哈尔滨工业大学

**<<信息检索>>**

**实验报告**

**(2019年度春季学期)**

|  |  |
| --- | --- |
| **姓名：** | **王丁子睿** |
| **学号：** | **1183710211** |
| **学院：** | **计算机学院** |
| **教师：** | **张宇** |

实验三 企业搜索系统的设计与实现

## 实验目的

本次实验目的是对企业有所系统的设计与实现过程有一个全面的了解。

本次实验设计的内容包括：

* 对数据建立索引，实现文档的搜索，并对检索结果排序。
* 实现企业搜索中的分权限访问。

## 二、实验内容

1. 建立检索系统

本次实验使用到的数据是实验 1 中爬取的网页数据，首先要对 1000 个网页的网页内容建立索引，其次也要对爬取到的所有附件文档建立索引。然后实现一个简单的检索系统，实现数据和文档检索，并且能够精确的对检索结果进行排序。

1. 分权限访问

定义多种不同的“企业角色”（至少 4 种），这些角色对数据或文档的访问权限不同，然后为每条数据增加访问权限。然后在现有检索系统的基础上加入分权限访问功能，使得不同角色的用户在使用检索系统时，只能看到自己具有访问权限的那部分内容。

## 三、实验过程及结果

1. 建立检索系统

数据的处理与检索部分见src/data.py。

对网页和文档的处理一起执行，网页的检索内容为网页全文；文档由于文本较复杂，故采用docx库提取且仅提取文档的文本，作为检索内容。

检索的实现沿用实验二中的方法：首先对每个检索内容进行分词，然后将每个句子变为tf-idf向量，检索时，查询tf-idf向量的余弦相似度最大的句子作为检索的结果。

由于每个句子的tf-idf向量存储占用空间太大，故每次启动检索系统时才会计算每个句子的向量表示。

1. 分权限访问

为了体现分权限访问，首先给每个检索内容随机设置了一个检索权限，只有用户的权限大于等于检索内容的检索权限时，才会在检索系统中显示。

初始页面如下所示：

图形用户界面

描述已自动生成

初始界面为游客模式，检索时会显示相应的内容（设置为五条检索结果），结果如下：

图形用户界面, 应用程序

描述已自动生成

切换用户可以点击右上角的按钮，输入相应的用户名和密码：

图形用户界面, 表格

描述已自动生成图形用户界面, 应用程序

描述已自动生成

切换后，重新输入上述检索内容，发现结果与之前并不同，表明分权限检索有效：

图形用户界面, 应用程序

描述已自动生成

## 四、实验心得

* 了解了检索系统的基本结构，初步实现了一个简单的检索系统。
* 初步尝试了分权限检索系统的实现。