

DOI: 10.19452/j.issn1007-5453.2017.04.030

基于 Solr 的飞机故障异构信息检索系统设计

印奇^{1,*}, 李青¹, 黄鹏²

1. 北京航空航天大学 机械工程及自动化学院, 北京 100191

2. 成都飞机工业(集团)有限责任公司, 四川 成都 610092

摘要: 针对企业异构信息在资源数量上和载体种类上不断攀升的情况, 为了实现对全部异构信息资源进行统一检索并满足检索效率要求, 设计了一个基于 Solr 的飞机故障异构信息检索系统。该系统首先针对异构信息特点设计索引结构, 实现统一检索; 然后将各类异构数据按其对应索引结构生成索引文件, 并导入至索引信息库, 实现信息融合; 最后基于 SolrJ API 进行系统开发, 实现各类复杂搜索功能, 并结合专业词汇权重优化排序结果。实验结果证明, 该系统能满足多种复杂搜索需求, 有效解决了对企业异构信息资源的检索问题, 为企业做出快速决策提供了技术支持。

关键词: Solr; 异构信息; 索引; 检索系统

中图分类号: TP311

文献标识码: A

文章编号: 1007-5453 (2017) 04-0030-07

在信息技术飞速发展的今天, 随着业务的发展, 企业信息量不仅在数量上以量级规模增长, 在信息载体类型上也是日益增多, 以航空飞机为研究对象的航空制造企业更是如此。随着主机厂在各个服务点的相关飞机故障信息的持续累积, 故障信息条数已经达到数十万条, 这些信息一部分被保存在数据库表中以结构化数据的形式存储, 另一部分则包含众多与故障信息相关的其他非结构化数据, 如技术文档、故障信息通报、备件求援单等相关信息。这些新增的异构信息资源往往不能及时转化为结构化信息保存至企业信息库中, 用户在检索相关信息过程中需要针对各个信息源分别进行交互操作, 导致无法统一检索和检索效率低下。为了解决此类信息融合的问题, 消除源数据之间在类型上的差异和冲突, 利用索引技术, 通过构建统一的索引结构来保存关键信息用于搜索, 同时提高检索效率。

Solr 就是在企业普遍的信息检索需求中脱颖而出的开源搜索引擎的代表, 在其基础上开发人员可以快速高效地开发出功能全面的企业级搜索引擎系统。Solr 是开源搜索服

务器, 主要基于超文本传输协议 (HTTP) 和语义检索工具类 Lucene 来实现, 其通过定制索引的方式组织信息, 通过倒排索引和相似度评分机制来搜索信息^[1]。Solr 提供了网站 (Web) 服务器, 支持通过 HTTP 来更新索引、重建索引、查询等功能, 它通过对外提供类似 Web-service 的应用程序界面 (API) 接口, 形成了一个高效独立的企业级搜索引擎应用服务器, 从而保证了其低耦合性。Solr 提供了比 Lucene 更为丰富的查询功能, 同时, 实现了开发系统的可扩展、可配置性能, 并优化信息查询的性能。Solr 支持多种输出格式 (XML、XSLT 和 JSON), 实现 Solr 索引定制的方法是通过 HTTP POST 方法向 Solr 服务器发送一个描述所有域 (Field) 及其类型和属性的 XML 文档, 发送 HTTP GET 请求就能实现搜索定制, 然后再对 Solr 返回的信息进行数据处理和整理布局^[2]。

目前, 周敬才等研究利用各类信息抽取工具分别对异构资源和数据库表信息进行信息抽取, 并使用 Lucene 这一开源工具的 API 进行二次开发来构建索引文件, 从而构建

收稿日期: 2016-12-16; 退修日期: 2017-01-17; 录用日期: 2017-03-10

* 通讯作者. Tel.: 18510372180 E-mail: yinqiyq@163.com

引用格式: YIN Qi, LI Qing, HUANG Peng. Design of aircraft fault heterogeneous information retrieval system based on Solr[J]. Aeronautical Science & Technology, 2017, 28 (04): 30-36. 印奇, 李青, 黄鹏. 基于 Solr 的飞机故障异构信息检索系统设计 [J]. 航空科学技术, 2017, 28 (04): 30-36.

