Software Engineering and Applications 软件工程与应用, 2020, 9(5), 441-446
Published Online October 2020 in Hans. http://www.hanspub.org/journal/sea
https://doi.org/10.12677/sea.2020.95051

Hans汉斯

垂直搜索引擎系统关键技术研究及其在通信行业招标领域的示范应用

李正军

贵阳高新数通信息有限公司，贵州 贵阳
Email: 276792749@qq.com

收稿日期：2020年10月8日；录用日期：2020年10月23日；发布日期：2020年10月30日

摘　要

本文首先对通信行业招投标领域信息获取的弊端进行分析，总结出用户当前存在三大需求，然后结合对比人工目录分类、文本处理和智能搜索引擎等技术，提出采用垂直搜索引擎系统实现行业应用的设计路线；并对大数据垂直搜索引擎系统的体系架构、信息采集、索引建立进行了详细设计，最后实现该系统并在通信行业招标领域进行推广使用。结果表明，本文提出的面向行业垂直搜索引擎系统具有良好的适用性，能够满足通信及其他行业的招标信息获取、精准分发和推送需求。

关键词

垂直搜索引擎，招投标，信息采集，索引创建

Research on Key Technologies of VSE System and Its Application in the Bidding Field of Communication Industry

Zhenjun Li

Guiyang Hi Tech Data Communication Co., Ltd., Guiyang Guizhou
Email: 276792749@qq.com

Received: Oct. 8th, 2020; accepted: Oct. 23rd, 2020; published: Oct. 30th, 2020

Abstract

This paper first analyzes the disadvantages of information acquisition in the field of bidding in the

文章引用: 李正军. 垂直搜索引擎系统关键技术研究及其在通信行业招标领域的示范应用[J]. 软件工程与应用, 2020, 9(5): 441-446. DOI: 10.12677/sea.2020.95051

communication industry, summarizes the three current needs of users, and then proposes the design route of using vertical search engine system to realize the industry application by comparing the technologies of artificial directory classification, text processing and intelligent search engine. The architecture, information collection and index establishment of big data vertical search engine system are designed in detail. Finally, the system is implemented and used in the field of communication industry bidding. The results show that the industry-oriented vertical search engine system proposed in this paper has good usability and can meet the requirements of bidding information acquisition, accurate distribution and push of communication and other industries.

Keywords

Vertical Search Engine, Bidding, Information Collection, Index Creation

1. 引言

随着社会进步及大数据时代的到来，各个行业都带来迎来了新的机遇和挑战，其中如何运用新一代信息技术提升行业工作水平成为迫切需求。以通信行业中的招投标为例，目前依法应招项目的招标率超过90%，但是随着招投标的范围和规模的扩大，招投标效率低、周期长等问题越来越显著，而传统的招投标方式由专人查询发布在各省政府、公共资源交易中心、泛运营商官网的招标信息，在众多各类的标讯中守候，难以从海量信息中精准定位并获取到高适应度的相关信息，在对招标信息公告守候、垃圾信息有效过滤的过程中，造成了人力、时间、财力的巨大浪费，总体上存在着“难以科学确定采购需求、信息搜寻成本较高、数据价值开发率低”三方面问题。

大数据技术应用到招标行业可快速打破相关局限，结合搜索引擎技术，将分散在各地、各运营商网站上的所有招标资讯和中标查询信息整合在数据库里，并按各类关键词比如地区、行业、金额、时间来加以分类，使得用户不用逐一守着多个网站信息就能获得全国各类标讯，还可有效通过关键词来主动获取分类后的查询和订阅信息推送。

本文围绕如何从百亿级的海量网页和文档中精准定位到所需内容，满足用户对专业的、有深度的知识的需求；并具体到通信行业的招投标信息，研究如何实现全面的数据采集、清洗、聚类，并快速反馈到分类客户的搜索引擎是核心。本文通过研究垂直搜索引擎关键技术，实现在面向通信行业招投标领域的网页信息采集，结合非结构化内容到结构化数据的数据解析技术，实现精准全面的全文索引和联合检索技术，帮助用户快速定位到想要的搜索结果。

