

搜索引擎大实验

《校园搜索》设计文档

计 23 黄必胜 2012011307

计 23 鲁逸沁 2012011314

计 23 谢晓晖 2012011315

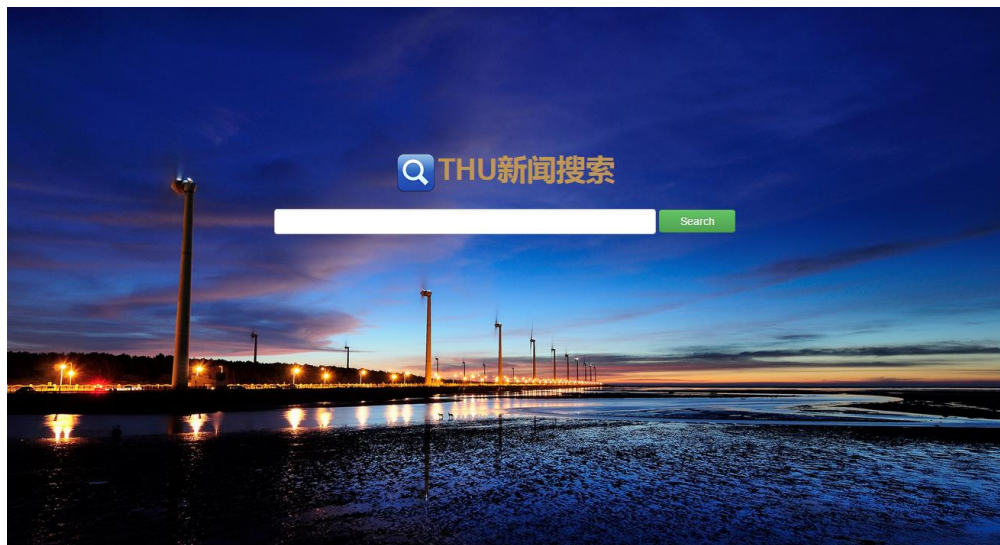
一、 项目介绍

1. 项目目标

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，用于 <http://news.tsinghua.edu.cn> 的网页内容搜索。

- 使用 Heritrix 抓取“清华大学新闻网”站点的所有内容
- 实现基于概率模型的内容排序算法；
- 实现基于 HTML 结构的分域权重计算,并应用到搜索结果排序中；
- 实现基于 Page Rank 的链接结构分析功能,并应用到搜索结果排序中；
- 采用便于用户信息交互的 Web 界面：
 - 文本补全；
 - 相关查询词推荐；
 - 查询词纠错；
 - SERP 中的每一条返回结果基本带有配图；
- 设计简洁美观的网站界面，争强用户的搜索体验。

2. 项目大致效果



从图中可以看到，我们的网站界面简洁美观，只用一个搜索框即可完成所有搜索任务。下方列的搜索结果，图文并茂，直观清晰，用户可以更快获得需要的信息。同时我们具有自动补全功能，在输入框中显示下拉菜单，方便和加速用户查找。此外还有纠错功能，在搜索框下方给出其他可能的关键词，帮助用户进行模糊查找。特别的，我们在右边栏中提示了相关词汇，帮助用户进行信息的拓展和知识的拓展，为搜索提供更多可能性。

二、 框架构建

1. Lucene 框架的使用

本实验中我们使用了以下开源工具，

- 数据抓取工具 : Heritrix 3.1
- 搜索引擎框架 : Lucene 4.0
- 中文分词软件 : IKAnalyzer V2012
- Html 解析工具 : Jsoup
- PDF 解析工具 : pdfbox 1.8.9
- M.S Office 解析 : poi 3.1.2
- 服务器端 : Tomcat 8.0.22
- 前端编写 : JSP

Lucene 是一个优秀成熟的、简单易用，用于全文检索和索引的 Java 开源工具包，在最近的几年广受欢迎。Lucene 的最新版本 5.2.0 支持很多新特性，例如内置了更复杂的概率模型和提供了更方便的评分定制接口等。但是由于 Lucene 的最新版本尚缺乏与之兼容的中文分词软件，最终我们在实验中选择了当前分词软件所能支持的较高版本 4.0.0。

