

2014 年 6 月 4 日

# 校园搜索引擎报告

《搜索引擎技术基础》课程

小组成员：

2011011237 张宏辉

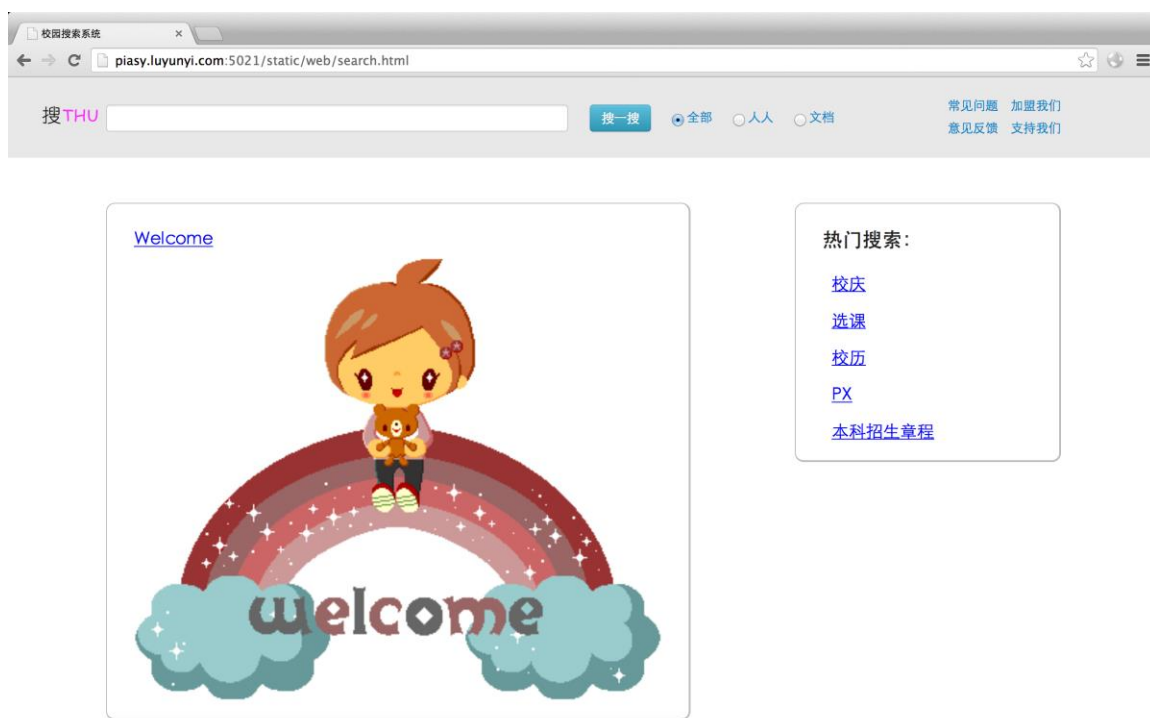
2011011238 许建林

## 目录

一、实验目标 .....	2
二、搜索数据处理 .....	2
2.1 官方网页爬取 .....	2
2.2 人人公共主页爬取 .....	3
2.3 文档解析与文档索引 .....	3
三、搜索引擎后端 .....	5
3.1 BM25 多关键词多域查询框架 .....	5
3.2 GBRT 训练查询参数 .....	5
3.3 特色功能实现 .....	6
四、搜索引擎前端 .....	7
4.1 前后端完全分离 .....	7
4.2 前端功能实现 .....	7
五、搜索引擎效果展示 .....	8
5.1 欢迎界面 .....	8
5.2 搜索关键字补全 .....	8
5.3 基本搜索结果 .....	9
5.4 垂直搜索结果 .....	10
5.5 指定域搜索 .....	11
六、实验总结 .....	12
七、文件说明 .....	12

