基于 Solr 的大规模标准文献可视化分析系统①

张 震[1], 甘克勤[2]

[1](北京航空航天大学 计算机学院, 北京 100191)
[2](标准化研究院, 北京 100191)

摘 要: 国家标准馆是唯一的国家级标准收藏机构, 建成了规模庞大的标准文献题录数据库、全文数据库. 但是面对海量的数据资源, 标准文献研究人员在没有计算机相关知识的情况下难以对相关数据进行全面的了解和研究, 传统的研究方式也无法实时并直观的对统计数据进行展示. 本文基于这些问题, 开发了大规模标准文献可视化分析系统, 设计和实现了可以自由定制的数据统计功能以及对标准文献的起草人、起草机构的关联分析功能. 本系统为标准文献研究领域的研究人员提供了一个对标准文献资源高效便捷的可视化分析工具, 研究人员对统计数据进行定制就能够获取到有效的数据, 大幅提升了标准文献资源的分析效率.

关键词: Solr; 数据分析; 标准文献; 关联分析

Standard Documents Resources Visual Analysis Systems Based on Solr

ZHANG Zhen[1], GAN Ke-Qin[2]

[1](School of Computer Science & Technology, Beihang University, Beijing 100191, China)
[2](China National Institute of Standardization, Beijing 100191, China)

Abstract: China National Institute of Standardization currently has large bibliographic databases and full-text databases. Facing the massive data, researchers without computer-related knowledge are difficult to conduct a comprehensive understanding of the relevant data. Thus this paper proposes the standard documents resources analysis system based on Solr. This paper designed and implemented the customized statistical features and relation analysis of the drafters and organizations of the standard documents. The system allows researchers to freely analyze the standard documents data, which greatly enhances the efficiency of the data analysis and promotes the development of related research.

Key words: Solr; data analysis; standard documents; relation analysis

国家标准馆[1]是我国唯一的国家级标准收藏机构, 收藏了全部的国内标准, 包括国家标准、行业标准、地方标准, 以及世界主要国际标准化组织、发达国家和重要标准学协会的标准, 建成了规模庞大的标准文献题录数据库、全文数据库, 记录了多年的标准使用情况, 目前标准文献题录数据库量已达 160 万余条.

面对海量的数据, 标准文献研究人员往往需要对整体标准数据进行研究, 包括趋势分析, 数量对比等[2-5]. 在实际中, 部分研究人员使用 SQL 语句对数据库中的数据进行分析, 但是这种方法需要相关研究人员具有较高的计算机水平, 对相关数据库结构有深入的了解, 并且在未经过优化的情况下 SQL 语句的执行也要耗费几分钟到几十分钟不等的时间, 无法迅速地看到相关数据结果. 同时, 数据库中提供的表格形式的数据也缺乏表现力, 无法让研究人员对数据整体有直观的了解.

为解决这些问题, 本文提出了基于 Solr 的大规模标准文献可视化分析系统. 将不同资源库中的标准文献数据进行统一集成, 并在此基础上实现高效的数据统计. 本文采用 Solr 的主要原因是其在信息检索领域属于比较成熟的解决方案[6-10], 同时针对大规模数据也拥有高效快速的统计方案, 也能够为将来使用该系

① 基金项目:中央基本科研业务费支撑项目(252015Y-4003)
收稿时间:2015-06-30;收到修改稿时间:2015-08-20

统对标准文献进行进一步文本分析做好基础.

1 Solr

Solr 是基于 Java 的高性能全文检索工具, 在企业级应用中被广泛的使用. Solr 基于 Lucene 构建, 在 Lucene 之上提供了较为完整的搜索功能, 主要原理是 Solr 服务器接受客户端的 HTTP 的数据请求, 然后通过 XML、JSON 等格式来返回数据结果. Solr 能够将来源于不同类型数据库、不同格式的数据统一建立索引, 并进行全文检索.

