

项目说明文档

● 2013.12.15

目录

1. 引言	3
1.1 编写目的.....	3
1.2 项目背景.....	3
1.3 术语定义.....	3
1.4 引用文档、参考资料及注释.....	4
2. 系统架构设计与实现.....	5
2.1 系统分层架构设计.....	5
2.2 系统模块设计与实现.....	6
2.2.1 系统总体模块架构.....	6
2.2.2 各模块设计与实现.....	6
2.3 数据库分析与设计.....	18
3. 项目特色	19
4. 小组成员分工	19
5. 致谢	19

1. 引言

1.1 编写目的

此文档为了通过描述项目的内容、架构、项目开发中的问题及解决方法、项目开发成果以及项目成员分工等,来对整个项目进行详细的说明,使读者可以更好理解项目的整体内容,并且使项目开发和维护更加有条理。

预期读者:项目组成员以及老师。

1.2 项目背景

项目名称:京东商城笔记本电脑商品深度 web 数据挖掘展示系统

项目任务提出者:第四小组成员讨论提出

项目开发者:第四小组团队

用户群:京东商城、笔记本电脑厂商、需要购买笔记本电脑商品的用户。

1.3 术语定义

缩写、术语	解 释
Web Crawler	是一种以某种规律、自动的浏览互联网的计算机程序 (spider), 这一浏览过程又可以被称为抓取 (spidering)。
URL	Uniform Resource Locator, 即统一资源定位符。是对可以从互联网上得到的资源的位置和访问方法的一种简洁的表示, 是互联网上标准资源的地址。互联网上的每个文件都有一个唯一的 URL, 它包含的信息指出文件的位置以及浏览器应该怎么处理它。
AJAX	是动态 JavaScript 和 XML 技术 (Asynchronous JavaScript and XML) 的缩写, 用于使用一系列和 Web 开发相关的技术在客户端创建异步的应用。
JavaScript/js	是一种基于对象和事件驱动并具有相对安全性的客户端脚本语言。同时也是一种广泛用于客户端 Web 开发的脚本语言, 常用来给 HTML 网页添加动态功能, 比如响应用户的各种操作。
JSON	(JavaScript Object Notation) 是一种轻量级的数据交换格式。简单说就是 javascript 中的对象和数组, 通过这两种结构可以表示各种复杂的结构

JQuery	是一个优秀的 Javascript 框架。它是轻量级的 js 库，它兼容 CSS3，还兼容各种浏览器。jQuery 使用户能更方便地处理 HTML documents、events、实现动画效果，并且方便地为网站提供 AJAX 交互。它还能够使用户的 html 页面保持代码和 html 内容分离，也就是说，不用再在 html 里面插入一堆 js 来调用命令了，只需定义 id 即可。
DOM	Document Object Model，即文档对象模型，DOM 可以以一种独立于平台和语言的方式访问和修改一个文档的内容和结构。换句话说，这是表示和处理一个 HTML 或 XML 文档的常用方法。
Html	超文本标记语言，标准通用标记语言下的一个应用。“超文本”就是指页面内可以包含图片、链接，甚至音乐、程序等非文字元素。超文本标记语言的结构包括头部分（Head）、和主体部分（Body），其中头部（head）提供关于网页的信息，主体（body）部分提供网页的具体内容。
Xml	可扩展标记语言，是标准通用标记语言的子集，用于标记电子文件使其具有结构性的标记语言，可以用来标记数据、定义数据类型，是一种允许用户对自己的标记语言进行定义的源语言。它非常适合 Web 传输，提供统一的方法来描述和交换独立于应用程序或供应商的结构化数据。
CSS	即级联样式表，它是一种用来表现 HTML（标准通用标记语言的一个应用）或 XML（标准通用标记语言的一个子集）等文件样式的计算机语言。
jsoup	是一款 Java 的 HTML 解析器，可直接解析某个 URL 地址、HTML 文本内容。它提供了一套非常省力的 API，可通过 DOM，CSS 以及类似于 jQuery 的操作方法来取出和操作数据。
Highcharts	是一个用纯 JavaScript 编写的一个图表库，能够很简单便捷的在 web 网站或是 web 应用程序添加有交互性的图表，并且免费提供给个人学习、个人网站和非商业用途使用。HighCharts 支持的图表类型有曲线图、区域图、柱状图、饼状图、散状点图和综合图表。

