# 搜索引擎原型系统设计与实现

## [要求]

实现一个搜索引擎的原型系统（demo版本）。检索（即搜索）的范围可以是某个网站的网页（比如南京大学的内部网页），也可以针对某个行业做垂直搜索（比如学术论文检索），或者针对社交媒体（比如微博）进行检索。

**基本功能**应包括：

1. 基于爬取的文档集合，构建倒排索引（inverted index）[1]。可以利用课程提供的文档集合进行构建，下载地址：<http://cs.nju.edu.cn/lwj/course/ds/dataset.rar>；也可以利用爬虫程序自己进行文档爬取 [2]。
2. 实现布尔检索（Boolean retrieval）功能 [1]，至少得支持“与（AND）”和“或（OR）”操作。
3. 对布尔检索返回的结果（即文档）进行排序，排序的标准为每个文档被其他文档链接（或者引用）的次数（即有向图中的入度），入度大的排在前面。

**额外加分的功能**（在大作业基本成绩基础上最多加20%）：

1. 利用爬虫程序 [2]，自动爬取相关的文档集合。例如，南京大学（http://www.nju.edu.cn）网站下的网页。
2. 实现PageRank排序算法 [3]，用来对布尔检索返回的结果（文档）进行排序。
3. 搭建一个完整的搜索引擎，包括用户界面设计、外部排序与搜索等。可以使用开源的框架，如Lucene、Sphinx等，也可以从头开始自己实现。但上面“基本功能”部分的核心代码必须自己实现。

**设计要求：**

1. “基本功能”部分必须以C++语言实现，并且不能调用其他已有的库（例如STL库）。
2. 必须构造图的数据结构，并实现统计入度信息的算法，提供函数接口。
3. 界面友好，函数功能要划分好。
4. 程序要加必要的注释。
5. 要提供程序测试方案。
6. 给出书面报告。

## [参考资料]

1. <http://cs.nju.edu.cn/lwj/course/wsm/lecture2-boolean.ppt>
2. <http://www.yildiz.edu.tr/~aktas/courses/CE-0114890/chapter-8.pdf>
3. <http://cs.nju.edu.cn/lwj/course/mmds/lecture7-LinkAnalysis.ppt>
4. <http://cs.nju.edu.cn/lwj/course/wsm.html>
5. Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. Introduction to Information Retrieval. Cambridge University Press, 2008.

## [检查方式]

每位同学单独完成，检查方式分两部分：现场演示与书面报告。

1. 2016年1月10日前提交系统实现的代码（包括源代码和可执行程序），并在机房现场演示（具体演示时间另行通知）。
2. 提交一份书面报告：小四字体，1.5倍行距，单栏，页数不得少于3页。内容包括问题描述、数据结构设计、算法设计、复杂度分析、实现模块、关键功能与代码、测试流程等。以pdf格式提交（书面报告可于期末考试结束后提交，具体时间另行通知）。
3. 所有代码和文档压缩打包，在相应的截止日期前发送给助教。