基于Elasticsearch的纪检数据搜索引擎的设计

摘要

# 第1章 绪论

## 1.1 研究背景

随着信息技术的迅速发展,各行各业的数据量呈爆发性的增长，人们对信息检索的需求越来越强烈，如何从海量数据中高效、准确的挖掘出人们需要的信息[1]成为热议的学术话题，搜索引擎由此应运而生。

传统搜索引擎的服务对象是是大众用户，随着信息的急剧扩张，其采集收录的信息内容不断扩大，搜索结果也逐渐覆盖到各行各业，无法为某一专业领域内的用户提供专一化的信息查询[2]。因此，近些年来，人们又对搜索引擎提出了新的需求，要求可以针对某一领域检索更为专业的信息，这就需要更加专业化和更具针对性的搜索引擎，由此发展出了面向特定领域的垂直搜索引擎[3]。

垂直搜索引擎是搜索引擎发展过程中的一个重要里程碑，其只收录特定专业领域的信息数据，向其提供深入、垂直的搜索功能，不属于该领域的信息均不收入。相比通用搜索引擎，垂直搜索引擎大大降低了搜索过程中无关信息的干扰现象的发生，提升了搜索的效率。同时，与通用搜索引擎使用网络爬虫采集数据的策略相比，垂直搜索引擎通常具备专门的信息获取方式，如直接从该领域数据库中采集数据，这不仅降低了数据更新的成本，也使得数据更新的周期大幅缩短[4]。

## 1.2 研究意义

社会信息化的不断发展，使我国的纪检监察信息化的建设取得了快速进步。纪检部门以“大数据”理念为指导，加大信息资源的整合力度，汇总其直属及各地方机构的信息纳入其监察体系，使得像过去“跑部门、做调研、要信息”的情况逐渐减少。这对实现数据资源的规范查询、问题线索的追踪溯源具有重要意义。

随着数据量级的不断增长，于纪检人员来说，如何高效地从海量数据中获取自己需要的信息，如何高效的针对一条问题线索顺藤摸瓜、举一反三，成为纪检人员的关注重点。

针对检索的需求，我们以沈阳市纪委为实验对象，利用Elasticsearch，针对其纪检监察中的民生资金、招投标项目数据设计了纪检数据的垂直搜索引擎。目前，系统已稳定上线近一年，极大方便了纪检人员在以上领域的查案效率。

截至当前，沈阳市纪委拥有从市、区县（市）两级共计835家政府职能部门[5]采集的数据22亿余条，涉及数据库总量达上千余个。数据的进一步扩充向我们提出了新的挑战：面对数千数据库数万数据表的大规模结构化数据，如何设计搜索引擎的索引使其可以高效的容纳这些数据，如何组织搜索使得搜索引擎可以兼顾性能与准确率，如何设计搜索结果的展示方式使其在大规模库表下仍保持保持良好的易读性，这都是需要考虑的问题。

面对这样的挑战，需要对原有系统的架构进行重新设计，使其能够兼容大规模的结构各异的数据。同时，针对数据扩充造成的系统运维管理不便的问题，还需设计一个便捷的搜索引擎管理工具，将复杂的索引操作向用户屏蔽，提供易用的交互界面供用户操作，使系统成为一个可以产品化的系统。

## 1.3 主要研究内容

本文旨在通过对搜索引擎相关技术的学习，结合纪检数据的特点与需求，利用Java、Elasticsearch等技术设计一个纪检监察大数据垂直搜索引擎系统。系统提供了数据字典的管理、搜索引擎的管理、数据同步、数据搜索与展示等功能，并提供友好的交互界面简化对上述模块的管理。同时，利用Spring Cloud技术，将系统以微服务的形式进行功能划分，使其成为一个高可用、高性能、高扩展性的分布式搜索引擎系统。

## 1.4 论文组织结构

本论文章节安排如下：

第一章 绪论：本章介绍了本文的研究背景、意义及主要研究内容。

第二章 搜索引擎概述：

第三章 纪检大数据搜索引擎系统需求分析

第四章 纪检大数据搜索引擎系统架构

第五章 纪检大数据搜索引擎系统详细设计

第六章 总结与展望

、 第七章 致谢

# 第2章 搜索引擎概述

## 2.1 搜索引擎的概念与发展历程

搜索引擎(Search Engine)是一种可以从互联网或其他媒介上搜集信息，按照一定的策略将信息进行组织和存储，并根据用户的需要向用户提供检索服务，将相关的内容排序后展示给用户的系统。搜索引擎是用户获取数据的核心工具。