1.4 引用文档、参考资料及注释

项目相关文件：《视图范围文档》，用来阐释项目背景、目的、需求等问题

参考资料：

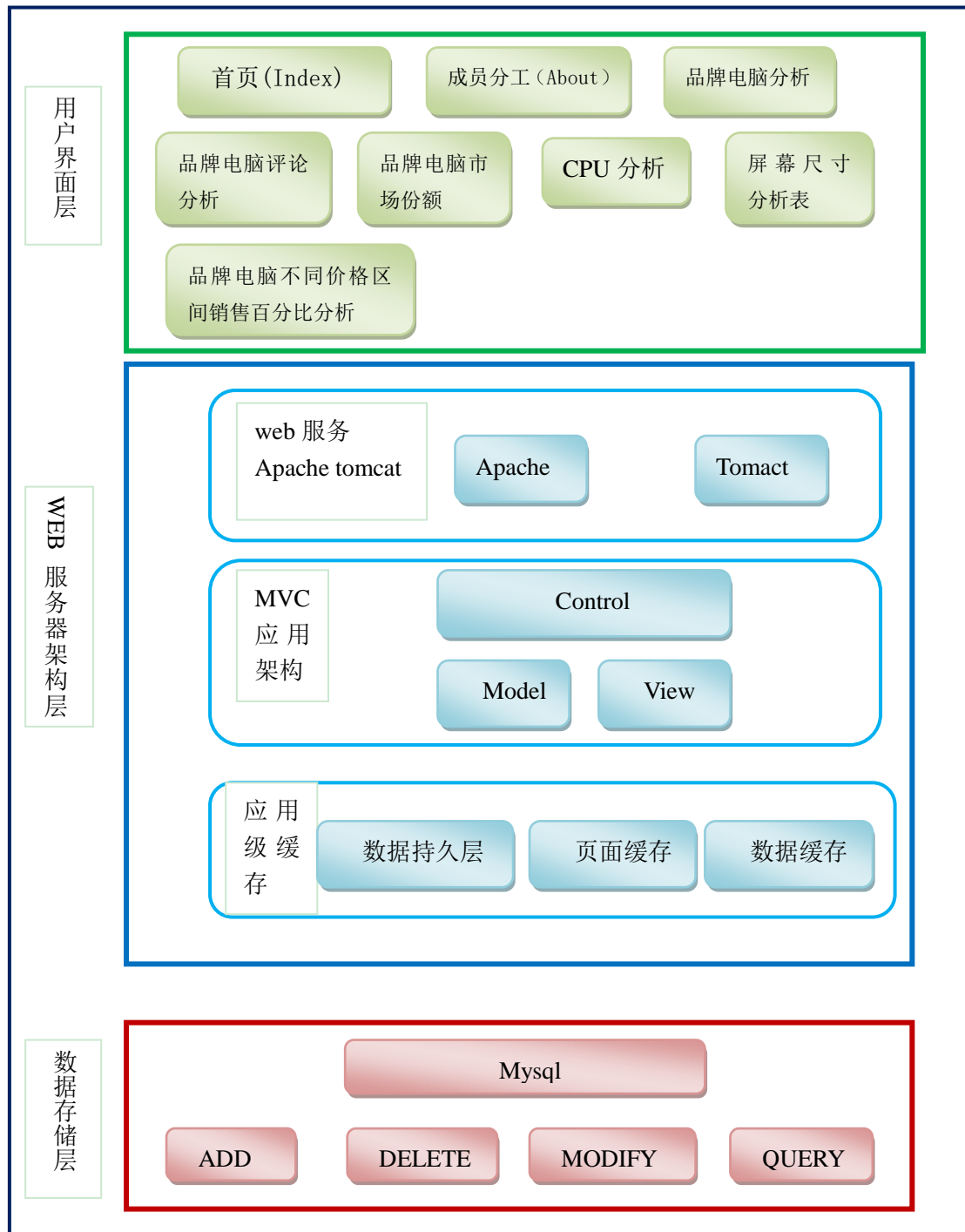
《深度 web 采集系统的设计与实现》

注：以下文档中电脑代指“笔记本电脑”

