

信息检索大作业

2212410 刘俊彤

本项目基于Elastic Search框架搭建了一个面对南开大学校内资源的简易Web搜索引擎，实现了站内查询、短语查询、文档查询、通配查询四种查询方式，并且对登录的用户提供查询日志、个性化查询和个性化推荐功能。

项目文件结构：

```
NKDEX/
|
├─ crawl.py                # 爬虫
├─ pagerank.py             # 链接分析
├─ buildIndex.py           # 构建索引
|
├─ Search/                 # 应用程序文件夹
|   └─ app.py              # 后端代码
|   └─ templates/          # 前端文件夹
|       └─ site_search.html # 站内查询
|       └─ file_search.html # 文档查询
|       └─ phrase_search.html # 短语查询
|       └─ wildcard_search.html # 通配查询
|       └─ login.html       # 登录页面
|       └─ register.html    # 注册页面
|       └─ user_page.html   # 用户主页
|   |
|   └─ static/              # 静态资源文件夹
|       └─ styles.css       # 自定义CSS文件
|
├─ result.json             # 网页信息
├─ user.json               # 用户信息
└─ 网页文件 ...
```

一、网页抓取

完整代码见 `crawl.py`

使用 `BeautifulSoup` 按照域名进行爬取网页，对于每个域名的网页信息按照 [{ `title`, `url`, `content`, `referenced_urls`, `html_filename` }] 存储在 `json` 文件中，并将 `html` 文件存储到域名对应的文件夹中。

```
def crawl_website(start_url):
    """爬取整个网站"""
    visited = set()
    queue = [start_url]
```

```

data = [] # 用于保存网页数据的列表
while queue:
    current_url = queue.pop(0)
    if current_url not in visited:
        print(f"Crawling: {current_url}")
        visited.add(current_url)
        links = get_links(current_url)
        for link in links:
            if link not in visited and 'zfx.y.nankai.edu.cn' in link and not
(link.endswith('.doc') or link.endswith('.pdf') or link.endswith('.docx') or
link.endswith('.rar') or link.endswith('.zip') or link.endswith('.xlsx') or
link.endswith('.xls') or link.endswith('.ppt') or link.endswith('.pptx') or
link.endswith('.jpg') or link.endswith('.png')): # 筛选文档网页
                queue.append(link)
        try:
            response = requests.get(current_url, timeout=10)
            response.encoding = chardet.detect(response.content)['encoding']
# 检测并设置编码
            if response.status_code == 200:
                soup = BeautifulSoup(response.text, 'html.parser')
                title = soup.title.string if soup.title else 'No Title'
                content = soup.get_text(strip=True, separator=' ')

                html_filename = save_html_to_file(current_url,
response.text)

                # 提取网页中引用的URL
                referenced_urls = [a['href'] for a in soup.find_all('a',
href=True) if is_valid_url(a['href'])]
                # 保存数据到列表
                data.append({
                    'title': title,
                    'url': current_url,
                    'content': content,
                    'referenced_urls': referenced_urls,
                    'html_filename': os.path.basename(html_filename)
                })
                with open('nankai_zfxy.json', 'w', encoding='utf-8') as f:
                    json.dump(data, f, ensure_ascii=False, indent=4)
                # 模拟请求之间的延迟
                time.sleep(0.5)

        except requests.exceptions.Timeout:
            print(f"Request timed out for {current_url}")
        except Exception as e:
            print(f"Error processing {current_url}: {e}")

```

最初爬取遇到了两个问题。第一个是爬下来的很多网页内容是乱码的，发现是字符集问题，故添加了编码的检查和设置；第二个问题是由于爬取了学院网，有很多文档下载url，会导致下载了大量的文档，空间占用很大，所以对于带有文档后缀的url进行了过滤。

共爬取了南开大学新闻网、南开大学报以及13个学院官网，共120729个网页。

二、链接分析

完整代码见 `pagerank.py`

具体步骤如下：

- 读取所有 json 文件获得所有网页信息 data
- 将 data 根据 url 进行去重，保证没有重复网页
- 使用 `networkx.DiGraph()` 根据 `referenced_url` 字段构建图模型
- 使用 `networkx.pagerank()` 进行计算每个网页的 pr 值
- 网页数据添加 pr 字段
- 将所有网页数据写入 `result.json` 文件

```
# 构建图模型
def build_graph(data):
    graph = nx.DiGraph()
    for item in data:
        graph.add_node(item['url'], pr=0)
        for referenced_url in item['referenced_urls']:
            graph.add_edge(item['url'], referenced_url)
    return graph

# 计算PageRank
def calculate_pagerank(graph):
    pageranks = nx.pagerank(graph)
    for node, pr in pageranks.items():
        graph.nodes[node]['pr'] = pr
    return graph
```

完成链接分析后计算了一下网页总数以及pr值情况。输出如下：

```
PS F:\NKU\大三上\信息检索系统原理\lab4-大作业\大作业> & D:
Total number of data: 120729
Total number of unique data: 120729
PS F:\NKU\大三上\信息检索系统原理\lab4-大作业\大作业> & D:
Total number of unique PageRank values: 546
```