文档评分机制是搜索引擎中极为重要的一部分，决定了各个文档在页面中的出现顺序。本实验中采用了结合了 html 分域权重和 Page Rank 的 BM25 概率模型。我们首先离线计算出各个页面的 Page Rank 值，作为文档 i 的一个权重增益 boostDocument_i ；同时将文档的标签进行分类，根据类别划分为多个域，通过小规模的数据集测试定下各个域的权重系数，将对应的系数作为域 j 的权重增益 boostField_j ，然后综合二者得出各个域的权重为 $\text{boostDocument}_i * \text{boostField}_j$ ，然后在索引时对其赋值；最终在索引时，对每个文档的各个域进行 BM25 概率模型的评分，得出文档的分数。

从 Lucene 4.0.0 版本开始，BM25 概率模型就已经内置于该工具包中，采用该模型只需要在索引时指定 Similarity 为 BM25 Similarity 即可。实验中模型参数采用常用的默认值， $k = 1.2$ ， $b = 0.75$ 。以下将分别介绍分

域权重和 Page Rank 的具体细节。

2. 分域权重

数据抓取的种子网站为清华新闻网 [new.tsinghua.edu.cn](http://news.tsinghua.edu.cn)，所以网页的内容以新闻内容为主。对数据进行抓取之后，我们首先观察了清华新闻网的 html 标签特点，发现其 html 标签及内容的对应大致如下：

- `<title>`：对应新闻标题内容。该信息是作为区分文档定位内容的最重要信息。但是对于英文网站的区分度不高，因为英文版网站的 title 都统一为“News Of Tsinghua University”。
- `<h1>-<h6>`：大多数情况下对应新闻的标题，但是在有些情况下能包含 title 没有的内容，例如当新闻含有多个小标题或者为英文新闻时。该信息的重要性仅次于 title。
- `<p>/<P>`：对应新闻正文内容，也是该页面较重要的信息。
- `<a>`：对应该页面所指向的其他新闻的标题。对于定位当前页面的价值应该较小。

其余的标签如`<keyword>`，``，``等出现频率较小，利用价值不高。综合以上信息我们为每一个文件索引对应建立了以下域(Field)：

- path :文档的 url 信息，是文档在数据集中的相对目录位置
- title : `<title>`内容
- keyword : `<h1>-<h6>`内容
- content : `<p>/<P>`内容
- link : `<a>`内容

而对于 pdf 文件和 MS.Office 文件，由于难以提取出类似于 html 这样有层次类别的内容组织形式，简单处理直接将文档名称置为 title，文档内容置为 content。

在分域权重的分配上，我们将 title、keyword、content、link 的权重增益分别设置为 100，10，5，1。该分配经过了小数据集上的测试，能将 title 含有检索词汇的页面排在更前边，但是也不至于导致 title 部分相关的文档得分高于 content 完全相关的文档得分。

3. Page Rank

Page Rank 的计算公式为

$$\text{PageRank}^{(k)}(n) = \alpha \cdot \frac{1}{N} + (1 - \alpha) \cdot \sum_{i \rightarrow n} \frac{\text{PageRank}^{(k-1)}(i)}{\text{Outdegree}(i)}$$

其中， N 为网页总数， $\text{Outdegree}(i)$ 为网页 i 链接到的网页个数。对于没有出链接的网页，我们将其的出链接设为所有网页。

本次实验中，我们使用的参数 $\alpha = 0.15$ ，迭代次数为 30 次，获得有效网页的 Page Rank 值共 54078 个。

获得网页的 Page Rank 值后，我们将其与 Lucene 中对某个查询词的默认网页打分相乘，获得最终的评分值，并将其排序显示在最终的搜索页面中。

另外，由于清华新闻网的特殊性和局限性，在用程序求出 Page Rank 后，我们发现其中有一些网页的 Page Rank 值有较大的缺陷。例如网页 <http://news.tsinghua.edu.cn/publish/news/mobile/4204/index.html>，其 Page Rank 值非常高的原因是因为入度特别大。深究其原因，则是因为它是移动清华新闻网的固定网页头上的链接，即所有移动清华新闻网的子网页上都有这个链接。但是这个页面显然评分不应该是非常高的，因为其内容仅仅是新闻的索引，并不能比真正新闻内容的 Page Rank 值高上几个数量级。因此我们手工将排名靠前的网页一一检查过滤，将符合上述描述的网页的 Page Rank 调低，以优化搜索结果。

