搜索引擎原型系统设计与实现

[要求]

实现一个**中文**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。

功能应包括：

(1) 利用爬虫程序 [1]，自动爬取相关的中文文档集合（文档之间必须有有向链接关系）。例如，南京大学（http://www.nju.edu.cn）网站下的网页。可以自己从头编写爬虫程序，也可以采用**开源**爬虫软件，例如雅虎爬虫软件 Anthelion 的开源版本：https://github.com/yahoo/anthelion，也可以采用其他开源爬虫软件。

(2) 设计中文分词算法，实现分词。可以自己从头编写分词算法，也可以采用开源软件，例如 nlpir：http://ictclas.nlpir.org/docs，或者 jieba：https://github.com/fxsjy/jieba 。

(3) 基于爬取的文档集合和分词结果，构建倒排索引（inverted index）[2]。

(4) 实现布尔检索（Boolean retrieval）功能 [2]，至少得支持“与（AND）”和“或（OR）”操作。

(5) 基于文档之间的有向链接图，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。另一种对布尔检索返回的结果进行排序的方法是基于每个文档被其他文档链接的次数（即有向图中的入度），入度大的排在前面。实现基于入度的排序算法，并跟 PageRank 排序算法得到的排序结果进行对比。在布尔检索返回的结果中，同时显示文档、文档对应的 PageRank 值和入度。

(6) 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。

设计要求：

- 上述“功能”部分的(1)和(2)两项功能中，最多只允许其中一项功能采用开源软件，至少得有一项功能是自己从头实现（实现语言不限）。
- 上述“功能”部分的(3)、(4)和(5)三项功能都得自己用 **C++语言**从头实现，并且**不能**调用 STL 库中的如下容器：vector, list, stack, queue, priority queue, set, multiset。
- 必须构造图的数据结构，并实现统计入度的算法和 PageRank 算法，提供函数接口。
- 上述“功能”部分的第(6)项是**额外加分的功能**（在大作业基本成绩基础上**最多**加 20%）。
- 界面友好，函数功能要划分好。
- 程序要加必要的注释。
- 要提供程序测试方案。
- 给出书面报告。

[参考资料]

[1] http://www.yildiz.edu.tr/~aktas/courses/CE-0114890/chapter-8.pdf

[2] http://cs.nju.edu.cn/lwj/course/wsm/lecture2-boolean.ppt

[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. （网上可以下载电子版）

[检查方式]

自行组队完成大作业，每个队伍最多 3 人。检查方式分两部分：现场演示与书面报告。

(1) 2016 年 12 月 15 日之前，完成组队，每个队伍选一个联络员 。

(2) 2017 年 1 月 4 日下午在机房现场演示，每个队伍演示 3 分钟，2 分钟提问。演示完后，现场提交系统实现的代码至课程网站（http://cslabcms.nju.edu.cn/，包括源代码和可执行程序）。

(3) 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。**在书面报告中必须明确每个成员的分工**。以 pdf 格式提交，提交截止日期为 2017 年 1 月 8 日晚上 23:55。