## 一、实验目标

以清华各类官方网站与相关人人公共主页数据为搜索源，实现清华校园搜索引擎，使搜索效果尽可能理想。同时尽量完善搜索引擎功能，使搜索引擎功能更全面更友好。



## 二、搜索数据处理

### 2.1 官方网页爬取

本次实验需要抓取清华校内不包括图书馆的网页资源，采用 heritrix-1.14.4 进行抓取。抓取的种子：

<http://student.tsinghua.edu.cn>

<http://news.tsinghua.edu.cn>

从 student 出发，可以扩展出清华几乎所有的网站（509 个），最终抓取了近 76 万个文档，总共 67GB，但由于很多都是 php, jsp 等界面，而 Heritrix 在保存网页时命名方式和实际 URL（当有 get 参数时）不一样，所以即便把它们进行索引，同样无法获取有效 URL，所以这些页面我都没有进行索引，另外很多 pdf, doc 文档因为加密以及不符合规范，解析会失败，所以我把这些文档删除了（删除的部分并未包括在 76 万个文档中）。

## 2.2 人人公共主页爬取

我们搜集了和清华的学习生活相关的 80 个人人公共主页(请见 data/renren.txt), 利用人人网开放的 API 获取这些公共主页的状态和日志, 把这些数据进行了索引, 补充了官方网站的数据源不足。最终我们获取了 8000 多条状态和 5000 多篇日志。

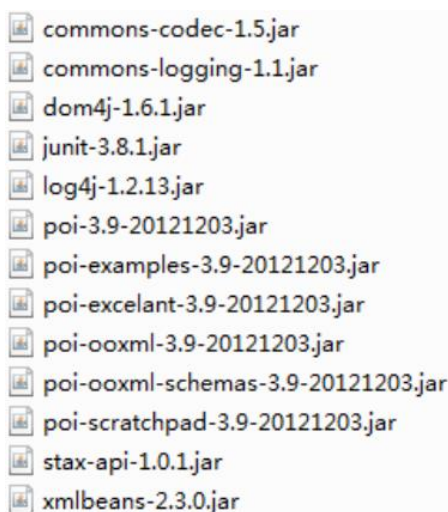
值得指出的是, 在获取公共主页的日志时, 由于人人网服务器的问题, 总是发生 Internal Error, 所以我们采取了累进式抓取, 失败后接着抓取。



1	600638900	学生清华
2	600435535	清华社团
3	601538374	清华舞协
4	600376271	清华大学学生会
5	600907735	清华经管 家园
6	601062142	清华创业
7	601303385	新清华学堂
8	601097635	清华职协
9	601676828	IF清华
10	601017285	清华就业
11	600806598	清华大学清新时报
12	600008302	清华大学社会实践
13	601718308	清华表白墙
14	600992694	清华电视台
15	600704941	清华明理人
16	600633037	清华大学经济管理学院
17	600725003	清华学生心理协会
18	600781251	清华项目管理协会
19	600977841	清华大学学生科协
20	600657173	清华大学物理系

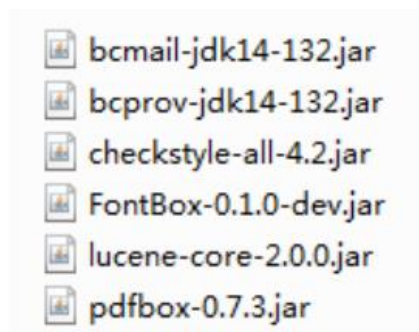
## 2.3 文档解析与文档索引

doc 解析使用 Apache POI, 使用包如下. 由于 doc 与 docx 格式不同, 所以需要分开进行解析:



- commons-codec-1.5.jar
- commons-logging-1.1.jar
- dom4j-1.6.1.jar
- junit-3.8.1.jar
- log4j-1.2.13.jar
- poi-3.9-20121203.jar
- poi-examples-3.9-20121203.jar
- poi-excelant-3.9-20121203.jar
- poi-ooxml-3.9-20121203.jar
- poi-ooxml-schemas-3.9-20121203.jar
- poi-scratchpad-3.9-20121203.jar
- stax-api-1.0.1.jar
- xmlbeans-2.3.0.jar

pdf 解析使用的 pdfbox-0.7.3。使用包如下。其中 pdfbox-0.7.3.jar 主要用于解析 pdf, 其他包与 pdf 中文编码乱码问题相关。



我利用了 jsoup 解析图片所在原始网页, 加入到索引内容中。之所以使用 Jsoup, 这是因为它一直在维护中, 并且 Jsoup 支持类似 JQuery 的 CSS 选择器语法获取 DOM 对象, 使用方便, 而且之前已经用过很多次了, 比较熟了。