我们还发现，一些差不多的新闻网页，其 Page Rank 值相差近 100 倍（一个是 10^{-4} 级别一个是 10^{-6} 级别），这会导致 Lucene 中其他内置的评分都会被忽略不计。针对这一特点，我们对 Page Rank 值进行了优化使得其不至于主宰页面的顺序，新的 Page Rank 值为

$$\text{newPageRank} = 16 + \ln(\text{PageRank})$$

这样，新的打分指标 newPageRank 将会至多产生 10 倍左右的区分度，从而不影响其他打分标准的效果。

三、 特色功能

1. 相关词推荐

相关词推荐，我们想做成类似百度搜索右侧的那部分，为用户提供与其该次搜索相关的关键词，以便用户进行迭代式搜索找到理想网页，或获取其他意想不到的信息。

首先定义相关性。我们使用的方法比较简单：如果两个关键词在同一个网页中出现，则认为它们是相关的。

假设查询词为 q ，定义关键词 p 和查询词 q 同时出现的文档个数为 $co(q, p)$ 。如果直接使用 $co(p, q)$ 作为相关度的度量，那么会出现所有查询词的相关词都是“的”、“是”等没有信息量的词，因为他们几乎出现在所有的网页中。

这时候可以借鉴一下 TF/IDF 模型，将 $co(q, p)$ 作为 TF 的值，通过求 TF/IDF 的值来平衡常用词带来的干扰。但是这时又会出现一个新的问题，如果一个很生僻的词恰好和查询词在某一个网页中共同存在，那么它的 TF/IDF 值就是 1。然而根据定义，TF/IDF 的值不会超过 1，因此这个生僻词就会成为最佳答案，然而其实它和查询词并不相关。

因此我们对 TF/IDF 公式做了一个修正，最终关键词 p 相对于查询词 q 来讲其相关度定义为

$$r(p, q) = \frac{co(p, q)}{\sqrt{idf(p)}}$$

这个公式中，即修正了常用词频繁出现的问题，也解决了生僻词被选为最佳答案的问题，总体效果不错。

下面是我们搜索引擎的一些效果。

THU新闻搜索

1. 转环保微博送口罩清华学生关注环保公益 - 清华大学新闻网



转环保微博送口罩清华学生关注环保公益 清华新闻网3月29日电 3月26日，由清华大学研究生会主办，“导航犬”软件团队赞助的“环保青春，任我导航”大型公益活动活动在清华紫荆园、桃李园、观畴园食堂门口

2. 原国家环保局局长曲格平与清华学子面对面谈环保 - 清华大学新闻网

【新闻中心讯 研通社通讯员 吕淼】10月31日下午，原国家环保局局长、全国人大环资委主任委员曲格平先生作客清华环境论坛，在中意清华环境节能报告厅向现场师生深入浅出地讲述了改革开放三十年来的中国

相关词汇：

污染
节能
环保部
环境保护
环保局
环境
排放
能源

THU新闻搜索

你要查找的是不是：[二氧化](#) [二氧化](#)

1. 胡鞍钢：中国有望2030年前达到二氧化碳排放峰值 - 清华大学新闻网

胡鞍钢：中国有望2030年前达到二氧化碳排放峰值 来源：中国新闻网 2015-3-3 马德林 清华大学国情研究院院长胡鞍钢3日在北京表示，中国有可能在2030年以前达到二氧化碳排放峰值，提前完成

2. 胡鞍钢：中国有望2030年前达到二氧化碳排放峰值 - 清华大学新闻网

胡鞍钢：中国有望2030年前达到二氧化碳排放峰值 来源：中国新闻网 2015-3-3 马德林 清华大学国情研究院院长胡鞍钢3日在北京表示，中国有可能在2030年以前达到二氧化碳排放峰值，提前完成

