

# 校园搜索引擎实验报告

吴昊哲 2015011297 张正彦 2014010515

## 实验环境

- apache-tomcat-7.0.86
- paoding-analysis-2.0.4-beta
- IDEA ULTIMATE
- win10

## 实验内容

综合运用搜索引擎体系结构和核心算法方面的知识，基于开源资源搭建搜索引擎，具体包括如下几点：

1. 抓取清华校园网内绝大部分资源，并且进行预处理；
2. 基于Lucene实现校园搜索引擎——**太强搜索**；
3. 加入关键词纠错、查询提示、语音搜索、相关推荐功能，以提高**太强搜索**的体验；
4. 美化Web界面，实现关键词高亮、快速预览等功能；
5. 完成对于**太强搜索**的性能评价。

## 实现过程

### 抓取校园网资源并处理

使用 Heritrix 抓取工具，抓取 HTML，PDF，M.S.Word 格式的文件28万份，共计31GB。编写 Python 脚本处理抓取到的数据，解析成 json 文件：首先遍历所有抓取到的文件，为每一个文件分配一个 ID，文件与 ID 一一对应，ID 用于之后PageRank的计算。获取文件的标题、文本 (docContent)、标签(h1~h6)、加粗(strong)信息等。使用 BeautifulSoup 库解析 HTML 文件内容，获取其中的超链接，为抓取到的整个数据包构建图结构，根据图结构计算网页的 PageRank，使用pdfminer库解析pdf文件，使用docx2txt库解析word文件。我们发现实际抓到的html文件给出的charset有时是错误的，因此使用了chardet自动判断网页的编码，这样我们便可以处理几乎所有的编码。

解析得到的文件json示例如下：

```
[
  {
    "title": "逾450亿投入新一轮“985工程”高校建设",
    "url":
      "www.tsinghua.edu.cn/publish/thunews/9669/2012/20121228135924483970631/20121228135924483970631_.html",
    "id": "228055",

    "docContent": "逾450亿投入新一轮“985工程”高校建设 来源：中国新闻网2012-12-27马海燕 记者27日从中国
```

教育部获悉，超过450亿元人民币将投入新一轮“985工程”高校建设。其中，中央财政专项资金投入264.9亿，地方协议配套资金投入186.33亿，分别比“985工程”一期时增长102%和93%。从1999年起，中国开始实施“985工程”，重点支持部分高校创建一流大学和高水平大学。前两期投入分别达到227.7亿元和225.83亿元，共支持32所重点高校建设。据了解，这32所重点高校分布于中国16个省市，目前教育部已和相关地方政府完成全部签约。其中，11个省市达到或超过1:1配套，即中央财政投入多少，地方配套多少或超过中央投入。教育部高等教育司司长张大良表示，通过实施“985工程”重点共建，调动了各方资源，特别是争取地方政府加大投入力度，体现了中国特有的体制优势，在增强办学实力方面发挥了重要作用；同时也推动了相关学校转变观念，主动服务国家战略需求，加大协同创新、科技成果转化力度。对于许多高校近年多幢大楼拔地而起、硬件大幅改善，相关资金是否出自“985工程”经费的疑问，中南大学校长张尧学表示，“985工程”的钱从未用来建过大楼，前两期的投入主要用于科研平台的建设和重点科研人才的引进。对于地方大量配套资金投入是否会导致“985工程”高校在招生上向当地倾斜的疑问，张大良表示，与地方政府签订的协议不存在地方保护问题，也不存在与招生挂钩的问题，目前国家的高校招生政策主要是向西部倾斜。教育部鼓励高校与地方探索建立长期、稳定的共建合作载体，建立资金使用监管机制，确保专项资金使用的规范和安全。“树立诚信品质，恪守学术道德，弘扬诚信美德，勇担社会责任……”在6日举行的高校研究生科研诚信研讨会上，来自北京大学、清华大学等全国29所“985工程”高校的研究生代表庄严宣誓，并签署了首份《中国研究生科研诚信公约》。今天，教育部在中国人民大学召开“985工程”高校章程建设工作交流会，交流高校章程建设所取得的进展、经验，深入研究分析形势与问题，旨在加快章程建设步伐，进一步推进现代大学制度建设。教育部副部长郝平出席会议并讲话。清华大学“985工程”三期绿色大学校园实验室项目总结研讨会举行清华新闻网1月6日电2013年12月24日，清华大学“985工程”三期本科生教育综合改革项目绿色大学实验室（CAL）项目组召开了项目总结研讨会。来自教务处、水利系、土……我校二期“985工程”能源等六个科技创新平台启动评审会召开【新闻中心讯】近日，校学术委员会在召开二期“985工程”环境、建筑、机械、能源、公共安全、工物等六个科技创新平台启动评审会。副校长康克军、相关领域的校学术委员会委员、特邀专家共20余人出席会议。与会专家在听取了6个科技创新平台负责人的汇报后，进行了热烈的讨论，提出了许多建设性的意见和建议。这些……我校进行二期“985工程”科技创新平台和哲学社会科学创新基地启动动员【新闻中心讯】4月1日上午，我校“985工程”办公室召开了二期“985工程”的科技创新平台、哲学社会科学创新基地负责人和院系负责人会议。康克军副校长就我校二期“985工程”的平台、基地启动工作进行了动员和布置。康克军回顾了一年多来我校按照教育部等上级领导部门要求进行的二期“985”规……无标题文档从985工程看高层次人才培养光明日报2003年7月24日本报记者丰捷朱宏伟，清华大学机械系1998级直博生。在2002年5月3日出版的《科学》杂志上，他以第一作者的身份发表了题为《直接合成超长单壁碳纳米管》的论文。姜开利，清华大学2001级物理系在职博士生。2002年10月24日，他同样以第一作者的身份在著名的《自然》杂志上发表了论文《……》，