了企业搜索引擎系统^[3]；梁艳等通过将所有异构资源进行解析成统一结构的 XML 文件,从而实现统一检索的研究方法^[4]；谢文亮等专门针对 Web 网页资源的企业搜索引擎,提出了具体的企业搜索引擎系统框架,并对其系统进行设计,该系统实现了基于特定行业的垂直搜索引擎^[5]。本文利用 Solr 定义与数据库的连接方式,直接与数据库信息交互来生成索引,从而解决了结构化数据的索引操作,同时,利用相应的 SolrJ 接口方法实现针对非结构化数据的索引操作。在对 Solr 服务器应用研究的基础上,结合 JavaEE 应用框架 SSH (Spring+Struts+Hibernate),采用 MVC 模式 WEB 应用模型,基于飞机故障领域结构化数据和非结构化数据的特点,设计了基于 Solr 的异构信息检索系统。

1 系统总体设计

1.1 系统体系架构

图 1 为基于 Solr 的飞机故障异构信息检索系统的体系架构。

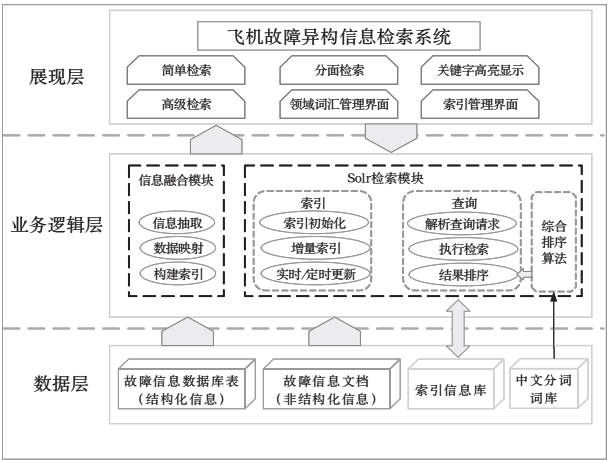


图 1 系统体系架构

Fig.1 System architecture

基于 Solr 的飞机故障异构信息检索系统的体系架构包含系统数据层、业务逻辑层和用户应用层,具体如下:

(1) 最底层为系统数据层:主要包含企业保存的故障信息数据库表、故障信息文档文件、保存的索引信息库以及用于优化排序的领域词汇库。

(2) 核心业务层:主要包含信息融合模块和 Solr 检索模块对各项检索参数进行详细配置和完善。信息融合模块主要针对各类异构数据进行信息抽取和构建索引,实现统一检索;Solr 检索模块则包含索引、查询操作以及优化排序结果,针对用户的查询配置中文分词等解析操作、执行检索过

程中的条件过滤、返回结果的高亮显示以及最终结果的综合评分排序处理。

(3) 用户应用层:是系统基于 Solr 进行二次开发的重点,主要提供用户的查询界面,实现用户与系统的交互功能,支持用户灵活选择查询方式,对检索结果支持分面检索分类、关键字高亮显示、检索结果的综合排序优化以及索引文件的管理界面。

1.2 数据处理流程

基于 Solr 的飞机故障信息检索系统的功能结构如图 2 所示。

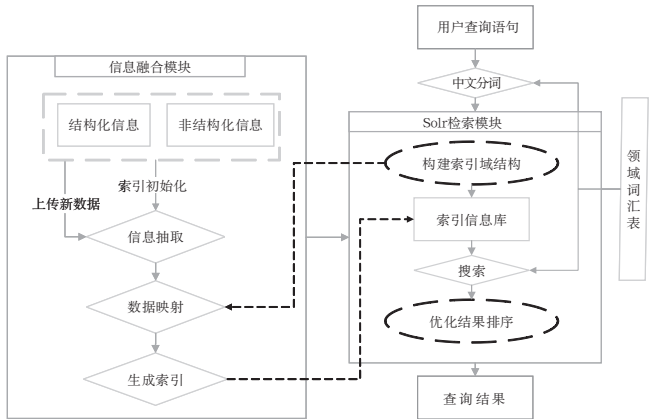


图 2 系统功能模块

Fig.2 System function module

系统功能模块主要包含信息融合模块和 Solr 检索模块。信息融合模块实现对所有异构数据的信息融合,针对结构化信息和非结构化信息进行信息抽取,并通过构建统一索引结构对异构信息集合进行数据映射整合,最终生成索引文件保存在索引信息库中;而 Solr 检索模块中,用户输入查询语句后,通过查询索引库信息,结合飞机故障领域词汇表优化排序,得到相关检索结果。