3. 郝吉明：中国亟需制定细颗粒物控制策略 - 清华大学新闻网

相关词汇：

氧化硫
化硫
二氧化
二氧
氮氧化物
燃煤
氧化
酸雨

THU新闻搜索

你要查找的是不是：[张学良](#)

1. 第十六届清华校园歌手大赛圆满落幕 - 清华大学新闻网

最后，来自法学院的齐汇摘取了“最佳男歌手奖”，他以浑厚的嗓音完美诠释了张学友的《回头太难》。来自经管学院的苏小博凭借意大利歌剧《阿玛利丽》获得“最佳女歌手奖”，她扎实的美声唱法功底令到场评委无不为之折服

2. 温家宝诗作成北航校歌 - 清华大学新闻网

歌手张学友与北京奥运会开幕式《歌唱祖国》的原声杨沛宜带领下，500名小朋友在“香港同胞庆祝中华人民共和国成立60周年文艺晚会”中齐声高歌。（关庆丰）链接 仰望星空 我仰望星空，它是那样寥

相关词汇：

苏芮
周华健
bagle
李宗盛
庾澄庆
glee
抢下
焦愁

从效果中我们可以看到，搜索“环保”的时候，我们能搜出“污染”、“节能”、“环境保护”、“排放”、“能源”等与环境保护极其相关的内容；搜索“二氧化硫”，我们能搜出“燃煤”、“酸雨”等与二氧化硫危害有密切关系的内容；搜索“张学友”，我们能搜出诸如“苏芮”、“周华健”、“李宗盛”、“庾澄庆”等歌手。由此可见，在清华新闻的数据集上，我们的相关词推荐算法在一些词上具有很好的性能，为用户提供了更大的搜索空间。

2. 文本补全

文本补全的原理是寻找到与当前查询词有相同前缀的查询词汇。在具体实现时我们利用了 Lucene 中 suggest 模块的部件 FSTCompletion。具体的做法是在建立好索引之后，将 content 域中的每个 token（即文本经过分词后得到的词汇）取出，并将每个查询词的重要性置为其在各个文档中的出现频度。然后交由 FSTCompletion 模块进行处理。该模块的原理是对输入的词汇构造有限状态自动机。在进行查询时，在查询词的字自动机中寻找终态，每找到一个终态就把当前路径对应的词汇添加进备选词汇列表中。在获得足够的备选词汇后，按照其重要性及其字母序进行排序，然后按顺序高低将指定数目词汇返回作为补全结果。实践中也发现这样的算法具有不错的效果。

从效果图中可以看到，无论中文、英文，一个字符、多个字符，我们的搜索引擎都能为用户提供，以用户键入的序列为前缀的，频度较高的词。这种设计使得用户在输入的时候不必输全所有的内容，就能获得搜索的关键词，简化了用户的搜索步骤，提高了搜索的效率。

下面是我们搜索引擎的一些效果。





我们可以看到，搜索引擎会实时更新用户键入的词作为前缀的，频率较高的搜索词。这里我们使用了 Bootstrap 中的自动补全插件来显示，数据来自于我们算好的候选词。可见这些文本自动补全的功能方便了用户输入，增强了搜索的效率和体验。

3. 查询词纠错

查询词纠错，我们想做的就是类似大多数搜索引擎在输入栏下面一行的效果，当用户打查询词时打错一部分，仍能用数据帮用户找到其想要的查询词。

我们使用的方法是基于 Q-Gram 的倒排列表的方法。

首先定义编辑距离。假设两个字符串能通过 k 次修改、添加、删除一个字符的操作互相转化，则称这两个串的编辑距离为 k 。

求串 a 和串 b 的编辑距离可以使用动态规划的方法。用 $f[i, j]$ 表示串 a 的前 i 个字符和串 b 的前 j 个字符之间的编辑距离。那么：

- a 进行添加操作 (b 进行删除操作)，转移到 $f[i+1, j]$ ；
- a 进行删除操作 (b 进行添加操作)，转移到 $f[i, j+1]$ ；
- a 进行修改操作 (或不操作)，转移到 $f[i+1, j+1]$ 。

