实验报告

2016010524 软 73 金昕祺

1. **实验目标**

   本实验希望实现一个基于现有网页数据库的检索与推荐系统

2. **实验环境**

   语言：C++11

   开发环境：Visual Stuido2017 + Qt5.12

3. **抽象数据结构说明**

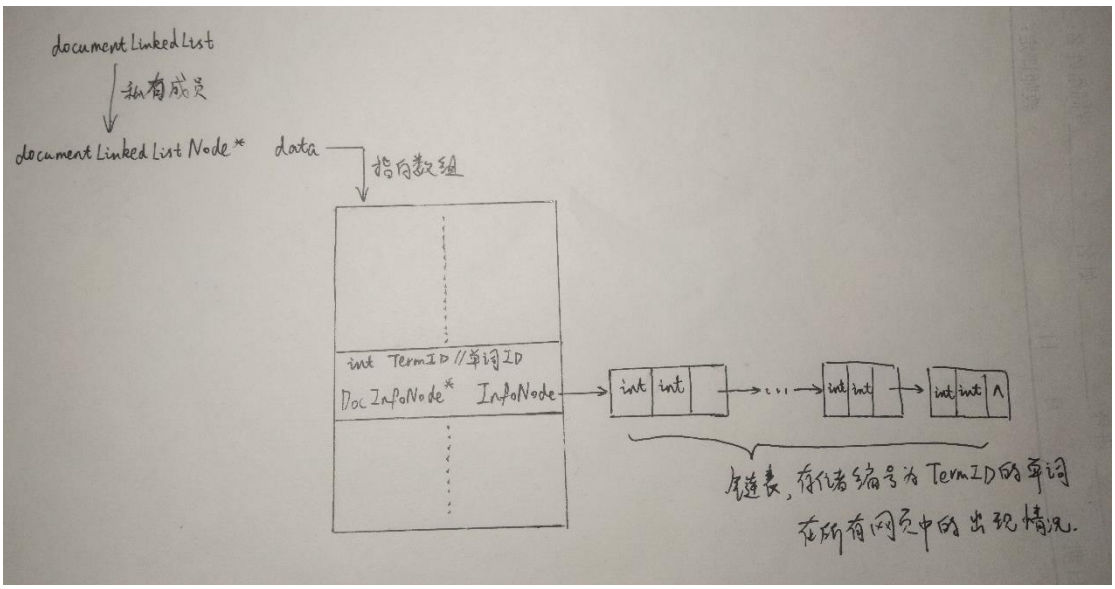
   1.平衡二叉树

   结构：定义为一个 struct，数据成员有 AVLNodeData data（自定义 struct，存储单词的信息如单词 ID、单词出现在多少篇文章中及总出现次数）、int bf（平衡因子）、指向左孩子和右孩子的指针。

   功能：添加节点、查找节点、调整二叉树使其平衡、转化二叉树为广义表达、释放二叉树内存空间。

   2.文档链表

   结构：定义为 class documentLinkedList,私有数据成员 data 指向一个动态的 documentLinkedListNode 数组，数组里的每一个 documentLinkedListNode 对象对应一个单词的全部记录。documentLinkedListNode 有一个成员 TermID 表示单词 ID，和一个成员 DocInfoNode*指向关于该单词的一条记录。DocInfoNode 也是自定义结构体，数据成员有 int DocID(表示该条记录对应的文档 ID)、int Times(表示单词在该文档的出现次数)、DocInfoNode* next 指向关于该单词的下一条记录。



功能：
①添加文档：给定单词 ID、文档 ID、出现次数，向文档链表添加记录
②搜索文档：给定单词 ID、文档 ID，返回指向相应记录的 DocInfoNode*指针。
③修改文档：给定单词 ID、文档 ID、单词在该文档出现次数的变化值、修改模式（分为增大出现次数和减少出现次数两种模式），对相应的 DocInfoNode 进行修改。
④删除某文档：删除文档链表内所有单词关于该文档的记录。
3.哈希表
结构：采用链地址法的哈希表。
功能：将单词插入哈希表、判断哈希表内是否已经存储某单词。
4.其他自定义数据结构：
字符串、字符串链表、链栈

