校园搜索引擎构建

马　也　2013011365 计 34
刘政宁　2013011362 计 34

June 17, 2016

1 实验要求和内容

本项目要求综合运用搜索引擎体系结构和核心算法方面的知识，基于开源资源搭建校园搜索引擎，掌握开源搜索引擎的运行流程。具体要求为：

1. 抓取清华校内绝大部分（30 万左右）网页资源及大部分在线文本资源（如 office 文档、pdf 文档等）
2. 实现基于 BM25 概率模型的内容排序算法，要求对查询进行分词；
3. 实现基于 HTML 结构的分域权重计算（content/title/h1-h6），并应用到搜索结果排序之中，并建立小规模测试集合，进行参数调节；
4. 实现基于 Page Rank 的连接结构分析功能，离线计算 Page Rank 值，并应用到搜索结果排序之中；
5. 采用便于用户信息交互的 Web 界面，实现查询扩展、查询纠错等功能。

2 实验功能

本项目在全部完成基础要求的同时，最终完成如下扩展功能，具体功能描述及效果参见实验成果一节。

- 基于 HTML 结构的分域权重计算（包含 content/title/h1-h6 等），不同域权值不同
- 实现了条件查询功能，即支持不同查询语句的 AND、OR、NOT 组合
- 实现了模糊查询功能，即支持自动纠正错误输入，并得到正确查询结果
- 实现了通配符查询（正则表达式）功能，即支持 * ？ 等~ 通配符进行查询
- 实现了查询特定范围功能，即支持查询某一网域下的相关网页
- 实现了查询特定类型功能，即支持查询某一或某几类型的资源
- 实现了查询结果界面的高亮处理，即找到最佳高亮位置，显示到摘要之中
- 实现了对特定 HTML 结构的查询，即支持仅在标题、h1 等结构中查询

3 实验框架及运行环境

3.1 框架概要

本项目主要分为五个模块：爬取模块、预处理模块、索引模块、查询解析模块以及 Web 模块。主要流程关系参见图3.1，具体阐述如下：

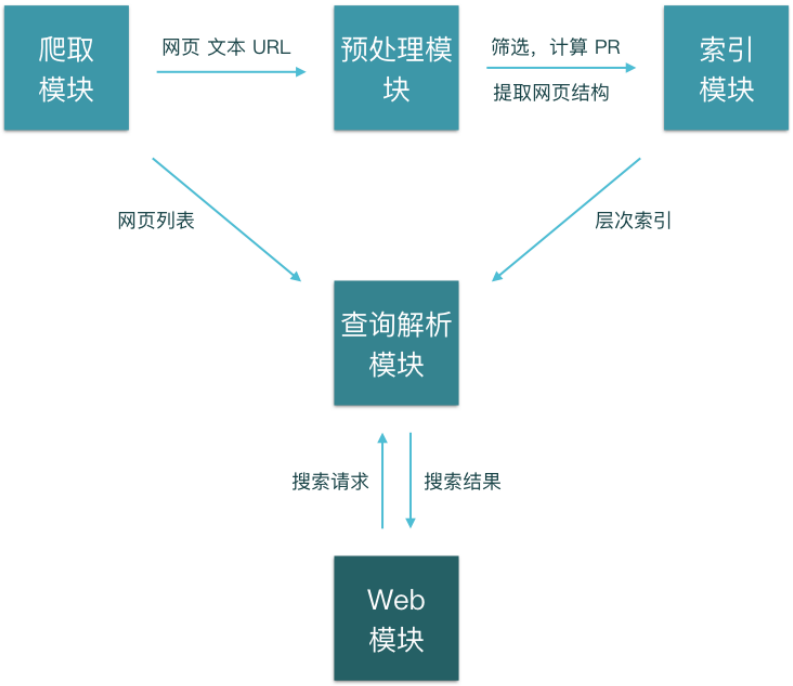

Figure 3.1: 实验框架

爬取模块从清华校内爬取网页和文本数据，并按照网页结构（url）分文件夹存放在本地，对应搜索引擎中的数据抓取子系统；

预处理模块分析和处理爬虫爬取到的数据，筛选高质量网页，提取网页 HTML 结构到纯文本文件之中，交由索引模块使用，并分析网页链接结构，计算 Page Rank 值，其对应搜索引擎中的链接分析子系统和内容索引子系统的一部分；

索引模块则利用预处理模块提取好的文本文件，使用分词器构建多层次索引，存储各层结构的文本信息到索引数据库中，以备查询模块使用，其对应搜索引擎中的内容索引子系统；

查询解析模块利用建立好的索引，根据查询条件和要求构建相应的查询语句，分拆高级查询语句为基本查询语句的组合，并在索引数据库中进行查询，返回查询到的文档列表，其对应内容检索子系统；

Web 模块代表了搜索引擎所对应的网站，其负责接收用户请求，发送给查询解析模块，收到结果后以合理的格式与结构返回给用户。

3.2 爬取模块

爬取模块使用第三方库 Heritrix 3.2.0 完成，Heritrix 可以对抓取的对象进行精确的控制，很好地符合我们校园搜索的要求。需要指出的是，最早项目使用课件上使用的版本，但其爬取速度太慢，且正则表达式过滤模块有隐含的 BUG，所以最终替换为更新的版本，因此配置和课件上的配置有所不同。主要配置了如下内容：

