第三周大作业《清华新闻网信息检索系统》设计文档

计52 于纪平2015011265

2016年9月

Contents

1 功能简述 **2**

2 用户界面 **2**

3 爬虫 **5**

4 HTML解析 **6**

5 分词处理 **6**

6 HTTP服务器 **6**

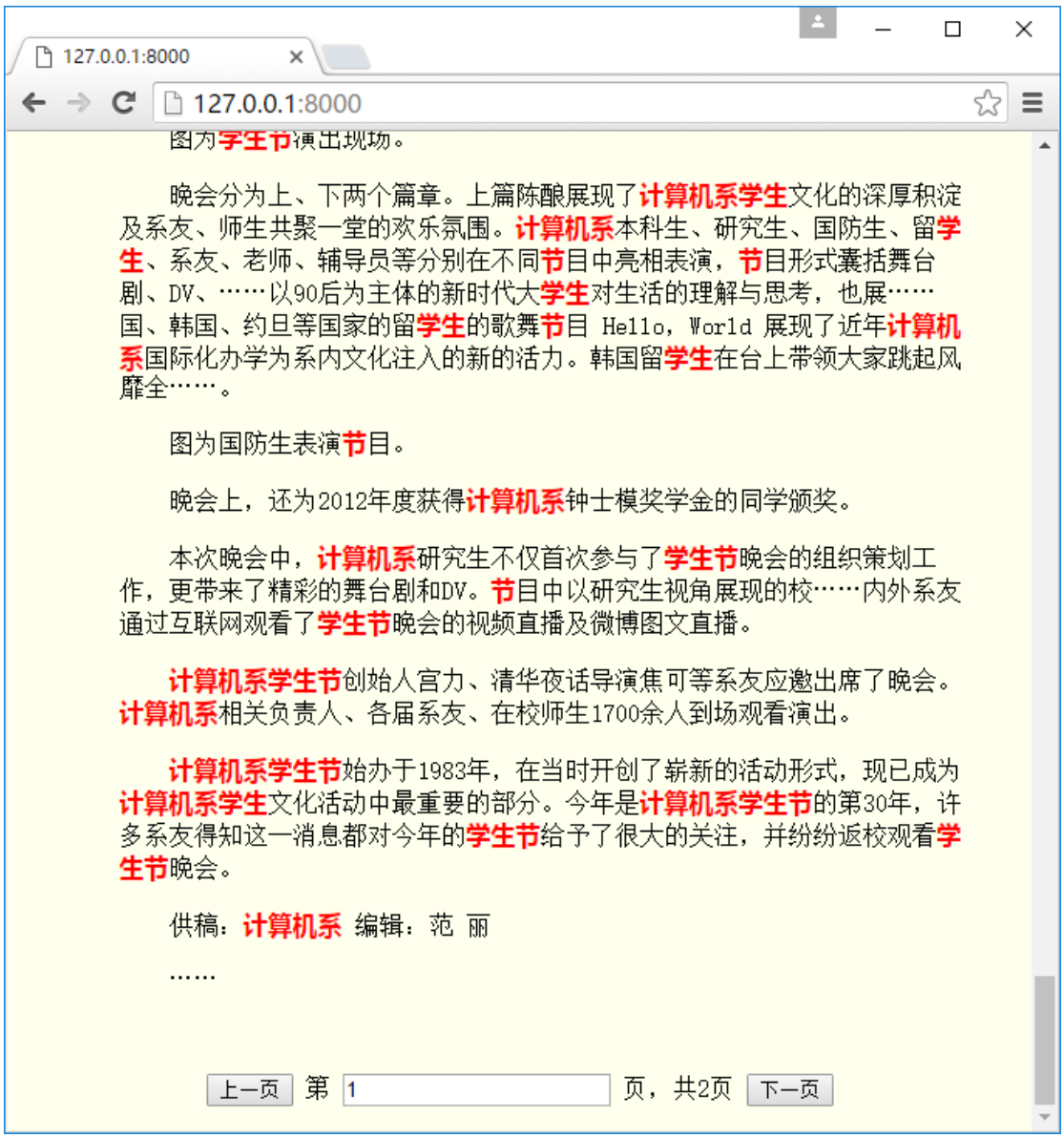
1 功能简述

- 开启网页服务器后，可供浏览器连接，进入搜索页面。
- 搜索时可以指定关键词和查询时间（不限，当月，当年，过去某年等）搜索。
- 用户可通过点击搜索页面的超链接，查看这篇新闻的全文。
- 实现了分页显示功能，在搜索结果页面提示总页数，提供了“上一页”“下一页”按钮，用户也可手工输入页码进行跳转。
- 在搜索结果页面也提供了对文中含有关键词的部分的预览，并将关键词标红；其余部分用“……”省略。
- 使用了CSS美化页面，包括设置页面背景色、文字大小、字体、颜色等。