4. 算法说明：

1.检索算法：
根据构建的倒排文档，检索每个网页出现几种目标关键词以及出现的目标关键词的总次数，对每个网页的优先级进行排序。比较优先级时，优先比较出现的目标关键词的种类，其次比较出现的目标关键词的总次数。根据排序后的结果依次输出，作为检索结果。

2.推荐算法:
判断输入的标题是否为数据库内网页的标题，如果是则根据提前准备的分词结果文件获取该网页中出现的全部关键词。以这些关键词作为输入，调用检索算法，并排除输入的标题对应的网页，便得到推荐的网页。

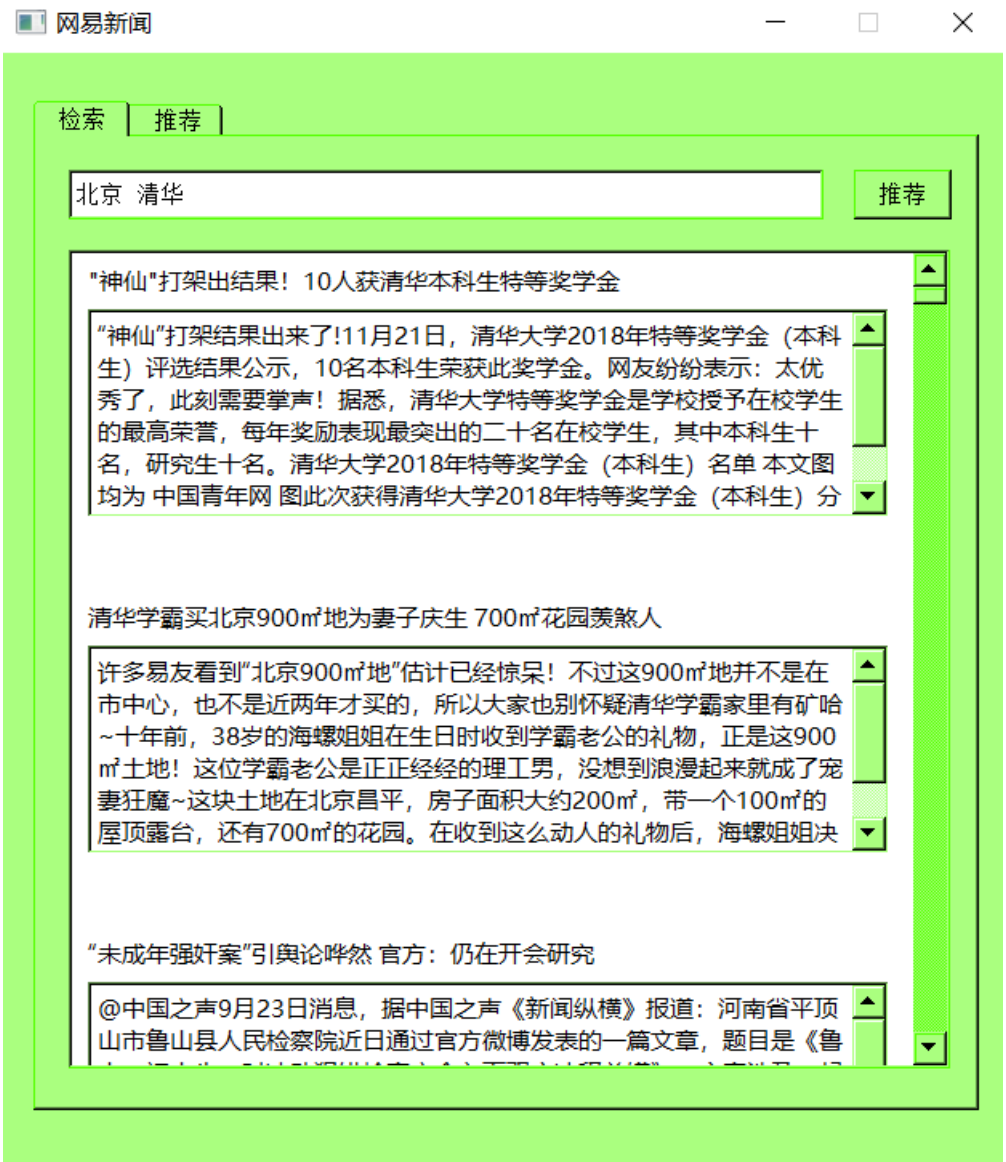
5. 实验流程

无论运行的是 GUI.exe 还是 query.exe，前面四步基本相同：
Step 1：预先完成解析和分词操作(用时约 35 分钟)。
Step 2：读取可执行文件同级目录下的配置文件。
Step 3：读入词库，构建存储词库的平衡二叉树。
Step 4：利用分词结果文件和构建的平衡二叉树构建文档链表。
Step 5：若运行的是 GUI.exe,则根据用户的交互进行检索和推荐操作；若运行的是 query.exe，则根据 query1.txt 执行检索操作，然后再根据 query2.txt 执行推荐操作。