- 种子资源：http://news.tsinghua.edu.cn/
- 接受网页总规则：以 tsinghua.edu.cn 结尾，且不以 lib.tsinghua.edu.cn 结尾，不属于 166.111.120 网段
- 拒绝类型规则：js|JS|axd|AXD|mso|tar|txt|asx|asf|bz2|mpe?g|MPE?G|tiff?|gif 等，从略

除此之外，还设置了爬取速度，最大条数等具体配置，最终爬取了 40 万以上的文件，删除掉其中大于 32M 的较大文件，最终剩余 35 万左右。

3.3 预处理模块

预处理模块基本使用 Python 3 完成，计算 Page Rank 使用 C++ 完成，代码在 preprocess 文件夹内，脚本文件及对应功能如下：

- **parse_log.py**：分析 Heritrix 日志，建立文件名到网页 URL 的双向映射关系
- **get_id.py**：分析链接结构，筛选有效网页，给每个网页分配独立 id，并得到链接图谱
- **append_id.py**：对 get_id 的补充
- **prepare_pr.py**：将 Python 结构按照规则写入文件，为 C++ 程序准备输入
- **page_rank.cc**：计算 Page Rank，写入文件中，每行一个 id 和对应的 Page Rank
- **get_text.py**：提取网页文本内容，删除 script、css 以及 HTML 标记，每个网页存入一个文本文件中，文件名为网页 id
- **get_title.py**：提取网页标题（<title> 域），写入文件中，每行一个 id 和对应的 title
- **get_file_text.py**：提取 PDF、DOC、DOCX 等格式的全部文本，每个文件存入一个文本文件中
- **get_file_title.py**：提取 PDF 文件的标题，写入文件中，每行一个 id 和对应的 title
- **get_docs_title.py**：提取 WORD 文件的标题，写入文件中，每行一个 id 和对应的 title
- **get_h1.py**：提取网页结构，（h1-h6），分别写入文件中

需要注意的是，由于 Heritrix 在抓取带 GET 请求的网页时，存储文件的文件名和网址 URL 并不能一一对应（其去掉了问号，挪动了文件类型的位置），且单从文件名并不能找到对应的 URL，所以第一步分析 Heritrix 爬取日志是必要且是必须的。通过分析日志，得到了 URL 到文件名的双向映射，同时删除了 404 网页，将网页个数减少到 32 万左右。

以上处理中对于 HTML 网页的处理使用 Beautiful Soup 完成，其负责提取链接关系，提取文本信息，提取 title 和提取 h1-h6 域等，部分编码混乱的网页被直接丢弃，将网页个数再减到 30 万左右。

另外，由于绝大多数下载附件网页的文件名都是数字编号或无意义符号，所以提取 PDF 和 DOC 文件的标题也是十分重要的。提取 PDF 内容使用了 pdftotext 程序，提取 PDF 标题使用了 pyPdf 库，通过 PDF 文件的 Meta 信息能够得到大部分 PDF 的准确文件名，提取 DOC 文件使用了 antiword 程序，提取 DOCX 文件使用了 docx 库。

3.4 索引模块

索引模块使用 Lucene 5.5 完成（MyIndexer.java），主要存储了网页 id、网页标题、网页内容（content/h1-h6 等）、网页类型（HTML/WORD/PDF 等）、网页 Page Rank 值到索引之中，其中网页标题、内容进行了分词处理，分词时使用了效果更好的 jcseg 第三方库，可以有效地区分数字、人名、专有名词等，分词准确率更高。

索引时使用的评分模块为 BM25 模型的改版（MySimilarity.java），BM25 模型在 Lucene 5.5 官方提供的 BM25Similarity 的基础上进行了重构，加入了 Page Rank 的计算，修改了最终得分公式，将 BM25 的得分与 Page Rank 的 0.5 次方乘起来得到最终得分，最终评分公式为：

$$\text{Score} = \text{idf} \cdot \frac{(k+1) \cdot \text{freq}}{k \cdot (1 - b + b \cdot \frac{|d|}{\text{avgLen}}) + \text{freq}} \cdot \sqrt{\text{PageRank}}$$

这里使用根号的原因是减少 Page Rank 对于结果的影响，因为 Page Rank 范围变化较大（10^{-3} 到 10^{-7}），如果直接将 BM25 结果和 Page Rank 结果相乘，会导致 Page Rank 高的网页，即使只出现一次也会排名非常靠前，这就让 BM25 算法失去了意义。

3.5 查询解析模块

查询解析模块使用 Lucene 5.5 完成（MySearcher.java），主要实现了普通搜索（search()）、获得文本高亮（getHighlight()）、高级搜索（searchComplex()）、通配符和模糊搜索（searchFuzzyOrWildCard()）等功能。

需要注意的是，如果直接使用 QueryParser，其生成的 Query 语句的分词间是 SHOULD 的关系，即只要有一个出现即可，这不太符合大多数搜索的要求，如搜索“清华大学计算机系”时可能出现只含“清华大学”而不含“计算机系”的文档，所以要使用 QueryBuilder 类的 createMinShouldMatch 方法构建 Query，这样保证分词后的每个词都必须出现，而不是出现一个即可。

