文章编号：1005-8451（2016）10-0053-04

Solr在乐龄易购网站中的应用

蔡宇晶[1]，孙玫肖[2]，朱建军[2]
（1.中国铁道科学研究院，北京 100081；2.中国铁道科学研究院 电子计算技术研究所，北京 100081）

摘 要：为了提高乐龄易购网站的搜索效率，对不同搜索服务器进行了功能性对比，选择了拥有快速高效全文搜索功能的Solr作为网站搜索服务器，对其原理和特性进行了分析，并对其在乐龄易购网站中的配置和应用进行了介绍。通过配置Solr建立索引文件结构，改变了网站从数据库直接访问数据的查询形式，解决了网站用户搜索缓慢的问题，为电商平台实现高效搜索服务提供了一种模式。

关键词：Solr搜索服务器；数据；索引
中图分类号：TP39 **文献标识码：**A

Application of Solr in Lelingyigou Website

CAI Yujing[1], SUN Meixiao[2], ZHU Jianjun[2]
(1. China Academy of Railway Sciences, Beijing 100081, China;
2. Institute of Computing Technologies, China Academy of Railway Sciences, Beijing 100081, China)

Abstract: In order to improve the search efficiency, through the functional comparison of different search server, this article chose the Solr as the website search server. The Solr processes the efficient and comprehensive full-text search capabilities. The article analyzed the principles and characteristic of the Solr, introduced the configuring and application of the Solr in Lelingyigou Website, By deploying the Solr to index file structure, it was changed the website's inquiry form that accesses the database directory, and solved the problem of low searching speed, and provided a model for the e-commerce platform to implement high efficient search service.

Key words: Solr search server; data; index

乐龄易购网站是一个老年旅游电子商务平台，作为一个电商网站，用户需要在网站主页根据关键字进行产品搜索。传统的搜索方式是直接从数据库中实现字段查询，这种查询方式效率较低，当用户搜索量较大时会造成搜索速率下降，影响用户的使用体验。为了解决这一问题，提高网站的搜索效率，本文进行了探讨并给出解决方案。

1 网站现状

1.1 乐龄易购网站概述

乐龄易购网站拥有老年旅游和异地（旅游）养老等核心产品，同时提供新闻资讯、订单处理、在线支付、旅游点评、会员中心等内容。乐龄易购网站使用 Broadleaf 作为项目开发框架，Broadleaf 是一个开源的电子商务平台，基于 Spring 框架开发，提供一个可靠、可扩展的架构，能够进行深度的定制和快速开发。

收稿日期：2016-01-22
作者简介：蔡宇晶，在读硕士研究生；孙玫肖，研究员。

1.2 网站面临的问题

随着互连网的迅速发展，Web 信息量成指数倍增长，如何快速、准确地从如此庞大的信息库中获取自己需要的信息，成为互联网用户面临的一个重要问题。

乐龄易购网站作为一个电商网站，需要用户在主页进行产品搜索。产品数据一开始是直接从数据库访问的，这样设置可以实现功能需求，但搜索时效不高，用户需要等待的时间很长，会影响用户使用网站的整体感受。提高搜索效率，改善用户体验，增加网站的可用性是必须要解决的问题。

1.3 搜索服务器的选择

使用搜索服务器能够提高搜索效率，当前主流的搜索服务器有 Solr、Elasticsearch 和 Sphinx 等。

3 种搜索引擎在技术上各有优缺点，而网站搜索服务器的选择，除了要充分考虑其技术成熟度和搜索效率外，还需权衡电商平台搜索数据规模量以及网站开发语言与搜索服务器的兼容度。

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

Sphinx 是基于 SQL 的全文检索服务器，支持比数据库本身更加专业的搜索功能，能够方便地与 SQL 数据库进行集成，其索引无法进行动态添加。在中文分词方面，Sphinx 目前只支持 sphinx for chinese 和 mmseg3 两种分词。

