

上海交通大学《电类工程导论 C》

## 大作业实验报告



### 《题目》 学术搜索引擎

组长：钱堃

学号：5120309700

组员：武治宸

学号：5120309702

组员：韩建华

学号：5120309707

组员：王玫

学号：5120309691

日期：2014-1-10

## 一、 分工简述

钱堃：交大视频搜索、后期工作整合、前期任务分配

武治宸：elsevier 搜索、排序算法的设计与完成、图片搜索

王玫：图灵社区搜索、ppt 制作

韩建华：界面制作、代码整合

## 二、 网站爬取与建立索引

我们分别对三个网站进行了各自的网站结构分析,三个人的工作内容虽然不同但是过程类似,我们以图灵社区的搜索为例简述一下我们在这部分的工作内容。

### 网站爬取：

(1) 爬取限定范围：我们每个人均在自己需要爬取的网站中添加域名限制,因此便将爬取的 url 限制在自己需要的网站内。例如：我们把图灵图书的域名限制设定为 `http://www.ituring.com.cn/book`, 而非仅仅是图灵网站。初次按图灵网站爬取时发现“图片链接、下载链接、排序链接”等等大量重复无用 url, 耗费人力物力并且降低运行效率。

(2) 抓取按需索求：我们需要根据建立索引所需的内容来找网页, 避免冗余与繁琐。我们建立学术索引需要的内容有：书名, 作者, 封面图片 url, 简介; 对排序和优化有益的方面有：投票, 阅读量和标签。以图灵社区为例, 找到一个 hint 便是只需爬 `"/book/tagged/..."` 或 `"/book/collected/..."` 中的内容, 即可涵盖索引所需的全部信息。

**建立索引时**：我们需要将爬取下来的网页中格式化的 html 代码中爬出所需信息, 即书名, 作者, 封面, 简介和标签。

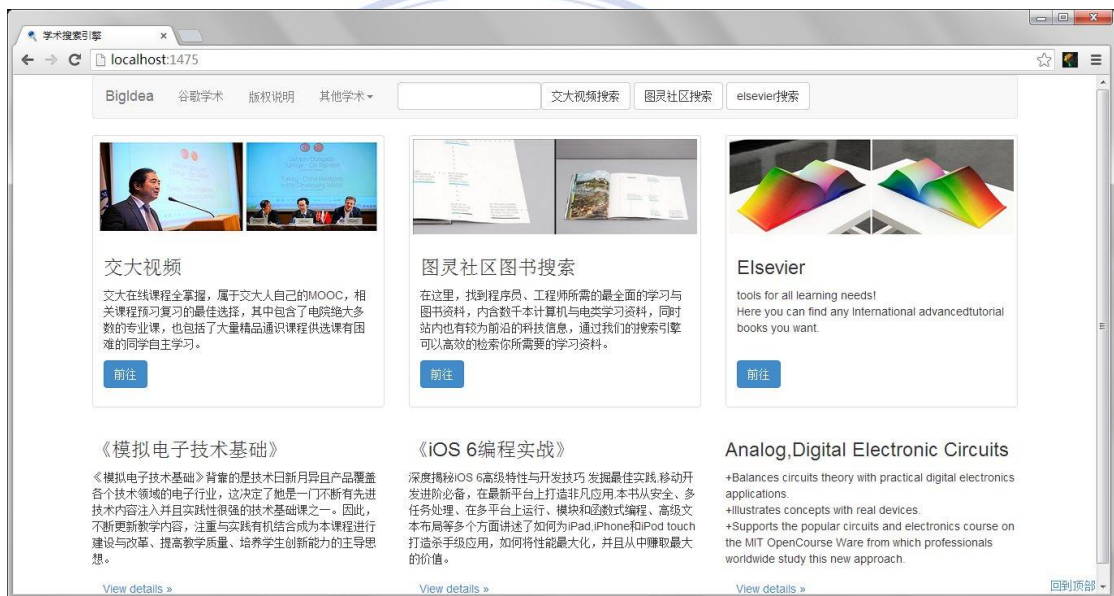
(1) 两个中文网站编码问题兼容：由于此次网页性质唯一, 编码仅为一种, 图灵社区

网站的编码为' utf-8' ,对 python 有很好的兼容性,因此避免之前大规模爬取网页建立索引时遇到的各种网页编码不同带来的问题,交大视频所采用的的方法是  
通过 BeautifulSoup 将 html 的内容转成一个 string 以避免 charset 的操作

(2) 采用标准分词:学术搜索与娱乐社论搜索的一个重要不同在于其科学严谨性,因此我们可以借助 jieba 分词,并采用 StandardAnalyser 用 lucene 建立索引。