得到了基础的 Query 之后，需要根据要求再不同域上进行搜索（content、title、h1、type 等），并且需要使用 BoosQuery 提供不同域的不同权值。如果进行条件搜索，则需要 BooleanQuery 对其进行组合，BooleanQuery 中的 SHOULD、MUST、MUST_NOT 分别对应 OR、AND 和 NOT。

对于通配符搜索和模糊搜索，则可以使用 WildCardQuery 和 FuzzyQuery，按照类似的要求进行组合，得到最终的 Query 查询。

文本高亮也是在这一模块实现的，主要使用 Lucene 中的 Highlighter 类，对于一个特定的 Query、特定的 field 和和内容，可以找到最佳的文本段，其出现特定关键词的次数最多、得分

最高，将其返回给前端，就可以实现类似于 Baidu、Google 等搜索引擎的文本高亮了。

3.6 Web 模块

该模块使用 Tomcat 服务器完成（IndexServlet.java 和 ResultServlet.java），主要实现了 myIndex.jsp 和 myResult.jsp 两个动态页面，其负责显示搜索结果，进行分页处理等操作，并接受用户高级搜索的输入，转化为查询解析模块的函数调用。

4 实验成果与分析

4.1 主要功能与效果图

4.1.1 主界面

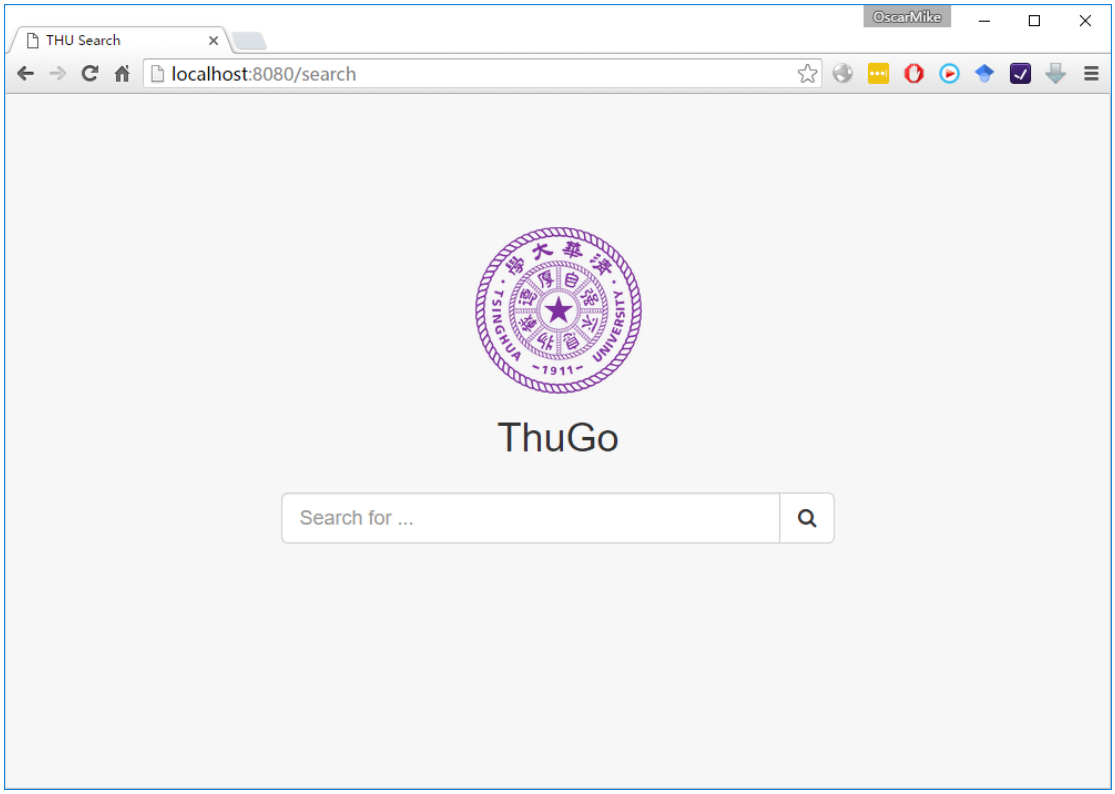

Figure 4.1: 主界面

如图4.1所示为简单大方的主界面。

4.1.2 搜索结果页面

如图4.2所示为搜索“清华大学”的结果，可以看出标题和摘要中都带有高亮处理，可以让人对网站内容有一个初步的了解。

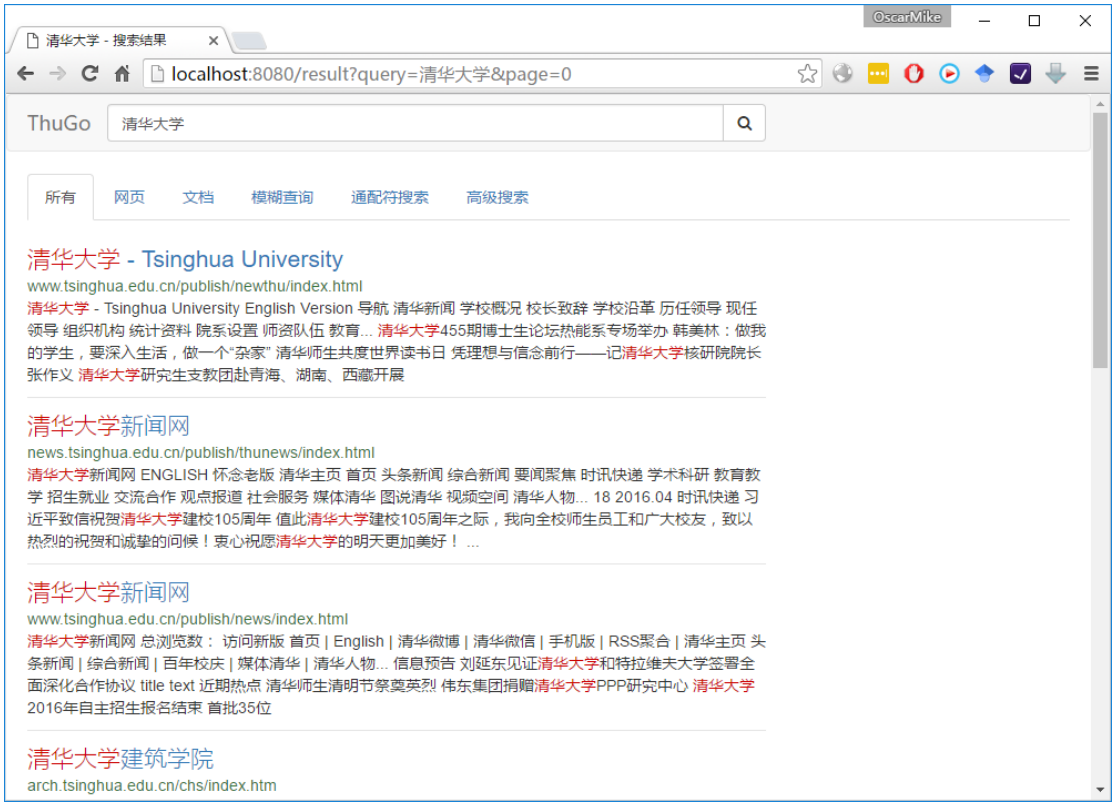

Figure 4.2: 搜索结果

Figure 4.3: 仅搜索文件功能

4.1.3 搜索文件功能

如图4.3为搜索“清华大学”，并且只要非 HTML 格式的文件，返回的搜索结果，可以看到非 HTML 的标题提取也都基本正确，并且摘要信息基本符合要求。

4.1.4 模糊搜索功能

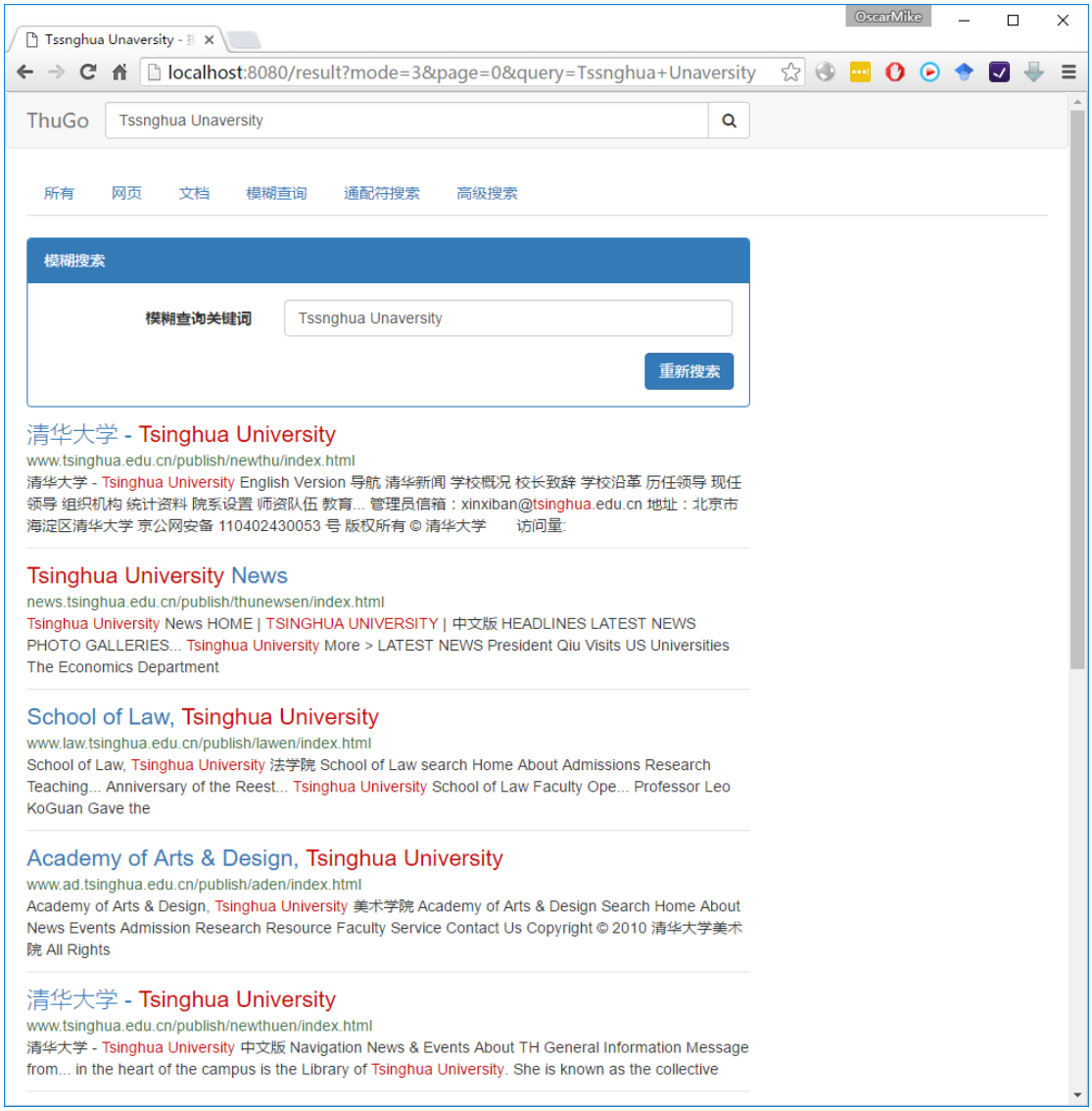

Figure 4.4: 模糊搜索功能

