**北京师范大学本科生毕业论文（设计）开题报告**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 学生姓名 | 许宏旭 | | 学 号 | 201411212027 | |
| 所在部院系 | 信息科学与技术学院 | | 专 业 | 计算机科学与技术 | |
| 指导教师姓名 | 付艳 | 指导教师职称 | 副教授 | 指导教师单位 | 北京师范大学 |
| 毕业论文题目 | 北京师范大学校园网搜索引擎 | | | | |
| 开 题 报 告 内 容 | | | | | |
| 选题依据及研究内容（国内、外研究现状，初步设想及突破点；研究目标、预期成果，及可行性论述等） | **问题提出**  目前，北京师范大学校园网内信息分散、查询不便。校园网内目前提供的几种信息检索接口，包括官网搜索栏、数字京师等，均不具备一个现代搜索引擎应该具备的信息检索能力，无法满足在校师生快速、准确检索校园内部信息的需求。  校园网现有的搜索功能没有应用先进的排序算法，搜索关键词返回的结果相关性较差；且该搜索基于校园内各站点的文章数据库进行关键词匹配检索，而非借助大规模爬虫获取信息，因此可检索信息不够全面；同时，返回结果信息展现形式不直观不美观，用户接口不友好。  经过初步的用户调查，拥有一个集中的、检索能力强的、信息全面的信息检索入口是十分必要的；另外调查数据显示用户对诸如附件搜索等特殊功能有较大需求。  基于这些问题，我计划研究并开发北京师范大学校园网搜索引擎。拥有独立的校园网搜索引擎可以满足这些针对性的搜索需求，并解决外网访问校园网不便导致通用搜索引擎无法检索等问题。同时，松耦合模块化的具备日志系统的搜索引擎也可以作为科研实验数据平台。  **国内外研究现状**  互联网内容丰富而庞大，且日益增长，对于信息获取而言是一项挑战。搜索引擎的出现，就是为了解决这一信息获取的问题，提供一个友好的人机接口，实现便捷的信息检索和获取。  互联网搜索引擎主要分为两类：目录搜索引擎和全文检索搜索引擎。  目录搜索引擎需要专业编辑进行人工分类。随着互联网的不断发展，分类工作已经很难进行，而且用户检索时也需要按照编辑的标准，通过分类来检索，造成使用困难。常见的目录搜索引擎有早期的Yahoo!，国内的搜狐、新浪也可以属于这种搜索引擎。  全文检索搜索引擎要求用户键入搜索关键词，通过倒排索引和排序机制提供最相关的结果页面。检索数据来源通过大规模的网络爬虫获得，能够与日益增长的互联网体积相适应，对用户的搜索能力要求不高。因此，目前市面上主流的搜索引擎均为全文检索搜索引擎。常见的全文检索搜索引擎包括Google、Bing、现代的Yahoo!，国内有百度、搜狗等。  校园网搜索引擎属于内网搜索引擎或自定义搜索引擎，即针对特定网络范围内的网络页面进行索引的搜索引擎。目前有ElasticSearch、Apache Lucene/Solr等开源搜索引擎项目，可用于这一目的。但这些搜索引擎框架功能强大，体积庞大，架构复杂，更适于企业级生产环境，而非校园内附加科研用途的场景。校园网内站点模式较为单一，更适合搭建针对性的搜索引擎；这样不仅可以做到架构简单，还能方便实验研究。  **初步设想及突破点**  初步设想利用C#编程语言，从零架构一个校园网的搜索引擎，其中包括：爬虫模块、存储模块、文件分析模块、检索模块、用户接口模块和日志模块。各模块之间做到松耦合，提高维护性和扩展性。  初步设想在最初架构时，首先在各模块中使用成熟的技术，例如存储模块首先应用数据库系统进行结构存储，文件分析模块中涉及的分词技术等也首先使用业界常用的分词工具实现。这样可以实现基本架构的快速构建。  由于模块之间的松耦合性，未来可以根据功能扩展的需要，方便地进行功能加强和更新，做到快速版本迭代。例如，最初采用的索引结果排序算法预期将使用业界成熟的PageRank算法[2]，在架构成熟后，则会继续增加基于用户点击路径日志的排序优化机制等，提高搜索引擎的检索能力；前文提到的数据存储以及分词技术也可以在后期改进中做出优化调整。  这种松耦合、高扩展性的架构设想，可以作为本课题的突破点，有助于课题项目的顺利完成；也能够在此基础上，进行各个模块相关的前沿实验探究。  **研究目标、预期成果**  从基础架构开始，完成初步设想的爬虫模块、存储模块、文件分析模块、检索模块、用户接口模块和日志模块，搭建一个松耦合高扩展性的便捷可用的校园网搜索引擎。  预期搜索引擎搭建完成之后，可以为师大在校师生提供一个用户友好的信息检索人机交互接口，为广大师生提供便利；同时，搜索引擎带来的巨大的站点数据、用户使用习惯数据等，也同样可以作为实验室其他实验研究的实验数据，为将来在信息检索、数据挖掘、自然语言处理、知识图谱、机器问答等领域的相关研究做基础。  **可行性论述**  校园网具备良好的网络环境，站点之间关联性较强；搜索引擎初步设想是全文检索搜索引擎，存储信息基本为文本数据，校园网内信息量级对于一般服务器而言可以接受，存储资源要求不高；检索过程中涉及诸多算法均有成熟案例，且架构过程中解耦各个模块，可以逐个实现，对个别复杂度较高的模块首先采用业界成熟的解决方案，后续再做针对性优化实验，这样可以显著提高本课题的可行性；搜索引擎用户接口简洁清晰即可，复杂度一般。 | | | | |