共计 120729 个网页，有 546 个不同的pr值。

三、文本索引

完整代码见 `buildIndex.py`

使用 `elasticsearch` 框架构建了名叫 `webpages` 的索引，使用 `ik` 分词器进行分词。

```
index_mapping = {
    "settings": {
        "analysis": {
            "analyzer": {
                "ik_analyzer": {
                    "type": "custom",
                    "tokenizer": "ik_smart" # 或者使用 "ik_max_word"
                }
            }
        }
    }
```

```

    }
  },
  "mappings": {
    "properties": {
      "title": {"type": "text", "analyzer": "ik_analyzer"},
      "url": {"type": "keyword"}, # URL作为关键字处理
      "content": {"type": "text", "analyzer": "ik_analyzer"},
      "referenced_urls": {"type": "keyword"}, # 作为关键字处理
      "html_filename": {"type": "keyword"},
      "pr": {"type": "float"} # PageRank值, 用浮点数表示
    }
  }
}

```

ik 分词器有两种模式,

- `ik_max_word` (最大词语粒度): 尽可能将文本切分为更多的词汇, 倾向于最大化词语的数量, 适用于尽可能多包含可能相关的词汇。
- `ik_smart` (智能分词): 尝试对文本进行较为合理的切分, 倾向于减少无意义的短语和单字, 提供更加简洁的分词结果, 适用于精确的分词。

本项目中选择使用 `ik_smart` 模式, 希望可以精确分词。

对 `title` 和 `content` 字段进行的分词处理, 用于后续查询。

四、查询服务

使用 `elasticsearch` 框架实现查询, 使用 `flash` 框架实现web服务

(一) 站内查询

完整后端代码见 `Search\app.py`, 前端代码见 `Search\templates\site_search.html`

```

search_query = {
  "query": {
    "function_score": {
      "query": {
        "multi_match": {
          "query": query,
          "fields": ["title", "content"],
          "type": "best_fields"
        }
      },
      "functions": [
        *[
          {"filter": {"match_phrase": {"content": term}}, "weight": 2}
          for term in user_boost_terms
        ],
        *[
          {
            "script_score": {
              "script": {
                "source": "if (doc['url'].value == params.url) {
return _score * 0.5 } else { return _score }",

```

```

        "params": {"url": url}
    }
}
}
for url in url_decrease_scores
]
],
"score_mode": "sum",
"boost_mode": "multiply"
}
},
"collapse": {
    "field": "url"
},
"sort": [
    {"_score": {"order": "desc"}},
    {"pr": {"order": "desc"}}
],
"size": 50
}

```

使用 `multi_match` 对 `title` 和 `content` 两个字段进行查询，使用 `best_fields` 匹配结果，实现站内查询，不要求用户输入必须作为一个词出现。查询结果首先按照相关度降序，其次按照 `pr` 值降序。

```

resultItem.innerHTML = `
    <a href="${hit.url}" class="result-link" data-url="${hit.url}">${hit.title}
</a>
    <p class="url">${hit.url}</p>
    <p>${hit.content.substring(0, 300)}...</p>
    ${hit.folder && hit.html_filename ? `<a
href="/snapshot/${hit.folder}/${hit.html_filename}" target="_blank"
class="snapshot-link">网页快照</a>` : ''}
`;

```

前端对每个查询结果，显示网页标题（点击跳转网页），网页 `url`，网页内容的前300字，网页快照。

查询效果如下：

NKDEX

站内查询

[站内查询](#)[文档查询](#)[短语查询](#)[通配查询](#)[用户主页](#)[登录](#)[Search](#)

《大师》——张伯苓-专题报道-南开大学

<http://news.nankai.edu.cn/ztbd/system/2016/04/05/000276621.shtml>

《大师》——张伯苓-专题报道-南开大学 首页 南开要闻 媒体南开 南开校史 光影南开 南开故事 南开大学报 视频 广播 您当前的位置：南开大学 >> 专题报道 《大师》——张伯苓 来源：南开新闻网发稿时间：2016-04-05 14:20 编辑：刘慕鑫 微信往期推送 更多... ..

网页快照

南开大学和张伯苓-南开大学

https://weekly.nankai.edu.cn/index/article/articleinfo.html?doc_id=10145

南开大学和张伯苓-南开大学 南开大学和张伯苓——大学和校长的特色（节选） 期次：第885期 阅读：23 □吴大猷南开大学在学术上的成就，或可由下举事见之。民国三十七年（1948年）中央研究院举行第一届院士选举，首由各大学校院、专门学会、研究机构及学术界有资望人士，分科提名候选人，约400余人，继...

网页快照

《重读张伯苓》出版-南开大学

https://weekly.nankai.edu.cn/index/article/articleinfo.html?doc_id=2844

《重读张伯苓》出版-南开大学 《重读张伯苓》出版 期次：第984期 阅读：12 中国著名教育家、南开系列学校创始人张伯苓先生诞辰130周年之际，由沈卫星编著的《重读张伯苓——纪念张伯苓先生诞辰130周年》由光明日报出版社出版，以缅怀这位在中国近代教育史上有着伟大功绩的教育家。 该书图文并茂，分...