Elasticsearch 是一个实时的分布式多用户搜索引擎，用于全文搜索及分析，基于 Lucene 实现。Elasticsearch 更适用于数据规模较大的检索，其优势在于实时导入搜索，但在传统搜索数据方面性能较差。

Solr 拥有快速全面的搜索功能，其开发基于 Java 语言，可以很方便的嵌入到 Java 应用程序中（本文所阐述的乐龄易购电商平台基于 Java 语言开发），扩展性强。Solr 提供 field collapsing（分组查询）来避免类似搜索结果的重复性，而 Sphinx 做不到。Solr 是提供极快的搜索目录的行业领导者，在对已有数据进行搜索时，Solr 搜索效率明显强于 Elasticsearch。Solr 基于开放接口的标准，能够提供相对丰富的查询功能，除了基本的查询，还包括分组，排序及统计等功能。除此之外，Solr 拥有高级的全文搜索功能，能够在网络流通量相对大的情况下进行优化，并且提供监控日志，极大地提高了网站站内搜索的效率，能够快速查询到用户所输入的查询信息。

基于上述原因，乐龄易购网站选择了 Solr 作为网站的搜索服务器。

2 Solr简介

Solr 是一个开源的搜索服务器，它能够提供类似 Web 服务的 API 接口。用户按照相应格式要求配置 xml 文件，通过 http 请求将配置好的 xml 文件提交到服务器，从而创建索引。Solr 开发通过 Java 语言，框架源于 Lucene 并在其基础上进行了扩展。存储在 Solr 中的资源均以 document（文档）为对象，每一个 document 由一系列的 field（字段）构成，每一个 field 可以用来代表所存储资源的属性。Solr 目前是一个比较成熟稳定的搜索引擎，已经被一些大型的网站所使用。

2.1 Solr特性

Solr 具有高效、灵活的缓存功能，能够高亮显示搜索结果，特性包括以下几点：

（1）高级的全文搜索功能；

（2）基于开放接口的标准；

（3）可伸缩性－能够有效地复制到另外一个 Solr 搜索服务器；

（4）使用 xml 配置达到灵活性和适配性。

2.2 Solr原理

对于数据索引的增删改查，Solr 提供了一种可实现的通用 http 接口。具体原理如下：用户向部署在 servlet 容器中的 Solr Web 应用程序发送 http 请求，启动索引和搜索，Solr 接受请求并确定所需的 SolrRequestHandler，处理相关请求后 http 以相同的方式返回响应。

3 Solr在乐龄易购网站的配置与应用

3.1 配置方法

Solr 保持高效运行需要考虑很多因素，通常认为 Solr 里最重要的配置文件就是 schema.xml，solrconfig.xml。

3.1.1 模式配置schema.xml

schema.xml 类似于数据表配置文件，是业务逻辑的核心。定义一个文档会包含什么字段、字段如何索引等都是在 schema 中完成的。模式配置主要分为 types 部分、fields 部分以及其他一些缺省设置。

（1）types 配置

<types> 部分是一些常用的可重用定义，用于声明 Solr 如何处理 field，以及这个 field 如何处理查询。

需要在 types 节点内定义一个 fieldType 子节点，同时声明 name,class 等一些参数。name 是一个标识，即这个 fieldType 的名称，<field> 中的 type 引用的就是这个 name；class 确定此 fieldType 的实际行为。

在对 fieldType 进行定义时，定义这个 fieldType 类型的数据在生成索引和搜索查询时所需使用的分析器 analyzer 十分重要，使用不同的 analyzer 可以对这个 fieldType 的数据进行不同处理，比如中文切词、删除空白等。分析器需要定义分词器和过滤器，analyzer type="index" 表示生成索引时使用的分词器及过滤器，analyzer type="query" 表示搜索查询时所