2. 搜索引擎系统简介

2.1. 搜索引擎的发展历程

通用搜索引擎的出现很大程度上提升了互联网信息查找的便捷性，但通用搜索引擎已不能满足行业用户的个性化信息检索服务需求，因此面向特定领域的垂直搜索引擎便应运而生。

搜索引擎技术经历了三代显著的技术发展，第一代是以Yahoo为代表的人工目录分类导航技术，但存在实际检索结果的相关性、排列序的合理度严重不足弊端；第二代是以Google为代表的文本处理技术，

并在检索呈现层面引入排序优化方法，在检准率、检全率和检索速率方面较第一代获得较大提升；第三代是以 Baidu、搜狗、Wolfram Alpha、Google 为代表的智能化搜索引擎，通过综合运用人工智能、数据挖掘、模糊匹配、神经网络、数理分析技术，实现了对目标用户的实际使用需求的更精确满足，获取到良好的综合效益，其中垂直搜索引擎技术是第三代技术的核心[1]。

2.2. 垂直搜索引擎的发展

垂直搜索引擎(Vertical Search Engine, VSE)是针对某一个行业的专业搜索引擎，是搜索引擎的细分和延伸，是对网页库中的某类专门的信息进行一次整合，定向分字段抽取出需要的数据进行处理后再以某种形式返回给用户，也被称为专业搜索引擎或主题搜索引擎[2]。相对通用搜索引擎，垂直搜索通过针对某一特定领域、某一特定人群或某一特定需求提供的有价值的信息和服务，其特点就是“专、精、深”，且具有行业色彩。垂直搜索引擎涉及信息索引、机器学习、数据挖掘、自然语言处理等多领域的知识及技术，综合性强、专业化程度高，已在我国各行各业得到了广泛应用，但在接口管理、数据挖掘和共享方面与国外存在较大差距。而优秀的垂直搜索引擎，不仅需要技术方面的专业知识背景，更需要行业领域的相关经验。

未来垂直搜索引擎技术的发展方向集中在如何提高信息检索结果的精确度、基于智能代理的信息过滤和个性化服务、综合相关信息搜索、与分布式体系结构的结合运用、面向民族和国家的本土化研究、多语言的搜索应用等方面。

2.3. 垂直搜索引擎系统设计思路

本文以项目组织知识体系(PMBOK)的规范词表系统为基准，将涉及到招标概念规范的 PMBOK 词表与本体，通过资源描述框架(Resource Description Framework, RDF)进行存储利用，采用 MySQL 作为数据仓储，支持查询、推理及应用服务[3]。

本文侧重于通信行业招标领域的项目管理知识体系，提出招标开放引擎系统(Bidding Open Engine System, BOES)开放式知识组织引擎。BOES 总体框架包括存储与索引层、查询与推理功能层、BOES API 层以及开放查询和推理接口层，采用语义仓储、索引、查询、推理、接口技术，构建存储索引体系、语义查询与推理内核，支持实现招标行业各类元素检索、浏览、关联、导航等功能。在 BOES 基础上，依据 BOES 数据特性，开发 APP 应用服务平台，构建高性能、可靠的知识存储索引体系和 BOES 检索查询与语义推理内核引擎，支持信息推送服务，并提供封装的 API 接口供第三方系统使用。

3. 基于大数据的垂直搜索引擎设计

3.1. 垂直搜索引擎系统架构设计

垂直搜索引擎系统架构分为表示层、逻辑业务层和数据访问层，各层间数据信息的传递依靠接口完成[4]，其中，表示层位于最外层，直接与用户进行交互操作，负责接收用户输入的搜索信息并显示搜索结果；业务逻辑层是整个系统的核心部分，实现对筛选数据、爬取网页信息、建立索引、管理系统等功能，它处在数据访问层与表示层中间，在数据信息的交换中有着承上启下的作用；数据访问层对主题网页信息数据库、用户及管理员信息数据库等进行访问，为逻辑层业务及表示层给予数据支持[5]。