这里我从每个 html 文件的内容中选择了 5 个域加入到索引内容中, 分别是:

title, 即 html 的<title>标签的文本;

content, 文档内容, 由于 html 的复杂性, 实现很难达到满意的效果。实验中, 我将所有 p, span, td, th, div, li, a, pre, code, em, strong, b, i 等可见标签的文本组合起来作为内容;

subtitle, 即 h1~h6。由于几乎所有网页只包含 h1~h6 的一个, 因此我没有把它们拆开;

anchor\_out, 网页中包含所有超链接的文本。

anchor\_in, 即指向该网页的锚文本。

这里需要特别指出的是, 网页文件的编码方式问题, 我是采取了这样的解决方案: 因为基本上所有的网页文件都会在其 head 标签中指出编码方式, 而这部分内容是英文的, 所以可以首先采用 utf-8 编码解析出这部分内容, 然后利用字符串查找, 找到编码方式, 然后再以正确的编码方式对该文件进行解析。当然, 如果这个查找过程失败, 就丢弃这个网页, 因为基本上所有的网页都会指出其编码方式, 所以这部分是可以忽略的。处理过程中发现只有极少数网页这一步骤出了问题, 都是什么 gb2313, utf\_8, GB, gb2312, gb2132, gb\_2312-80 等低级的问题, 由此可见互联网的世界是多么肮脏黑暗!

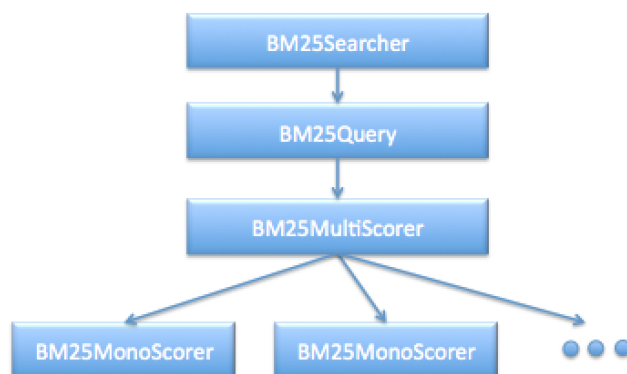
解析完成之后直接使用 Lucene 的索引工具即可, 建立完索引之后就是进行查询了。

## 三、搜索引擎后端

### 3.1 BM25 多关键词多域查询框架

图片搜索实验中,我们在 Lucene 框架下实现了简单的 BM25 查询框架。在此基础上我们重写了评分框架,支持 BM25 的多关键字多余查询。

查询流程图:



上述流程图中每个图块都对应 src/search 下面的一个文件。BM25Searcher 是接口程序:构造函数载入索引文件,loadGlobals()载入全局参量,BMSearch()为搜索的接口函数。当搜索时会调用 BM25Query 获得 BM25Weight,其中的 score()会生成 Scorer,用于生成所有符合条件的文档并计算分值。score()函数对分词后的每个 term 先调用单关键词查询 BM25MonoScorer,然后进行多关键组合 BM25MultiScorer,这里的多关键字指目标文档必须同时包含这多个关键字。这两个 Scorer 的实现与图片搜索类似,应用到的机器学习方法将在下文 3.2 中说明,具体可以参考实现代码。

### 3.2 GBRT 训练查询参数

BM25MonoScorer 的评分结果是搜索引擎搜索效果的重点。评分时我们需要涉及五个域的 BM25 Score (包括 title, h, anchor\_in, content, anchor\_out),此外还需要涉及 PageRank, Time 两大因素。我们选择用 GBRT (Gradient Boost Regression Tree) 来学习这 7 个参量之间的参数。

```
train.txt
1 #title h anchor_in content anchor_out time pagerank score
2 1.0328724 0.0 0.0 2.165893 0.0 1396571841309 7.683168405492324E-6 5
3 #选课 学堂在线选课总人次突破10万 http://news.tsinghua.edu.cn/publish/
4 news/4217/2014/20140304133206950428807/20140304133206950428807_.html
5 1.0124726 0.0 0.0 2.8554556 0.07934865 0 6.239399112928368E-7 4
6 #选课 对学生选课制度的思考 http://student.tsinghua.edu.cn/topic/Lesson/Lessons.htm
7 1.1178254 0.0 0.0 0.0 0.0 1393855567000 1 4
8 #选课 人人选课状态 http://page.renren.com/601045217/fdoing/5149226663
9 0.0 0.0 0.0 2.0670345 0.0 1384993861682 1.1213090829187422E-6 3
10 #选课 清华公开课上线 首日超万人选修 http://news.tsinghua.edu.cn/publish/
11 news/4207/2013/20131021142817774472895/20131021142817774472895_.html
12 0.0 0.0 1.5861807 2.2207792 0.0 0 7.445471510436619E-7 3
13 #选课 专业限选课程—选课举例 http://www.ee.tsinghua.edu.cn/publish/
14 ee/3732/2013/20130530134029436894297/20130530134029436894297_.html
```