2 关键技术

2.1 信息融合模块设计

2.1.1 索引结构设计

索引结构设计是实现异构信息资源统一检索的重要途径,是实现信息融合的前提基础。针对不同类型的异构信息类型分别设计其对应的索引域,然后在查询过程中对所有异构资源的域进行搜索,从而实现查询可以覆盖到全部的异构信息。

Solr 的索引信息库结构主要包含域 (Field)、域类型 (FieldType) 等参数的定义。其中,域和域类型描述检索系

统索引的字段及其属性特点,在域类型中还可以对中文分词等功能进行配置。域是添加至索引文件中字段的定义,域类型用来定义域的分类、属性及处理方式^[6]。

飞机故障领域的异构信息资源主要包含两部分:关系型数据库表和各类文档,如以“故障编号”、“故障日期”、“故障性质”、“故障位置”等为列名的表,以及以 WORD、EXCEL、PDF、TXT 为格式的技术文档、故障报告等。索引结构设计如图 3 所示,针对数据库表类信息,由于其本身是结构化数据可以直接按其列名映射成索引的域名,并保存在索引库中。而针对文档类信息则选择其文档名称、文档创建日期和文档内容等信息作为其索引的域项。

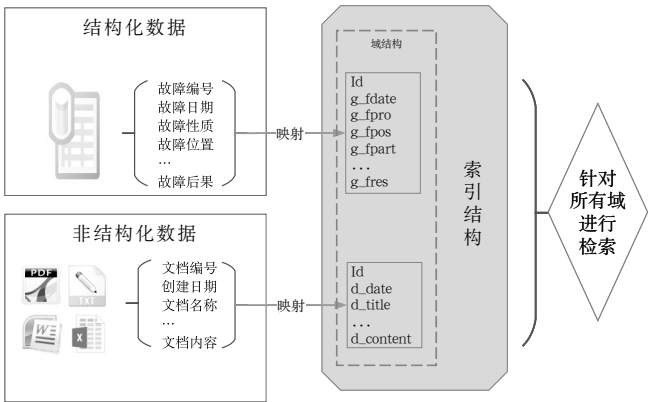


图 3 索引结构设计
Fig.3 Design of index structure

本系统基于故障信息表的基本结构来构建其中一部分域,以故障信息数据库表和文档的部分字段为例,如表 1 和表 2 所示,这些异构信息的域项在检索过程中都设置为被查询项,使得系统能检索到全部异构信息,从而实现统一检索需求。

表 1 数据库表信息与域名称映射表
Table 1 Map of columns of database table and field names

表字段名称	中文名	类型	域名称
faultId	故障编号	String	id
faultDate	故障日期	Date	g_fdate
faultProperty	故障性质	String	g_fpro
faultPosition	故障位置	String	g_fpos
faultPart	故障件名称	String	g_fpart
faultCode	故障模式	String	g_fcode
faultReason	故障原因	String	g_frea
faultResult	故障后果	String	g_fres

表 2 文档信息与域名称映射表
Table 2 Map of documents and field names

文档信息名称	中文名	类型	域名称
docId	文档编号	String	id
docDate	文档日期	Date	d_date
docTitle	文档名称	String	d_title
docContent	文档内容	String	d_content