则最后 $f[|a|, |b|]$ 就是串 a 和串 b 的编辑距离。


但是如果每次对于一个查询词，那所有串与其去匹配，效率太低了。但是考虑到两个串如果相似，那么必定存在很多字符能互相匹配上。因此考虑取出串 a 的所有连续长度为 Q 的子串 Q-Gram，那么如果串 b 和串 a 相似，串 b 一定也包含很多串 a 的 Q-Gram。

基于这个思路，我们将所有串的 Q-Gram 建立一个倒排列表。当一个查询词 q 到来时，拿串 q 的 Q-Gram 去倒排列表里搜索，取出所有和串 q

拥有至少 T 个公共 Q-Gram 的串，那么这些串才是很有可能与串 q 相似的。接着对这些串逐一去使用动态规划求出编辑距离，再将所有编辑距离小于 ED 的词按 idf 值排序，就能得到用户最有可能输错的词是哪几个。


实际操作过程中，由于长的词需要纠错的字符个数大，短的词需要纠错的字符个数小，因此 ED 值会设为查询词长度*0.2 这样动态的值，以保证纠错的合理性。

下面是我们搜索引擎的一些效果。

 THU新闻搜索


>

你要查找的是不是：[tsinghua](#) [tsinghua](#) [tsinghuax](#)

 THU新闻搜索

>

你要查找的是不是：[administration](#) [administration](#) [dministration](#)

 THU新闻搜索

>

你要查找的是不是：[一丝不苟](#) [一丝不挂](#)

从效果上来看，我们的搜索引擎能为用户揪出用户打错的地方，为其推荐正确的关键词。无论用户是多打、少打、漏打，基本上只要用户打错的比例占总的 20% 左右或以下，我们的搜索引擎都有能力为其提供正确的关键词。这种设计提高了搜索的鲁棒性，用户可以进行模糊搜索，从而不会因为打错或记错的原因错过应该搜索的信息，搜索的良好体验也会大大增加。

4. 网页图片显示

搜索引擎在搜索过程中，会将一些网页的图片放在搜索结果旁边，以增加用户更直观的感受，方便用户更快速的筛选出自己想要的内容。我们正想添加这一特性。

基于这个想法，我们想在有搜索结果中，尽可能的加上网页的图片信息。那么问题就变成了，如何在一个网页的众多图片中，挑选出最具有代表性的网页。



例如上面的网页，选取开头“清华大学新闻网”这张图片肯定毫无信息量，需要选取的是文章内容中的这张合影。

和从文章中提取关键词的方法一样，从网页中提取特征图片也可以使用 TF/IDF 的方法。将图片的路径作为哈希值统计其在一个文档中出现的次数 TF 和出现过的文档总数 IDF，这样一个网页就能取出 TF/IDF 最大的网页进行显示。

当然，有些网页中只有类似上面“清华大学新闻网”这种图片，因此需要设一个 TF/IDF 阈值把没有信息量的图片滤去。最终，相当多的网页会仍然找不到对应的图片，但是一些找到图片的网页，看起来的效果都还可以。

下面是我们搜索引擎的一些效果。

THU新闻搜索

奶茶

Search

1. 【迎新花絮】“奶茶MM”清华报到 - 清华大学新闻网



清华新闻网8月17日电（记者 高 原）清华大学综合体育馆前，人声鼎沸。许多等待办理入学手续的新同学们正在有序排队，一个美丽宁静、身着紫裙连衣裙的女生吸引了众人的目光。她就是因为一张手捧奶茶的

2. 【迎新花絮】“奶茶MM”清华报到 - 清华大学新闻网



清华新闻网8月17日电（记者 高 原）清华大学综合体育馆前，人声鼎沸。许多等待办理入学手续的新同学们正在有序排队，一个美丽宁静、身着紫裙连衣裙的女生吸引了众人的目光。她就是因为一张手捧奶茶的

3. 清华出特刊“招”人气 - 清华大学新闻网

昨天新鲜出炉的清华大学《清新时代》的招生特刊封面，五个在校欢迎的手势让人心向往之。奶茶妹妹童瑶天、参加《非你莫属》的博士李一舟、录制了《天天向上》的校园歌手马瑞男，一个个清华的网络红人