网页快照

(二) 文档查询

完整后端代码见 `Search\app.py`，前端代码见 `Search\templates\file_search.html`

文档查询的查询语句构建与站内查询相同。但因为这种查询会筛选 `referenced_urls` 中含有文档链接的网页作为最终的查询结果，所以将直接查询的返回条数增加到了1000。

```
referenced_files = [ref for ref in hit['_source'].get('referenced_urls', [])
                    if any(ref.endswith(ext) for ext in ['.docx',
'.docx', '.pdf', '.rar', '.zip', '.xlsx', '.xls', '.ppt', '.pptx', '.jpg',
'.png'])]]
if referenced_files:
    result = {
        'title': hit['_source'].get('title', 'No title'),
        'url': url,
        'content': hit['_source'].get('content', '')[:300] + '...', # 截取前300个
        'html_filename':html_filename,
        'folder':folder,
        'files':referenced_files
    }
    results.append(result)
```

这样就可以保证只返回带有文档链接的网页，并在前端将文档链接显示，用户点击即可查看或者下载。

查询效果如下：

NKDEX

文档查询

[站内查询](#)[文档查询](#)[短语查询](#)[通配查询](#)[用户主页](#)[登录](#)

计网

Search

<http://222.30.45.189/ueditor/net/upload/file/20171023/6364434796547746757405331.docx> ↓
[【研究生助管】关于做好2017-2018年度第二学期研究生助管考核、评优及夏季学期工作的通知-人工智能学院](#)
<https://ai.nankai.edu.cn/info/1020/1444.htm>
【研究生助管】关于做好2017-2018年度第二学期研究生助管考核、评优及夏季学期工作的通知-人工智能学院 中文 | EN 首页
学院概况 学院概况 发展愿景 组织机构 学科学术分委员会 控制科学与工程学位评定分委员会（含运筹学与控制论） 专业技术职
务评审委员会 学院领导 学院直属办公室 党组织 工会组织 团组织 重大奖项 新闻中心 最新动态 学院公告 学生之窗 科研信息
本科生教学 党团园地 研究生招生 研究生教学 就业信息 国际交流 学科建设 一级学科 二级学科 师资队伍 各类人才 教授(研究员)
副教授(副研究员) 讲 师 实验教学队伍 博士后 兼职教授 荣休教师 招聘与人才引进 信息...
网页快照
<http://222.30.45.189/ueditor/net/upload/file/20180528/6366310948286944978715663.docx> ↓
<http://222.30.45.189/ueditor/net/upload/file/20180528/6366310989187043999134133.xls> ↓
<http://222.30.45.189/ueditor/net/upload/file/20180528/6366310951272886837838556.docx> ↓

文档链接点击即可下载或者查看对应文档。

（三）短语查询

完整后端代码见 `Search\app.py`，前端代码见 `Search\templates\phrase_search.html`

短语查询与站内查询的差别只在查询语句的构建：

```
search_query = {
    "query": {
        "function_score": {
            "query": {
                "multi_match": {
                    "query": query,
                    "fields": ["title", "content"],
                    "type": "phrase"
                }
            },
            "functions": [
                *[
                    {"filter": {"match_phrase": {"content": term}}, "weight": 2}
                    for term in user_boost_terms
                ]
            ]
        }
    }
}
```

```

        * [
            {
                "script_score": {
                    "script": {
                        "source": "if (doc['url'].value == params.url) {
return _score * 0.5 } else { return _score }",
                        "params": {"url": url}
                    }
                }
            }
        ],
        "score_mode": "sum",
        "boost_mode": "multiply"
    }
},
"collapse": {
    "field": "url"
},
"sort": [
    {"_score": {"order": "desc"}},
    {"pr": {"order": "desc"}}
],
"size": 50
}

```

使用 `phrase` 查询类型，对于用户输入的查询词作为一个完整词汇进行查询，且要保证用户输入的多个词在网页内容中出现的顺序一致。

`phrase` 查询语句中有 `slop` 参数，表示查询词之间最多可以有几个词的距离。代码中没有使用这个参数，使用的默认值0，也就各词之间必须是相连的。

查询效果如下：

NKDEX

短语查询

[站内搜索](#)[文档查询](#)[短语查询](#)[通配查询](#)[用户主页](#)[登录](#)[Search](#)

会 国际合作项目 出国指南 宣传材料 联系我们 学生发展 校园...

网页快照

念奴娇 -南开大学

https://weekly.nankai.edu.cn/index/article/articleinfo.html?doc_id=530

念奴娇 -南开大学 念奴娇悼念数学大师陈省身 期次: 第910期 阅读: 15 □曾涤 大师谢世, 同悲恸, 一代巨星陨落。遥望西方, 数十年, 荣登科学巅峰。青年立志, 追求真理, 卷起归衣衫, 漂洋过海, 满怀豪情救国。 回忆当年南开, 大学初兴建, 立夫主数, 毓泰论理。一时间, 科学殿堂初立。叶落归根, 数学国际化, 谈...