互联网的蓬勃发展，使得各种网络信息资源不断产生，其中最主要的有WWW(World Wide Web)信息资源、FTP(File Transfer Protocol)信息资源以及Blog信息资源等。这些资源具有信息量大、类型多样、分散无序的特点，为了满足对这些信息的检索请求，搜索引擎开始出现并发展起来。

* Archie——搜索引擎的鼻祖

Archie是由加拿大麦吉尔大学(McGill University)的三位学生Alan Emtage、Peter Deutsch、Bill Wheelan在1990年发明。在当时，WWW(万维网)还没有出现，人们主要通过FTP服务来共享文件，而大量的FTP服务器使得很多文件散布在多个不同的地方，文件的查询变得十分不方便。因此，Alan Emtage等三人便提出了一种可以直接查询出文件所在FTP地址的工具Archie。Archie可以自动对互联网上的FTP文件信息进行索引存储，用户通过相应文件的名称即可查询到该文件的FTP地址。Archie被称为所有搜索引擎的祖先。

* Gopher——菜单式搜索工具

1991年，在Archie的启发之上，美国内华达大学(University of Nevada)的“系统计算服务”小组发明了Gopher搜索工具。这是一种菜单式的信息浏览工具。用户选择某项菜单时，系统会将该项菜单的查询结果在用户终端上显示出来。如果要查询的信息在当前服务器上不存在，系统则会自动向另一台服务器发送相同的查询请求，直至检索到用户所需的信息。

* Yahoo Directory——目录式搜索引擎的代表

1994年，两位美国斯坦福大学（Stanford University）的博士生创办了Yahoo公司。Yahoo Directory是目录式搜索引擎的一种，其通过人工手动筛选，将各种类别的网站收集在一起。用户可以通过树形的分级目录逐级查找自己感兴趣的网页，也可以通过关键词的形式检索相关的网页。当以关键词形式检索时，系统会根据分类目录及网站与关键词的匹配度对网站进行排列，匹配度高的排在前面；当以目录形式检索时，系统会以字典序对网站进行排列。

* Metacrawer——元搜索引擎

1995年，美国华盛顿大学(University of Washington)的两名硕士生[Eric Selberg](http://web.archive.org/web/20010407110524/www.cs.washington.edu/homes/speed/home.html)和[Oren Etzioni发明了Metacrawer。这是一种新的搜索引擎形式——元搜索引擎(Meta Search Engine)。Metacrawer共支持12个独立的搜索引擎。当用户提交查询请求时，Metacrawer会将该次请求转换处理后提交给12个搜索引擎系统，并将这些系统的搜索结果经过汇总、筛选、排序后返回给用户。](http://www.cs.washington.edu/homes/etzioni/)

* Google——Web搜索引擎巨头

Google是Web搜索引擎的代表，其源自拉里·佩奇(Larry Page)的一个研究项目——BackRub。1998年，拉里·佩奇和另一位Google创始人[谢尔盖·布林](https://baike.baidu.com/item/%E8%B0%A2%E5%B0%94%E7%9B%96%C2%B7%E5%B8%83%E6%9E%97" \t "_blank)(Sergey Brin)提出了PageRank(网页排名)算法，这是一种分析网页的重要性的算法，其以网页之间超链接的个数和链接的网页的质量作为主要评估因素对网页进行排名。1998年9月，Google以私营企业的形式创立。2004年8月，Google在纳斯达克正式上市。现如今，Google已成为最流行的搜索引擎系统之一。

## 2.2 搜索引擎的分类

搜索引擎按照其工作原理的不同主要可以分为三种类型，分别为目录搜索引擎(Directory Search Engine)、全文搜索引擎(Full Text Search Engine)和元搜索引擎(Meta Search Engine)。

1. 目录搜索引擎

目录搜索引擎是一个按照目录分类的网站链接列表，其数据库大多是由专职编辑人员来建立和维护。编辑人员在查看某个Web网站后人工撰写一段该网站的概要，并将该网站的地址和概要信息置于一个事先确定的类别中。用户可以通过关键词检索到相关的网页，也可以通过分类目录的形式逐级的找到自己需要的信息。国内的新浪、搜狐以及国外的雅虎(Yahoo)等都属于目录搜索引擎。