4. 缤纷灿烂男生节 - 清华大学新闻网



“游园会活动在紫荆学生综合服务楼前开幕，本次嘉年华活动包括精心设计的游戏环节、丰富精美的礼品、现煮的暖暖的奶茶，趣味游戏看似简单其实蕴含着各种能力的考验，丰厚的礼品更是增加了大家的参与热情，暖暖的奶茶香气

5. 清华出特刊“招”人气 - 清华大学新闻网

昨天新鲜出炉的清华大学《清新时代》的招生特刊封面，五个在校欢迎的手势让人心向往之。

相关词汇：

护膝
张震岳
武汉站
真谈
声嘶力竭
睡
网络红人
搅客

THU新闻搜索

搜狗

Search

1. 计算机系召开“清华—搜狗搜索技术联合实验室”创新精神研讨会 - 清华大学新闻网



计算机系召开“清华—搜狗搜索技术联合实验室”创新精神研讨会 清华新闻网5月14日电（通讯员 蔡英明）5月12日下午，“清华—搜狗搜索技术联合实验室”创新精神研讨会在东主楼召开。北京市经信委科技

2. 搜狗CEO王小川获互联网行业“2013年度风云人物”大奖 - 清华大学新闻网



搜狗CEO王小川获互联网行业“2013年度风云人物”大奖 来源：techweb 2013-12-20 12月20日消息，国内领先的TMT社交媒体DoNews.com举办的2013DoNews

3. 清华大学新闻网



搜狗CEO王小川获互联网行业“2013年度风云人物”大奖 来源：techweb 2013-12-20 12月20日消息，国内领先的TMT社交媒体DoNews.com举办的2013DoNews

4. 清华大学新闻网

：成功创建有道并成为网易最年轻高级副总裁的周枫；从输入法开始进军互联网的搜狗首席执行官王小川；创办了点点、啪啪、乌鸦等互联网公司的许朝军；学习谷歌广告模式创建了浪淘金的周杰；联合创力团购导航网站团800

5. 迈向创新大国从何处起步——一个两栖实验室的创新解读 - 清华大学新闻网

不断出现，使中国得以迈向创新大国。“清华 - 搜狗搜索技术联合实验室”管委会主任、搜狗公司CEO王小川近日在联合实验室成果汇报会上说。 搜狗与清华大学计算机系2007年共建了搜索技术实验室，开展网络信息

6. 【校庆报道】揭秘中国第一代互联网人：一个不得不提的群体 - 清华大学新闻网

：成功创建有道并成为网易最年轻高级副总裁的周枫；从输入法开始进军互联网的搜狗首席执行官王小川；创办了点点、啪啪、乌鸦等互联网公司的许朝军；学习谷歌广告模式创建了浪淘金的周杰；联合创力团购导航网站团800

相关词汇：

infoseek
tumblr
敢点
几无二致
aws
电话本
等星级
利益输送

你要查找的是不是：[教师](#)、[教师应](#)

相关词汇：

节
节前
优秀教师
人民教师
教师节
教育工作者
为人师表

1. [刘延东教师节看望清华大学教师 - 清华大学新闻网](#)



刘延东教师节看望清华大学教师 清华新闻网9月11日电（记者 刘蔚如）9月10日，第30个教师节的下午，中共中央政治局委员、国务院副总理刘延东来到清华大学，亲切看望慰问清华大学教师。校党委书记陈

2. [清华大学举行教师节专场音乐会慰问教师 - 清华大学新闻网](#)



清华大学举行教师节专场音乐会慰问教师 清华新闻网9月10日电（记者 王冰冰）9月7日晚，在我国第28个教师节来临之际，清华大学在新清华学堂举行“书香之夜”清华大学教师节专场音乐会——中央民族

3. [清华大学教师节表彰教师先进集体和个人 - 清华大学新闻网](#)

【新闻中心讯 记者 周襄楠 摄影 郭海军】9月10日，在第23个教师节来临之际，清华大学在大礼堂举行了庆祝2007年教师节大会，170余个先进集体和个人在大会上获得表彰。大会由校党委常务副