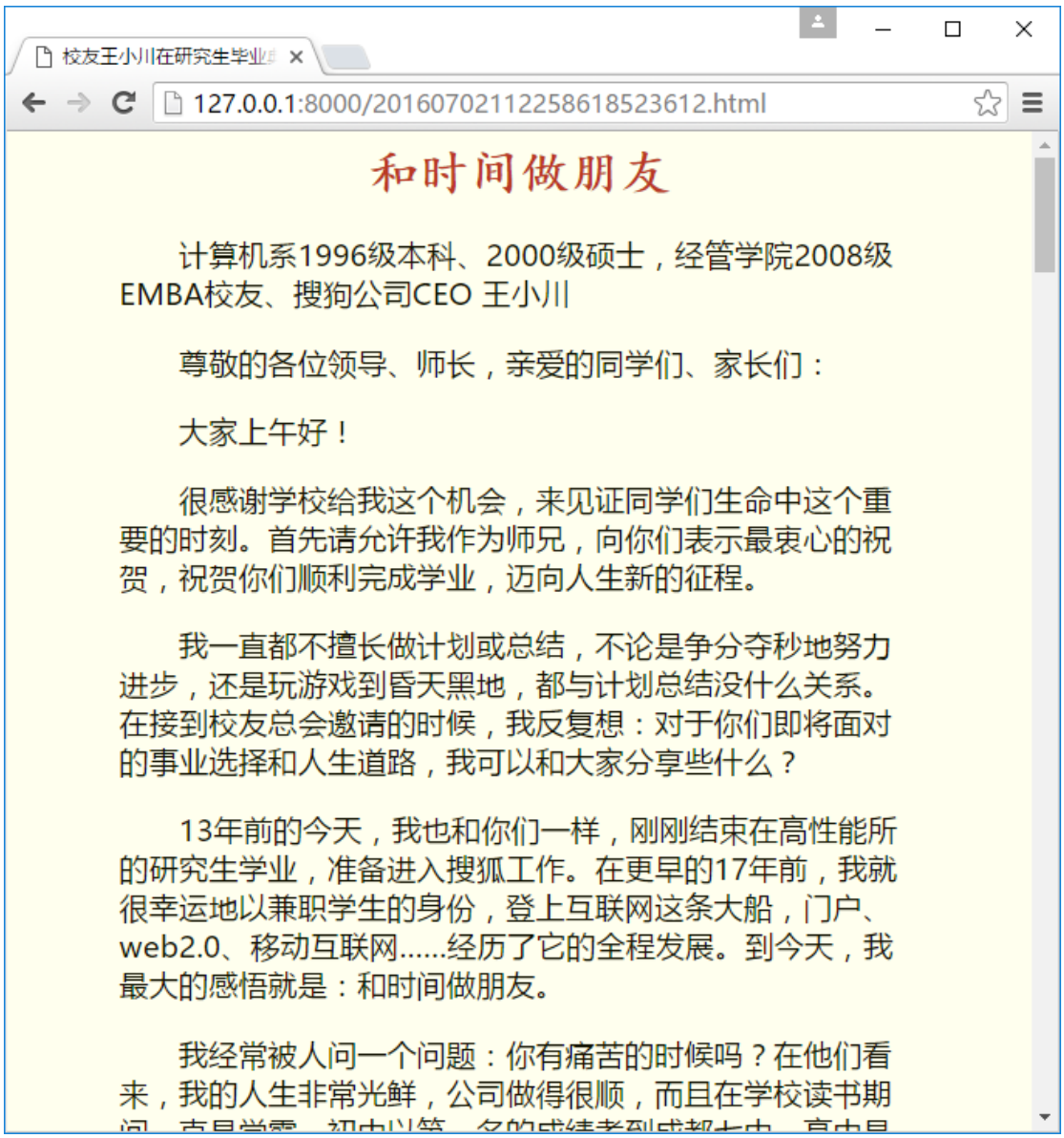
2 用户界面

页面顶端（搜索框、时间范围选项、搜索按钮）：

页面底端（翻页按钮与页码输入框）：


127.0.0.1:8000
127.0.0.1:8000
图为学生节演出现场。
晚会分为上、下两个篇章。上篇陈酿展现了计算机系学生文化的深厚积淀及系友、师生共聚一堂的欢乐氛围。计算机系本科生、研究生、国防生、留学生、系友、老师、辅导员等分别在不同节目中亮相表演，节目形式囊括舞台剧、DV、……以90后为主体的新时代大学生对生活的理解与思考，也展……国、韩国、约旦等国家的留学生的歌舞节目 Hello，World 展现了近年计算机系国际化办学为系内文化注入的新的活力。韩国留学生在台上带领大家跳起风靡全……。
图为国防生表演节目。
晚会上，还为2012年度获得计算机系钟士模奖学金的同学颁奖。
本次晚会中，计算机系研究生不仅首次参与了学生节晚会的组织策划工作，更带来了精彩的舞台剧和DV。节目中以研究生视角展现的校……内外系友通过互联网观看了学生节晚会的视频直播及微博图文直播。
计算机系学生节创始人宫力、清华夜话导演焦可等系友应邀出席了晚会。计算机系相关负责人、各届系友、在校师生1700余人到场观看演出。
计算机系学生节始办于1983年，在当时开创了崭新的活动形式，现已成为计算机系学生文化活动中最重要的部分。今年是计算机系学生节的第30年，许多系友得知这一消息都对今年的学生节给予了很大的关注，并纷纷返校观看学生节晚会。
供稿：计算机系 编辑：范 丽
……
上一页
第
1
页，共2页
下一页

新闻内容页面：



3 爬虫

