionary-based Chinese word segmentation approach in Lucene[C]//Proceedings of 2010 international conference on services science, management and engineering (Volume 1). Tianjin:[s. n.],2010:363-366.

[9] Zhao Xu,Xu Wenbo,Chai Zhilei. Adding synonym query for Chinese language to Lucene search engine[C]//Proceedings of 2008 international symposium on distributed computing and applications for business engineering and science. Dalian:[s. n.],2008:426-432.

[10] Liu Tianyuan,Song Meina,Zhang Xiaoqi. Research of massive heterogeneous data integration based on Lucene and Xquery[C]//Proceedings of 2010 IEEE 2nd symposium on Web society. Beijing:IEEE,2010:648-652.

[11] Song Jia,Zhu Yunqiang,Liu Runda. Enhanced full text retrieval kit based on Lucene[J]. Computer Engineering and Applicatlons,2008,44(4):172-175.

[12] 马景娣,田 稷. 基于 J2EE 的科技查新综合信息系统的设计与实现[J]. 现代图书情报技术,2004(8):77-78.

[13] 阳沛湘,柏立嘉,吴曙霞,等. 军队医药卫生科技查新管理系统的设计与实现[J]. 军事医学科学院院刊,2009,33(6):564-566.

[14] 胡 伟. 科技查新综合业务管理系统的设计与实现[J]. 图书情报工作网刊,2011(4):55-59.

[15] 马小雨. 基于 AHP 的煤炭科研项目评价系统的设计与实现[D]. 北京:北京邮电大学,2007.

基于Solr的科技成果查新系统的构建研究

作者： 温慧明， 宫晓辉， WEN Hui-ming， GONG Xiao-hui

作者单位： 温慧明,WEN Hui-ming(煤炭科学研究总院 经济与信息研究分院,北京,100013)， 宫晓辉,GONG Xiao-hui(国网电力科学研究院 中电普华信息技术有限公司,北京,100096)

刊名： 计算机技术与发展 ISTIC

英文刊名： Computer Technology and Development

年，卷(期)： 2014(6)