4. [清华大学教师节表彰教师先进集体和个人 - 清华大学新闻网](#)

【新闻中心讯 记者 周襄楠 摄影 郭海军】9月10日，在第23个教师节来临之际，清华大学在大礼堂举行了庆祝2007年教师节大会，170余个先进集体和个人在大会上获得表彰。大会由校党委常务副

5. [顾秉林校长等教师节看望一线教师 - 清华大学新闻网](#)

□□【新闻中心讯 记者 魏磊 张宪昀】在我国第21个教师节来临之际，校长顾秉林等先后到院系和教师家中看望一线教师。9月9日下午，顾秉林校长来到电机系看望了正在工作的孙宏斌副教授，对他获得

6. [顾秉林校长与清华附中教师共度教师节 - 清华大学新闻网](#)

□□【新闻中心讯 记者 曹朋伦 通讯员 高峰 摄影 崔航】9月10日下午，清华附中全体教职员丁欢聚一

从效果上看，我们的搜索引擎能为部分页面提供图片信息，但是大部分页面还是没有图片支持的。从“奶茶”、“搜狗”、“教师节”这三个词的查询效果来看，图片还是和具体网页内容紧密相关的，也非常能体现网页的主要信息，为用户提供更快捷更直观的了解网页信息的渠道。目前设的阈值可能比较高，使得带图片提示的网页较少，随着阈值的下降，带图片提示的网页会增多，但是可能也有更大概率会出现图片与内容无关的现象，具体效果需要进一步尝试。

四、 实验总结

本次实验让我们熟悉了多种方便开发且功能强大的开源软件，例如用于数据抓取的工具 Heritrix，搜索引擎整体框架的搭建 Lucene，HTML 解析工具 Jsoup，PDF 解析工具 pdfbox，M.S Office 解析工具 poi 等等。当然在使用这些工具中我们也遇到了一些配置上的小问题，例如 heritrix 参数的选择等，在此要感谢助教对于部分开源工具的详细讲解与指导，节省了我们很多的时间。

巩固了 BM25 概率模型, Page Rank 算法等相关知识, 并学会将这些理论相结合应用于实际的工程中, 并在实际工程的开发中, 结合抓取数据的独特性, 对原始算法内容进行改进以更好地适应数据的情况。除此之外, 我们在完成了基础功能之后, 对数据进行了全面的分析, 结合之前课程所学知识, 独立设计了包括相关词推荐、文本补全、查询词纠错在内的一系列扩展功能。例如使用了 TF/IDF 模型实现了相关词推荐, 并结合数据库的独特性, 不断对原始设计的 TF/IDF 公式进行尝试和调整, 最后获得了不错的效果。在算法的设计、实现和应用中我们遇到了许多小问题, 在不断地讨论, 查找相应资料中一步步地进行了解决, 通过这一次的实践, 也让我们再一次深刻地认识到理论指导实践的同时, 从理论到实践依然是一个富有挑战性的过程, 这一过程也是整个过程中最有成就感的部分。

在实现了相应功能之后, 对于如何将这些功能很好地融合进我们的 SERP (搜索结果页面) 中, 争强用户的搜索体验, 我们进行了讨论和设计, 并认真研究了包括 “baidu”、“google”、“sogou” 在内的流行商业搜索引擎的结果页面反馈与布局形式。选择将查询词纠错功能置于搜索框之下, 相关词推荐置于页面右边, 文本补全使用 JS 在用户输入文本时, 在下拉菜单中进行实时显示。当然这一反馈页面仍然存在改进的空间, 例如相关词推荐的配图问题, 实践证明, 有配图的相关词推荐能够让用户更好地找到自己感兴趣的内容。

在整个工程的开发中, 我们巩固课程知识, 尝试设计算法; 遇到困难, 解决困难; 在相互讨论中收获知识, 在项目取得阶段性成果时欢欣鼓舞。作为本科生学习中最后一个小组一起开发的项目, 我们十分珍惜这一次作业的机会, 也许作为计算机系的同学, 一起完成大作业的经历将是最为难忘的回忆。

感谢老师和助教在这次大作业中提供的建议和指导!