实验报告

软件 71
唐建宇

一、 实验目标

1. 读取分词结果建立倒排索引。
2. 利用倒排索引进行检索与推荐。
3. 实现友好的交互界面。

二、 实验环境

编译器：Visual Studio 2012；操作系统：Windows 10；语言：C++
GUI 为 VS 中的 MFC 应用程序。

三、 抽象数据结构说明

1. 二叉平衡树

类 AVL_Tree

结构：指向树根节点的指针。每个节点包括一个指向文档链表的指针，平衡因子以及指向左右孩子的指针。

接口：

void Insert(CharString *word,int urlcode);插入节点

void Adjust(treenode *&currentnode); 调整平衡树使之平衡

FileLinkList *Search(CharString &word);

以上为本实验中用到的三个接口。

另外实现了二叉平衡树模板类（代码也一并附上了），在模板类中实现了以上三个以及

void Remove(T x); 删除节点

void Edit(T currentkey,T newkey); 将关键字为 currentkey 的节点替换为关键字为 newkey 的节点

2. 文档链表

类 FileLinkList

结构：关键词 key，包含的文档总书 doc_num 以及指向链表头节点的指针 head。链表节点包括文件编号，该词语在文件中的出现次数以及指向前后节点的指针。

接口：

void Add(int code);加入编号为 code 的节点

linknode *Search(int code);搜索编号为 code 的文档的节点，返回指向节点的指针

void Edit(int oldcode,int newcode);将编号为 oldcode 的文档替换为编号为 newcode 的文档

void Remove(int code);将编号为 oldcode 的文档从链表中删除

四、 算法说明

检索算法：

对于每一组查询首先分割所给定的关键词，对每一个关键词在正文的倒排索引中进行查找并返回一个文档链表。将所得到的每一个文档链表进行归并，归并时，包含不同的关键词更多的文档节点靠前，包含不同关键词数量相同时，包含所有关键词总数最多的靠前。由此得到一个归并后的文档链表，将这个文档链表中的节点按格式依次输出。

推荐算法：

对于给定的标题，首先判断是否在数据库中：对标题进行分词，将这些词在标题的倒排索引中查找，与检索算法中相似地进行归并，如果最靠前的文档节点包含所有给定标题中的关键词，则认为标题在数据库中，反之则认为不在数据库中。（之所以采用先分词再比较而非直接遍历所有新闻标题进行匹配，是为了防止因为标点、开头或末尾空格等情况导致匹配不成功）

推荐采用 TF-IDF(Term Frequency-Inverse Document Frequency)算法，即词频-反向文档频率算法。目标是选出文档中相对重要且有代表性的词汇作为关键词进行检索即可得到相似的新闻。

词频 tf 即某个词汇在这个文档中出现的总次数除以该文档中所有出现的词汇次数之和，数值越大说明出现次数越频繁，反映了这个词汇对于这个文档的重要性。

反向文档频率 idf 即整个数据库中的文档数量与某个词汇所出现的所有文档数的比值的对数。这个值越大说明这词汇在较少的文档中出现，即这个词比较独特，换言之更有代表性。

对某个新闻做推荐时，给该新闻中所有词进行赋分值 P

$$\mathrm{p} = \mathrm{tf} \times \mathrm{idf} = \frac{n_{i,j}}{\sum_m n_{i,m}} \times \log_{10} \frac{d_j}{\sum d}$$

其中，$n_{i,j}$为词汇 j 在新闻 i 中出现次数，分母即该新闻中所有词汇数之和，d_j为包含词汇 j 的文档数，$\sum d$为数据库中所有文档总数。

为该新闻中出现的每一个词赋分后，按照分值从大到小排序，取前三个词汇，认为他们是文档中兼具重要性与代表性的词汇。再将这三个词作为关键词进行一次检索，即得到了推荐新闻的结果，取前五位按格式输出。

五、 预分词的说明

本次实验中提前进行了网页的解析与分词操作，为方便检索与推荐，我对标题和正文分别进行了分词并保存，即每一个网页得到了三个文件，分别为 xxx.info(信息提取)，xxx.txt(正文分词)以及 xxxt.txt(标题分词)，放在了 input 文件夹中一并提交。