网页快照

难忘大运会 -南开大学

https://weekly.nankai.edu.cn/index/article/articleinfo.html?doc_id=8421

难忘大运会 -南开大学 难忘大运会 期次: 第1171期 阅读: 11 ①8月中旬, 我校啦啦队35名成员刻苦练习、精心排演。她们在大运会期间以精彩的表演展示了体育文化。 ②9月3日, 南开大学赛区志愿者举行誓师大会。 ③9月6日, 我校迎来了第九届全国大学生运动会女排第一批参赛队。...

网页快照

图片新闻 -南开大学

https://weekly.nankai.edu.cn/index/article/articleinfo.html?doc_id=8963

图片新闻 -南开大学 图片新闻 期次: 第1180期 阅读: 12 著名学者、南开大学教授宋玉柱先生因病医治无效于2012年11月1日上午10时在天津逝世, 享年79岁。 宋玉柱, 1933年4月生, 河北昌黎人。1953年考入南开大学中文系,

(四) 通配查询

完整后端代码见 `Search\app.py`, 前端代码见 `Search\templates\wild_card_search.html`

通配查询与站内搜索的差别也在于查询语句的构建:

```
search_query = {
    "query": {
        "function_score": {
            "query": {
                "bool": {
                    "should": [
                        {
                            "wildcard": {
                                "title": {
                                    "value": query,
                                    "boost": 1.0
                                }
                            }
                        },
                        {
                            "wildcard": {
                                "content": {
                                    "value": query,
                                    "boost": 1.0
                                }
                            }
                        }
                    ]
                }
            }
        }
    }
}
```

```

        }
    }
    },
    ],
    "minimum_should_match": 1
  }
},
"functions": [
  # Boost terms from user's search history
  *[
    {"filter": {"match_phrase": {"content": term}}, "weight": 2}
    for term in user_boost_terms
  ],
  # Decrease score for URLs that have been clicked using
script_score
  *[
    {
      "script_score": {
        "script": {
          "source": "if (doc['url'].value == params.url) {
return _score * 0.5 } else { return _score }",
          "params": {"url": url}
        }
      }
    }
    for url in url_decrease_scores
  ]
],
"score_mode": "sum",
"boost_mode": "multiply"
}
},
"collapse": {
  "field": "url"
},
"sort": [
  {"_score": {"order": "desc"}}, # 首先按照相关度排序
  {"pr": {"order": "desc"}}     # 其次按照PR值降序排序
],
"size": 50 # 限制返回条数为50
}

```

因为通配查询 `wildcard` 不支持多字段匹配，不能像前面的三种查询直接使用 `multi_match`，所以这里使用 `should` 语句，对 `title` 和 `content` 字段分别进行通配查询，要求两个查询至少满足一个，即 `"minimum_should_match": 1`。

这实现了支持通配符 `*` 和 `?` 的查询，但这种查询需要对索引中的词条进行模式匹配，查询性能远不如其他查询。

查询效果如下：

NKDEX

通配查询

[站内查询](#)[文档查询](#)[短语查询](#)[通配查询](#)[用户主页](#)[登录](#)[Search](#)

2013年职工第一次医疗费用申报的温馨提示

<https://ibs.nankai.edu.cn/n/1486.html>

2013年职工第一次医疗费用申报的温馨提示 新网站入口 教工登录 南开大学 English 首页 | 学院简介 | 科学研究 | 师资队伍 | 学科专业 | 教学教务 | 实验教学 | 对外交流 | 学生工作 | 教育培训 | 校友工作 | 人才招聘 首页 学院动态 学术活动 通知公告 教学招生 教授声音 新闻专题 精品课程 本科精品课程 研究生精品课程 2013年职工第一次医疗费用申报的温馨提示 浏览次数: 11445 更新时间: 2013年04月01日 各位参保职工: 2013年4月份医疗费用(门急诊、门特二次报销、住院二次报销)申报时间如下: 4月1日(星期一)——4月3日...

[网页快照](#)

关怀入宿舍, 温情暖人心——商学院辅导员进行宿舍走访

<https://bs.nankai.edu.cn/2020/0927/c9481a313186/page.htm>

关怀入宿舍, 温情暖人心——商学院辅导员进行宿舍走访 English | 南开大学 EN 内网入口 南开管理评论 图书资料 南开大学 导航 学院概况 学院介绍 使命与愿景 学院领导 系所介绍 VI标识与规范 新闻中心 学院动态 媒体商学院 通知公告 教授声音 视频 师资队伍 人才通告 教师团队 管理团队 招聘信息 学术团队 学术与科研 学术机构 科学研究 学术活动 招生与教学 本科教育 硕博教育 专业学位 产教融合 产教融合发展中心 合作·交流 媒体·南开 通知·公告 学术·科研 国际交流 新闻通告 国际化战略 战略顾问委员会 国际合作项目 出国指南 宣传材料 联系我们 学生发展...

