

# HW6: 南开百事通

陈雨农

更新: November 17, 2020

本次作业的要求是针对南开校内网站构建一个 Web 搜索引擎, 为用户提供南开信息的查询服务乃至个性化推荐。本次作业可以借助各种工具和包, 希望大家善于利用以减少工作量。

## 1 具体实现

实现这次作业主要有网页抓取、文本索引、链接分析、查询服务、个性化查询几个步骤, 个性化推荐为扩展内容。

### 1.1 网页抓取

对南开大学各网页内容进行抓取。

### 1.2 文本索引

对网页及其锚文本构建索引, 可以按锚文本、网页标题、URL 等域构建索引。

### 1.3 链接分析

使用 PageRank 对链接进行分析, 评估网页权重。

### 1.4 查询服务

使用向量空间模型并结合链接分析对查询结果进行排序, 为用户提供站内查询、文档查询、短语查询、通配查询、查询日志、网页快照等高级搜索功能。更多的内容可以参考百度 (图 1) 或谷歌 (图 2) 的高级搜索功能。

### 1.5 个性化查询

个性化查询为不同的用户提供不同的内容排序。可以实现一个账号登录系统, 通过用户完善的学院专业等个人信息为其呈现不同的查询结果; 或者是记录用户的查询历史, 通过历史查询来提供个性化的查询结果。在 google 的查询中就会通过这些手段来优化用户的查询体验 (图 3)。

搜索设置

高级搜索

×

搜索结果:

包含全部关键词

包含完整关键词

包含任意关键词

不包括关键词

时间: 限定要搜索的网页的时间是

全部时间

文档格式: 搜索网页格式是

所有网页和文件

关键词位置: 查询关键词位于

☒ 网页任何地方
☐ 仅网页标题中
☐ 仅URL中

站内搜索: 限定要搜索指定的网站是

例如: baidu.com

高级搜索

图 1: 百度的高级搜索功能

高级搜索

使用以下条件来搜索网页。

以下所有字项:

与以下字项完全匹配:

以下任意字项:

不含以下任意字项:

数字范围: 从

到

在搜索框中执行以下操作。

输入重要字项: 玉山帆影

用引号将需要完全匹配的字项引起来: "帆影"

在所需字项之间添加 or: 松发 or 特价

在不需要的字项前添加一个减号: -山大、-帆影

在数字之间加上两个句号并添加度量单位: 10、30 斤, 300、500 元, 2010、2011 年

然后按以下标准缩小搜索结果范围...

语言:

任何语言

查找使用您所选语言的网页。

地区:

任何国家和地区

查找在特定地区发布的网页。

最后更新时:

任何时间

查找在指定时间内更新的网页。

网站或域名:

搜索某个网站 (例如 wikipedia.org), 或将搜索结果限制为特定的域名类型 (例如 edu、.org 或 gov)

字项出现位置:

网页上任何位置

在整个网页、网页标题、网址或网页内容所查找网页的链接中搜索字项。

安全搜索:

显示含有醒目色情内容的搜索结果

查找安全搜索筛选过裸露内容的色情内容。

文件格式:

任何格式

查找采用您所指定格式的网页。

使用权限:

不限附件可过或

查找可自己随意使用的网页。

高级搜索

图 2: 谷歌的高级搜索功能

### 私人搜索结果

私人搜索结果可帮您找到相关度更高的内容, 包括只有您可以看到的内容和社交关系。

- ☒ 使用私人搜索结果
☐ 不使用私人搜索结果

### 结果打开方式

- ☐ 在新的浏览器窗口中打开所选的每条搜索结果

### 搜索记录

搜索记录有助于 Google 根据您搜索过的内容、点击过的搜索结果以及其他信息向您提供更相关的搜索结果和推荐内容。您可以随时关闭或修改自己的搜索记录。

图 3: 谷歌的个性化查询服务

## 1.6 个性化推荐

本次作业的扩展内容为个性化推荐，个性化推荐系统通过用户的个人信息和查询历史获取用户可能的兴趣点，在用户查询时给用户推荐相关领域的其他内容。比如在百度上搜索 **iphone**，其会在查询结果的右侧为你推荐 **ipad**、**iMac** 等相关产品（图 4）。



图 4: 查询 **iphone** 时百度查询右侧的推荐

## 2 作业提交

在这个学期剩下的时间里，大家还需完成包括这次作业在内的两次作业。这次作业的截止日期为 12 月 13 日（周日），因为时间比较长，为了帮助大家合理规划时间，要求大家每周日发送邮件简单总结一下本周完成了哪些内容，最后在截止日期前将代码、文档、演示视频打包（命名“学号\_姓名\_hw6”）发送到 [nkulixuy@163.com](mailto:nkulixuy@163.com)。