3.2. 信息采集模块设计

本模块完成网页信息的采集是整个系统的基础和重点[6]，包含 9 个子模块协同实现整体功能：

(1) 主题词库子模块：负责行业及领域主题词的挑选，并建立形成主题词库，如图 1 所示；

Figure 1. Submodule diagram of thesaurus
图 1. 主题词库子模块图

(2) 链接种子集合子模块：负责得到多个与主题相吻合链接，作为数据爬取的开端，如图 2 所示；

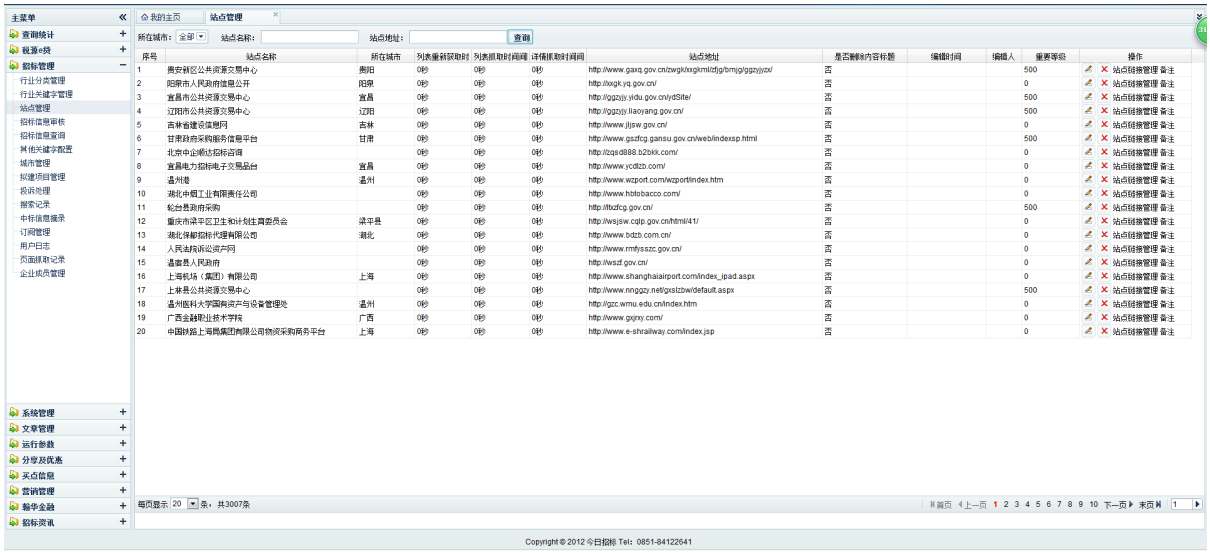

Figure 2. Link seed aggregation sub module diagram
图 2. 链接种子集合子模块

(3) 网页下载子模块：负责完成网页的下载工作；

(4) 网页解析子模块：负责完成爬虫的配置、网页内容的提取和链接定位的工作；

(5) 内容相关性判断子模块：负责完成对分类器进行训练，然后利用训练好的分类器对爬取的网页内容进行筛选，过滤掉与人工智能主题无关的网页[7]；

(6) 主题相关性评估子模块：负责利用 PageRank 算法计算出链接的 PR (链接拥有价值的高低)值，保留 PR 值大的作为继续爬取的网页链接，使得爬虫工具每次都能爬取到有价值的网页信息；

(7) 链接管理子模块：负责抓取链接的管理及去重；

(8) 数据保存子模块：负责将满足主题要求的抽取信息保存到数据库中；

(9) 爬虫启动子模块：负责创建 Spider 对象，并启动爬虫。

3.3. 索引建立模块设计

本模块主要负责中文分词和建立索引，并能够按需提供智能化处理功能，如自动分类、自动聚类、自动标引、自动排重、文本挖据等。本文采用 Solr 实现了可配置、可扩展，并对索引、搜索性能进行了优化。Solr 作为 Apache 下的顶级开源项目，是基于 Lucene 的全文搜索服务器，并提供了比 Lucene 更为丰富的查询语言，可独立运行在 Jetty、Tomcat 等 Servlet 容器中。本模块具体过程分为客户端用 POST()方法向 Solr 服务器发送一个描述 Field 及其内容的 XML 文档，Solr 服务器根据 xml 文档添加、删除、更新索引的创建索引过程，和客户端用 GET()方法向 Solr 服务器发送请求，然后对 Solr 服务器返回 Xml、json 等格式的查询结果进行解析的搜索索引过程。