2. 系统架构设计与实现

2.1 系统分层架构设计

系统的分层架构设计：



图表 1 系统的分层架构设计图

2.2 系统模块设计与实现

2.2.1 系统总体模块架构

根据需求，系统将划分为 3 个模块，包括数据采集模块、数据分析模块和界面展示模块，其中数据采集模块主要负责对目标数据源爬数据，并将获取到的数据存入数据库；数据分析模块主要负责从数据库中提取数据进行分析，并将分析和处理结果写入文件；界面展示模块主要负责将由数据分析模块所得到的分析结果展示在用户界面上。



图表 2 系统总体模块架构图

2.2.2 各模块设计与实现

2.2.2.1 数据采集模块

（1）数据采集工作原理

信息采集部分主要完成了对京东商城中的笔记本电脑数据的采集。从京东商城爬取到所有的正在销售的笔记本电脑的品牌，型号，用户评论等信息。采集部分将采集到的数据存入到数据库中，供系统后续分析和挖掘。

Web 上绝大多数的数据是以 HTML 页面的形式存储在网站的服务器上的，这些页面有数据库和相关的网页模版来生成，具有较高的结构性和局部数据快代码结构重复性。同一个网站通常有统一的风格。

采集系统利用分析到的抽取规则，可以将这些网页中的特定的信息抽取出来，进行存储利用。对于一个信息采集系统，抽取信息的规则是核心，他影响到采集结果的准确性。

系统使用的采集系统中，分析京东商城的商品列表页面，商品简介页面，商品详细页面和商品评价页面的特定信息。使用 Jsoup 将分析得到的数据选取出来存储到数据库中，供系统进行后续分析和挖掘。

（2）数据采集功能要求与设计

本项目旨在对京东商城的笔记本电脑的相关数据进行分析 and 挖掘，因此数据采集部分主要需要对可能需要进行分析的数据进行挖掘。这一部分的信息主要包括：

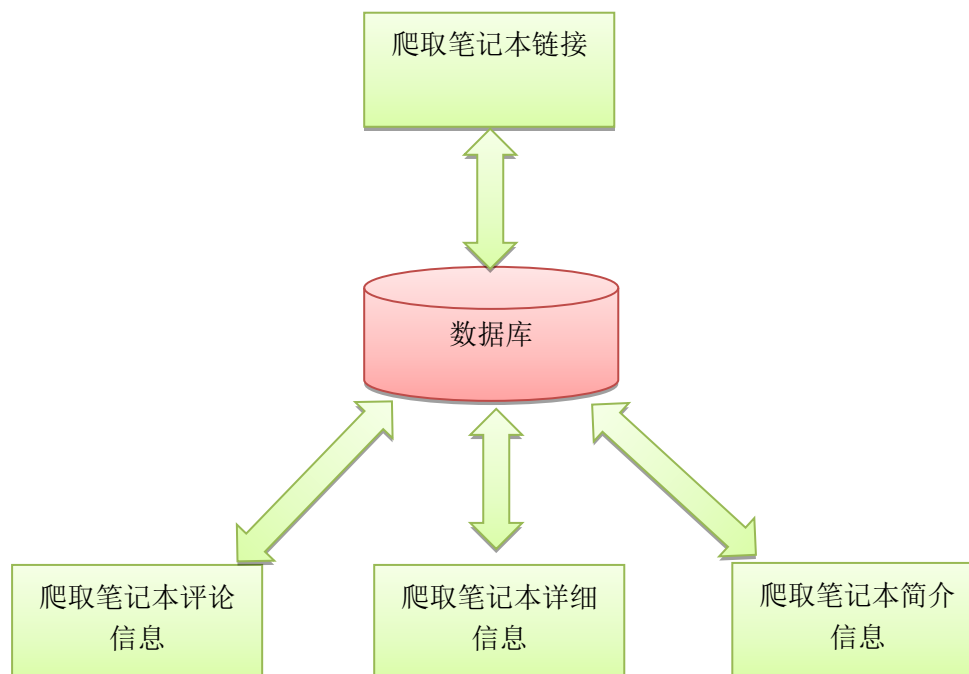
表格 1 数据采集信息汇总表

价格
标题
品牌

型号
颜色
平台
操作系统
Cpu 型号
内存
硬盘
光驱
显示器
尺寸
重量
好中差评数
评价内容

数据爬取部分通过对输入的源网址向下进行广度优先搜索，爬取到符合规则的 URL 插入到数据库中。爬取的另外三个模块爬取笔记本评论信息，爬取笔记本详细信息，爬取笔记本简介信息从数据库中取出相应的 URL 进行分析，获取所需的信息插入到数据库中。数据采集模块可以分成四个工作部分：爬取笔记本链接，爬取笔记本评论信息，爬取笔记本详细信息，爬取笔记本简介信息。这四个部分分别由一个线程执行，在数据爬取部分工作的工程中，这四个部分并行工作。