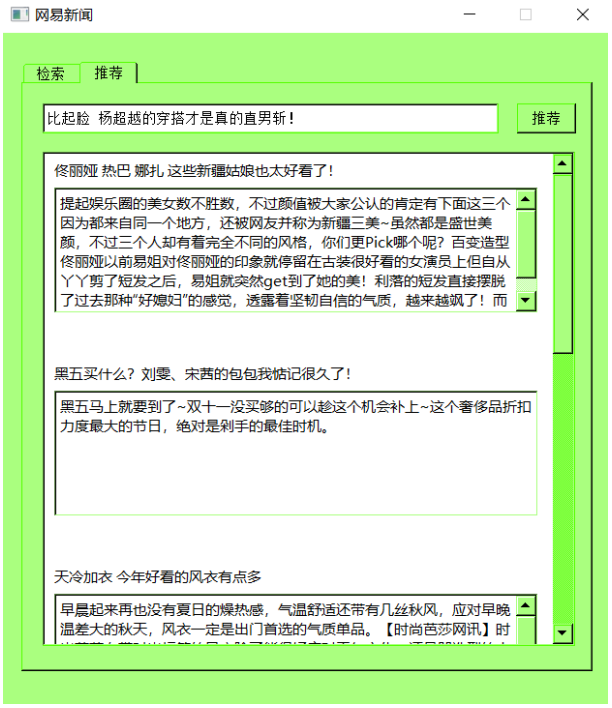
6. 操作说明

1 运行要求：
①query1.txt、query2.txt 和词库文件均应以\n 结尾；query1.txt 每一行最后一个关键词后不可有空格，一行内相邻关键词之间有且仅有一个空格。具体可参考提交的作业内 exe 目录下的 query1.txt 和 query2.txt。
②output 目录（和可执行程序同级）下应提前放入分词结果文件和解析网页结果文件，文件的命名为*.info/*.txt（*均为自然数）。

③与可执行程序同级的 Qt5Core.dll、Qt5Gui.dll、Qt5Widgets.dll、platforms 文件夹必不可少，否则无法运行 GUI.exe。
④由于网络学堂提交作业时的容量限制，可执行程序同级的 input 文件夹为空。实际运行 GUI.exe 时，为保证双击标题能打开相应的 html 文件，需要在 input 文件夹内放入 781 个网页文件。
2.对于 query.exe,直接双击运行即可，在本人机器上，对于大小为 3337KB 的 query1.txt 和大小为 34KB 的 query2.txt，约 10 秒即可运行完毕。
3.对于 GUI.exe,双击即可运行。具体交互方式如下：
①检索界面：
输入关键词，得到检索结果，若检索失败会得到提示，若检索成功，显示出检索到的每一个网页的标题和摘要，单击标题能够打开相应的网页。



②推荐界面：输入标题即可显示推荐结果，最多推荐五条，单击标题能打开相应的网页



7. 功能亮点

1. 实现了友好的交互界面，既可以显示检索和推荐得到的网页的摘要，还可以通过单击打开网页。
2. 运行效率较高，在本人机器上进行测试，对于大小为 3337KB 的 query1.txt 和大小为 34KB 的 query2.txt，约 10 秒即可运行完毕。
3. 程序鲁棒性较好，无论是命令行程序 query.exe 还是图形用户界面，对于非法输入都会给予相应提示。
4. 实现了哈希表，但由于时间限制，未能比较哈希表和平衡二叉树的效率。

8. 实验体会

1. 通过使用实验一预留的接口，显著提升了开发效率，这使我感受到了 OOP 编程方法的优点。
2. 通过使用 Visual Studio 的性能探查器，不断发现程序性能的瓶颈并有针对性地进行优化，这使我感受到了熟练使用 IDE 的重要性。