如图4.4为搜索 Tssnghua Unaversity 这一错误查询时返回的结果，可以看到其自动更正为 Tsinghua University，并且正确的进行了高亮。

4.1.5 通配符搜索功能

如图4.5为搜索 T?inghua Univ*ty 这一带通配符查询时返回的结果，可以看到其匹配到了 Tsinghua University，并且正确的进行了高亮。

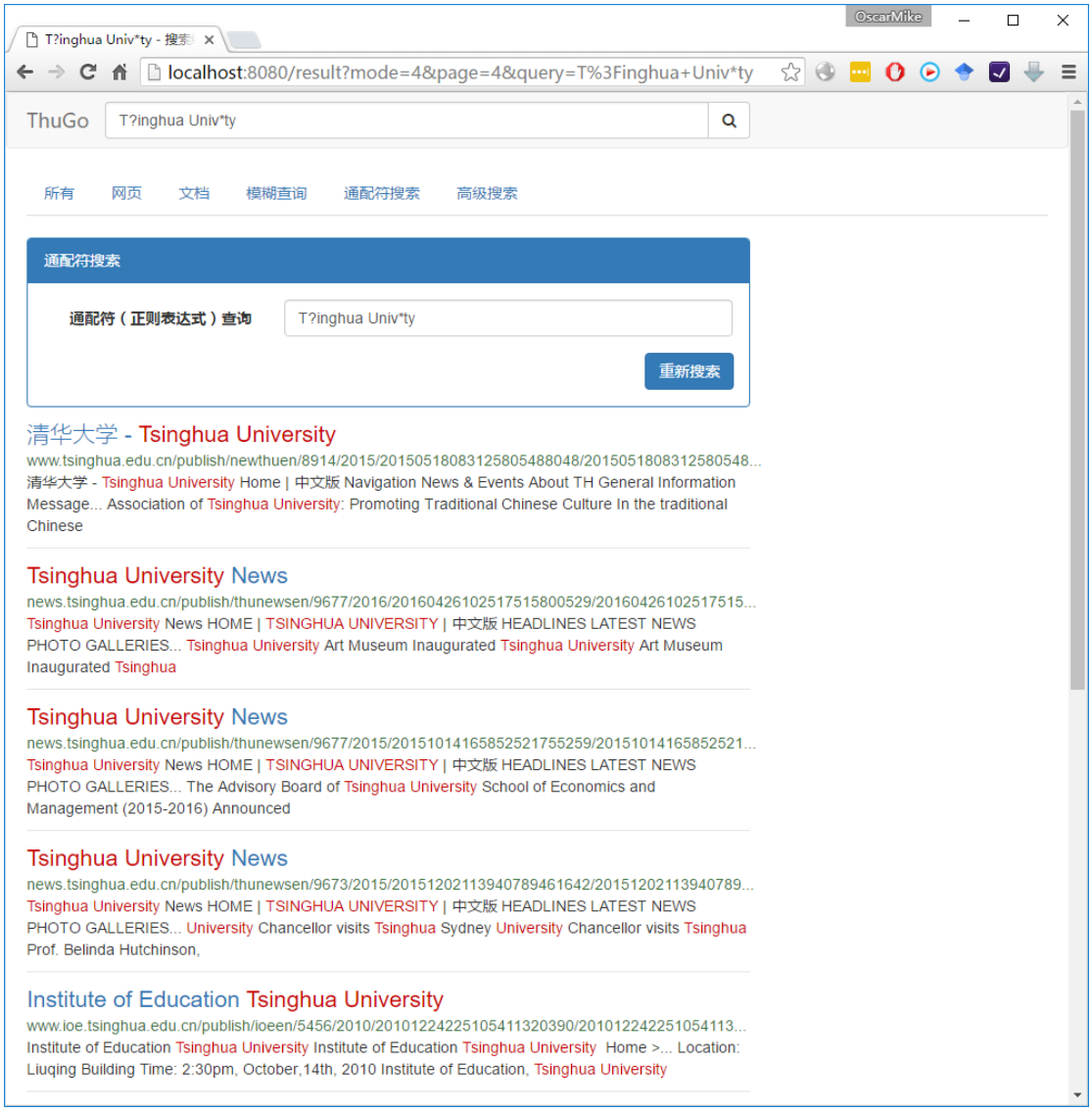

Figure 4.5: 通配符搜索功能

4.1.6 条件搜索功能

如图4.6为高级搜索的界面，其支持如下功能：

1. 包含**全部**关键词的搜索（即每个词都要出现）
2. 包含**任意**一个或多个关键词的搜索（即所有词至少出现一次）
3. **不包含**任何一个关键词的搜索（即所有词都不能出现）
4. 限定搜索所在的**网域**，如只搜索 www.cs.tsinghua.edu.cn 网域下的所有网页
5. 限定搜索文档的**格式**：PDF、DOC、DOCX、HTML 等
6. 限定关键词在文档中出现的**位置**：全部、仅出现在标题、仅出现在 h1 等
7. 以上 6 点搜索的任意组合