确定好需要检索的数据库表字段以及文档字段与相关域的映射关系后,即可以在 Solr 的配置文档文件中定义相关域属性,如图 4 所示。

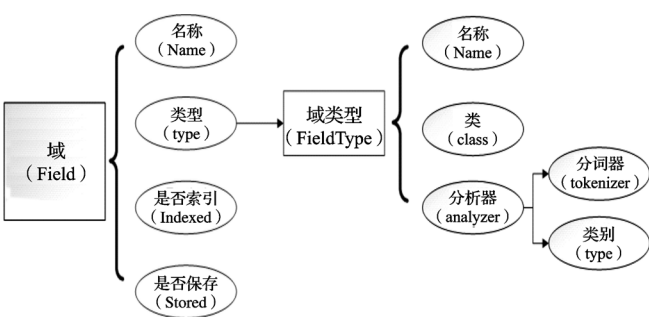


图 4 域及域类型组织关系
Fig.4 Relations of field and fieldType

域定义内容包含域的名称 (name)、类型 (type)、是否索引 (indexed)、是否保存 (stored) 等属性信息,其中,域的类型本身包含很多参数及属性设定,需要额外详细的配置定义,本系统域类型设置了索引和查询时均通过 IKAnalyzer 中文分词器处理。域类型的定义除了指定域类型的名称和加载类,还分别定义了针对域进行索引和查询时的操作设置。在查询操作配置部分,除了配置 IK 中文分词功能外,还配置了相似词查询功能,通过指定自定义相似词集合来完善搜索结果。

2.1.2 索引文件生成

索引域结构的设计主要是为了建立现有各类异构信息的数据项和索引项的映射关系,完善的索引结构能完整覆盖异构信息的各项信息。而信息融合模块则是在完成了对异构信息索引结构的组织后,以异构信息为对象建立索引文件的过程,将各类异构信息转变成统一的索引文件。

在建立索引过程中,系统主要处理两方面的问题:一是在系统数据库中已经存在的各类异构信息,需要实现这些信息的索引建立过程,即索引初始化过程;二是针对在系统使

用过程中不断新增的各类异构信息,需要建立完整的信息处理过程,自动建立其对应的索引文件,系统每次新上传数据时,根据上传数据的类型采用对应的索引生成方案,即索引实时新增过程。

在具体的映射过程中,针对结构化数据(数据库表等)和非结构化信息(各类格式文档文件)分别采用 SolrDIH 和 SolrJ 两种方式。其中, Solr 数据导入模块 (Solr Data Import Handler, SolrDIH) 是指 Solr 的数据导入处理器,主要用来建立目标数据库的连接,并定义数据库表字段与对应索引域的映射关系;而 SolrJ 是使用 Java 语言连接 Solr 进行查询检索和索引更新维护的 jar 包,开发者可以利用其类接口方法在业务代码层实现对索引和查询的控制处理。

Solr DIH 索引方式和 SolrJ API 索引方式在处理百万级数据量时,前者的执行速度比后者快约 15%,但是后者在处理文档数据时更加灵活可靠,系统根据最优化原则使用 Solr DIH 处理数据库表信息,使用 SolrJ API 处理文档文件信息。具体处理流程如图 5 所示。

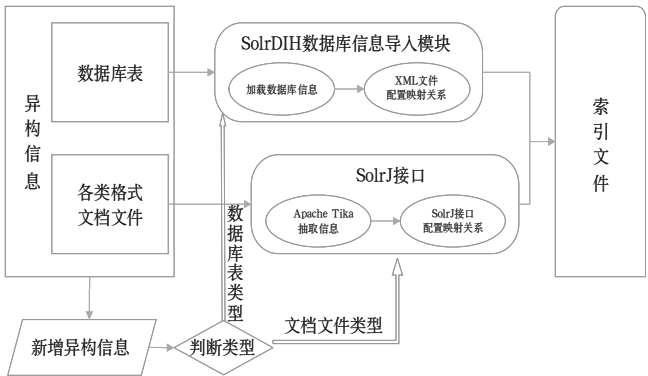


图 5 信息融合模块流程设计
Fig.5 Design of information fusion module

(1) 数据库表信息导入: Solr 的 DataImport 模块 (Solr DIH) 利用 XML 文件描述数据库连接信息和数据库表字段与域名的映射信息,通过在 Solr 实例的配置文件夹下新建 data-config.xml 文件来定义,主要分为数据源 (dataSource) 配置和文档域映射 (fieldMapping) 两部分,数据源配置部分主要描述数据库连接信息,文档域映射部分主要描述索引域名称与数据库字段的映射关系,其中, field 元素中的 name 属性表示域名称, column 属性表示对应的数据库表字段名称。这样就建立了数据库端表跟索引中域的映射关系,实现了数据库表信息的导入,并生成了对应的索引文件。