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

使用的分词器及过滤器，以 analyzer type="inder" 为例：

<fieldType name="text_general"class="solr.TextField"positionIncrementGap="100">

<analyzer type="index">

<tokenizer class="solr.StandardTokenizerFactory" /> -- 定义分词器

<filter class="solr.LowerCaseFilterFactory" /> -- 定义过滤器

</analyzer>

（2）fileds 配置

在 fields 节点内定义具体的字段，即模式的 <fields> 部分，是添加到索引文件中出现的属性名称。

field 是固定的字段设置，声明 field 包含 name、字段类型名；type 即之前定义过的各种 fieldType；indexed 代表是否需要建立索引以方便之后对该字段进行查询，该值置为 true 时表示数据应被搜索排序；stored 代表是否需要随索引存储该字段内容，该值置为 true 时表示在正常情况下搜索结果中会包括这个字段。

当有太多字段无法一一进行设定时，需要使用动态字段 dynamicField 属性，即不制定具体名称，仅定义字段名称的规则，如下：

<dynamicField name="*_i" type="int" indexed="true" stored="true" />

* 号是通配符，*_i 表示定义所有以 _i 做结尾的字段。

（3）其他配置

schema.xml 文件最后部分声明了一些其他配置。

在设计数据库的表结构时会为表建立唯一索引来标识表中数据的唯一性，同样 schema 配置也需要一个唯一标识符 uniquekey，这个字段必须填写，否则索引可能会报错；schema 配置时一般会配置搜索字段，如果没有搜索字段时就需要用到默认搜索字段 defaultSearchField；copyField 复制字段，可以将多个字段值复制至一个字段，当多种方式索引相同内容时可以使用复制字段。

3.1.2 solrconfig.xml 配置

solrconfig.xml 文件大部分的参数是用来配置 Solr 本身的，指定了 Solr 如何处理索引、突出显示、分类、搜索以及其他请求。solrconfig.xml 分为索引配置和查询配置。

3.2 在网站中实现全文检索

Solr 在实现全文检索时一般分为两大流程：（1）创建索引（Indexing）；（2）搜索索引（Search）。传统的数据库都是由文档 id 来映射文档中的关键字，而 Solr 采用的是反向索引，反向索引就是从关键字到文档的映射过程，关键字可以在内容分析时随时添加、删除或修改。

3.2.1 创建索引

原始文档需要被交给分词组件进行中文分词，用户使用网站在前台进行搜索时，搜索时间很慢会影响网站用户体验，因此需要合适的分词组件。Solr 默认提供的分词组件进行中文分词时效果十分不理想，当 fieldType 选择“text_general”（即 Solr 默认分词组件）为分词依据时，输入中文“选择一条旅游线路”进行分词，观察到分词效果如图 1 所示。

此时的分词是将句子按字依次隔开，分词模式效率很低，搜索时需要一个个去索引包含这些汉字的全部文档，做交叉“与”运算，即包含这些字并且位置连续的文档才是能满足需求的结果，搜索时效将会十分缓慢。运行这样的索引结构，用户搜索时需要等待很长时间，而且对硬件和算法都是极大的挑战。运用中文分词检索可以解决这个问题，词是汉语中最小的一种语义单位，当搜索引擎索引的 key 值以词为单位时，会大幅度的提高该引擎搜索结果的准