[网页快照](#)

温延龙

<https://cc.nankai.edu.cn/2021/0323/c13619a551346/page.htm>

温延龙 首页导航 首页 学院概况 学院概况 学院领导 学院党委 组织结构 学科学术分委会 学位评定分委会 教代会、工会委员会 专业技术职务评审分委员会 学科建设 学科结构 计算机科学与技术学科 师资队伍 教授/研究员 副教授/副研究员 讲师 实验教师队伍 博士后 兼职教授 人才培养 本科教育 研究生教育 硕士生导师 博士生导师 科学研究 系所中心 学术团队 国际交流 合作院校公

(五) 网页快照

爬取网页信息时, 我记录的网页对应的 html 文件名 `html_filename`, 存储在域名对应的文件夹中。比如 url 为 `https://ibs.nankai.edu.cn/n/1486.html` 对应的 html 文件就存在 `nankai_ibs` 文件夹下。这样, 根据查询到的网页的 `url` 和 `html_filename` 就可以找到对应的 html 文件, 实现网页快照。

```
results = []
for hit in response['hits']['hits']:
    html_filename = hit['_source'].get('html_filename', '')
    url = hit['_source'].get('url', '')
    folder = ''
    if 'ai.nankai.edu.cn' in url:
        folder = 'nankai_ai'
    elif 'bs.nankai.edu.cn' in url:
        folder = 'nankai_bs'
    elif 'cc.nankai.edu.cn' in url:
        folder = 'nankai_cc'
    elif 'ceo.nankai.edu.cn' in url:
        folder = 'nankai_ceo'
    elif 'cs.nankai.edu.cn' in url:
        folder = 'nankai_cs'
    elif 'cyber.nankai.edu.cn' in url:
        folder = 'nankai_cyber'
    elif 'finance.nankai.edu.cn' in url:
```

```

        folder='nankai_finance'

    ...
    else:
        folder=None
    result = {
        'title': hit['_source'].get('title', 'No title'),
        'url': url,
        'content': hit['_source'].get('content', '')[:300] + '...',
        'html_filename':html_filename,
        'folder':folder
    }
    results.append(result)
return jsonify({'hits':results})

```

后端将网页快照链接到对应的文件路径就可以点击查看对应的快照文件。

```

${hit.folder && hit.html_filename ? `<a
href="/snapshot/${hit.folder}/${hit.html_filename}" target="_blank"
class="snapshot-link">网页快照</a>` : ''}

```

查询效果如图：

站内查询

站内查询
文档查询
短语查询
通配查询
用户主页
登录

✕

Search

每日新报：南开大学建校一百周年纪念邮票今发行（图）-媒体南开-南开大学

<http://news.nankai.edu.cn/mtnk/system/2019/10/20/030035893.shtml>

每日新报：南开大学建校一百周年纪念邮票今发行（图）-媒体南开-南开大学 首页 南开要闻 媒体南开 南开校史 光影南开 南开故事 南开大学报 视频 广播 您当前的位置： 南开大学 >> 媒体南开 每日新报：南开大学建校一百周年纪念邮票今发行（图） 来源： 每日新报2019年10月17日第3版发稿时间：2019-10-20 19:11 新报讯【记...

网页快照 ➡ 网页快照

今晚报：南大百年纪念邮票明发行-媒体南开-南开大学

<http://news.nankai.edu.cn/mtnk/system/2019/10/20/030035881.shtml>

今晚报：南大百年纪念邮票明发行-媒体南开-南开大学 首页 南开要闻 媒体南开 南开校史 光影南开 南开故事 南开大学报 视频 广播 您当前的位置： 南开大学 >> 媒体南开 今晚报：南大百年纪念邮票明发行 来源： 今晚报2019年10月16日第5版发稿时间：2019-10-20 17:36 本报讯（记者刘桂芳 通讯员李聪琮）10月17日，南开大...

网页快照

建校100周年纪念邮票发行 -南开大学

https://weekly.nankai.edu.cn/index/article/articleinfo.html?doc_id=16552

建校100周年纪念邮票发行 -南开大学 建校100周年纪念邮票发行 期次：第1391期 阅读：9 本报讯（记者 聂际慈）为庆祝南开大学百年华诞，中国邮政集团公司与我校合作，于10月17日发行南开大学建校一百周年纪念邮票，以国家名片的形式展现巍巍学府百年芳华，为新中国成立70周年献礼。 邮票首发仪...

网页快照

中国新闻社：百年南开登上“国家名片”-媒体南开-南开大学

<http://news.nankai.edu.cn/mtnk/system/2019/09/22/030035368.shtml>

中国新闻社：百年南开登上“国家名片”-媒体南开-南开大学 首页 南开要闻 媒体南开 南开校史 光影南开 南开故事 南开大学报 视频 广播 您当前的位置： 南开大学 >> 媒体南开 中国新闻社：百年南开登上“国家名片” 来源： 中国新闻社9月20日发稿时间： ▾

点击即可显示文件：



五、用户功能

（一）注册与登录

实现了简单的用户注册和登录功能。因为本项目的重点不是这一部分，所以没有连接数据库，直接将用户信息按照 `[{username,password,search_history,clicked_links}]` 记录在 `users.json` 中。来实现用户的查询日志，个性化查询和个性化推荐功能。