爬虫模块的作用是从清华大学新闻网主页开始尝试打开所有不超出范围的网页。

内部的HTML解析模块只负责处理超链接。对于访问的每一个（不重复的）网页，解析其中所有的超链接并尝试扩展。如果这是一个新闻页面则将其保存下来。

共爬取39028个新闻页面，总大小1.85G，压缩后73.0M。

2011年2月26日及以前的页面（可能是由于导入旧版本新闻网的问题）有大量重复页面或编码有问题的页面，2011年2月27日及以后共17418个页面，总大小809M，压缩后32.1M。

4 HTML解析

这个模块负责将提取新闻网页的标题和文本。

由于清华大学新闻网的格式设计十分合理，故可以直接抓取标题（<title>），而文章内容可以从元素（<article>）中的文本内容获得。

解析之后将每个html文件转化为txt文件，总大小63.0M。

5 分词处理

这个模块将标题和正文分词，并建立倒排索引。

分词使用了jieba工具。倒排索引按照“词语：网页，网页，……”的方法存在文本文件中，大小为150M。

6 HTTP服务器

这个模块利用了Django框架。共设计了两个函数，分别处理搜索页面和每个新闻页面。

由于不能直接返回一个txt文件，新闻页面函数对该txt进行了处理（设计样式，转换空格、回车等字符）然后返回。

对于搜索页面，设计了一个表单（如以上【用户界面】所示），并且设计了一个隐藏项目表示总页数（以便浏览器端检查翻页按钮是否合法）。服务器收到浏览器的请求后先解析请求，将查询文本进行分词，在倒排索引中找到这些文本取交，返回结果。

由于数据量不大，可以预先读入完整的倒排索引和所有网页的内容，每次需要访问时直接在内存中寻找即可，搜索时间可以接受。