GBRT 的源代码可以在网上下载, 放置在 `src/gbt` 里, 但我们需要足够的训练集来训练模型。训练集见 `data/train.txt` 文件, 我们手工标注了 110 条搜索结果, 用 1-5 进行了打分。GBRT 的训练与模型的载入与导出都在 `BM25MonoSearcher` 里面实现, 导出的训练结果文件见 `data/ScoreGbrt.txt`。

### 3.3 特色功能实现

为了满足前端特色功能的实现, 后端在 `BM25Searcher` 中实现了 `genAbstract()` 获得摘要, `getRelated()` 获得历史记录中的相关搜索 (Json 格式), `BM25Search()` 获得搜索结果 (Json 格式, 包含一系列特殊信息)。

关于 Json 格式返回信息做如下说明:

#### `BM25Searcher()`

- `result`: 搜索结果, 包括 `num`, `title`, `text`, `url`, `unique`
- `related`: 相关搜索
- `picSpecial`: 图片型垂直搜索结果, 包括 `text`, `url`, `pic`
- `textSpecial`: 文字型垂直搜索结果, 包括 `title`, `content`

#### `getRelated()`

- `related`: 相关历史搜索记录

#### `getTop()`

- `result`: 返回热门搜索词

## 四、搜索引擎前端

### 4.1 前后端完全分离

为了减少合作开发过程中的耦合,我们采用了前后端完全分离的架构,后端通过 `http server` 向前端提供信息,全部采用 `get` 请求的方式。而前端则是 Python Django 框架提供一个 `web server`,在用户浏览器端,采用 JQuery 向这个 `web server` 发起查询。实际上这个 `web server` 只是一个服务器端的代理,它将和后端进行数据交换,转发用户的 Query 以及后端返回的结果。

这样设计的好处不但减小了开发过程中的耦合,而且使得前后端可以分别部署到不同的服务器,而后端则可以提供 REST API 的形式为其他开发者提供一个 `web service`。

### 4.2 前端功能实现

前端的实现主要采用 JQuery 库,通过 js 脚本来实现数据通信和数据跳转,使得整个搜索界面都在一个网页内,这样做可以大大简化用户 cookie 等数据的传递开销,使得进一步的开发更加容易。而 UI 的设计则是采用了流行的 bootstrap 框架,使得 UI 更加友好,而不是只有一个框,一个钮,其余部分全是空白。

垂直搜索是通过后端给出的 API 来实现的,每次查询的时候,后端返回的结果都可能包含一个垂直搜索结果,通过特判就可以获知是否有垂直搜索结果,有的话就进行特殊展示。

搜索关键字补全,后端会提供一个相关搜索的 API,返回与当前用户输入的字符串前缀匹配的 Query,然后前端会在搜索框下面对用户进行提示,至于发送这个 API 请求的时机,则是采用了目前很成熟的方式,通过回调函数,当用户输入暂停一定时间之后,在这个回调函数内发送 API 请求,这样可以减轻服务器压力,但是并不影响用户体验。

页面去重处理,后端返回的结果中,会包含一个标签,显示这个结果是否重复,如果重复则先将其 `div` 设置为不可见,当用户点击继续查看后将这些 `div` 设置为可见即可。