"anchor": " ENGLISH 清华主页 首页 头条新闻 综合新闻 要闻聚焦 时讯快递 学术科研 教育教学 招生就业 交流合作 观点报道 社会服务 媒体清华 图说清华 视频空间 清华人物 校园写意 广角透视 校园生活 文化漫谈 清华史苑 高教视点 专题新闻 新闻排行 新闻合集 分享 首页 校园写意 高教视点 全国首份研究生科研诚信公约发布 “985工程”高校章程建设工作交流会举行 清华985三期绿色大学校园实验室项目总结… 我校二期“985工程”能源等六个科技创新… 我校进行二期“985工程”科技创新平台和… 从985工程看高层次人才培养 更多> 网站地图 关于我们 友情链接 清华地图”，

"h1": "",

"h2": "相关新闻 更多>图说清华 最新更新",

"h3": "全国首份研究生科研诚信公约发布 “985工程”高校章程建设工作交流会举行 清华985三期绿色大学校园实验室项目总结… 我校二期“985工程”能源等六个科技创新… 我校进行二期“985工程”科技创新平台和… 从985工程看高层次人才培养”，

"h4": "",

"h5": "",

"h6": "",

"strong": "逾450亿投入新一轮“985工程”高校建设”，

"pr": "1.6498809498263923e-06"

},

其中 pr 表示网页的 PageRank 的值。我们生成了多个json，减轻一次入读过大文件给程序造成的压力。

## 分词处理

在检索时使用了庖丁解牛分词器 PaodingAnalyzer 作为 Parser 的 analyzer，支持分词检索。另外，使用 IKAnalyzer 以支持检索结果的关键词高亮。

## 词表建立

为了便于后续的查询提示和查询纠错，我们需要建立一个包含中英文常用词的词表，词表的构成如下：

- 10000个英文单词
- 56065个常用中文词语
- 50000个新闻网页分词得到的词频最高的词语

## 结果排序

使用 `Lucene. MultiFieldQueryParser` 实现对关键词的跨域搜索。Lucene提供了基于向量空间模型的排序算法，评分公式为：

$$score(q, d) = coord(q, d) \times queryNorm(q) \times \sum_{t \in q} (tf(t \text{ in } q) \times idf(t)^2 \times t.getBoost()) \times norm(t, d)$$

另一种常用的相关度计算方法为 BM25 算法，评分公式为：

$$score(q, d) = \sum_i^n idf(q_i) \frac{f_i(k_1+1)}{f_i + k_1(1-b+b \frac{dl}{avgdl})}$$

实验中对两种算法均进行了尝试，发现两者在检索效果上差异不大，但 BM25 算法速度明显偏慢，因此最终采用了 Lucene 默认的排序算法，并对于参数做了以下的调整：

- PageRank的影响

由于 PageRank 变化范围较大( $10^{-3} \sim 10^{-7}$ )，将 PageRank 值扩大  $10^4$  倍后乘以得分，与原始分数相加，即：

$$score(q, d) = score(q, d) \times (1 + PR(d) \times 10^4)$$

从而提高 PageRank 较大的页面的排名，同时也能防止 PageRank 较大但相关度不高的页面排在前面。

- 文档文件的处理

对于 doc、pdf 文件，由于仅包含 title、url、content 这三个域，在相同条件下评分会偏低，因此对其 score 加倍处理。

- index 页面的处理

通常情况下网站主页的质量比其他页面更高，因此对于以“index.html”结尾的页面，对其 score 加倍处理。

- url 检索

使用正则表达式匹配判断检索关键词是否为 url，如果是，则仅在 url 一个域内进行检索，提高此类情形下的检索准确率。例如搜索“info.tsinghua.edu.cn”，查询结果第一位为清华信息门户页面。