本系统在通过在 Solr 中添加中文分词项目，整合 IK Analyzer 分词器实现对数据表信息的中文分词，并将 Solr 与数据库连接创建数据表倒排序索引。完成对索引的建立以后，通过向发送包含查询关键字、语法版本、返回结果的条数等参数的 HTTP 请求进行查询，Solr 收到请求后以 XML 的形式响应结果。

4. 面向招投标领域的应用

本文结合垂直引擎系统关键技术、查询优化与推理策略关键技术、面向招标行业的综合服务平台研发技术[8]，研发融合行业特点、定位企业精准需求、企业金融服务于一体的综合移动服务平台——“今日招标”，如图 3 所示。“今日招标”招标大数据平台系统通过记录和分析招标采购过程中的各类数据，建立数据分类模型并深度计算数据形成数据集，在此基础上，构建招投标大数据资源，让招标采购行业数据从封闭系统走向开放的平台，实现自动化推送服务。

Figure 3. Big data platform of “Today’s Bidding” diagram
图 3. “今日招标”招标大数据平台

“今日招标”招标大数据平台招投标采购信息覆盖全国 95%以上政府采购及招标网企业招标平台，每日新增招标采购信息 50000 条以上，超过 1000 万用户在这里寻找商机。通过使用智能排队和分配，实现了抓取服务器的分布式部署，易于整个系统的部署与维护、负载均衡；实现了招标信息的基于行业的自动智能分类。平台存储的网页快照和相关附件有近两千万条左右，能高效的对数据库进行存储及查询

进行各种优化，保证高并发下同一个数据的二十万次上锁执行释放锁的操作，平台的稳定运行稳定，反响热烈，截止 2020 年 7 月，累计安装超过 1851 万次，如图 4 所示。

Figure 4. Application example of bidding in communication industry diagram
图 4. 通信行业的招投标领域应用示例图

5. 总结

为解决有效招标领域难以科学确定采购需求、信息搜寻成本较高、数据价值开发率低等问题，本文重点研究基于 BOES 的垂直引擎系统、基于 Solr 的全文检索索引、面向通信行业的招标综合服务平台等关键技术，并以此为基础，研发出融合行业特点全面覆盖、企业需求精准定位推动、企业金融服务于一体的综合移动服务平台。平台系统通过记录和分析招标采购过程中的各类数据，建立数据分类模型并深度计算数据形成数据集，在此基础上，构建招投标大数据资源库，让招标采购行业数据从封闭系统走向开放的平台，实现自动化推送服务，取得良好效果。

参考文献

[1] 姜琨, 朱磊, 宋省身, 等. 倒排索引压缩算法研究综述[J]. 小型微型计算机系统, 2020, 41(4): 715-723.
[2] 王露. 垂直搜索引擎中 PageRank 改进算法的研究与应用[D]: [硕士学位论文]. 昆明: 云南大学, 2019.
[3] 董珊. 新政策下对工程招标代理机构发展与转型的思考[J]. 安徽建筑, 2019(11), 33-35.
[4] 翟霞. 基于 Lucene 的面向大数据主题的垂直搜索引擎研究[J]. 科学技术创新, 2019(31): 96-97.
[5] 苏鹏. 基于“大数据”的电子招标投标平台建设研究和应用[D]: [硕士学位论文]. 青岛: 青岛理工大学, 2019.
[6] 乔柱, 刘伊生. 大数据背景下我国电子招投标监管研究[J]. 工程管理学报, 2019, 33(1): 1-5.
[7] 吴丹, 唐源. 搜索引擎结果页面(SERP)研究述评[J]. 情报学报, 2018, 37(2): 220-230.
[8] 陈立东. “互联网+”招标采购提升信息通信业监管服务水平[J]. 招标采购管理, 2018(1):12.