数据采集部分的主要框架如下：



（3）数据采集所用技术总结

相关技术包括以下几种：

- Jsoup 解析器

Jsoup 是一款 Java 的 HTML 解析器,可直接解析某个 URL 地址、HTML 文本内容。Jsoup 的主要功能如下:

1. 从一个 URL, 文件或字符串中解析 HTML;
2. 使用 DOM 或 CSS 选择器来查找、取出数据;
3. 可操作 HTML 元素、属性、文本;
4. 可依据一个安全的白名单过滤用户提交的内容, 以防止 XSS 攻击。Jsoup 是基于 MIT 协议发布的, 可以免费使用。

一般一个 HTML 由节点名, 属性和文本三部分构成, Jsoup 提供了一套非常简洁优雅的 API, 可通过 DOM, CSS 以及类似于 jQuery 的操作方法来取出和操作数据。在元素的选取方面, Jsoup 几乎无所不能; Jsoup 提供了一种选择器, 选择器能够帮助我们迅速地导航到所需要的 HTML 元素、属性、文本等数据进行操作。另外这种选择器很容易和 chrome 的开发者工具进行结合

在数据采集部分中使用 jsoup 进行了 html 页面的获取, 并且对获取的页面进行解析, 提取出需要的数据存入到数据库中。使用了对网页数据进行获取, 代码片段如下:

```
String pageSelect = "div.page>a[href]";
Elements pageUrls = doc.select(pageSelect);
```

图表 4 对网页数据获取代码片段

其中选择字符串的获取可以从 chrome 中非常方便地获取:



图表 5 选择字符串示意图

■ MySQL

MySQL 是一种关联数据库管理系统, 关联数据库将数据保存在不同的表中, 而不是将所有数据放在一个大仓库内, 这样就增加了速度并提高了灵活性。MySQL 所使用的 SQL 语言是用于访问数据库的最常用标准化语言。

MySQL 在过去由于性能高、成本低、可靠性好, 已经成为最流行的开源数据库, 因此被广泛地应用在 Internet 上的中小型网站中。随着 MySQL 的不断成熟, 它也逐渐用于更多大规模网站和应用, 比如维基百科、Google 和 Facebook 等网站。非常流行的开源软件组合 LAMP 中的“M”指的就是 MySQL。

数据爬取部分爬取获得的 URL 和经过分析页面抽取到的信息均存储在 MySQL 数据库中。

■ 多线程

由于数据采集模块在上面提到要分为四个部分并行执行, 因此要用到多线程技术。

Java 中实现多线程有两种途径: 实现 Runnable 接口和继承 Thread 类。系统中使用的是继承 Thread 类这种方法, 通过覆写 run 方法实现多线程:

```
public class DetailAnalyzer extends Thread{
    private DBManage dbManager;
    public DetailAnalyzer(DBManage dbmanager){
        this.dbManager = dbmanager;
    }
    public void run() {
        Connection conn = dbManager.getConnection();
        while (true) {
```



```
// get detail url
```

图表 6 实现多线程代码片段图

■ URL 去重方案选择

在爬虫启动工作的过程中，我们不希望同一个网页被多次下载，因为重复下载不仅会浪费 CPU 机时，还会为系统增加负荷。而想要控制这种重复性下载问题，就要考虑下载所依据的超链接，只要能够控制待下载的 URL 不重复，基本可以解决同一个网页重复下载的问题。

非常容易想到一个方案，即在系统中建立一个全局的专门用来检测，是否某一个 URL 对应的网页文件曾经被下载过的 URL 存储库。

然而遇到一个新的 URL，将其放于 URL 存储库中进行检测，要求去重工作更加高效，在检测时应将已被下载的 URL 库加载到内存中，在内存中进行检测一定会比直接从磁盘上读取速度快很多。