除了强大的检索功能外, Solr 对于已经索引的文档能够进行各种类型的数据统计. 包括按照多维度统计文档数量, 统计独特值数量, 数据求和等. 本系统使用 Solr 实现全文检索功能, 并使用其统计功能完成系统的高效数据统计.

2 系统概述

本系统总体分成用户管理模块、数据管理模块、模板管理模块、预设管理模块、关联分析模块、主题分析模块以及标准检索模块. 系统总体模块图如图 1 所示.

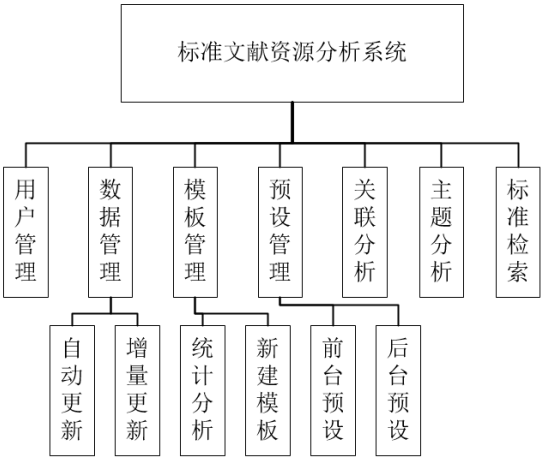

图 1 系统总体模块图

用户管理: 系统用户分为管理员和普通用户, 并拥有不同的系统权限.

数据管理: 系统定时将原始 Oracle 数据库中的信息导入到 Solr 中, 提供每天定时的增量导入和管理员用户的手动导入.

模板管理: 每一个用户都可以建立自己的统计模板, 运用自己的统计模板统计数据获得结果.

预设管理: 管理员用户能够事先预设好统计模板, 未登录系统或者新用户都可以查看一些预设的统计图.

标准检索: 用户可以通过关键词检索, 查看标准的基本信息以及关联分析和主题分析的结果.

关联分析: 对于每份标准, 计算出其起草人、起草机构之间的关联关系.

主题分析: 对于标准文献全文, 自动计算出文献的关键词列表.

3 关键模块及关键技术

3.1 数据管理

标准文献数据库中主要分为题录表、电子资源表两大部分, 每部分又都有一些相关的数据表, 原始的数据都是存储在 Oracle 数据库中. 为了方便统计需求, 将其导入 Solr 中后建立了 9 个 Solr 的 collection, 每一个 collection 代表了一个类型的数据.

数据的导入是通过配置 Solr 的 DataImportHandler, 解决了原始数据处于不同数据库、不同机器以及不同数据库(Oracle 和 MySQL)的问题.

部分数据导入配置文件如下:

```
<dataConfig>
<dataSource type="JdbcDataSource"
driver="oracle.jdbc.OracleDriver"
url="jdbc:oracle:thin:@****"
user="****"
password="****"/>
<document>
<entity name="resource"
query="select t.*,DBMS_LOB.SUBSTR(t.A330) as
DES from T_PRODUCT_STANDARD t where
A104='CN-GB'"
deltaQuery="select A001 from
T_PRODUCT_STANDARD t where
update_date >
to_date('${dih.last_index_time}','yyyy-mm-dd
hh24:mi:ss')"
deltaImportQuery="select
t.*,DBMS_LOB.SUBSTR(t.A330) as DES from
T_PRODUCT_STANDARD t where A104
```

='CN-GB' AND t.A001='${dih.delta.A001}'">
</entity>
</document>
</dataConfig>

DataSource 节点中包含了连接的基本信息，包括数据库驱动，用户名，密码等，Document 节点中包含了导入数据所使用的 SQL 语句，以及增量导入数据的判定及导入语句.

数据导入系统后，为了保证系统数据的实时性，需要经常对系统数据进行更新，本系统分为自动更新与手动更新两部分. 自动更新是指系统在每天指定时间根据原始数据表中 update_time 字段来对上次更新后修改过的数据进行增量更新. 手动更新是指管理员通过系统手动控制更新的进行. 增量更新减少了每次更新的数据量，从而在几分钟至十几分钟就能完成当日索引的更新.

