实验二 报告书

一.实验目标

上次实验进行了，对网页信息的提取。这次通过提取的内容来构建倒排文档与索引机制，根据输入的单词来查找所要的内容，提高学生的编程能力。

二、开发环境

操作系统：Windows 10

IDE：Visual Studio 2012

编程语言：c++

Windows 10 专业版 64位

CPU:Intel(R) Pentinum(R) CPU 4405U @ 2.10GHZ 2.11GHz

内存：4.00GB

三、抽象数据结构说明

单词及文档链表(Term,Doc)

我把文档链表Doc作为了一个单词(Term)的变量，如果在帖子中分出来了一个单词，那么在词典中给它一个ID，并添加出现该单词的url序号，并加入AVL中，如果下次又有该单词出现，那么就再添加一个链表。文档链表的成员有 DoicID,DocTime分别指该链表在该单词的url序号和该单词出现次数。它的成员函数有

add:添加链表 Edit:修改链表 search:查找链表 remove：删除链表

Term的变量有 DF，ID，occur，height分别指共出现在几个url中，单词的编号，共出现次数，和AVL中节点的高度。

Term – DocID1 - DocID2 – DOCID3,差不多是这种形式的

平衡二叉树（AVL）：

平衡二叉树的成员有

insert,search,adjust,remove,edit函数和链表head，head是用来构建平衡二叉树的一个元素。

本次实验最让我头疼的是AVL的实现。

AVL需要保持左树和右数的高度差小于一，而在插入或删除节点的时候，无法避免这一个规律被打破，每次这个规律被打破的时候，我们都要手动把它调整回来，调整回来的思路大概如下我们把调整的过程

简单的分成LL,RR,LR,RL这四个过程,用递归算法，每个节点都比较左右的高度，根据具体情况旋转，使AVL平衡。

insert:添加节点的函数　search:查找AVL中单词的函数　adjust:调整AVL的函数

remove:删除节点的函数　edit:修改AVL中节点的函数

单词的查找与修改，利用了二叉树左孩子比右孩子大的规律，当查找的单词比节点大时，向右移动，反之向左移动。在AVL中自定义了另外两个函数，一个是MAX(a,b)用来返回最大值，另外一个是IFNULL函数，用来排除NULL干扰的。

四、算法说明

用哈希表引入词库：当从字典读入一个单词时，根据该单词的首字母在单词数组中放入，如果遇到了两个相同首字母的单词，那么用链表接着保存该单词的信息。

给单词编号：如果我们给每一个词库中的单词一个编号，那么这个会浪费太多空间，如果分词时建立另外一个链表来判断该单词是否有ID，那么这样AVL就没有多大的意义，因为反正都要遍历一遍还不如直接里用链表。所以我给每一个单词变量给了一个 int *ID，并且把这一个变量设为NULL，如果分词时，字典中该单词的ID如果是NULL那么说明它是第一次出现的，就给它ID。

查找单词：输入我要查找的单词时，有可能输入多个单词，那我们就按空格的个数来判断我总共输入了多少个单词，并把它们保存在二维字符串组中，一个词一个词的查找。先从词库中找出该单词的ID是多少，如果该单词的ID是NULL那么说明该单词在这些url中没有出现过。如果有id存在，那么根据它的id在AVL中一个一个单词的比较，如果该单词的id比节点的id大那么向右移动，反之向左移动。

url的排序：我们在输出结果的时候，要根据输入的词语在这些url中出现次数的大小来输出的，所以我们要把每一个单词的文档链表保存下来，当前面单词的文档链表中有相同的DocID存在，那么把出现次数DocTime加起来，否则就接着后面保存该文档链表。在查找单词结束后，根据DocTime的大小排序文档链表（此时只改变值不该表链表），并一个链表一个链表的输出他的DocID和DocTime。

五、流程概述

载入词库->读取网页->提取相关信息->分词并制作AVL树->输入单词->输出结果

六、输入输出及操作相关说明

批量搜索:input文件夹中放入url，在exe同级目录中有query.txt，在文档里面输入想要查中的单词即可。

GUI：GUI界面的使用方法为，先运行“批量搜索”的exe文件，它会帅选出一些信息保存在 contetns.csv 文件当中。再运行GUI界面，在界面运行之后，有一只绿色的怪物会不停的跳，等它跳完了会变成一个start按钮，点击之后就可以开始查询，看左上方会有三个按钮，指的是三个页面，这三个按钮是为了方便，有利于同时查找许多单词。在框子里输入想要查找的单词，点击 search按钮即可。

图中，三个按钮指的是三个界面，绿色的界面为当时界面，可以点击不同页面来同时查找单词

七、实验结果

完成了实验目标，能完美的查找出每一个单词存在的url序号以及出现次数。

八、功能亮点说明

使用了哈希表。

GUI界面中，使用了多线程，在载入信息时界面中有了等待提示。

GUI界面中创建了多个窗口，可以同时查找不同的单词。

九、实验体会

一个学期大作业终于结束了，每次做大作业就有灵魂被掏空的感觉，不过这次第二次实验和第一次实验比较工作量太少了一点，感觉明年开始第二次实验可以再增加一点任务，第一次和第二次实验工作量相差太大不太好。