### 三、 前台界面

#### 1. 结构整齐, 美观



2. 简洁易懂，所需操作简单、
3. 细节特点

弹出框可以告知用户版权信息

首页上的人性化链接，如：谷歌学术链接、热点推送、搜索来源网站链接等

返回顶部的小按钮以及保留上次搜索结果的人性化设计

4. 搜索图片部分代码分析

1. 获取照片绝对位置

```
<div class="file-box">
  <form action="/tu" method="GET" enctype="multipart/form-data">
    <input type="file" name="keyword" class="file" id="fileField" size="28"
      onchange="document.getElementById('textfield').value=this.value" />
    <button type="submit" class="btn btn-default" >图片搜索</button>
  </form>
</div>
```

2. 将位置变为可以识别的位置，在 opencv2.2 中 ‘\’ 不被识别需用 ‘\\’ :

```
user_data = web.input()
a=""
for i in user_data.keyword:
    if (i=="\\"):
        a+="\\"
    else:
        a=a+i
```

3. 先将图片带入图片匹配系统再将匹配结果图片名称放入搜索框切最后

只选择输出第一个;

```
res=SearchImgs.SearchImgs_elsevier(a)
vm_env.attachCurrentThread()
user_data.keyword = res
.....
b=1
```

## 四、 排序、 图片搜索以及亮点功能

- 1、 排序 :

我们在建立索引时给不同的搜索范围加上了初始权重(通过 setBoost 函数实现)。



根据网上资料，搜索时给出的默认排序为匹配程度\*权重，而不加指示时默认初始权重都为 1.0。通过初始权重的分化，我们可以使搜索结果有等级地排序，例如先按书名、作者，再按介绍，最后按科目。

## 2、图片搜索：

首先我们在建立索引时使用 `urllib.urlretrieve` 函数将需要的图片下载下来，为之后 LSH 图像匹配做准备，之后我们改进了 LSH 程序，在对图库中图片处理之后我们存储下每张图片的 title、Hash 值(舍去了 Hamming 码转化部分以提高匹配速度)、12 维颜色向量，这样在以图搜图时只需要对目标图片 Hash 一下就能直接进行对比。

图书搜索有其自己的特点，我们要做的只是匹配上原图，不少图书的封面颜色比例类似 LSH 如果不对原图搜索的话精度不高。

## 3、领域分类

我们在搜索对每本图书加上链接使得用户可以通过搜索结果找到更多相关科目的图书。

# 五、 个人心得

## 1、专项搜索的利弊

本次我们团队大作业实现学术搜索引擎，与之前小作业的一个不同之处在于大作业应用性强，针对性强，搜索内容方面固定。这样在完成时既有易处，同时也带来一些难题。

(1) 易处：限定搜索内容方向后，有固定对的网页范围爬取，在网页编码解码问题，爬取网页范围，索引分词方法等方面减少了工作量。

(2) 难题：然而为实现学术搜索的便利与准确，需要额外研究网页成分构成、索引内容解析与排序算法优化上面的问题，为达到最优效果着实费了一番功夫。

## 2、排序算法优势

自定义权重赋值相关度结合网评参数的排序算法,实现了机械的相关性与人文的热度的有机结合。顺序大方向由搜索项出现位置决定,当搜索词项出现在同一位置时,再按照相关度排序,排序准确。

## 3、实验收获与感想

本次大作业团队项目最终能够顺利完成,离不开组内 4 名同学每个人的付出与努力。我们都疑惑过、苦思过、通宵过,经历一段难忘的合作学习探索搜索引擎的经历后,在完成的时候感到欣慰与满足,并在学习的过程中收获颇丰。

- (1) 团队合作能力
- (2) 自主探索能力
- (3) 索引排序优化提升
- (4) 从学习到实践
- (5) 由一般到特殊的拓展能力

在完成本次大作业后,本学期电类工程导论 C 多媒体搜索引擎的课程也随之告一段落,本学期也画上了一个圆满的句号。感谢老师一个学期以来的讲解与指导,为我们多媒体检索的知识打开了门窗,扩宽了我们的视野,培养自学互学的能力与兴趣,为今后追赶时代浪潮,探索神秘的科技世界打好积淀。


## 六、 结果展示


## Big idea

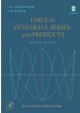
学术搜索引擎 localhost:1475/e?keyword=english


BigIdea 图片搜索 谷歌学术 其他学术 - english 交大视频搜索 图灵社区搜索 elsevier搜索

我们找到了与elsevier相关的"english"20条消息。

 **Writing and Presenting in English**  
作者: Pezey Young  
介绍: The Rosetta Stone of Science is a useful and practical guide to presenting scientific research in the English language. It is written specifically for scientists who would like to improve the effectiveness with which they use the English language and improve their communicative  
• 相关领域: Chemistry  
网址: [http://textbooks.elsevier.com/web/product\\_details.aspx?isbn=9780444521187](http://textbooks.elsevier.com/web/product_details.aspx?isbn=9780444521187)