登录页面如图：

NKDEX

用户登录

站内查询

文档查询

短语查询

通配查询

用户名:

请输入用户名

密码:

请输入密码

登录

没有账号? [注册](#)

注册页面如图：

NKDEX

用户注册

站内查询	文档查询	短语查询	通配查询
------	------	------	------

用户名:

密码:

确认密码:

注册

已有账号? [登录](#)

(二) 查询日志

新添加了一个用户主页页面，在主页页面会显示用户的查询历史和访问过的网页。

```
# 捕捉用户点击的url
@app.route('/update_click', methods=['POST'])
def update_click():
    print("点击触发")
    if 'user' in session:
        url = request.json.get('url')
        if url:
            update_clicked_links(session['user'], url)
            return jsonify({"status": "success"}), 200
        else:
            return jsonify({"status": "failed", "message": "URL not provided"}),
400
    else:
        print("未登录")
        # return jsonify({"status": "failed", "message": "Not logged in"}), 403
        return jsonify({"status": "success"}), 200
```

添加新路由，在用户点击时调用保存用户访问的url。

```
if 'user' in session:
    update_search_history(session['user'], query)
```

在原本的查询最后，如果现在的回话中有登录的用户，则将查询词添加到用户的查询历史中。

```
// 监听点击行为
document.addEventListener('DOMContentLoaded', function() {
    document.body.addEventListener('click', function(event) {
        if (event.target.classList.contains('result-link')) {
            console.log('Result link clicked:',
event.target.getAttribute('data-url')); // 调试信息
            handleResultClick(event, event.target.getAttribute('data-url'));
        }
    });
});
```

```

    }
  });
});

function handleResultClick(event, url) {
  event.preventDefault();
  console.log('Handling result click for URL:', url); // 调试信息
  fetch('/update_click', {
    method: 'POST',
    headers: {
      'Content-Type': 'application/json'
    },
    body: JSON.stringify({url: url})
  })
  .then(response => response.json())
  .then(data => {
    if (data.status === 'success') {
      console.log('Click recorded successfully'); // 调试信息
      window.location.href = url;
    } else {
      console.error('Failed to update click:', data.message);
    }
  })
  .catch(error => {
    console.error('Error updating click:', error);
  });
}

```

前段添加 `<script>` 标签监听用户的点击行为，与后端对接后，就可以在用户查询和点击链接时保存历史到用户json文件中。最终在用户主页显示用户的查询日志 `search_history` 和访问历史 `clicked_links` 即可。

```

<h3>搜索历史</h3>
<ul id="search-history" class="history-list">
  {% if user.search_history %}
    {% for query in user.search_history %}
      <li>{{ query }}</li>
    {% endfor %}
  {% else %}
    <li>暂无搜索历史</li>
  {% endif %}
</ul>

<h3>访问记录</h3>
<ul id="clicked-links" class="history-list">
  {% if user.clicked_links %}
    {% for link in user.clicked_links %}
      <li><a href="{{ link }}" target="_blank">{{ link }}</a></li>
    {% endfor %}
  {% else %}
    <li>暂无点击记录</li>
  {% endif %}
</ul>

```

主页显示效果如下：



(三) 个性化查询

个性化查询是根据用户的查询历史和点击过的网页来提供不同的查询结果排序，基本思路是将用户查询过的term的权重提高，保证能够优先提供用户感兴趣的内容；同时将用户访问过的网页的权重降低，保证用户不会反复查询到相同网页。

在构建查询语句时：

```
"functions": [  
    # Boost terms from user's search history  
    * [  
        {"filter": {"match_phrase": {"content": term}}},  
    ],  
    "weight": 2}  
    # Decrease score for URLs that have been clicked  
    * [  
        {  
            "script_score": {  
                "script": {  
                    "source": "if (doc['url'].value ==  
params.url) { return _score * 0.5 } else { return _score }",  
                    "params": {"url": url}  
                }  
            }  
        }  
    ],  
    for url in url_decrease_scores  
]
```


],

对于用户查询历史中的term，每个的权重值+2，对用户访问过的url，相关度得分降为原本的一半。

查询效果如下：

用户初次站内查询“考试”得到的结果如下，<https://sfs.nankai.edu.cn/7305/list.htm>这个网页为相关度第一的结果。

站内查询

站内查询

文档查询

短语查询

通配查询

用户主页

登录

考试

×

Search

GRE考试

<https://sfs.nankai.edu.cn/7305/list.htm>
GRE考试 外国语学院 旧版入口 English 首页 学院概况 学院简介 历史沿革 学院领导 机构设置 委员会 师资队伍 英语 翻译 俄语 日语 法语 德语 西班牙语 葡萄牙语 意大利语 阿拉伯语 公共外语教学部 学生培养 本科生 硕士研究生 博士研究生 格拉斯哥 MTI 招生信息 本科生 硕士研究生 博士研究生 格拉斯哥 MTI 学术科研 科研动态 科研团队 科研项目 科研奖励 学术成果 国际交流 院级交流项目 南开外国语学院-威尔士三一圣大卫大学 其他交流项目信息 出国出境 考试中心 通知公告 IELTS考试 TOEFL考试 GRE考试 GMAT考试 日语考试 韩语考试 德福考试 诚聘英才...