Figure 4.6: 高级搜索功能

4.2 实验结果分析

4.2.1 入链接、出链接分布

首先，对抓取到的网页进行入链接、出链接统计，可以得到它们的分布情况，如下图所示：

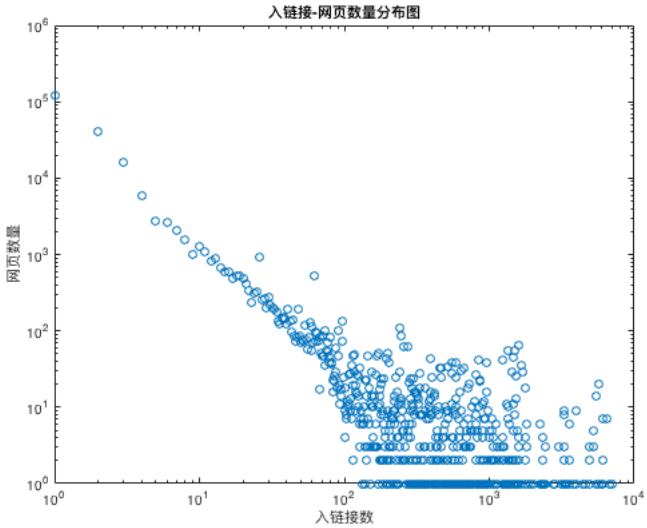

Figure 4.7: 入链接分布情况

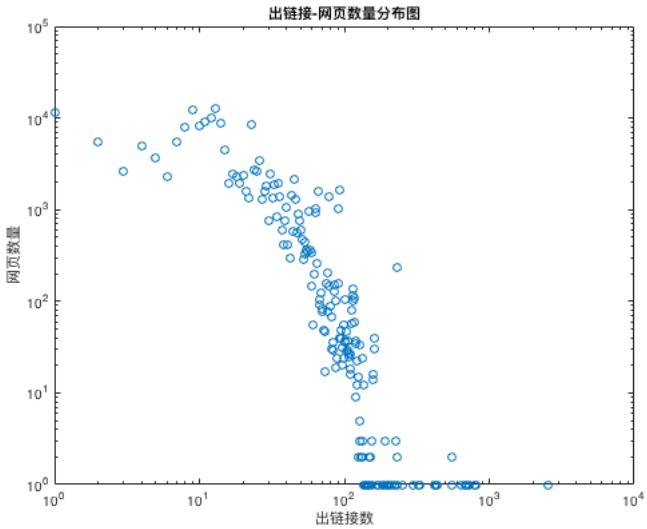

Figure 4.8: 出链接分布情况

从中我们可以看出，对于入链接个数分布情况来说，较好的满足了幂律，即入度与出现次数满足指数函数关系；而对于出链接个数分布来看，其在低频区不能很好的满足幂率，但是依然有着相似的分布。

4.2.2 PageRank 算法结果

接着，我们计算爬取所有网页的 Page Rank，并把前 20 名列举如下：

内容	入度	Page Rank
环境资源能源法学院 - 评论栏	2007	0.004005
清华启航网 - 岗位列表	6518	0.003324
清华启航网 - 基地列表	6518	0.003324
清华启航网 - 首页	6518	0.003298
环境资源能源法学院 - 评论栏	2008	0.003085
清华启航网 - 联系方式	6518	0.002955
清华启航网 - 企业服务	6518	0.002955
清华启航网 - 平台介绍	6518	0.002955
清华启航网 - 学生实践反馈	6518	0.002955
环境资源能源法学院 - 评论栏	3	0.001818
清华大学研究生院 - 首页	2744	0.001672
清华大学深研院院报 - 首页	3144	0.001305
清华大学建筑学院 - 首页	529	0.001280
清华大学新闻网 - 首页	7034	0.001261
清华大学美术学院 - 首页	4402	0.001208
清华大学新闻网 - 英文版首页	6765	0.001179
清华大学美术学院 - 英文版首页	4405	0.001131
清华大学新闻网 - 旧版首页	6752	0.001085
实验室管理系统 - 首页	3253	0.001077
清华大学新闻网 - 首页（url 重定向）	1758	0.001058

Table 4.1: 前 20 名 Page Rank 结果

从结果中，我们发现，尽管像清华新闻网首页和一些院系首页进入了前 20 名，但是排名靠前的却都是一些内容极其差的网页，例如环境资源能源法学院评论栏中都是些奇怪的文本，且评论栏有上千页写死到网站 HTML 之中，所以入度较高；又如清华启航网，他的网页个数极多，且都是测试使用的网站。由此可见，这些垃圾网页都是靠着极多的互相链接关系，形成一个很大的闭环，并不代表网站质量很高。基于以上考虑，最终部署时，已将这些垃圾网页去除。