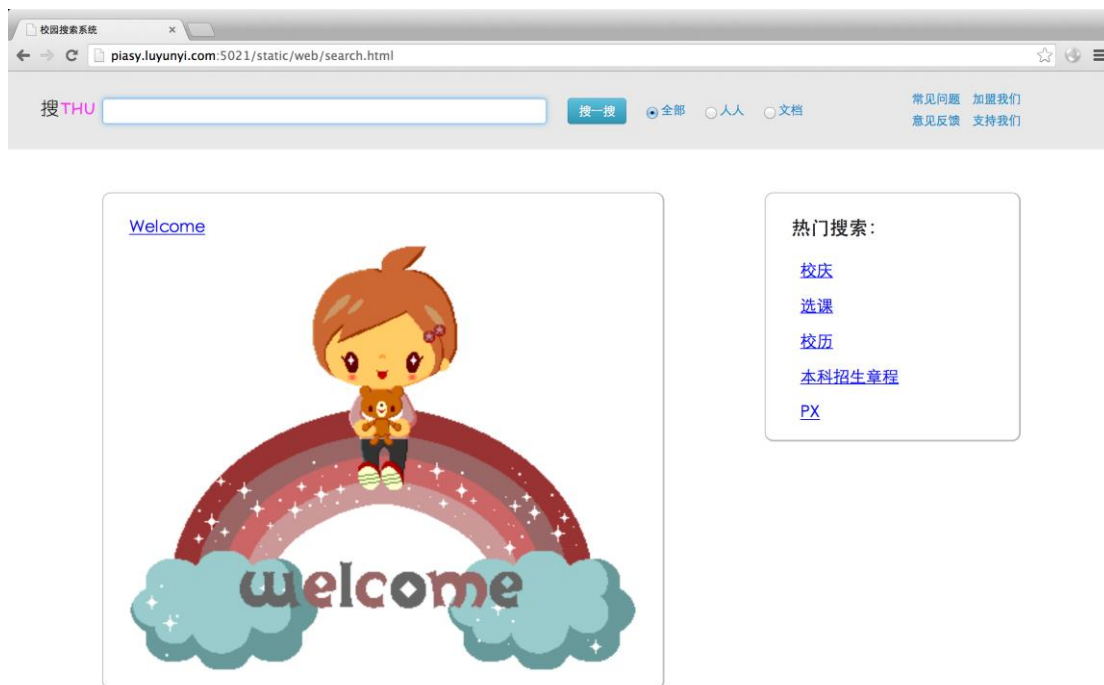
指定域搜索则是后端在处理 Query 时,对 `type:xxx` 开头的 Query 会进行类型过滤,所以当用户指定搜索类型之后,只需要在其输入的关键词之前加上相应的 `type` 即可。

