**南京大学软件学院研究生学位论文开题报告格式**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **南京大学软件工程硕士学位论文开题报告** | | | | | | | | |
| 导师1姓名 | 顾庆 | | 研究生姓名  （学号） | MF1232104 | | | 方向 | 应用软件工程 |
| 导师2姓名 |  | |
| 论文题目 | 基于xPlore的统一资源检索平台AIOS的设计与实现 | | | | | | | |
| 论文选题来源及研究的目的和意义（500字左右）：  AIOS项目源于公司业务模型：出售技术支持为主，软件为辅。公司每天接受客户数以万计的请求，而随着公司规模的扩大，需要处理的请求急剧增加，导致了研发人手紧缺，成本急剧上升。事实上，一方面，大部分的请求处理的问题是类似的，导致重复劳动；另一方面，请求对应于产品各个不同的版本，每个版本代码都需要下载，当需要解决一个用户请求时，需要从代码管理服务器上下载总量约为3G的代码，严重影响了研发人员的工作效率，同时也提高了Bug定位的难度。针对产品相关的文档搜索时，并不能集中搜索，必须从各大内部网站上进行信息收集，而且内部网站并不提供精确搜索，很多文档无法获取。AIOS基于全文检索平台xPlore，将产品相关的代码、文档和报告等内容收集管理，实现资源整合与信息检索。  AIOS统一资源检索平台的主要功能是定期抓取与更新产品相关信息，如Jira的Bug记录（Bug记录帮助定位客户请求是否已经在某个发布版本中解决），EMC Powerlink，EMC在线支持文档，产品各个版本的代码，利用公司现有的全文检索平台xPlore实现内容检索与相关内容推荐，并作为xPlore的应用方案向外推广。  AIOS利用公司现有搜索平台xPlore，实现统一检索平台，可以帮助公司降低运营成本，提高生产效率，避免重复劳动，促进公司全球职员的交流，减少用户请求处理延滞时间，提高用户满意度。同时AIOS易于扩展，可用于公司其他产品的运维活动。 | | | | | | | | |
| 国内外在该方向的研究现状概述（1000字左右）：  AIOS主要实现了信息获取，内容检索与相关信息推荐。涉及的内容包括信息获取，全文检索，基于自然语言处理的相似文章推荐引擎。  信息获取:指网页爬取，称为Web Spider。Web Spider用于更新网页内容或爬取更多的内容。网页爬取首先访问一组URL链接，通过分析这些网页中所包含的URL，将它们作为新的内容获取源，迭代进行直到没有新的URLs出现结束。该方向主要问题有：web规模不断增长，信息量巨大，占用大量存储空间；URL/html可能存在语法错误，解析出错；有些网站禁止Spider的访问；采用flash架构的网站内容无法爬取。  全文检索，使用公司的全文检索平台xPlore，支持多种文档类型的内容提取，支持多种语言的分词与索引。竞争的产品包括开源的Solr，Baidu，Google和Bing等。该方向主要问题为中文分词算法的设计与实现，索引的存储结构与效率之间的折中。自2003年Bakeoff分词评测开展以来，中文分词技术获得了长足的进步。通过“分词规范+词表+分词语料库”的方法，使中文词语得到了可计算的定义。基于手工规则的分词方法在评测中不敌统计学习方法，在Bakeoff数据上的估算表明，未登记词造成的分词精度损失至少比分词歧义大5倍以上，因此能够提高未登记词识别性能的分词方法必将带动分词性能的提升，例如基于字标注的统计学习方法。Bakeoff评测数据证明，基于字标注的分词系统优于以往的基于词(或词典)的分词系统。由于自然语言的模糊性和复杂性，对于任何语言对象都应当有一种可计算的定义，因此当前分词算法的研究之一在于寻求一种可计算的模型。大部分的文档搜索引擎是基于Lucene，其存储结构难以完成局部更新，严重影响系统的性能，需要设计可更新的紧凑型高性能的数据结构。  相似文章推荐主要涉及自然语言处理方向，AIOS集成关键字提取算法。关键字提取算法从简单的TF-IDF，朴素贝叶斯算法到基于语义的算法。相似文章推荐的主流算法包括基于元相似度，语义概念相似度等算法。本文拟设计和实现RAKE（Rapid Automatic Keyword Extraction）关键字提取算法和基于关键字内容推荐算法实现基于xPlore的相似文章推荐引擎。 | | | | | | | | |
| 论文拟采用的主要技术路线、研究思路和实现方法（1000字左右）：  本项目涉及前端的用户操作界面与结果展示，后端的数据获取与存储，全文检索与相似文章推荐。前端采用Twitter Bootstrap；前后端通信采用Spring Web框架，缓存采用Redis；内容索引与检索采用公司现有产品xPlore；相似文章推荐引擎拟采用基于关键字的内容推荐算法，各部分的关系如下图    实现规划：  1 信息获取组件 Crawler：功能包括支持多种数据源抓取，定期更新，定期执行。  数据源包括perforce服务器上产品各个版本的源代码，Powerlink 的文档，Jira的bug报告，在线技术支持文档support notes。此组件拟采用Java8的新特性Akka，基于消息的分布式框架，使其具有更高的可扩展性，支持更多数据源。  支持多种数据源：数据源使用XML和DTD 与JAXB描述，通过XML的配置信息，实现java类的自动生成，为每一类数据源编写对应的Actor，由Supervisor统一管理，负责各个Actor 的实例化与调用。整个Crawler 的流程如下图    2 Crawler与xPlore 的集成。  通过调用xPlore Client API 或者REST API 实现数据的索引与存储。  3 相似文章推荐引擎的实现  调查与研究关键字提取算法，设计并实现算法，将此部分算法封装为独立可供调用的模块。拟采用RAKE算法提取关键字并基于关键字实现内容推荐引擎。Rapid Automatic Keyword Extraction是一个无监督，领域独立，语言独立的关键字提取算法。  