- 不同域的权重

| Field   | Boost | Field  | Boost |
|---------|-------|--------|-------|
| title   | 5.0   | h3     | 0.8   |
| content | 0.2   | h4     | 0.6   |
| url     | 5.0   | h5     | 0.4   |
| anchor  | 1.0   | h6     | 0.2   |
| h1      | 1.2   | strong | 0.8   |
| h2      | 1.0   |        |       |

### 查询提示

根据输入框中的内容，利用Lucene的 `SpellChecker` 在本地词库中进行匹配，并按照长度从小到大的顺序给出不超过 6 个查询提示。使用 js 实现下拉框前端，支持使用键盘上下键选择检索条目。

### 查询纠错

使用 Lucene 的 `SpellChecker` 类进行查询纠错。将关键词在本地词库中进行查找，若关键词不存在则调用 `suggestSimilar` 函数获取与之最相近的词作为纠错建议。对于一个 query，根据空格将各个关键词分开，对每个关键词分别进行检查纠错处理，最后合并得到纠错建议。

### 关键词高亮

对于检索结果的摘要部分，使用 Lucene 的 `Highlighter` 类，定义相应的高亮格式 `Formatter`、评分类 `QueryScorer`，实现关键词高亮。其中使用了 `IKAnalyzer` 进行分词处理。摘要的长度上限设为 100 字符。

### 语音输入

在 js 中创建 `webkitSpeechRecognition` 对象，识别用户的语音输入。

### 快速预览

检索结果提供预览功能，当鼠标悬停在“快速预览”标签上时，页面的右部将动态加载页面，实现预览。

### 相关推荐

在许多成熟的搜索引擎都提供了相关推荐服务，以此来提供更加多样的结果，并且增长用户的使用时间。很多人都不断的在相关搜索之中跳转，然后忘了自己到底是在搜索什么的经历。下面是谷歌（上图）和百度（下图）的相关推荐（搜索词：LaTeX）：

用户还搜索了

查看另外 10+ 项

Share



ShareLaTeX

MiKTeX



Texmaker



Microsoft  
Word



TeXstudio

相关软件

展开



[maple](#)

数学和工程  
计算软件



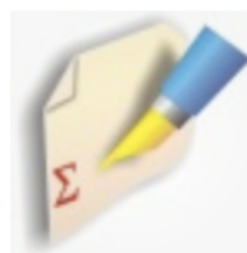
[公式编辑器](#)

PC平台公式  
处理软件



[texpad](#)

适用Mac操  
作系统软件



[winedt](#)

兼用型文本  
编辑器

从中可以发现，谷歌的相关推荐应该是基于用户大量的搜索日志，然后去发现共现的query，而我们在实验中不能够获得大量用户的使用数据，因此这条路看起来行不通。观察百度的相关推荐，我们猜测它应该是一个基于规则的推荐，先判断出LaTeX是一种文字软件，然后返回与其同一分类的软件，这些软件不一定和LaTeX有关。对比两个结果，谷歌的推荐相关度更高，但相比百度可能缺少一点多样性。

我们打算借鉴百度的思路，根据额外的数据（知识图谱）来进行相关的推荐。在这里我们选择[中文通用百科知识图谱](#)作为额外数据，CN-DBpedia主要从中文百科类网站（如百度百科、互动百科、中文维基百科等）的纯文本页面中提取信息，经过滤、融合、推断等操作后，最终形成高质量的结构化数据，供机器和人使用。

具体的算法流程如下：

- 先对于预料进行分词，构建基于预料的词表
- 再使用知识图谱中的entity与词表进行匹配，找到出现的entity的集合
- 使用entity的类别信息构建一个entity-类别的二部图
- 对于二部图，使用网络表示学习的技术[PTE](#)，学习每个entity的embedding
- 对于query，先分词，找到出现的entity，构建query的embedding，再找到最近的entity作为推荐返回

我们使用embedding的技术来绕过编写规则的繁琐操作。我们可以看一下基于规则应该要如何进行推荐：首先，找到entity属于的多个类别，然后通过规则确定最吻合、范围最小的类别，然后在类别中寻找相关推荐。例子：复旦大学，属于 大学、985大学、上海大学 分类，这时候选择类别就比较头疼。