页面上所有的关键字高亮都是在前端通过匹配结果的摘要及标题中的查询词,然后调节 `css` 样式来实现的。

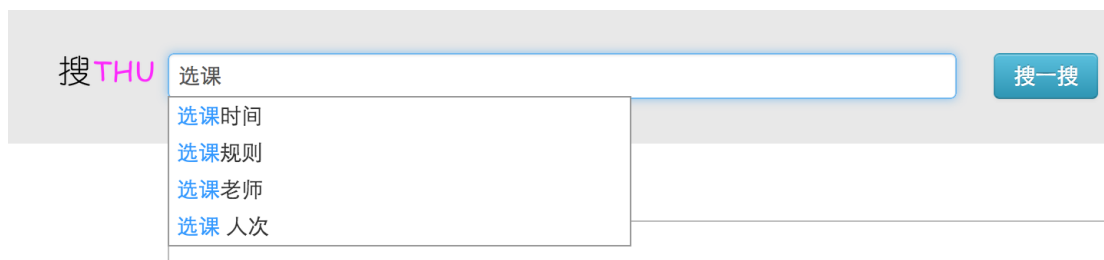


## 五、搜索引擎效果展示

### 5.1 欢迎界面

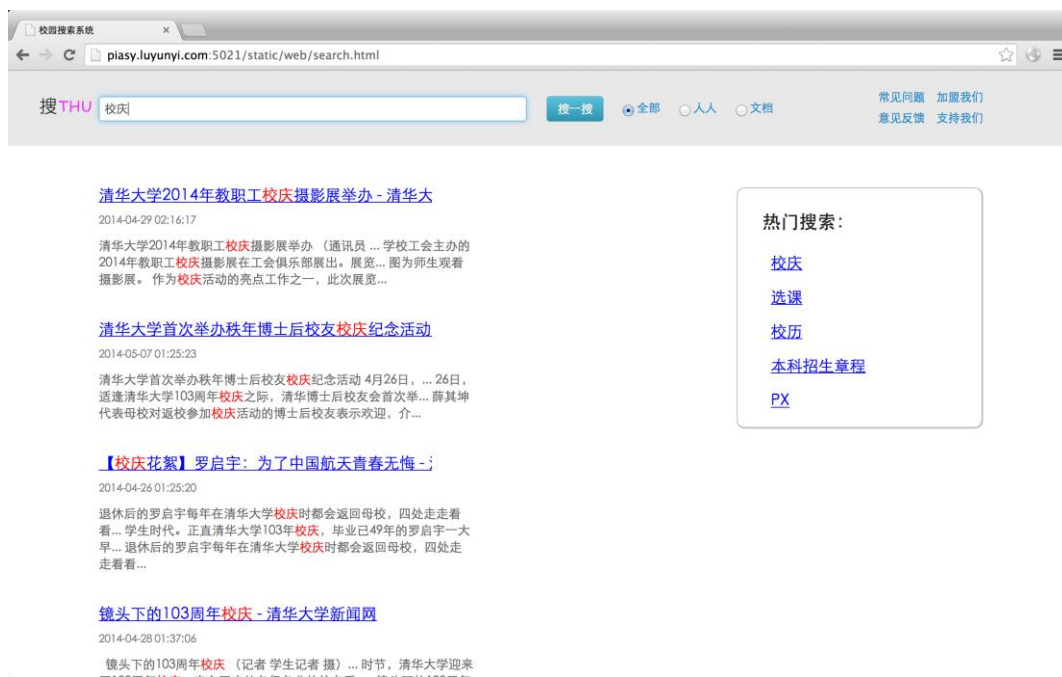


### 5.2 搜索关键字补全



## 5.3 基本搜索结果

搜索“**校庆**” (搜索效果理想, 时间符合, 页面去重实现):



校园搜索系统

← → ↻ piasy.luyunyi.com:5021/static/web/search.html

搜THU 校庆 搜一搜 全部 人人 文档 常见问题 加入我们 意见反馈 支持我们

[清华大学2014年教职工校庆摄影展举办 - 清华大](#)  
2014-04-29 02:16:17  
清华大学2014年教职工**校庆**摄影展举办 (通讯员 ... 学校工会主办的2014年教职工**校庆**摄影展在工会俱乐部展出。展览... 图为师生观看摄影展。作为**校庆**活动的亮点工作之一, 此次展览...

[清华大学首次举办秩年博士后校友\*\*校庆\*\*纪念活动](#)  
2014-05-07 01:25:23  
清华大学首次举办秩年博士后校友**校庆**纪念活动 4月26日, ... 26日, 适逢清华大学103周年**校庆**之际, 清华博士后校友会首次举... 薛其坤代表母校对返校参加**校庆**活动的博士后校友表示欢迎, 介...

[【\*\*校庆\*\*花絮】罗启宇: 为了中国航天青春无悔 -](#)  
2014-04-26 01:25:20  
退休后的罗启宇每年在清华大学**校庆**时都会返回母校, 四处走走看看... 学生时代, 正直清华大学103年**校庆**, 毕业已49年的罗启宇一大早... 退休后的罗启宇每年在清华大学**校庆**时都会返回母校, 四处走走看看...

[镜头下的103周年\*\*校庆\*\* - 清华大学新闻网](#)  
2014-04-28 01:37:06  
镜头下的103周年**校庆** (记者 学生记者 摄) ... 时节, 清华大学迎来了103周年**校庆**...

热门搜索:  
[校庆](#)  
[选课](#)  
[校历](#)  
[本科招生章程](#)  
[PX](#)

已经为您隐藏部分相似度极高的结果, 如需要继续查看, 请点击[这里](#)

搜索“**PX**” (搜索效果理想, 人人信息补充官网信息):



校园搜索系统

← → ↻ piasy.luyunyi.com:5021/static/web/search.html

搜THU PX 搜一搜 全部 人人 文档 常见问题 加入我们 意见反馈 支持我们

[\[视频\]毒性之争引发的PX“词条保卫战” - 清华](#)  
2014-04-08 01:27:33  
[视频]毒性之争引发的PX“词条保卫战” 来源: CC... 亮 周博 最近一段时间, “PX”成为广受关注的一个热词。什... 为广受关注的一个热词。什么是“PX”呢, 如果你通过百度的百科词...