分词的过程用时 30 秒以内（如果网页加载进了内存（例如第二次运行程序）则 10 余秒即可，I/O 和屏幕输出为主要消耗），内存消耗 300MB。

六、 流程概述

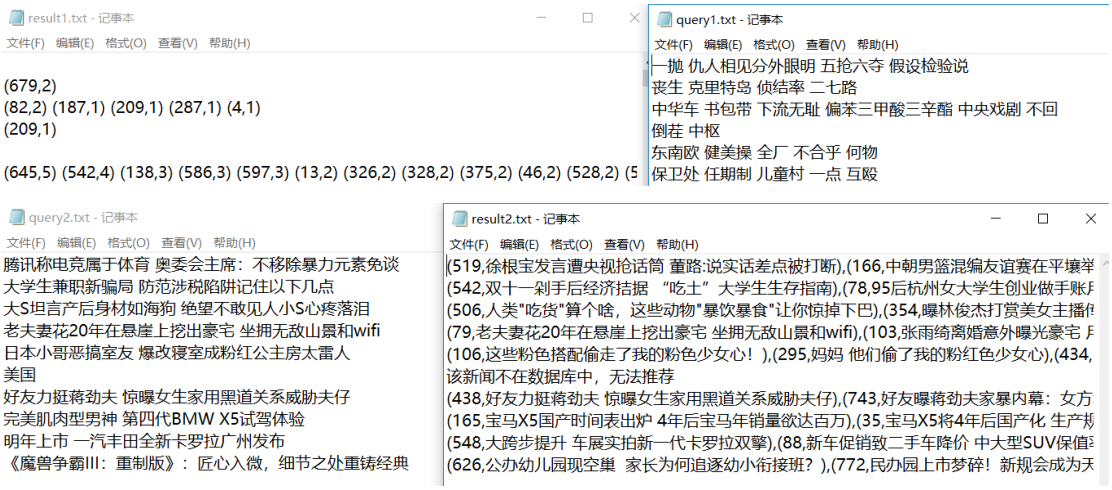
进入主函数 main，首先调用 initdictionary 加载词库；
接着在 input 目录下搜索后缀为.txt 的文件，每搜到一个，调用 init_InvertedFileIndex 将文件中的所有词汇加入倒排索引。本次实验建立标题和正文的两个倒排索引。
接着进入批量检索 bulksearch 函数。按行读取关键词，对每一组关键词按算法说明进行检索操作。
再进入批量推荐 bulkrecommendation 函数，按行读取标题，对每一个标题调用 fuzzy_search 模糊搜索函数将标题分词后与数据库中标题进行模糊匹配，接着调用 TF_IDF 函数对每一个词进行赋分并排序，最后对排名最高的前三个词调用 bulksearch2 进行搜索。

七、 输入输出及相关操作说明

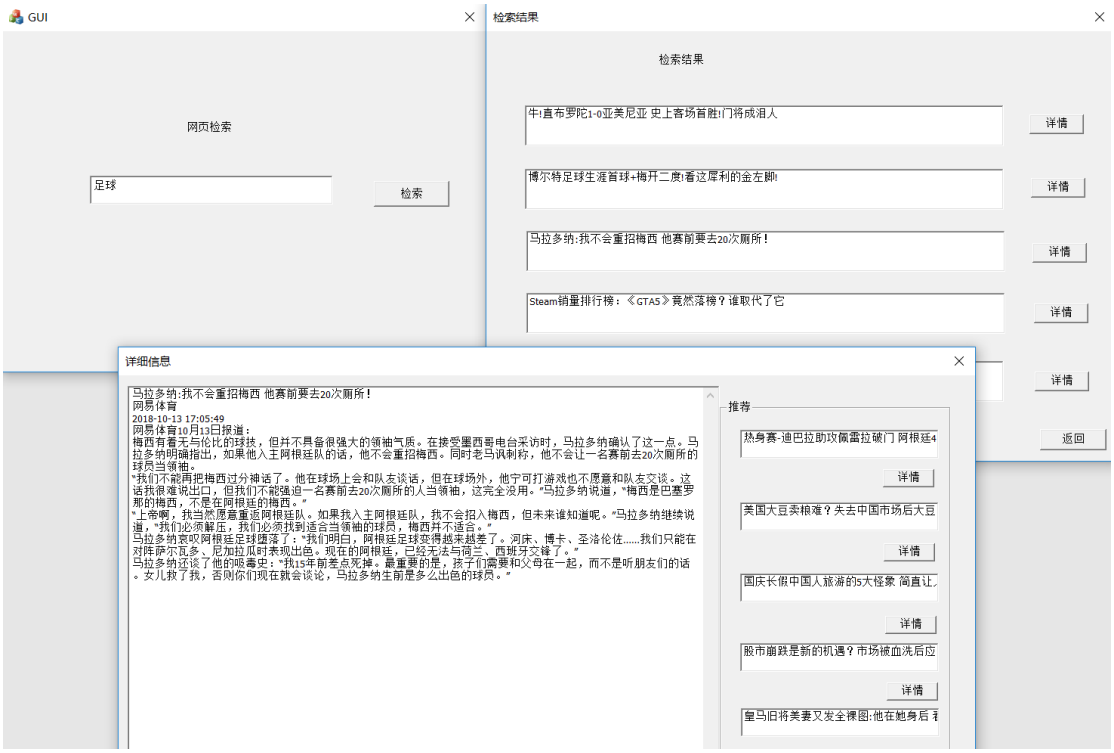
输入：
批量检索与推荐程序只需助教配置 query1.txt 与 query2.txt 即可运行，所需要的配置文件都已放在 input 和 exe 同目录下。
GUI:
进入界面后，在编辑框中输入想要检索的关键词，关键词间以空格隔开，点击检索键即可进入检索结果页面。检索结果页面输出 5 个检索结果的标题，每个标题旁边有详情键，点击可进入详情页面，包含新闻全文与推荐结果，推荐结果显示方式与检索页面类似，五个标题与对应的按键。
输出：批量检索与推荐程序屏幕输出为建立倒排索引时所读取的文件名，文件输出为符合要求的输出。
GUI 均为屏幕输出。

八、 实验结果

批量检索与推荐：



GUI:



九、 实验体会

MFC 不好用，debug 很痛苦。
平安夜还算愉快。
也祝助教圣诞愉快！