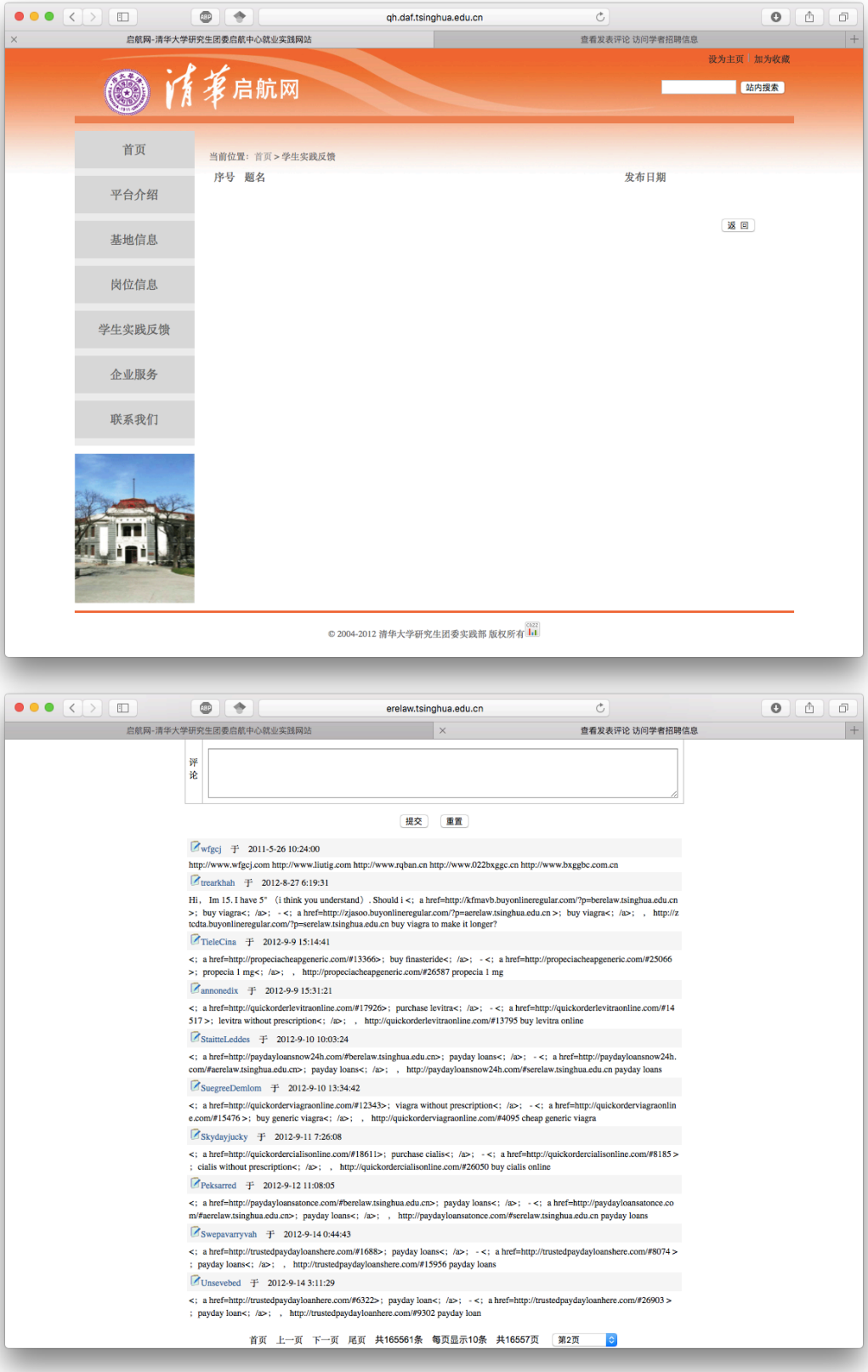

Figure 4.9: 部分垃圾网页截图

4.2.3 PageRank 结果分布情况

接着，我门分析了 Page Rank 结果的分布情况，绘制如下图表：

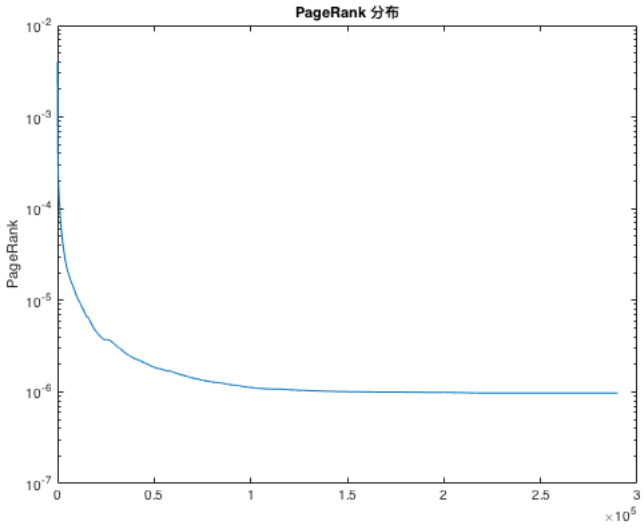

Figure 4.10: PageRank 结果分布情况

其中横轴为倒序排布后网站编号，纵轴为 Page Rank 值的以 10 为底的对数值。从图中我们可以发现，Page Rank 值的衰减是非常非常快的。

4.2.4 PageRank 与入链接数的关联分析

如果画出 Page Rank 和入链接数的散点图，我们可以得到如下结果：

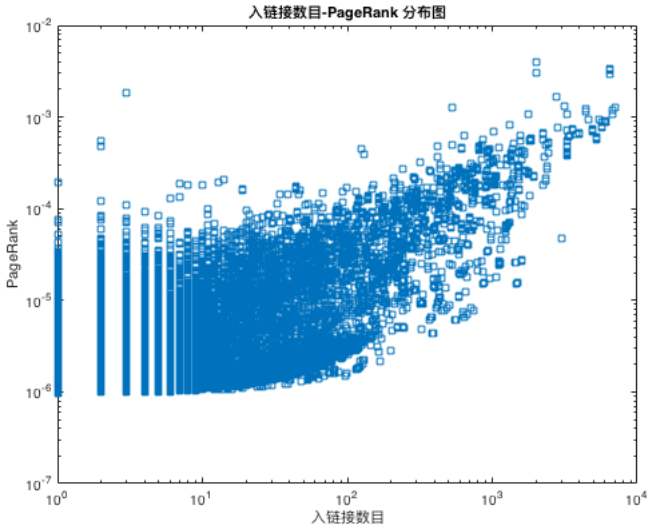

Figure 4.11: PageRank 与入链接数的关联分析

其中横轴是入链接数以 10 为底的对数坐标，纵轴是 Page Rank 以 10 为底的对数坐标。从图中我们可以发现，尽管二者并不是严格的正相关，但是随着入度的增加，Page Rank 是有增

加的趋势的。

4.2.5 搜索结果得分

在调试过程中，我们经常需要知道一个页面排名靠前的原因，因此得分的计算过程就显得尤为重要了。在我们实现的 MySimilarity 中，就提供了 explain 的接口可以解释得分的来龙去脉，以搜索“相声”为例，排名前 3 的网页如下所示：

Figure 4.12: “相声”搜索结果

接下来我们分别导出前三名网页的得分及其来源

- 0.09240038 = sum of:
  - 0.07716647 = weight(titleField:相声 in 620) [MySimilarity], result of:
    - 0.07716647 = score(doc=620,freq=1.0 = termFreq=1.0), product of:
      - 6.0 = boost
      - 11.227245 = idf(docFreq=3, maxDocs=263025)
      - 0.66136855 = tfNorm, computed from:
        - 1.0 = termFreq=1.0
        - 1.2 = parameter k1