(2) 文档信息的导入: 文档信息的导入利用 SolrJ 的类接口方法,通过 Java 代码直接操作 Solr 索引的建立过程。具体步骤是: 先提取文档文件的各项数据,系统使用 Apache Tika 模块,利用其现有的解析类库,从不同格式的文档中 (如 HTML、PDF、WORD) 侦测和提取出元数据和结构化内容。即通过 Tika 对导入文档进行解析,得到文档的创建日期、标题、格式类型、内容等信息,并利用 SolrJ 接口操作导入至 Solr 实例生成索引文件。

算法 文档信息抽取并生成索引文件算法

输入 各类文档文件路径 pathList

输出 Solr 索引文件

第 1 步 通过上传的文档路径,获取文档路径 pathList。

第 2 步 根据文档的 pathList,利用 Tika 解析类遍历文档并抽取出文档的创建日期、标题、格式类型和文档内容等信息。

第 3 步 创建索引文件类 docs,遍历上述文档信息并分别映射至域 d_date、d_title、d_content (这些域在 Solr 索引结构中已经建立),并保存至索引文件 docs 中。

第 4 步 创建 Solr 服务器 Server 类,执行添加索引文件 docs 操作,将生成的索引文件保存至 Solr 索引信息库中。

2.2 创建 Solr 检索服务器

建立系统的 Solr 检索模块,即建立 Solr 服务器是实现系统检索功能的核心部分, Solr 服务器是独立于主系统之外建立的,主系统通过 HTTP 通讯的方式向 Solr 服务器提交查询信息并以同样的方式获得返回的结果信息。这种检索服务结构极大地降低了检索模块的耦合性,使得信息检索模块和业务管理模块得以有效分离,也为后续系统扩展提供了便利。

飞机故障异构信息检索系统在 Tomcat 环境下部署,而 Solr 检索模块依赖于 Solr 语义检索服务器,因此,首先需要实现 Solr 服务器在 Tomcat 环境下启动。Solr 服务器搭建完成后需要新建一个 Solr 实例,实例包含具体的索引文件以及索引加载的参数。实例文件夹中包含“conf”和“data”两个子文件夹以及 core.properties 文件。其中,“conf”文件夹主要包含与索引相关的具体参数和属性等;“data”文件夹是用来保存系统生成的索引文件;core.properties 用来描述实例的名称和数据路径等参数^[7]。

建立好索引模块并生成索引文件后,系统通过 SolrJ 接口方法开发索引及查询业务逻辑,并实现简单搜索、高级搜索、关键字高亮显示功能和综合排序功能。

2.2.1 简单搜索和高级搜索

简单搜索是用户只需要输入自然语句,系统将对所有索引域进行搜索;高级搜索是用户可以选择一项或多项自己感兴趣的域进行搜索,缩小查询范围。两者的区别主要是查询域的设置,如选择只在“g_fpart”(“故障件名称”)这个域中进行查询,或者选择查询“故障件名称”为“液压阀”、“故障位置”为“前机身”的故障记录。

2.2.2 高亮显示及分面查询

Solr 服务器返回的高亮结果是通过在查询到的关键字两端添加 `<span></span>` 等 HTML 标签实现;分面查询功能是对搜索结果按某些字段进行分组统计,同样可以通过 SolrJ 接口代码实现,部分 Java 代码如下:

```
// 参数 key 是要查询的关键字
SolrQuery query = new SolrQuery (key);
// 设置高亮为 true,并指定 HTML 标签格式;
query.setHighlight (true).setHighlightSimplePre ("<span
class='highlighter'>").setHighlightSimplePost ("</span>");
// 设置分层查询功能可用;
query.setFacet (true);
// 设置分层查询域名称为“故障性质”,最小数目为 1;
query.addFacetField ("g_fpro");
query.setFacetMinCount (1);
```

2.2.3 综合优化排序

Solr 本身提供基于 Lucene 的查询结果评分排序机制,该评分机制的目的是用来衡量查询语句和对应匹配文档之间的相似程度^[8]。具体计算方式为查询语句 (q) 中每个项 (t) 与文档 (d) 的匹配分值之和,如式 (1) 所示:

$$\sum_q (tf(t \text{ in } d) \times idf(t)^2 \times boost(t, \text{field in } d) \times \text{lengthNorm}(t, \text{field in } d)) \times coord(q, d) \times queryNorm(q) \quad (1)$$

