学校编码：10384　　分类号_____密级_____
学号：X2013232352　　UDC _____

厦门大学

工 程 硕 士 学 位 论 文

基于大数据技术的公安智慧搜索平台设计与实现

Design and Implementation of Public Security Intelligence Search-Platform Based on Big Data Technology

何志昭

指 导 教 师：吴清强副教授

专 业 名 称：软 件 工 程

论文提交日期：2015 年 09 月

论文答辩日期：2015 年 11 月

学位授予日期：2015 年 12 月

指 导 教 师：_____

答辩委员会主席：_____

2015 年 09 月

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下，独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外，该学位论文为（　　）课题（组）的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成。（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。）

声明人（签名）：

年　　月　　日

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（　　）1.经厦门大学保密委员会审查核定的保密学位论文，于　　年　　月　　日解密，解密后适用上述授权。

（　√　）2.不保密，适用上述授权。

（请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：

年　　月　　日

摘 要

公安信息化的高速发展积累了丰富的业务数据，种类不断增加、结构不断异化、总量不断增长，截至 2014 年底，全国公安机关掌握的数据资源已达 100 多类上千亿条,信息数据已成为继警力资源、装备资源之后的新一类核心资源。搜索是公安信息获取和进行研判的基础手段。传统的全文搜索系统存在大量结果数据“堆积”、再次筛查工作量大、目标结果无法精准定位，使用效率低下等问题。利用分布式计算架构，融合了 SolrCloud、Zookeeper、高速缓存等技术手段，以人、车、案为对象重组数据，采用智能分析、对象全景档案等手段，帮助用户在海量数据资源中快速“秒”准目标，从可以获得的全息资源中检索出最符合用户期许的资源。因此，设计提供一个先进、高效、智能、安全的公安海量智慧搜索平台，满足大数据时代下公安警务实战工作的需求具有重要战略意义。

本文结合公安搜索实际应用需求，进行大数据海量索引模型及公安智慧搜索平台的研究设计。文章首先对平台系统涉及的轻量级 J2EE 体系、大数据海量索引等主要技术进行阐述和选型分析；其次对系统总体业务需求、功能需求分析以及非功能需求分析描述；再次就是系统总体设计、数据库设计、应用功能设计、系统安全设计、系统实现与测试；最后总结了系统在技术研究、设计和实现过程中的收获与存在的不足之处，展现系统了未来进一步完善方向，以及服务于全警更多实战场景的可能性。

该系统的成功研发和部署上线，从整体上提升了海量数据的检索性能及用户搜索体验，使系统具备智能化的特征，为各级公安机关提供便捷高效、高价值的数据搜索服务。

关键词：大数据；海量索引；智慧搜索

Abstract

The high speed development of public security informationization has accumulated rich business data，Species continue to increase,the structure of continuous alienation, the total amount of continuous growth. As of the end of 2014, the national public security organs to grasp the data resources has reached more than 100 kinds of hundreds of billions of dollars, information data has become a new type of core resources after the police resources, equipment resources. The search is the public security information acquisition and basic means of judge. The traditional full text search system has a large number of results of data "accumulation", once again the screening of large workload, the target can not be accurate positioning, the use of low efficiency and other issues. Using distributed computing architecture, the SolrCloud, Zookeeper, high-speed cache and other technical means, the human, vehicle, case as the object of the reorganization of data, the use of intelligent analysis, object panoramic files and other means, to help users in massive data resources in the second accurate target, can be obtained from the holographic resources to retrieve the most consistent user expectations. Therefore, the design provides an advanced, efficient, intelligent, safe and intelligent search platform to meet the needs of the public security police in the era of big data has an important strategic significance.

In this dissertation, the research and design of large data massive index model and public security intelligence search platform are carried out with the actual application of public security search. Firstly, this dissertation describes and analyzes the main technologies of J2EE system, such as the lightweight system, big data massive index, and the system's overall business requirements, functional requirements analysis and non functional requirements analysis. The system design, data base design, application function design, system security design, system implementation and testing analysis.

The successful development of the system and the deployment of on-line, from the whole to enhance the retrieval performance and user search experience of massive data, so that the system has the characteristics of intelligent, for all levels of public security organs to provide convenient and efficient, high value data search service.

Key Words： Big Data; Mass Index; Intelligent Search.

目 录

Contents

第一章 绪论

本章讨论公安智慧搜索平台的研究背景、研究意义，基于当前工程实践和分析研究基础，提出对于这个系统研究的目标与内容，最后介绍论文内容的组织结构。

1.1 研究背景与意义

随着公安信息化建设的不断发展，公安采集数据的种类越来越多（以某省公安厅数据中心为例，已汇集 1000 多个业务表），单表数据量越来越大（其中近 3 年采集的网吧数据达 28 亿条记录，国内旅客 6 亿条记录）。搜索系统适合作为公安信息获取和进行研判的入口，集成于各警种专业研判应用平台。当前，公安行业搜索类系统主要还是以单条记录为单位提供查询搜索，以单表的方式组织结果展示，告诉用户在什么表中找到什么记录，存在以下局限性：

- 输入条件少了，虽然可以很快响应搜索结果，但返回的结果记录太多，找出想要的数据要花更多时间；
- 输入条件多了，由于信息关联性不好，搜索不到结果；
- 一个对象的信息散落在各个表的记录中，无法获取全面、完整的信息；
- 数据体量的不断增大带来的检索性能及用户体验的逐步下降。

