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**）。

图 1: 百度的高级搜索功能

图 2: 谷歌的高级搜索功能

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。