# 数据结构课程实验报告

## ——基于文本内容的音乐检索与推荐

清华大学软件学院 李肇阳 2014013432

2015 年 11 月 3 日

## 目 录



## §1 概述

### §1.1 实验目的

解析HTML提取歌曲的若干字段信息，并对歌词执行分词。为编写歌曲检索和推荐系统做准备。

熟悉经典数据结构、相关算法及其应用，提高编程能力。

### §1.2 开发环境

集成开发环境Visual Studio 2012 Premium (MSVC++ 11.0)，操作系统Microsoft Windows 7 Ultimate。

## §2 实现说明

### §2.1 主要数据结构

#### §2.1.1 链表

模板类。双向不循环链表。实现了链表的若干核心功能，包括获取长度、添加元素、用下标查询。

未编写迭代器，但用下标查询时会缓存上一次查询结果，这样一来，用下标遍历也只需要线性时间。

为确保指针安全性以及避免内存泄露，全程采用深拷贝。

#### §2.1.2 堆栈

模板类。继承自链表。实现了堆栈的基本功能，包括压栈、退栈、取栈顶元素、检查栈是否为空等。

#### §2.1.3 字符串

动态管理的连续空间。实现了内存动态管理、KMP模式匹配、增改字符、截取子串、拼接等。

主要有3个成员变量，char[]数组首地址、字符串当前长度和数组容量。内存空间管理采用倍增策略。

实现了与std::string的双向转换。全程采用深拷贝。

#### §2.1.4 字符串链表

继承自链表以字符串进行的实例化。

### §2.2 主要算法

#### §2.2.1 超文本解析

大致分为三步：

- 切取热点区域。
- 解析HTML，将所有标签保存在一张链表中。
- 遍历上述链表，提取信息。

**第一步** 利用字符串模式匹配算法，根据硬编码的特征字符序列直接切取特点区域。

**第二步** 采用堆栈，查找“＜”、“＜／”、“＞”等具有特征的字符或字符序列以定位标签。具体流程叙述如下：

- 查找下一个“＜”，若找不到，结束。
- 检查紧跟“＜”的字符。若是“／”，说明遇到标签结束，取栈顶标签，设置标签的结束位置为当前位置，发送到解析结果，退栈并回到上一步。否则继续。

- 在“＜”之后寻找第一个空格或“＞”。若找到，说明遇到标签开始，取它们中间的内容为标签名称，记录当前位置为标签的起始位置，将标签入栈，并回到第一步。否则报错退出。

“标签”是一个自定义的结构，内含标签名称、标签在原HTML字符串中的起止位置。它是所用堆栈和所得链表中的元素。

在这一步得到的结果中，各标签间的层次关系已经丢失。

**第三步** 遍历上一步得到的链表，寻找特定名称的标签（“h2”、“li”、“textarea”），提取其中的歌曲信息。

**封装情况** 接受HTML字符串，返回解析好的歌曲信息结构。封装在HTML解析器类之中。

### §2.2.2 中文分词

采用正向最大匹配算法。需提供词库。这是一个很平凡、也很简单的算法，这里略去对它的描述。

对于英文单词（连续的ASCII字符）进行了特别处理。具体来说，遇到ASCII字符时，检查其下一个字符，如果是空格或非ASCII字符，则判定当前单词结束，发送到分词结果。

**封装情况** 接受字符串，返回字符串链表。以配置文件的文件名进行初始化。封装在分词器类之中。

## §2.3 总体流程

- 首先检查命令行参数合法性。
- 根据输入目录获取输入文件名列表。
- 初始化分词器，加载词典。
- 遍历输入文件名列表，对于每一个输入文件，依次读入全文，执行解析，执行分词，并将结果写入输出文件。
- 结束。

# §3 使用说明

## §3.1 使用方法

命令行：iMusic arg1 arg2 arg3

其中arg1为配置文件的文件名，arg2为输入文件所在目录，arg3为期待得到输出文件的目录。

配置文件只有一行，为词典文件的文件名。词典文件每行为一个词。

程序将遍历输入目录中的*.html文件，处理后输出相应*.info、*.txt文件。

## §3.2 注意事项

目录、文件名均应为绝对路径或者相对当前工作目录的相对路径。目录应是已存在的目录。

所有文件内容、文件名、目录名，均应统一采用GBK编码格式（Code Page 936）或其子集（GB2312、ASCII等）。程序不会读取字节序标记或HTML文档的charset元数据。

## §4 实验结果

助教提供的10个网页全部通过，自行构造了若干边界输入（空的、无实际内容的、标签错乱的等）全部通过。

解析结果符合要求，分词结果不存在明显不当。

经简单测试，初始化耗时秒级，单个网页处理耗时0.01秒级。经性能分析，不存在明显性能瓶颈。

## §5 得意之处

- 我是真·用C++写的！面向对象思想！
- “遍历输入目录得到输入文件名”这一步，只用了std::system函数，执行命令行“dir /b *.html > filelist.txt”输出重定向到文件然后去读取那个文件，可移植性肯定比调用一些下划线开头的函数要好得多吧，比如移植到linux下只需改为命令“ls -1”。
- 代码风格自认为不错（雾）假装写了单元测试（雾）

## §6 感想与体会

感觉坑很多。课程实验（大作业）是个很好的形式，可以让我把所学的知识串联起来、应用起来，加深理解，加强工程实践能力。

赶早“踩坑”，而后得以为其他同学提供经验和帮助，这一经历令我感到愉快。

作业刚刚布置时我曾对这种“重复发明轮子”的任务有些抵触，但完成之后，我感到对经典数据结构、经典算法、相关的工程上的问题的理解大大加深了。正所谓“绝知此事要躬行”，亲自动手实现一遍，比看书、看别人代码、直接调用标准库高到不知道哪里去了。总之是收获颇丰。

## §7 致谢

感谢陈凯助教及时回答我的疑问。感谢叶曦同学与我就本实验进行了不少有益的讨论。感谢git这一由开源社区提供的方便的版本控制工具。感谢ThinkPad品牌为我提供了很高的工作效率（小红帽大法好，Mac和其他一切都是异端!）。

（完）

## 参考文献