网页快照

日语考试

<https://sfs.nankai.edu.cn/7307/list.htm>
日语考试 外国语学院 旧版入口 English 首页 学院概况 学院简介 历史沿革 学院领导 机构设置 委员会 师资队伍 英语 翻译 俄语 日语 法语 德语 西班牙语 葡萄牙语 意大利语 阿拉伯语 公共外语教学部 学生培养 本科生 硕士研究生 博士研究生 格拉斯哥 MTI 招生信息 本科生 硕士研究生 博士研究生 格拉斯哥 MTI 学术科研 科研动态 科研团队 科研项目 科研奖励 学术成果 国际交流 院级交流项目 南开外国语学院-威尔士三一圣大卫大学 其他交流项目信息 出国出境 考试中心 通知公告 IELTS考试 TOEFL考试 GRE考试 GMAT考试 日语考试 韩语考试 德福考试 诚聘英才...

网页快照

TOEFL考试

<https://sfs.nankai.edu.cn/7304/list.htm>
TOEFL考试 外国语学院 旧版入口 English 首页 学院概况 学院简介 历史沿革 学院领导 机构设置 委员会 师资队伍 英语 翻译 俄

点击第一个网页后，用户主页的访问历史中添加该url，再次查询“考试”，得到结果：

站内查询

[站内查询](#)[文档查询](#)[短语查询](#)[通配查询](#)[用户主页](#)[登录](#)

日语考试

<https://sfs.nankai.edu.cn/7307/list.htm>

日语考试 外国语学院 旧版入口 English 首页 学院概况 学院简介 历史沿革 学院领导 机构设置 委员会 师资队伍 英语 翻译 俄语 日语 法语 德语 西班牙语 葡萄牙语 意大利语 阿拉伯语 公共外语教学部 学生培养 本科生 硕士研究生 博士研究生 格拉斯哥 MTI 招生信息 本科生 硕士研究生 博士研究生 格拉斯哥 MTI 学术科研 科研动态 科研团队 科研项目 科研奖励 学术成果 国际交流 院级交流项目 南开外国语学院-威尔士三一圣大卫大学 其他交流项目信息 出国出境 考试中心 通知公告 IELTS考试 TOEFL考试 GRE考试 GMAT考试 日语考试 韩语考试 德福考试 诚聘英才...

网页快照

TOEFL考试

<https://sfs.nankai.edu.cn/7304/list.htm>

TOEFL考试 外国语学院 旧版入口 English 首页 学院概况 学院简介 历史沿革 学院领导 机构设置 委员会 师资队伍 英语 翻译 俄语 日语 法语 德语 西班牙语 葡萄牙语 意大利语 阿拉伯语 公共外语教学部 学生培养 本科生 硕士研究生 博士研究生 格拉斯哥 MTI 招生信息 本科生 硕士研究生 博士研究生 格拉斯哥 MTI 学术科研 科研动态 科研团队 科研项目 科研奖励 学术成果 国际交流 院级交流项目 南开外国语学院-威尔士三一圣大卫大学 其他交流项目信息 出国出境 考试中心 通知公告 IELTS考试 TOEFL考试 GRE考试 GMAT考试 日语考试 韩语考试 德福考试 诚...

网页快照

日语考试

<http://sfs.nankai.edu.cn/7307/list.htm>

日语考试 外国语学院 旧版入口 English 首页 学院概况 学院简介 历史沿革 学院领导 机构设置 委员会 师资队伍 英语 翻译 俄语

可以看到 <https://sfs.nankai.edu.cn/7305/list.htm> 这个网页的相关度降低，再次查询时这个网页不再是第一个结果。

(四) 个性化推荐

个性化推荐的主要思路是对于用户查询历史最近的5条进行查询，并对用户的所有查询历史term权重提高2，访问过的url的相关度降低到原来的一半。

```
if has_search_history:
    user_boost_terms = [term for term in user['search_history']]
    url_decrease_scores = [url for url in user['clicked_links']]
    query = " ".join(user_boost_terms[-5:])
    search_query = {
        "query": {
            "function_score": {
                "query": {
                    "multi_match": {
                        "query": query,
                        "fields": ["title", "content"],
                        "type": "best_fields"
                    }
                },
                "functions": [
                    *[
                        {"filter": {"match_phrase": {"content": term}}],
                    "weight": 2]
                ],
                "for term in user_boost_terms
            ],
            *[
                {
```

```

        "script_score": {
            "script": {
                "source": "if (doc['url'].value ==
params.url) { return _score * 0.5 } else { return _score }",
                "params": {"url": url}
            }
        }
    }
    for url in url_decrease_scores
]
],
"score_mode": "sum",
"boost_mode": "multiply"
}
},
"collapse": {
    "field": "url"
},
"sort": [
    {"_score": {"order": "desc"}},
    {"pr": {"order": "desc"}}
],
"size": 100 # 推荐的数量
}
else: # 没有查询历史则直接返回权重最高的20个网页
    search_query = {
        "query": {
            "match_all": {}
        },
        "sort": [
            {"_score": {"order": "desc"}},
            {"pr": {"order": "desc"}}
        ],
        "size": 20 # 返回20个最高权重的网页
    }
}