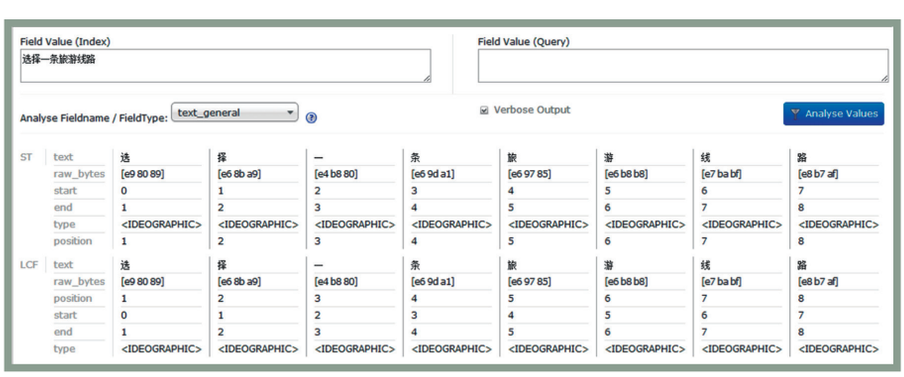

图1 Solr默认分词组件分词效果

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

确性，同时也可以保证在搜索过程中保持相对较小的计算量。比较常用的能够和 Solr 集成的中文分词组件有以下几种，mmseg4j、Paoding、IkAnalyzer 等等，这些分词组件各有各的特点，用户可以选择最适合自己的。本电商网站选用的组件是 mmseg4j。将其下载安装之后，需要修改配置文件 schema.xml，添加如下代码：

```
<!--Solr mmseg4j -->
<fieldType name="text_zh" class="solr.TextField "positionIncrementGap="100">
    <analyzer>
        <tokenizer class="com.chenlb.mmseg4j.solr.MMSegTokenizerFactory" mode="max-word" />
    </analyzer>
</fieldType>
```

配置完成后检查一下 mmseg4j 分词的效果，再次输入“选择一条旅游线路”，能够看见分词效果如图 2 所示。

现在的分词基本符合正常语义，较之前效果优化了许多，提高了搜索效率，用户搜索时的体验度会更好。

分好的词汇单元会被传给语言处理组件，语言处理组件对得到的词元做一些语言相关的处理，将得到的词传递给索引组件，索引创建完成。

3.2.2　搜索索引

索引创建完毕后用户可以找到需要的文档，如果找到的结果数量很庞大，让用户去一一筛选显然不可行。从用户体验感出发，需要运用搜索索引在大量的文档中找到最相关的文档进行排序。

搜索索引先对查询内容进行词法分析、语法分析、语言处理，得到一个查询树，之后通过索引存储将索引读到内存，再利用查询树来搜索索引，从而得到每个词的相关文档链表，对文档链表进行判断并得到结果文档，根据查询语句与文档的相关性，对查询结果进行筛选排序，使用户得到一个相对满意的结果。

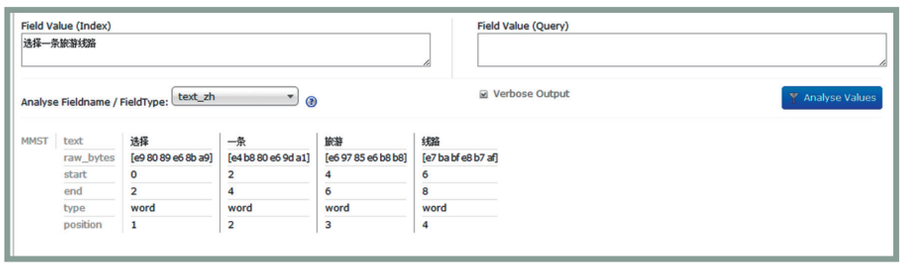

图2　中文分词组件配置后分词效果

4　结束语

Solr 搜索服务器拥有快速高效全面的全文搜索功能，能够与多种中文分词组件进行集成。在乐龄易购网站中，通过配置 Solr 建立一个索引文件结构匹配相关规则，根据文档的字段 id 来唯一标识文档数据，改变了网站从数据库直接访问数据的查询形式，解决了用户搜索缓慢的问题，实现了网站数据的高效搜索。

参考文献：

[1] Trey Grainger, Timothy Potter. Solr in Action[M]. Manning Publications, 2014.

[2] David Smiley, Eric Pugh. Apache Solr 3 Enterprise Search Server[M]. Packt Publishing, 2011.

[3] Wikipedia:Solr[DB/OL]. https://no.wikipedia.org/wiki/Solr.

[4] 刘建国，黄厚宽．使用分类和聚类提高搜索引擎的可用性 [J]. 铁路计算机应用，2006，15（3）：44-46.

[5] 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.

责任编辑　付　思

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net