[PX, 一场特殊的“科学保卫战”](#)  
2014-04-06 09:11:14  
strong>PX “PX即对二甲苯。可燃, 低毒化合物...

[《谁将PX妖魔化?》-工21孙念念, 贾斯然](#)  
2013-03-28 22:25:57  
dicd">摘要: PX项目在全国一波未平一波又起, ... 网络查询和采访大学教授, 分析了PX被妖魔化的原因, 总结了在流言... 关键词: PX 环保 剧毒 网络谣言 真相...

[南方周末的谣言是PX事件的推手 - 媒体 - 清华大学](#)  
2014-03-04 01:33:28  
、宁波等地一闹就停的成功后, 反PX行动在今天似乎已经具备了“政... 已经具备了“政治正确性”。近日PX项目就再次在昆明——安宁与成... 宁与成都——彭州激起了波澜。PX (对二甲苯) 明明是低毒 (与汽...

热门搜索:  
[校庆](#)  
[选课](#)  
[校历](#)  
[本科招生章程](#)  
[PX](#)

## 5.4 垂直搜索结果

搜索“**放假**”（出现清华校历供查询放假时间）：



校园搜索系统

piasy.luyunyi.com:5021/static/web/search.html

搜THU 放假

搜一搜 全部 人人 文档 常见问题 加入我们 意见反馈 支持我们

### 2013-2014学年度春季学期和夏季学期

2013-2014 学年度春季学期和夏季学期

日	星期	一	二	三	四	五	六	日
0	二	17	18	19	20	21	22	23
1	三	24	25	26	27	28	1	2
2	四	3	4	5	6	7	8	9
3	五	10	11	12	13	14	15	16
4	六	17	18	19	20	21	22	23
5	日	24	25	26	27	28	29	30
6	一	31	1	2	3	4	5	6
7	二	7	8	9	10	11	12	13
8	三	14	15	16	17	18	19	20
9	四	21	22	23	24	25	26	27
10	五	28	29	30	1	2	3	4
11	六	5	6	7	8	9	10	11
12	日	12	13	14	15	16	17	18
13	一	19	20	21	22	23	24	25
14	二	26	27	28	29	30	31	1
15	三	2	3	4	5	6	7	8
16	四	9	10	11	12	13	14	15
17	五	16	17	18	19	20	21	22
18	六	23	24	25	26	27	28	29
19	日	30	1	2	3	4	5	6
20	一	7	8	9	10	11	12	13
21	二	14	15	16	17	18	19	20
22	三	21	22	23	24	25	26	27
23	四	28	29	30	31	1	2	3
24	五	4	5	6	7	8	9	10
25	六	11	12	13	14	15	16	17
26	日	18	19	20	21	22	23	24

### 清华大学

#### 2013-2014 学年度校历

##### 春季学期(2014 年)

- 2月21日-23日, 本科生、研究生到校注册。
- 2月24日全校本科生、研究生开始上课。
- 清明节: 4月5日-7日放假调休3天。
- 校庆及五一: 4月26日、27日(校庆日) 教职工照常上班; 4月30日-5月4日放假调休5天。
- 端午节: 5月31日-6月2日放假公休3天。
- 第6周期中测验。第17周、18周期末考试。

##### 夏季学期及暑假(2014 年)

- 6月30日-9月21日(共12周)本科生夏季学期及暑假。
- 参加社会实践的研究生6月30日-8月10日(共6周)进行社会实践; 8月11日-9月7日(共4周)暑假。
- 7月3日下午学校学位评定委员会会议。
- 7月5日上午研究生毕业典礼, 7月6日上午本科生毕业典礼。
- 7月21日-8月17日, 从事一线教学和科研的教师、不参加社会实践的研究生暑假4周。

#### 热门搜索:

- [校庆](#)
- [选课](#)
- [校历](#)
- [本科生招生章程](#)
- [PX](#)

搜索“**搜索引擎上课**”（出现搜索引擎基础这门课程的基本信息）：



校园搜索系统

piasy.luyunyi.com:5021/static/web/search.html

搜THU 搜索引擎上课

搜一搜 全部 人人 文档 常见问题 加入我们 意见反馈 支持我们

### 2013-2014学年度搜索引擎上课信息

上课教师: 刘奕群

上课时间: 每周二下午第一大节(13:30-15:05)