2. 全文搜索引擎

全文搜索引擎是目前使用最广泛的搜索引擎，其代表有国外的Google搜索以及国内的百度、搜狗搜索等。全文搜索引擎通过爬虫等信息采集技术获取数据，使用倒排索引的数据结构将数据存储在磁盘，其能迅速地检索出与用户查询条件相符合的记录，并将搜索结果按照相关度排序后返回给用户。

3. 元搜索引擎

元搜索引擎本身并不存储数据，它相当于多个搜索引擎的代理。当用户向元搜索引擎发送查询请求时，元搜索引擎会将查询请求经过处理后同时发给其预先支持的多个搜索引擎，然后将这些搜索引擎的返回结果经过处理、筛选和排序后再返回给用户。图2.2.1展示了元搜索引擎的基本查询步骤。



图 2.2.1 元搜索引擎查询步骤

## 2.3 搜索引擎基本技术简介

2.3.1 数据获取——网络爬虫

2.3.2 数据存储与索引——倒排索引

2.3.3 数据检索——相关度模型

## 2.4 通用搜索引擎的缺陷与垂直搜索引擎的发展

通用搜索引擎的出现极大的方便了人们的生活，但由于其包罗万象的广泛性，在使用过程中仍然面临许多待解决的问题：

1.无法满足某一专业领域的个性化查询需求[2]：由于通用搜索引擎的服务对象是大众用户，其收录的数据源十分广泛，因此搜索结果通常会涉及到各行各业，导致数据的查准率较低，用户很难在其中找到自己所属领域的信息。

2.实时性差： 由于互联网数据规模的日益庞大，数据类型的日新月异，通用搜索引擎在技术上很难做到与互联网信息的及时同步，这使得搜索引擎的实时性变得很低。

3. 成本高：数据的迅速扩张使得搜索引擎对存储设备的需求越来越大，同时，数据的索引、检索和排序也需要消耗大量的计算资源。这导致通用搜索引擎的成本开销愈来愈高。

针对通用搜索引擎针对性差、实时性低、成本较高的缺陷，近些年来，面向某一专业领域或学科的垂直搜索引擎逐渐出现并发展起来。垂直搜索引擎，又称主题搜索引擎。与通用搜索引擎同，垂直搜索引擎只针对某一专门的领域或某一特定的需求提供定制化的信息检索服务[5]，其只收录特定专业领域的信息数据，向其提供深入、纵向的搜索功能，该领域外的信息均不收录。常见的垂直搜索引擎有面向商品的小商品搜索引擎、面向新闻、歌曲的搜索引擎等等。

## 2.6 本章小结

## 1.2 国内外研究现状

随着信息技术的迅速发展,各行各业的数据量呈爆发性的增长，人们对信息检索的需求越来越强烈，如何从海量数据中高效、准确的挖掘出人们需要的信息[0]成为热议的学术话题，搜索引擎由此应运而生。

搜索引擎的历史最早可以追溯至1990年，由加拿大麦吉尔大学的三位学生Alan Emtage、Peter Deutsch、Bill Wheelan发明的Archie，一种能根据文件名称直接查询其所在FTP地址的搜索工具[1]。1994年，Yahoo推出了Yahoo Directory，这是一种目录搜索引擎(Search Directory)，它由人工筛选网站，对每个网站加以简短的描述性信息，并将它们按照目录分类，人们通过分级的目录查找或关键词筛选即可定位到相关的网站。1995年，一种新的搜索引擎形式—元搜索引擎(Meta Search Engine)出现，它可以将用户的搜索请求转发给其预先选定的多个独立的搜索引擎，并将这些搜索引擎的搜索结果处理排序后再返回给用户。1998年开始，谷歌、百度、搜狗、360等全文搜索引擎开始兴起，其会从网页库中检索与用户的查询串匹配的相关记录，然后按照一定规则排序后将结果返回给用户，全文搜索引擎已占据了搜索引擎领域的半壁江山。

传统搜索引擎的服务对象是是大众用户，其搜索结果往往涉及到各行各业，无法为某一专业领域内的用户提供精确化的信息查询[2]。因此，近些年来，人们又对搜索引擎提出了新的需求，要求可以针对某一领域检索更为专业的信息，这就需要更加专业化和更具针对性的搜索引擎，由此发展出了面向某一领域的垂直搜索引擎。