RAKE(Rapid Automatic Keyword Extractor)算法的主要过程为：  1）依据stop words和句子的结束标志符完成分词，形成phraselist  2) 基于词频（freq(w)）与词度（deg(w)）计算词的分数,采用公式为  score = deg(w)/freq(w)  freq(w) = 该文本中除stop words 之外，每一个词的频数  deg(w) = 1）中phraselist所有包含该词的phrase的长度的总和，如  phraselist =[“hello world then good night”,”hello Lilei”,”hei Peter”]  deg(“hello”) = 5 + 2 =7  3） 由2）得到的wordscore 进行排序，获取前N个作为关键字。  4 编写UIMA插件，实现推荐引擎与xPlore的集成。  xPlore支持嵌入UIMA插件，按照xPlore的集成规范，编写SimilarSearchEngine类，并使其实现UIMAProcessFactory接口，修改xPlore配置文件，添加pipeline。  xPlore文档处理过程如下图所示:    5 用户界面UI。  采用Bootstrap实现。界面功能包括：用户输入时自动补全功能，采用Tie树实现; 全局信息搜索；Facet搜索，如只在Jira记录里搜索，也可以在Powerlink做搜索，用户可自行选择。对于搜索结果，用户可以将内容打开浏览，同时也显示相似文章。  6 测试用例编写 | | | | | | | | |
| 相关项目的应用前景（限200字）：  主要用于企业级的内容管理与检索，能快速检索并统计搜索信息。同时易于扩展，可用于其他产品的检索与管理，用户只需通过继承接口与简单配置即可支持新数据源的内容索引。此应用有广泛前景，如万方数据的资源统一管理平台，电力系统公网通信资源管理平台等,均是对公司资源的整合。 | | | | | | | | |
| 本人在相关项目中的扮演的角色和承担的工作：  需求分析，架构与设计，代码实现，测试，运维以及推广 | | | | | | | | |
| 论文的拟进行的主要工作（500字左右）：  主要工作有，  1 需求分析，确定项目范围,分析各个组件的功能需求，设计用例并文档化。  2 各种技术的学习与掌握，如Bootstraps + CSS + JavaScript, xPlore API，UIMA, NLP领域知识的学习。  3 内容推荐方法的调查与研究。  4 制定项目计划，前期需求分析与文档化，中期设计与实现，包括信息获取组件，推荐方法设计与实现，各组件与xPlore集成，用户界面设计与实现。  5 设计，实现与测试，包括前期单元测试，后期集成测试与系统测试，性能测试。  6 项目上线与推广。  7 后期的维护与更新，并从用户反馈中不断升级产品。 | | | | | | | | |
| 工作方案及进度安排，预期达到的目标：  11月底，需求分析基本完成，掌握相关技术  12月底，设计各个组件的接口，完成crawler组件。  1月底， 完成推荐引擎的设计与实现  2月底，完成各组件与xPlore集成  3月底，完成界面与后端服务器的集成  4月底，完成论文的编写 | | | | | | | | |
| 预计研究过程中可能遇到的困难和问题，以及应对措施：  1信息获取时，可能会被检测为spider而被网站禁止，可使用多个账号与切换IP地址，或者一次性获取全部内容。  2 SSL访问受限，无法通过多层SSL的验证，导致无法获取页面完整信息。  3与xPlore集成时无法顺利集成，因为xPlore 逻辑非常复杂，对代码的修改容易引入新的错误。针对这个问题，请教身边的资深工程师以防止不必要错误的发生。  3 内容推荐引擎技术难度大，而且与xPlore的集成规范多，需要解决的未知问题较多。  尽量使用xPlore的开发语言java,减少不必要的集成问题。  4 拟采用的基于关键字的内容推荐引擎，需要确定实现方案。 | | | | | | | | |
| 主要参考文献：   1. Naw Naw, Ei Ei Hlaing Relevant Words Extraction Method for Recommendation System 2. Struart Rose, Dave Engel, Nick Granmer and Wendy Cowley   Automatic keyword extraction from individual documents   1. UIMA Classification Annotator for xPlore <https://community.emc.com/docs/DOC-13296> 2. 电力系统公网通信资源统一管理平台   <http://www.letswireless.com.cn/cn/yyfa/show.asp?id=15>   1. xPath 入门<http://www.ruanyifeng.com/blog/2009/07/xpath_path_expressions.html> 2. xPlore Product Page <https://support.emc.com/products/9339> 3. UIMA <http://uima.apache.org/dev-quick.html> 4. JAXB <http://www.oracle.com/technetwork/articles/javase/index-140168.html> 5. Bootstrap <http://getbootstrap.com/getting-started/> 6. Akka <http://akka.io/> | | | | | | | | |
| 导师意见： | | | | | | | | |
| 审查时间 | | 年 月 日 时至 时 | | | 地点 |  | | |
| 审查组成员 | |  | | | | | | |
| 审查会上提出的主要问题及解决办法： | | | | | | | | |
| 学院备案意见：        年 月 日 | | | | | | | | |