式中: $tf(t \text{ in } d)$ 为项频率因子,文档 (d) 中出现项 (t) 的频率; $idf(t)$ 为项在倒排文档中出现的频率; $boost(t, \text{field in } d)$ 为域和文档的加权; $\text{lengthNorm}(t, \text{field in } d)$ 为域的归一化值,表示域中包含的项数量; $coord(q, d)$ 为协调因子 (Coordination factor),基于文档中包含查询的项个数; $queryNorm(q)$ 为每个查询的归一化值,指每个查询项权重的平方和。

加权因子 (boost) 用于对某个查询或者某个域的权重进行控制,域加权因子的默认值为 1.0。通过引入飞机故障领域词汇,将包含领域词汇的域的权重进行适当提高,

系统默认提高至 2.0,并可以由用户管理各个飞机故障领域词汇的权重值,使得返回结果更加有效,更具有参考价值。为了完善查询结果的排序机制,以 Solr 自身评价得分和信息发生时间两方面为权重值,重新得到综合评分排序结果。

3 系统实现与验证

通过 Solr 服务器创建检索系统的优势在于,一旦建立起针对异构信息资源进行索引创建和更新的机制后,就可以专注于系统功能和用户交互界面的设计、检索结果的解析和展示,以及系统整体功能的实现^[9]。

系统最终实现针对异构信息的索引结构组织、索引文件创建、简单查询和高级查询,并在结果展示中实现了关键字高亮显示和综合优化排序,初步实现了异构信息检索系统的总体功能,如图 6 所示,图 6 中输入查询语句“氧气管道有裂纹”,左下侧为查询出的数据库表信息结果,共 79 项相关;右下侧为查询出的文档文件信息结果,共 3 项相关。



图 6 检索系统演示界面
Fig.6 Interface of retrieval system

实验证明,在配置的服务器测试环境 (CPU 为 Inter Core i5 6500 3.2GHz 4 核;内存为 16GB;操作系统为 64 bit Windows 7) 下,针对数据库表数据和文本文档的索引和查询时间进行测试,具体的测试时间结果如表 3 所示,测试结果描述了数据库表信息和文件文档信息在不同数量情况下,索引和查询时所耗费的时间对比情况,其中,文本文档大小均为 10~100kB,测试结果时间不包含前台响应时间。数据表明,表信息和文本文档均是索引时间占了较大比重,查询时间在一定数量级内几乎均为毫秒级;随着信息条数的增加,索引时间也随之增加,而对查询时间的影响不大,系统在实际应用过程中,基本能满足用户的快速检索的需求。

表3 系统测试结果
Table 3 Results of system test

信息类型	信息条数	索引时间 /ms	查询时间 /ms
数据库表信息	1000	1492	2
	10000	3212	3
	100000	14780	5
文本文档 (DOC、PDF、TXT、 EXCEL)	10	1618	3
	100	3741	3
	1000	12835	4

4 结束语

本文基于 Solr 服务架构,快速搭建企业异构信息检索系统,以企业的飞机故障异构信息为源数据对象,针对数据库表信息等结构化信息和各类型文档等非结构化信息进行索引结构设计、索引文件的创建、搜索逻辑的定义,以及各种检索功能的扩展,如高亮显示、分面查询、优化排序等。在系统实现过程中重点关注 3 个方面:一是根据信息源特点设计索引结构,实现统一检索;二是设计智能的信息融合模块,自动将新上传的异构信息根据其类型快速地转换为索引信息,实现索引的自动更新和实时检索;三是根据用户检索需求实现各类检索功能,提高领域内词汇的权重,优化检索结果。总体来说,基于 Solr 的异构信息检索系统能够对目前企业累计的信息资源进行高效检索,为企业在信息化时代提供可靠的竞争力。

AST

参考文献

[1] 宋瀚涛. 异构信息源集成技术[J]. 北京理工大学学报, 2002, 22 (5): 533-536.
SONG Hantao. Integration technology of heterogeneous information sources [J]. Journal of Beijing Institute of Technology, 2002, 22 (5): 533-536. (in Chinese)