        - 0.75 = parameter b
        - 7.1060734 = avgFieldLength
        - 16.0 = fieldLength
      - 3.0E-6 = pageRank
  - 0.015233911 = weight(contentField:相声 in 620) [MySimilarity], result of:
    - 0.015233911 = score(doc=620,freq=16.0 = termFreq=16.0), product of:
      - 6.54909 = idf(docFreq=376, maxDocs=263025)
      - 1.3429809 = tfNorm, computed from:
        - 16.0 = termFreq=16.0
        - 1.2 = parameter k1
        - 0.75 = parameter b
        - 1487.8988 = avgFieldLength
        - 16384.0 = fieldLength
      - 3.0E-6 = pageRank

Figure 4.13: 排名第一的得分情况

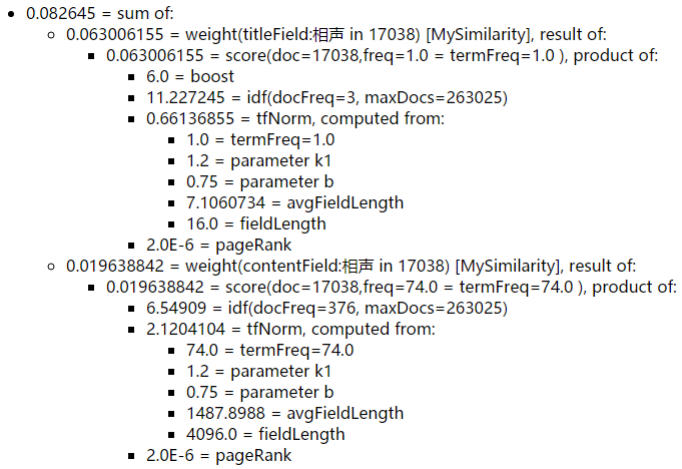

Figure 4.14: 排名第二的得分情况

Figure 4.15: 排名第三的得分情况

从得分中我们可以看出，排名第一、第二的网站之所以得分较高，是因为其在标题（titleField）中出现了关键词，且 titleField 设置的权值为 6，因此总得分较高；而由于第三名标题中没有出现关键词，尽管其正文中出现次数很多，正文得分高于前两名的正文得分，但总分却没有前二者高。当然，由于第三名网页 Page Rank 值较高（是前两名的近百倍），所以排名也比较靠前。

5 实验总结

通过这次实验，我们掌握了搜索引擎的基本架构和重要实现，通过自己完成搜索引擎四个子系统，我们对每个子系统的功能和架构有了更深的理解。在实验过程中，我们曾遇到爬取速度太慢、UTF-8 编码问题、得分公式不够优、搜索结果不好等问题，但最终都在同学的启发下得以克服。总之，我们在这次实验中收获颇丰，更好地理解了课堂上所讲的原理知识。