纪检数据搜索引擎在工作原理上属于垂直搜索引擎。与传统搜索引擎不同，垂直搜索引擎是针对某一专门的领域或某一特定的需求提供定制化的信息检索服务[5]，其只收录特定专业领域的信息数据，向其提供深入、纵向的搜索功能，该领域外的信息均不收录。同时，纪检数据搜索引擎的数据来源为各部门电子政务数据的上报汇总，其面向的是本地的结构化数据，因此，并不需要采用网络爬虫技术，将相关网站的数据进行爬取及整理。

# 第3章 系统需求分析

## 3.1 纪检数据的特征

（1）数据来源广泛

纪检业务的数据来源十分广泛。为了满足纪检业务的需要，纪检部门通常会收集汇总及其直属及下属各地区各部门的电子政务数据，而下属部门的数据又是汇总自其下属乡镇、街道委员会相关单位，整个数据来源根据职能单位的级别近似形成了一个树形的结构。以沈阳市纪委为例，截至目前，该部门拥有从市、区县（市）两级共计835家政府职能部门[1]采集的数据22亿余条，涉及数据库总量达上千余个。这些数据通过人工拷贝、专网接口、专线网闸等方式流入纪委内部数据系统。

（2）数据类型多样

纪检数据汇集了多类异构的数据，其既有关系性数据库的结构化数据，也有类似XM形式的半结构化数据以及图片等非结构化数据，数据类型多种多样。以沈阳市纪委为例，其现有数据的类型涉及MySQL、Oracle、Access、Microsoft SQLServer等几乎市面上常用的各种数据库以及Excel、Word、GIS、报表等非结构化数据。

（3）数据结构散乱

大多数政务部门并不具备传统互联网公司那样大规模且专业的技术团队，其数据库表、字段的命名方式多种多样，例如使用拼音首字母、使用字母数字编码等等，这给阅读数据的人员带来很大的阻碍。而且，很多数据都是由Excel汇总导入至数据表，由于Excel在数据校验方面的不足，使得大量数据未经检查便被导入数据库，数据的格式不能保证统一。例如“2019-03-01”写法有“2019年3月1号”、“3-1-2019”、“二零一九年三月一日”等等。这些数据在存储时，便不能按照数字类型、日期类型等去存储，只能使用字符串类型，这也增加了数据错误、数据重复现象出现的概率。

（4）数据更新方式单一

## 3.2 搜索引擎在纪检业务中的应用场景

## 3.3 本系统需要解决的问题

### 3.3.1 数据索引——大规模库表下的Elasticsearch索引设计

Elasticsearch的索引为分布式的分片结构，每个分片相对独立，因此搜索请求会广播到索引的每个分片，每个分片并发的查询自己分片内的数据。各分片的查询结果有Elasticsearch汇总后返回给用户。

在传统的互联网应用中，索引的设计方式多为针对每个实体类（数据表）建立一个索引，索引各字段的名称与实体类（数据表）中的字段名称一一对应。这种查询方式适合单索引的查询，即每次查询请求都是在某一个具体的索引下进行。

而纪检数据面对的是成千上万的数据表，且表与表之间的结构、字段各不相同。查询方式为全库全文搜索，即每次查询针对的是所有库表所有字段进行搜索。在这种情况下，若使用传统的“一表一索引”的方式，假设库中有5000个数据表，每个表2个分片。当查询请求到来时，Elasticsearch会将请求广播至10000个分片中，这10000个分片会去一同竞争使用处理器的资源，使得处理器将大量时间花费在线程上下文的过程中，从而严重拖垮处理器及Elasticsearch的性能。

面对这样的情况，需要设计高效、通用的索引结构，使其不仅能够容纳不同类型的数据库表，还要确保高效的数据查询效率。

### 3.3.2 数据同步——非规范数据表与搜索引擎之间的数据同步

纪检数据具有数据库表多、数据量大及结构不规范的特点。针对库表多、数据量大的特征，需要设计便捷的自动化数据同步方案，并结合优秀的开源数据抽取工具，提高数据读写效率，减少数据的丢失率。针对结构不规范的特点，需要充分考虑到数据库表中的主键缺失、格式不一的现象，尽可能降低搜索引擎中数据重复、数据错误情况的发生率。

### 3.3.3 数据搜索——海量数据下的搜索方案

纪检数据搜索引擎针对的是大规模库表的结构化数据，在设计搜索方案时，要需要考虑到方案的通用性，确保其适用于系统下的绝大多数库表结构。同时，数据各表中也存在一定数量的无用字段，例如业务的预留空字段、标志位字段、测试字段等等，这些字段在一定程度上会加重搜索的负担，影响搜索结果的准确率，因此，在搜索时，如何设计灵活的规则，使得搜索引擎可以跳过无需检索的字段，这也是一项需要解决的问题。