而我们的embedding方法能保证有相似类别的entity会有着很相似的embedding，比如复旦大学和上海交通大学都是 大学、985大学、上海大学，因此他们的embedding就离的很近，因此在做推荐的时候就能直接由复旦大学找到上海交通大学。

具体的相关推荐的实验结果见使用说明部分。

## 使用说明

### 主页

主页上方有一个清华大学的logo，可以通过在文本框中输入文本进行搜索，也可以通过点击右侧的语音按钮进行搜索。



搜索

### 查询推荐

查询推荐在搜索结果页面的右侧，类似于百度的 热搜栏

施一公的推荐结果:

施一公

搜索

为您找到相关结果约1000个结果,耗时242毫秒

[1. 健康达人施一公](#) - 快速预览

[tv.tsinghua.edu.cn/publish/video/9684/2014/20140610150948368...](http://tv.tsinghua.edu.cn/publish/video/9684/2014/20140610150948368...)

[2. 健康达人施一公](#) - 快速预览

[www.tsinghua.edu.cn/publish/video/9684/2014/2014061015094836...](http://www.tsinghua.edu.cn/publish/video/9684/2014/2014061015094836...)

[3. 往事施一公《回家》](#) - 快速预览

[www.tsinghua.edu.cn/publish/video/9682/2017/2017092711305663...](http://www.tsinghua.edu.cn/publish/video/9682/2017/2017092711305663...)

[4. 良师益友：施一公](#) - 快速预览

[tv.tsinghua.edu.cn/publish/video/10140/2014/2014100908125547...](http://tv.tsinghua.edu.cn/publish/video/10140/2014/2014100908125547...)

[5. 健康达人施一公](#) - 快速预览

[www.tsinghua.edu.cn/publish/video/10176/2014/201406101509483...](http://www.tsinghua.edu.cn/publish/video/10176/2014/201406101509483...)

[6. 良师益友：施一公](#) - 快速预览


为您推荐

- 1 王伯雄
- 2 许宁生
- 3 韩太元
- 4 许倬云
- 5 曹达人
- 6 吴德星
- 7 蒋树声
- 8 吴南轩
- 9 彭永臻
- 10 陈鼓应

上海交通大学的推荐结果:



上海交通大学



搜索

为您找到相关结果约1000个结果,耗时488毫秒

1. [DOCX] 环境学院2012年大事记1月上海交通大学.docx - 快速预览

环境学院2012年大事记1月上海交通大学授予顾夏声院士杰出校友成就奖11名硕士生（含8名在职工程硕士生）、6名博士生和6名留学生被授予清华大学学位2月中共中央国务院任命陈吉宁为清华大学校长胡锦涛等...

[www.env.tsinghua.edu.cn/publish/env/6316/6316/2012event.docx](http://www.env.tsinghua.edu.cn/publish/env/6316/6316/2012event.docx)

2. [DOCX] 环境学院2012年大事记1月上海交通大学.docx - 快速预览

环境学院2012年大事记1月上海交通大学授予顾夏声院士杰出校友成就奖11名硕士生（含8名在职工程硕士生）、6名博士生和6名留学生被授予清华大学学位2月中共中央国务院任命陈吉宁为清华大学校长胡锦涛等...

[www.tsinghua.edu.cn/publish/env/6316/6316/2012event.docx](http://www.tsinghua.edu.cn/publish/env/6316/6316/2012event.docx)

为您推荐

- 1 对外经济贸易大学
- 2 莫斯科大学
- 3 山东工艺美术学院
- 4 哥本哈根大学
- 5 大连海事大学
- 6 首都医科大学
- 7 上海师范大学
- 8 北京电影学院
- 9 天津大学
- 10 中国人民大学

校庆的推荐结果:



校庆



搜索

为您找到相关结果约1000个结果,耗时383毫秒

1. 清华视频-校庆日记 - 快速预览

今天是清华106周岁校庆日。为表达学子对母校生日的祝福，特献上重新录制的视频节目——配乐诗朗诵《半个世纪清华情》。这个新版本增添了诗文字幕，以及与诗相配的照片和老照片。经王钊熠老师配乐，声情并茂的...

[www.tsinghua.edu.cn/publish/video/10143/](http://www.tsinghua.edu.cn/publish/video/10143/)

2. 清华视频-校庆日记 - 快速预览

今天是清华106周岁校庆日。为表达学子对母校生日的祝福，特献上重新录制的视频节目——配乐诗朗诵《半个世纪清华情》。这个新版本增添了诗文字幕，以及与诗相配的照片和老照片。经王钊熠老师配乐，声情并茂的...