```

这样就可以根据用户的查询历史筛选出用户近期感兴趣的网页，并避免向用户重复推荐其访问过的网页。

在用户主页中增加推荐网页的显示：

```

<h3>为你推荐</h3>
<div id="results"></div>
<script>
    document.addEventListener('DOMContentLoaded', function() {
        fetchRecommendations();
        // 使用事件委托处理所有推荐链接的点击事件
        document.body.addEventListener('click', function(event) {
            if (event.target.classList.contains('result-link')) {
                console.log('Recommendation link clicked:',
event.target.getAttribute('data-url')); // 调试信息
                handleResultClick(event, event.target.getAttribute('data-
url'));
            }
        });
    });
}

```

```

});
function fetchRecommendations() {
  fetch('/recommendations', {
    method: 'GET'
  })
  .then(response => response.json())
  .then(data => {
    var recommendationsDiv = document.getElementById('results');
    recommendationsDiv.innerHTML = '';
    data.hits.forEach(hit => {
      var recommendationItem = document.createElement('div');
      recommendationItem.innerHTML = `
        <a href="${hit.url}" class="result-link" data-
url="${hit.url}">${hit.title}</a>
        <p class="url">${hit.url}</p>
        <p>${hit.content.substring(0, 300)}...</p>
        ${hit.folder} && hit.html_filename ? `<a
href="/snapshot/${hit.folder}/${hit.html_filename}" target="_blank"
class="snapshot-link">网页快照</a>` : ''
      `;
      recommendationsDiv.appendChild(recommendationItem);
    });
  })
  .catch(error => {
    console.error('Error fetching recommendations:', error);
  });
}

function handleResultClick(event, url) {
  event.preventDefault();
  console.log('Handling result click for URL:', url); // 调试信息
  fetch('/update_click', {
    method: 'POST',
    headers: {
      'Content-Type': 'application/json'
    },
    body: JSON.stringify({url: url})
  })
  .then(response => response.json())
  .then(data => {
    if (data.status === 'success') {
      console.log('Click recorded successfully'); // 调试信息
      window.location.href = url;
    } else {
      console.error('Failed to update click:', data.message);
    }
  })
  .catch(error => {
    console.error('Error updating click:', error);
  });
}
</script>

```

并保证用户点击推荐的网页的行为也会记录到访问历史。

用户主页的推荐如下：

为你推荐

我的“成长记” -南开大学

https://weekly.nankai.edu.cn/index/article/articleinfo.html?doc_id=1064

我的“成长记” -南开大学 我的“成长记” 期次: 第933期 阅读: 13 王子祯 我的“记者”生活 来到南开, 最大的成就莫过于我加入了《南开大学报》记者团。一进校园, 铺天盖地的社团招聘传单、海报令人眼花缭乱。我像所有大一新生一样没有目标, 只是觉得哪里都好, 什么都有兴趣, 直到一个小册子《吾爱吾团》被送...

网页快照

“以夏之名, 与你相约”——文学院第六届运动会成功举行

<http://wxy.nankai.edu.cn/2018/0515/c15915a197444/page.htm>

“以夏之名, 与你相约”——文学院第六届运动会成功举行 教室预定 | 校友合作 | 交流合作 | English | 登录 | 首页 学院概况 学院简介 各委员会 组织机构 教学机构 研究机构 学科专业 学院新闻 师资队伍 中国语言文学系 东方艺术系 艺术设计系 文化素质教学部 党政学工 学院办公室 教学办公室 研究生办公室 实验教学示范中心 图书资料中心 《文学与文化》编辑部 荣休教师名录 党务行政 党团机构 行政机构 师资人事 申请表格 规章制度 图书资料 学术研究 新闻通告 学术会议 学术讲座 研究项目 研究成果 文件下载 教育教学 本科教育 研究生教育 留学生教育 实验教学 访问进修 招...

网页快照

商学院举办第四届研究生趣味运动会

<https://ibs.nankai.edu.cn/n/429.html>

商学院举办第四届研究生趣味运动会 新网站入口 教工登录 南开大学 English 首 页 | 学院简介 | 科学研究 | 师资队伍 | 学科专业 | 教学教务 | 实验教学 | 对外交流 | 学生工作 | 教育培训 | 校友工作 | 人才招聘 首 页 学院动态 学术活动 通知公告 教学招生 教授声音 新闻专题 精品课程 本科精品课程 研究生精品课程 商学院举办第四届研究生趣味运动会 浏览次数: 9117 更新时间: 2011年11月15日 南开新闻网讯(通讯员 徐琳 刘翰 王玫) 11月11日, 南开大学商学院学工办主办、研究生会承办的第四届研究生趣味运动会在MBA大楼多功能厅举行, 商学院研...

网页快照

.....

退出