 **Sobotta Atlas of Anatomy Package, 15th ed., English/Latin**  
作者: Friedrich Paulsen  
介绍: THE exam atlas: learning and understanding anatomy The English-language Sobotta Atlas with Latin nomenclature is specifically adapted to the needs of preclinical medical students. Right from the start, the book and the Internet content concentrate on exam-relevant  
• 相关领域: Anatomy / Physiology  
网址: [http://textbooks.elsevier.com/web/product\\_details.aspx?isbn=9780723437314](http://textbooks.elsevier.com/web/product_details.aspx?isbn=9780723437314)

 **Table of Integrals, Series, and Products**  
作者: Alan Jeffrey  
介绍: - Fully searchable CD that puts information at your fingertips included with text - Most up to date listing of integrals, series and products - Provides accuracy and efficiency in work The Table of Integrals, Series, and Products is the essential reference for integrals in the English  
• 相关领域: Mathematics & Statistics  
网址: [http://textbooks.elsevier.com/web/product\\_details.aspx?isbn=9780123736376](http://textbooks.elsevier.com/web/product_details.aspx?isbn=9780123736376)

 **The Channels of Acupuncture**

回到顶部

## Elsevier 搜索结果

学术搜索引擎 localhost:1475/s?keyword=电子

BigIdea 图片搜索 谷歌学术 其他学术 - 电子 交大视频搜索 图灵社区搜索 elsevier搜索

我们找到了与交大视频相关的"电子"5条消息。

 **课程: 电磁场**  
作者: 周海朋  
介绍: None  
网址: <http://v.sjtu.edu.cn/course/showopencoursevideo.aspx?oid=167>

 **课程: 基本电路理论**  
作者: 张峰  
介绍: None  
网址: <http://v.sjtu.edu.cn/course/showopencoursevideo.aspx?oid=183>

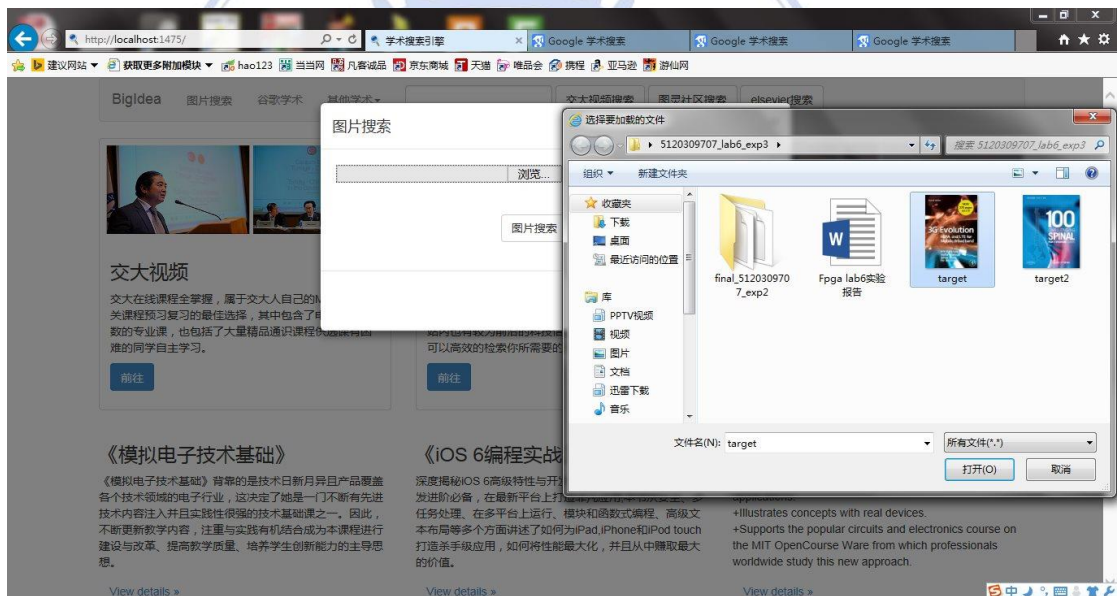
 **课程: 电工与电子技术**  
作者: 姚志红  
介绍: 《电工学》是我校"国家工科电工电子基础课程教学基地"的核心课程,作为全校非电类专业学生的一门重要基础课,它的任务是:通过本课程的学习,使学生熟悉电能在现代生产和生活领域中的应用,了解电气信息技术对现代科技事业发展的作用,能够综合运用所学的电工及电子技术基  
网址: <http://v.sjtu.edu.cn/course/showopencoursevideo.aspx?oid=125>

回到顶部

## 交大视频搜索结果



### 图灵社区搜索结果



### 图片搜索示例



## 图片搜索



C:\Users\John\Desktop\51203097 浏览...

图片搜索

Close

输入图片路径

The screenshot shows a web browser window with the following elements:

- Address bar: `http://localhost:1475/Au?keyword=C963A%5CUser...`
- Search bar: Contains the text "3G".
- Navigation buttons: "BigIdea", "图片搜索", "谷歌学术", "其他学术", "交大视频搜索", "图灵社区搜索", "elsevier搜索".
- Search results: A single result for "3G Evolution" by Erik Dahlman. The description includes details about the book's content and a link to the Elsevier website.

回到顶部

图片搜索结果