[tv.tsinghua.edu.cn/publish/video/10143/](http://tv.tsinghua.edu.cn/publish/video/10143/)

3. 校庆致辞 - 快速预览

四月的清华园，阳光明媚，春和景明，清华大学迎来了105岁的生日。我谨代表学校向海内外广大校友和全体师生员工致以亲切的问候和良好的祝愿，向多年来关心支持我校发展的各界人士和朋友表示衷心的感谢！过去一年...

[www.tsinghua.edu.cn/publish/bzw/7583/2016/201604221757395463...](http://www.tsinghua.edu.cn/publish/bzw/7583/2016/201604221757395463...)

为您推荐

- 1 民办教育
- 2 教职工
- 3 五年级
- 4 校历
- 5 校委会
- 6 封校
- 7 复旦大学
- 8 中欧国际工商学院
- 9 招生简章
- 10 夜校

可以看到施一公和上海交通大学的推荐结果都是较好的。

查询提示

在输入查询词时动态地给出查询提示:







## 查询样例

| 查询名称     | 类别  | 热度 | 需求                         |
|----------|-----|----|----------------------------|
| 校庆       | 信息类 | 热门 | 获取校庆的相关新闻                  |
| 教育改革     | 信息类 | 热门 | 获取教育改革的相关新闻                |
| 清华CMU    | 信息类 | 冷门 | 获取清华-CMU合作的相关新闻            |
| 大类招生     | 信息类 | 热门 | 获取大类招生的相关信息                |
| 小桥烧烤     | 信息类 | 冷门 | 获取小桥烧烤的相关新闻                |
| academic | 导航类 | 热门 | 找到academic.tsinghua.edu.cn |
| info     | 导航类 | 热门 | 找到info.tsinghua.edu.cn     |
| 双学位      | 事务类 | 热门 | 找到双学位相关介绍页面                |
| 四六级      | 事务类 | 冷门 | 找到四六级的相关页面                 |
| 选课       | 事务类 | 热门 | 找到选课的相关页面(academic, info)  |

## 构建Pooling

根据查询样例集合，抓取查询结果的前十位结果，构建Pooling

## 构建相关性标注集合

对Pooling中的结果进行相关性标注，1表示"答案"，0表示"非答案"。

| 查询样例     | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|----------|---|---|---|---|---|---|---|---|---|----|
| 校庆       | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1  |
| 教育改革     | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 1  |
| 清华CMU    | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0  |
| 大类招生     | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1  |
| 小桥烧烤     | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1  |
| academic | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0  |
| info     | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0  |
| 双学位      | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1  |
| 四六级      | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0  |
| 选课       | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 0  |

## 性能分析

在十个查询样例中，除了"清华CMU"外都搜索到了相关结果，虽然对"清华CMU"的搜索没有找到相关页面，但是对"清华-卡耐基梅隆"的搜索找到了较多的相关页面，经过分析，我们认为出现该问题的原因是分词器进行分词时，并未将"CMU"分为一个完整的词，导致清华CMU无法搜索得到相关结果。

由于对url的特殊处理，academic和info都在搜索结果的第一个出现，效果较好。

下面我们分别使用MAP，P@10，MRR对搜索引擎的性能进行定量分析：

| 评价指标 | MAP  | P@10 | MRR  |
|------|------|------|------|
| 全部查询 | 0.80 | 0.52 | 0.85 |
| 热门查询 | 0.79 | 0.64 | 0.79 |
| 冷门查询 | 0.50 | 0.20 | 0.67 |
| 导航类  | 1.0  | 0.15 | 1.0  |
| 信息类  | 0.63 | 0.58 | 0.7  |
| 事务类  | 0.95 | 0.67 | 1.0  |

性能分析可以看出来，热门查询比冷门查询的性能更好，导航类查询，由于对url所做的特殊处理，使其性能最好，随后是事务类查询，信息类查询的效果不够理想。

在我们的搜索引擎中，为了优先显示pdf，我们将pdf和docx的优先级设置的较高，导致很多无关的pdf和docx有着较高的score，但是当我们降低pdf和docx的权重时，几乎所有的pdf和docx都不能在第一页显示了，这一问题仍需要后续的研究和解决。

## 总结

这次大作业让我们对搜索引擎有了更加深入的了解，我们从爬虫，建立索引，搭网站，写扩展一步一步实现了一个完整的搜索引擎，获益匪浅！最后也感谢老师和助教一学期的辛苦付出！