| 理论和实践  意义 | 本课题需要从基础架构开始，搭建一个搜索引擎所需具备的几个基本模块：爬虫模块、存储模块、文件分析模块、检索模块、用户接口模块和日志模块。本课题是对信息检索领域方向的实践。  考虑到校园网体量相对互联网较小，本课题同时注重成果的实用价值和科研价值。本课题要求做到切实可用的校园网搜索引擎，可以方便广大在校师生在校内的信息检索任务。本课题的预期成果满足架构从简，模块间解耦，注重扩展性，强调日志模块的要求，能够实现功能扩展，并提供校园网内各类站点的数据以及用户使用习惯日志等。这些都可以为后续进一步的相关领域实验提供基础工作，使得本课题的成果搜索引擎成为我校信息检索领域的实验大本营。 | | | | |
| 论文撰写过程中拟采取的方法和手段 | 1. 借助中国知网、Google Scholar等学术搜索引擎查阅并引用相关领域文献，完善课题设计和实施过程； 2. 阅读《信息检索导论》《统计自然语言处理》等相关领域专业书籍，熟练掌握课题所需的基础知识； 3. 注重课题项目的架构，保持各模块之间松耦合，以实现各个模块相对独立的完成过程，并在论文中分模块进行撰写，陈述实现过程和原理等； 4. 基本架构完成后，对各个模块逐个探索优化关键点，进行拓展的前沿研究实验，逐步完善校园搜索引擎，并在论文中体现优化过程； 5. 与指导教师保持交流沟通，及时发现设计和实现中存在的问题，及时修正并完善课题实施过程以及论文的撰写； | | | | |
| 论文撰写  提 纲 | 1. 背景    1. 北京师范大学信息检索现状    2. 搜索引擎的背景    3. 内网搜索引擎    4. 设计目标       1. 高效便捷的信息检索入口       2. 科研实验基础数据提供者 2. 设计特点    1. 松耦合的模块化设计    2. PageRank算法    3. 附件检索能力    4. 友好的用户接口 3. 系统架构    1. 概览    2. 各模块设计、原理    3. 成果 4. 结论 5. 参考文献 | | | | |
| 计划进度  及其内容 | |  |  | | --- | --- | | 2017.11.1 – 2018.1.11 | 可行性调查、背景调查、立项开题。 | | 2018.1.12 – 2018.3.10 | 阅读文献和专业书籍；  搭建爬虫模块、文件分析模块、存储模块。 | | 2018.3.11 – 2018.4.10 | 搭建检索模块、日志模块；  完成论文初稿，寻求指导教师意见。 | | 2018.4.11 – 2018.5.10 | 搭建用户接口模块；  完善论文，准备答辩。 | | 2018.5.11 – | 答辩 | | | | | |
| 参考文献 | 1. Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press. 2008. 2. Brin S, Page L. Reprint of: The anatomy of a large-scale hypertextual web search engine. Computer networks. 2012 Dec 17;56(18):3825-33. 3. Croft WB, Metzler D, Strohman T. Search engines: Information retrieval in practice. Reading: Addison-Wesley; 2010 Feb. 4. 刘向辉. 专题性智能搜索引擎的研究与实现[D].昆明理工大学,2001. 5. 黄知义,周宁.几类搜索引擎的原理剖析、比较研究及发展趋势探讨[J].图书馆学研究,2005(03):61-64+67. 6. 李岩, 陈新中. 基于 Web 挖掘的智能门户搜索引擎的研究. 计算机工程与应用. 2002;38(4):34-6. 7. 胡骏, 李星. 校园网信息资源搜索引擎的研究与实现. 计算机工程与设计. 2006;27(24):4629-31. | | | | |
| 教师指导意见  （针对选题、研究方法、计划进度等的意见和建议） | 意见：  通过对北京师范大学校园网现状的调查，以及对互联网搜索引擎领域的深入调研，论文拟针对北京师范大学校园网设计并构建一个完整的搜索引擎系统，以期为校内用户及科研人员提供更好的信息与数据服务。  论文的选题具有较高的实际应用价值，也涉及多项关键技术的理论研究。论文的前期调研工作扎实可靠，拟定的研究方法切实可行，进度安排合理。综上，同意开题。  是否同意开题： （是、否）  指导教师签字： 20 年 月 日 | | | | |
| 部院系  意见 | 部院系毕业论文工作小组负责人（单位公章）：  20 年 月 日 | | | | |

注：此表格可延长加页。