3.2 统计分析

3.2.1 底层统计

通过对标准化研究院研究人员的调研，其日常统计数据使用到的数据库功能主要包括按照某个字段或多个字段进行 GROUP BY 操作，从而求取对应的文献数量或其他统计值，题录表与电子资源表进行 JOIN 操作以及与代码表进行 JOIN 获取代码对应的中文名称等.

这些基本操作都能够完全的对应到 Solr 的统计功能中. GROUP BY 操作通过使用 Solr 中的 Facet 与 Stat 功能，能够实现按照多个维度的统计操作. 如按照字段 A 和字段 B 进行统计操作，字段 A 取值为 A1，A2，字段 B 取值为 B1, B2, Solr 的 Facet 功能就能直接计算出，字段 A 为 A1，字段 B 为 B1 的文献数量有 100 篇，字段 A 为 A2，字段 B 为 B1 的文献数量有 200 篇等信息，从而实现了与 GROUP BY 的同样效果.

题录表与电子资源表的 JOIN 操作直接在数据导入时就使用 JOIN 语句进行实现，因为题录表与电子资源表是一对多的关系，在统计时，如果要统计文献数量就使用对题录表中的 id 求 distinct 值，如果要统计电子资源数量就直接统计全部数据数量.

与代码表的 JOIN 操作可以事先将代码表导入到系统中，在输出结果时直接将代码替换即可.

这样就在统计数据时实现了 Solr 与 SQL 语句的同样效果，从而实现等同的统计效果，但在速度上由于 Solr 对统计功能做了优化，时间从 SQL 语句的几分钟至几十分钟减少到了 5 秒钟之内.

3.2.2 可视化分析

用户在新建统计模板后，调用统计模板，系统首先使用底层统计中介绍的方法到 Solr 服务器中查询，然后将返回的结果存储到自定义的数据结构中.

用户可以对获取到的数据进行自由定制统计操作，只需要在系统界面将所需的统计维度选中，也可以选中在已有的统计维度基础上进行组合生成的新统计维度. 然后选择生成图的种类(包括饼图、直方图与折线图)，就能够实现实时的可视化统计，如图 2 所示.

图 2 可视化统计

为了实现可视化统计的实时性，系统在第一次接受用户请求时将原始统计数据放置到 Session 中，用户接下来的各种操作都在原始数据基础上生成新的数据拷贝来进行处理，避免了在更换统计维度或更换统计图种类时反复查询 Solr 服务器带来的时间延迟和访问压力.

同时在系统中也使用了 Google Guava 自带的缓存系统对相似的统计模板进行了缓存操作，如果两个用户拥有相似的统计模板，只要其中一个用户使用过这个模板，另一个用户在使用时就能够直接从缓存中读取，减少了系统的时间延迟.

3.3 关联分析

标准文献资源中包含了文献的起草人与起草机构等信息，从中可以挖掘出标准文献研究领域的人物关系与机构之间的关系.

本系统中主要处理了三种关系.

1) 通过某一篇标准文献出发，显示其起草人，再挖掘其起草人起草的其他文献，如图 3 所示.

2) 通过某一篇标准文献出发，显示其起草机构，

再挖掘其起草机构起草的其他文献.

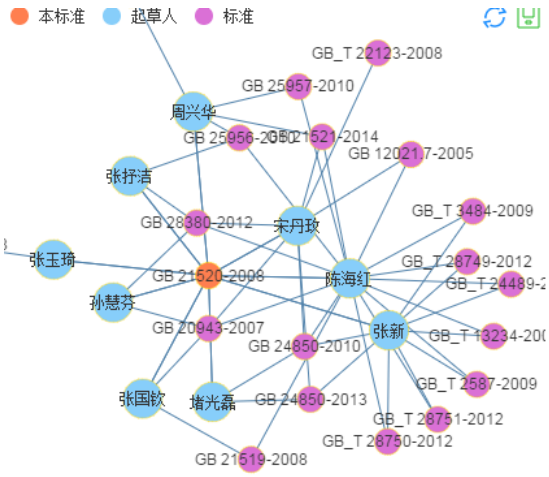

图 3　标准文献起草人图

3) 通过某一篇标准文献出发, 显示其起草人, 再挖掘其起草人起草的其他文献中的人物, 构建标准研究领域之间的人物关系.