从 2010 年大数据概念在国内首次被公众所知晓再到 2015 年大数据技术已经上升到国家战略层次，大数据在国内几乎完成了从起步到飞跃的快速脱变。在数据金山的诱惑下，各个机构纷纷开始探索从数据中提取洞见并指导实践的可能。而在这个需求的刺激下，从数据收集到处理，一直到数据可视化和储存，整个大数据技术生态繁花似锦。比如：用于数据收集的 Flume(NG)和 Sqoop，分布式消息队列技术 Kafka、RabbitMQ，用于数据可视化的 HighCharts、D3.js、Kibana、Echarts 等等。此外，加之 HBase、MongoDB、Redis 等 NoSQL，Lucene、Solr、ElasticSearch 等搜索技术，Docker 等容器技术，ZooKeeper 等分布式应用程序协调服务。

基于成熟的大数据技术框架和硬件资源的支撑，设计、研发一款具备智能化特征的公安搜索系统——公安智慧搜索平台，利用数据的相关性，将分散零乱的数据组织成有关系的数据集，并在查询数据结果的基础上，采用智能分析、全景档案等手段，帮助公安各警种用户在本地海量数据中心与异地共享数据资源中快速“秒”准目标，实现对人员、车辆、案情、物品等信息的深度查询，并对目标结果数据实现横向关联、纵向钻取，提供高效率、高价值的对象研判式搜索服务，同时，在架构上实现了海量数据存储的分布式、动态扩展，提升了检索性能及搜索体验。

1.2 研究目标和内容

本课题的目标是结合某省公安厅新型研判和搜索实战需求，研究基于大数据时代下公安智慧搜索平台的关键技术实现手段、大数据对象整合思路、应用模式及系统安全性设计。通过引入全文库、分布式搜索引擎技术，实现对海量数据进行快速的分布式全文索引和存储，提升索引和搜索体验；引入大表技术和全息对象整合思路，解决大量原始业务数据按照实体对象索引构建方法；利用语义分析方法、行业词库和基于理解技术，实现海量数据的智能化搜索及精确结果定位；分析了应用开放涉及到系统安全的几个层面设计。同时将以上研究目标进行产品原型孵化，完成了系统的设计与实现。

本文主要研究内容如下：

1、研究 lucene、Solr、Solrcolud、Zookeeper 技术框架、应用方法，提出基于 SolrCloud 和 Zookeeper 构建分布式搜索、索引存储和配置管理方案。

2、利用 HBase 大表来存储业务对象，其独立的索引构建过程基于 Hadoop MapReduce 编程模型实现，在 SolrCloud 上部署分层索引，从而实现支持传统的资源搜索和全息对象搜索。

3、研究全息对象索引数据的组织设计，通过将业务数据的表信息打散后，按照实体对象重新进行整合设计。同时，将数据整合思想引入搜索过程，对大量原始的资源结果按业务主题进行关联合并，从主题分类、主题档案等贴近用户业务的角度呈现数据。

4、研究语义分析方法，主要通过扩展/改进机械分词的 IK 分词器，支持最

细粒度和智能分词，支持分词结果重用，支持代码翻译，支持拼音索引，支持多音字；再者，构建一套公安行业云词库，通过持续收集、维护完善，进一步提升公安行业特定术语搜索和语义分析的精确度。

5、公安的数据和应用安全敏感度特别高，本文重点从应用软件安全和数据安全两个方面分析并阐述系统性安全设计，从而构建一个高安全的搜索应用平台。

1.3 论文内容组织结构

本论文研究的内容分为 7 章：

第一章为绪论，简要介绍公安行业大数据体系下研究快速、精准、智慧搜索应用的背景与意义，研究目标和研究内容、论文内容组织结构介绍。

第二章为系统需求分析，主要论述系统的总体业务需求、功能性、非功能性需求。

第三章为系统总体设计，包括基本设计概念、系统架构、系统角色、程序结构、数据库设计、开发与运行环境设计。

第四章为系统功能设计，介绍了基础模块设计、搜索引擎、搜索结果处理及智能语义分析、档案定位等核心模块的设计。

第五章为系统安全设计，阐述应用系统各层面安全设计要求。

第六章为系统实现设计与测试分析，主要介绍智能搜索、档案服务模块的功能实现逻辑，以及相应模块的健壮性、功能性及系统的安全性测试。

第七章为总结与展望。

第二章 系统需求分析

本章进入公安智慧搜索平台的系统需求分析阶段，明确系统总体需求、系统各部分的功能需求、非功能性需求的详细内容。

2.1 系统总体需求

2.1.1 总体业务需求分析

依托本地公安数据中心，按照智慧搜索平台标准数据模型和信息代码规范整合的数据仓库，具有主题分类支持的数据资源和数据服务，以及数据关联模型，结合全文索引技术，在基础搜索功能基础上，突出公安行业分词、用户环境上下文相关、实体（资源）自动识别、结果智能关联等智能化特征，在满足分布式基础设施的情况下，提供具有智能化的分布式全息检索功能。根据用户的请求，从全息数据资源中快速搜索最符合民警研判期许的资源，按照最有价值的信息关联方式展现给用户。

- 智能识别与语义分析：智能识别或是依托行业分词技术分析出用户输入的检索关键词的语义，进而执行相应的搜索行为和动作。
- 传统搜索：通过针对全息资源进行关键词检索，同时系统后台自动进行相关度计算，并按相关度分值从高到低排序返回相应结果给用户。
- 命令搜索：依据命令符识别，以及语义分析出的检索资源和条件，智能的针对特定资源按特定条件进行检索，并返回相应的结果数据给用户。
- 主题搜索：通过针对全息资源进行关键词检索，将结果信息按主题分类（人、车、案、文本、网页等）进行重组合并，并返回相应结果按主题分类展现关键词相关的信息给用户。
- 档案服务：依据实体特征识别（身份证号、车牌号、案件编号等），系统智能的将全息资源中与该实体相关联的信息按某类主题档案的要素分类进行归集，并以档案形式推送给用户。

系统总体业务管理结构如图 2-1 所示。