[2] 王雅坤, 成全. 信息检索相关性研究综述及发展趋势[J]. 图书与情报, 2012 (1): 88-96.
WANG Yakun, CHENG Quan. Research on the development trends and review of information retrieval relevance [J]. Library & Information, 2012 (1): 88-96. (in Chinese)

[3] 周敬才, 胡华平, 岳虹. 基于 Lucene 全文检索系统的设计与实现[J]. 计算机工程与科学, 2015, 37 (2): 252-256.
ZHOU Jingcai, HU Huaping, YUE Hong. Design and implementation of Lucene-based full-text retrieval system[J]. Computer Engineering & Science, 2015, 37 (2): 252-256. (in

Chinese)

[4] 梁艳, 刘双广, 劳定雄. 基于 Solr 的异构数据融合检索技术[J]. 无线互联科技, 2013 (5): 61-64.
LIANG Yan, LIU Shuangguang, LAO Dingxiong. The retrieval technology of heterogeneous data integration based on Solr [J]. Wireless Internet Technology, 2013 (5): 61-64. (in Chinese)

[5] 谢文亮, 王石榴. 基于语义 Web 的科技期刊网络信息检索及其应用[J]. 科技管理研究, 2015, 35 (2): 196-200.
XIE Wenliang, WANG Shiliu. Semantic Web-based network information retrieval for scientific periodical journal and its application[J]. Science and Technology Management Research, 2015, 35 (2): 196-200. (in Chinese)

[6] 王兰成. 论实现异构档案信息整合的信息组织与检索技术[J]. 档案信息化, 2011 (2): 55-59.
WANG Lancheng. Information organization and retrieval technology of integration of heterogeneous archival information[J]. Archives Science Study, 2011 (2): 55-59. (in Chinese)

[7] 高利雅. 基于 Solr 全文搜索引擎的研究与实现[D]. 成都: 电子科技大学, 2014.
GAO Liya. The research and implementation of full-text search engine based on Solr [D]. Chengdu: University of Electronic Science and Technology of China, 2014. (in Chinese)

[8] Michael M, Erik H, Otis G. Lucene 实战[M]. 北京: 人民邮电出版社, 2011.
Michael M, Erik H, Otis G. Lucene in action [M]. Beijing: Posts & Telecom Press, 2011. (in Chinese)

[9] 李戴维, 李宁. 基于 Solr 的分布式全文检索系统的研究与实现[J]. 计算机与现代化, 2012 (11): 171-176.
LI Daiwei, LI Ning. Research and implementation of distributed full-text retrieval system based on Solr [J]. Computer and Modernization, 2012 (11): 171-176. (in Chinese)

作者简介

印奇(1987—)男,硕士。主要研究方向:装备保障信息化。
Tel: 18510372180
E-mail: yinqiyq@163.com

李青(1961—)女,教授,博士。主要研究方向:知识管理,企业信息化。

黄鹏(1987—)男,工程师。主要研究方向:装备维修保障。

Design of Aircraft Fault Heterogeneous Information Retrieval System Based on Solr

YIN Qi^{1,*}, LI Qing¹, HUANG Peng²

1. School of Mechanical Engineering and Automation, Beihang University, Beijing 100191, China

2. Chengdu Aircraft Industrial (Group) Co. Ltd., Chengdu 610092, China

Abstract: In view of the status that the heterogeneous information of aircraft fault is rising on the quantity of resources and the kind of carriers, the aircraft fault heterogeneous information retrieval system based on Solr was put forward to realize the unified retrieval of the heterogeneous information resources and meet the retrieval efficiency requirements. Firstly, designed the index structure of according to the data characteristics to realize the unified retrieval. Then, generated the index file of heterogeneous information and imported into Solr platform to realize the information fusion. Lastly, realized several kinds of complex search functions through SolrJ API, and optimized the rank of results by the way of synthesizing the degree of the comprehensive correlation. The experimental results show that the system can meet the needs of a variety of complex search, effectively solve the problems of heterogeneous information resources of enterprises, and provide technical support for enterprises to make quick decisions.

Key Words: Solr; heterogeneous information; index; retrieval system

Received: 2016-12-16; **Revised:** 2017-01-17; **Accepted:** 2017-03-10

***Corresponding author. Tel. :**18510372180 **E-mail:** yinqiyq@163.com