上课地点: 六教6A211

## 5.5 指定域搜索

搜索“**校庆（指定人人域）**”：

The screenshot shows a web browser window with the URL `piasy.luyunyi.com:5021/static/web/search.html`. The search bar contains the text '校庆' and the results are filtered by the '人人' (People) domain. The search results list several items:

- [大数据来清华，高端论坛校庆献礼](#)  
2014-04-21 20:54:04
- [清华大学102周年校庆贺辞](#)  
2013-05-05 08:27:19
- [清华电视台](#)  
2014-04-26 09:32:08  
清华电视台103周年校庆之际重播30集电视剧《水木清华》：4月26日起，每天上午9:00和晚上20:05（因校庆特别报道可能会顺延）在校内有线电视网中播出（高清频道903，标清频道912），每次连播3集。也可通过清华大学网络电视（tv.tsinghua.edu.cn）同步收看。《水木清华》用散文般的电视语言和影像艺术独有的感染力，讲述“清华”这一百年学府建校初期的传奇故事。重点演绎了唐国安、周诒春、梅贻琦等几代校长的故事，他们筚路蓝缕、披荆斩棘，团结师生、不懈努力，使清华成为一代名校，培养了无数先贤达人。
- [清华大学推理协会](#)  
2014-04-26 19:46:48  
转自清华大学推理协会：【社团嘉年华】孩纸们有木有玩得很high啊，数独简单易玩走起来，有（keng）趣（die）的推理题也玩了，小伙伴们搬到社团部高大上的奖品了嘛？60分好难啊主页嘛已然弃

On the right side, there is a '热门搜索' (Hot Search) box with the following links:

- [校庆](#)
- [选课](#)
- [校历](#)
- [本科招生章程](#)
- [PX](#)

搜索“**校庆（指定文档域）**”：

The screenshot shows the same web browser window, but the search results are filtered by the '文档' (Document) domain. The search results list several PDF documents:

- [news20110330.pdf](#)  
2014-04-21 20:54:04  
-03-22 “清华百年校庆房地产高峰论坛”会议通知 ... 知 清华百年校庆 房地产(新格局新趋势)... 从? 作为清华百年校庆系列活动之一、由清华校友总会...
- [xqikmd.pdf](#)  
2013-05-05 08:27:19  
校庆捐款名单 (2011-2013...)
- [鏊糠籽璁“ 堉铤面派2013943355.pdf](#)  
2014-04-26 09:32:08  
后，我们还参加了新百年基金校庆交流酒会。第三次活动：... 聚餐中、在清华新百年发展基金校庆交流酒会等活动中，张 老师... 且在晚上 组织所有参与观者看校庆献礼话剧《马兰花 开》，体...
- [2010qhgqich.pdf](#)  
2014-04-26 19:46:48  
为主题，清华大学从99 周年校庆开始，举办为期一年多的“百年... 庆开始，举办为期一年多的“百年校庆年”活动，包括学术、文化... 华学堂等重点工程。我们要以百年校庆为契机，弘扬 清华光荣传统...

On the right side, there is a '热门搜索' (Hot Search) box with the following links:

- [校庆](#)
- [选课](#)
- [校历](#)
- [本科招生章程](#)
- [PX](#)

## 六、实验总结

总结本次实验, 校园搜索引擎主要亮点为如下三点:

### 索引数据量丰富

- ✓ 索引 26 万文档, 索引文件 2.02G
- ✓ 来源官方网站、人人公共主页等

### 搜索效果理想

- ✓ 考虑时间因素
- ✓ GBRT 学习参数

### 友好功能多

- ✓ 垂直搜索返回
- ✓ 关键字补全, 页面去重
- ✓ 热门搜索, 相关搜索, 指定域搜索

通过本次实验, 我们对搜索引擎构建的步骤有了更清楚的认识。再次感谢老师与助教在实验中给予的指导与建议!

## 七、文件说明

文件内容如下:

文件夹	文件(夹)	说明
doc/	校园搜索引擎报告.pdf	实验报告
	slides.pptx	展示 PPT
	校园搜索实验说明-2014.pdf	实验说明
THUSearch/	run.sh	Linux 下运行脚本
	THUSearch.jar	部署文件
	frontend/	前端代码
	src/	后端代码
	index/	索引文件
	data/	包括 GBRT 训练集等