### 3.3.4 数据展示——大规模库表下搜索结果的合理排布

搜索展示模块是与用户直接接触的模块，也是用户最为关心的一个模块。搜索展示主要需要考虑两部分：一为搜索结果的排布，二为搜索结果的易读性。

针对搜索结果的排布，需要充分利用库、表、字段信息的导向作用，将搜索结果进行合理的分类汇总。由于面对的是结构化的数据，若是像百度搜索那样以“标题—内容”的方式将搜索结果一条条的罗列，将会使页面十分散乱。下图3-1是本人所在部早期搜索引擎的查询结果截图，可以看到，在搜索关键词“项目”时，表中的“发放对象”、“资金种类”、“地点”、“发放时间”四个字段及其字段值被拼成一个文本作为一条搜索结果的标题，而这763页搜索结果没有经过合理的分类排布，只是单纯的罗列出来，这对后面的搜索定位造成了很大的苦难。



图 3-1 搜索结果示例图

二为搜索结果的易读性，纪检数据中，库、表、字段的命名几乎都是以拼音首字母命名，对于不熟悉数据业务的人来说，根本无从知晓这些属性所代表的意思。因此，对于这些晦涩的命名，还需要设计一个“数据字典”映射模块，存放这些命名计它们所对应的中文名称。在搜索引擎返回搜索结果前，先从字典模块做一次映射操作，将易读性好的名称展现给用户。

## 3.5 系统功能概述

基于3.4节的需求，按照模块化的思想，本系统可划分为四大功能模块：数据字典模块、搜索引擎索引管理模块、数据同步模块、数据搜索与展示模块。

### 3.5.1 数据字典模块

数据字典模块主要承担数据库、数据表、表字段的命名对照信息的管理，其功能用例图如图3-2所示



图 3-2 数据字典模块用例图

3.5.2 搜索引擎索引管理模块

搜索引擎管理模块是数据库与搜索引擎之间相接的“桥梁”，其负责搜索引擎索引的管理、数据表与搜索引擎之间的索引配置以及数据搜索的配置等。其功能用例图如下：

3.5.3 数据同步模块

### 3.5.4 搜索与展示模块

# 第4章 系统架构

4.1 系统架构图

4.2 数据库设计

# 第5章 系统关键技术设计

## 5.1 Elasticsearch索引设计

### 5.1.1 Elasticsearch基本索引结构

### 5.1.2 Elasticsearch默认索引结构的缺陷

### 5.1.3 基于表、字段二级映射的索引设计

## 5.2 数据同步方案设计

5.2.1 基于kafka、logstash的数据库全量同步

5.2.2 基于binlog、canal的数据删改同步

## 5.3 搜索与展示方案设计

5.3.1

5.3.2 基于库、表二级聚合的搜索结果展示

# 第6章 系统各模块实现

## 5.1 Eureka服务注册中心模块

## 5.2 Zuul服务网关模块

## 5.3 数据字典模块

## 5.4 搜索引擎索引管理模块

## 5.5 数据同步模块

## 5.6 搜索与展示模块

# 第7章 总结和展望

# 第8章 致谢

# 第9章 参考文献

[1]Kwang-I Yu,Shi-Ping Hsu,Peggy Otsubo.The fast data finder—an architecture for very high speed data search and dissemination[C].Intemational Conference on Data Engineering,1984(4):50-55.

[2] 张莉. 垂直搜索引擎中分词和排序技术的研究与应用 [D]; 西南交通大学, 2014.

[4] 陈翠婷. 无分类小商品搜索引擎关键技术研究 [D]; 中国科学院大学, 2016.

[4] 姜华. 基于Lucene面向主题搜索引擎的研究与设计 [D]; 华东师范大学, 2007.

[5]中国青年网.沈阳：用大数据深挖“不正之风” 管住身边腐败[EB/OL].https: //baijiahao.baidu.com/s?id=1618821229579984841&wfr=spider&for=pc,2018-12-03.

Hhhhhhhhhhhhhhhhhhh’

[4] 徐海. 基于Lucene垂直搜索引擎的研究与实现 [D]; 西安科技大学, 2009.

[5] 肖亮. 垂直搜索引擎的研究与实现[D]; 北京交通大学, 2007.