4　系统实现与应用

本系统基于 Java Servlet 开发, 数据库采用 MySQL 5.5,Solr 部分采用最新的 Solr 5.2.0, 应用容器采用 Tomcat 6.0.

4.1 数据分析功能实现

系统的分析功能主要是用 Solr 的 Facet 和 Stat 功能实现的, 举例一个最基本的查询语句为

http://localhost:8983/solr/order/select?
q=*%3A*&&stats=true&facet=true
&stats.field={!tag=piv1}id
&facet.pivot={!stats=piv1}A000,A001

其中, 前半部分为 Solr 服务器的地址, select 后面是相关的查询语句, 首先将 stat 和 facet 功能开启, 然后选择 stats 功能针对的字段是 id, facet 功能针对的字段是 A000 与 A001, 然后 stats 与 facet 通过 piv1 这个字段联系起来, 即对 A000 与 A001 字段做统计. 一个示意的统计结果如下:

"facet_pivot":{
　"A000,A001":[{
　　"field":"A000",
　　"value":"A",
　　"count":54061,
　　"pivot":[{
　　"field":"A001",
　　"value":"5884559",
　　"count":216,
　　"stats":{
　　"stats_fields":{
　　"id":{
　　"min":"100048",
　　"max":"99607",
　　"count":216,
　　"missing":0}}}},

结果中首先显示了 A000 字段取值为"A"的问的数量总共有 54061 条, 然后在 A000 字段取值为"A"的条件下, A001 字段取值为"5884559"的共有 216 条, 其中对 id 进行 stats 统计, 显示其中的最小值, 最大值, 由于这个例子中的 id 不是数值型字段, 最大最小值可能不太准确.

在实际系统实现中, 查询语句和参数可能还会再要复杂一些, 但原理与上述并无太大区别. 在接受用户选择的统计字段统计量后, 系统使用 Solr 进行查询, 将结果存储到相关的数据结构中, 然后将结果进行可视化.

4.2 可视化功能实现

可视化效果的实现主要使用百度的 Echarts 可视化控件, 通过传入不同的数据参数, 控制图表的显示, 系统中主要采用的图种类主要包括条形图, 柱状图和饼图. 数据示例如下, 该数据为出版日期为 2010-2013 年的中美标准数量:

{
"jxAxis":["2010","2011","2012","2013"]",
"jdata":"{" 美 国 ":["8658","5521","5000","2842"],"
中国":["6267","8081","6268","7152"]}",
"jseries":"["中国","美国"]"
}

可视化的数据主要包含 jxAxis, jseries, jdata 三部分, jxAxis 对应的是条形图, 柱状图的横轴, jseries 对应的是同一个图中同一横轴数值中不同的对比数据, jdata 对应的是详细的数字数据. 这三种数据也能够和 Echarts 本身的数据结构很好的结合在一起, 实现可视化效果. 图 4 和图 5 为上述数据的不同类型图展示.

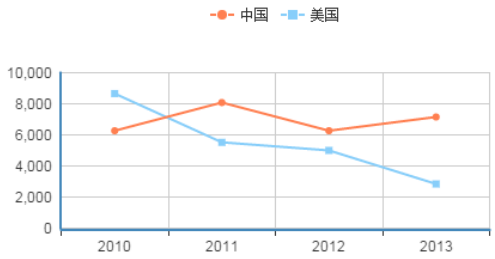

图 4 条型统计图

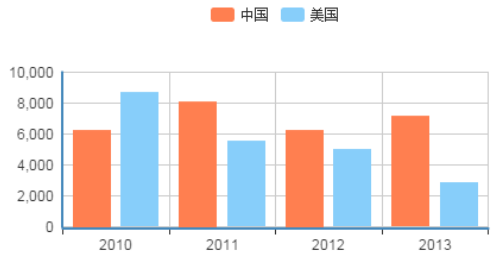

图 5 柱状统计图

4.3 系统应用

本系统已经成功部署在中国标准化研究院内部，供研究人员使用来对国家标准馆藏资源进行分析. 研究人员可以对 160 万馆藏题录数据与 80 万电子文献数据从文献状态、国别、ICS 分类、CCS 分类等维度对数据进行统计，并生成对应的统计图表. 从而在数据中发现规律，进行更一步深入的研究，也能利用统计数据来快速研究自己的研究假设.

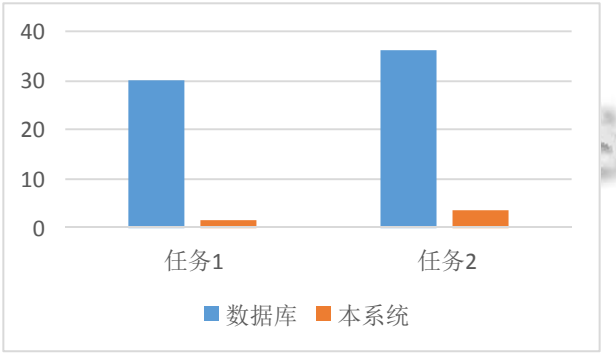

图 6 运行速度对比图

本系统性能优异，如图 6 为在数据库执行 SQL 语句与使用本系统执行两项任务的消耗时间对比. 任务 1 为按照文献状态进行统计，任务 2 为按照文献状态与国别进行统计. 可以看到，在大规模题录数据库的数据上，使用 SQL 语句两条查询的时间都在 30s 左右，而本系统中都在 4s 以内就能返回结果，大幅提高了统计效率.

5 结语

本系统为标准文献研究领域的研究人员提供了一个对标准文献资源高效便捷的可视化分析工具，研究人员不需要掌握数据库方面的知识，也不需要对资源库有深入的了解，只需要对自己想要了解的数据做一些模板定制就能够获取到有效的数据，大幅提升了标准文献资源的分析效率.

同时本系统也可以作为标准文献领域数据统计系统的基础，支持加入新的其他类型数据，从而使本系统的用途更加广泛.

参考文献

1 中国标准化研究院. http://www.cnis.gov.cn. 2015.
2 李景,汪滨,周洁,刘恬渊,李菁.以国家标准馆馆藏数据浅析我国标准制定的领域动态趋势.中国标准化,2007,(11):46–49.
3 李景,汪滨,周洁,刘恬渊,李菁.我国标准制定的领域动态趋势分析—基于国家标准馆 2006-2007 年度国内外标准文献新到馆藏.图书情报工作,2009,53(1):56–60,68.
4 李景,刘亚中.农业标准文献专业分布与热点领域的文献计量学分析—以国家标准馆 2006-2008 年度新到馆藏为例.图书情报工作,2009,53(18):44–4778.
5 任晨鸿.我国环境保护标准制定动态趋势的文献计量学分析—基于上海标准文献馆 2010-2012 年度新到馆藏.中国标准导报,2014,(4):47–50.
6 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.现代图书情报技术,2011,(6):51–58.
7 霍庆,刘培植.使用 Solr 为大数据库搭建搜索引擎.软件,2011,32(6):11–14.
8 李戴维,李宁.基于 Solr 的分布式全文检索系统的研究与实现.计算机与现代化,2012,(11):171–176.
9 温慧明,宫晓辉.基于 Solr 的科技成果查新系统的构建研究.计算机技术与发展,2014,(6):67–70.
10 郭利敏.基于 Solr 的 IPAC 书目信息系统整合.微型电脑应用,2013,29(3):4–7.