URL 去重有许多方案，选取不同的方案会直接影响到去重的效果和效率，方案的分析如下：

第一，基于磁盘的顺序存储。

这是一种最直观但最不可行的方案，就是指把每个已经下载过的 URL 进行顺序存储在磁盘文件中。你可以把全部已经下载完成的 URL 存放到磁盘记事本文件中。每次有一个爬虫线程得到一个任务 URL 开始下载之前，通过到磁盘上的该文件中检索，如果没有出现过，则将这个新的 URL 写入记事本的最后一行，否则就放弃该 URL 的下载。总体来说占据较大存储空间的同时，查找效率低下，淘汰这种方案。

第二，基于 Hash 算法的存储。

对每一个给定的 URL，都是用一个已经建立好的 Hash 函数，映射到某个物理地址上。当需要进行检测 URL 是否重复的时候，只需要将这个 URL 进行 Hash 映射，如果得到的地址已经存在，说明已经被下载过，放弃下载，否则，将该 URL 及其 Hash 地址作为键值对存放到 Hash 表中。这种方案，维护 URL 库相当于维护一个哈希表，如果哈希函数设计的不好，发生碰撞的几率很大，而对碰撞的处理也较为复杂。总体来说，需要占据较大存储空间，且去重的效果和效率依赖于哈希表的设计。

第三，基于 MD5 压缩映射的存储。

MD5 算法是一种加密算法，同时它也是基于 Hash 的算法。这样就可以对 URL 字符串进行压缩，得到一个压缩字符串，同时可以直接得到一个 Hash 地址。另外，MD5 算法能够将任何字符串压缩为 128 位整数，并映射为物理地址，而且 MD5 进行 Hash 映射碰撞的几率非常小，即使对于较大规模爬取都是可接受的，在爬虫进行检测的过程中，可以通过记录日志来保存在进行 MD5 时发生碰撞的 URL，通过单独对该 URL 进行处理也是可行的。

本项目组选用这种方案进行 URL 去重操作。

下面就是对字符串进行压缩的 MD5 方法，通过调用 java 的 security 包来获取 MD5 方法实例，并通过 MD5 方法将一个传入字符串先压缩为一个 128 位整数，再将其转为字符串返回：

```
class EncryptionByMD5 {
    public static String getMD5(String src) {
        byte[] source = src.getBytes();
        String s = null;
        char hexDigits[] = { '0', '1', '2', '3', '4', '5', '6', '7', '8',
            '9', 'a', 'b', 'c', 'd', 'e', 'f' };
        try {
```

```

        java.security.MessageDigest md =
java.security.MessageDigest.getInstance("MD5");
        md.update(source);
        byte tmp[] = md.digest();
        char str[] = new char[16 * 2];
        int k = 0;
        for (int i = 0; i < 16; i++){
            byte byte0 = tmp[i];
            str[k++] = hexDigits[byte0 >>> 4 & 0xf];
            str[k++] = hexDigits[byte0 & 0xf];        }
        s = new String(str);
    } catch (NoSuchAlgorithmException e) {
        e.printStackTrace();
    }
    return s;
}
}

```

图表 7 MD5 加密字符串代码片段图

将源 URL、压缩后的 URL、是否分析过等属性存入 URL 库中，如果当前 URL 分析完成，它的是否分析过属性由 0 变为 1，表示分析完成，在检测新的 URL 时可以非常快速的检测已压缩后的 URL 是否未存在库中或存在库中但还未分析这些情况。可以将压缩后的 URL 串作为 Key，而将 Boolean 作为 Value 进行存储，然后将工作中的 Map 在爬虫停止工作后序列化到本地磁盘上；当下一次启动新的爬虫任务的时候，再将这个 Map 反序列化到内存中，供爬虫进行 URL 去重检测。

本项目选取的通过 MD5 加密 URL 的方案，已经可以较好的实现 URL 去重。当然还有两种方案值得考虑，只是在学习和使用上较为复杂，留作日后探讨。

第四，基于嵌入式 Berkeley DB 的存储。

Berkeley DB 的特点就是只存储键值对类型数据，这和 URL 去重有很大关系。去重，可以考虑对某个键，存在一个值，这个值就是那个键的状态。使用了 Berkeley DB，你就不需要考虑进行磁盘 IO 操作的性能损失了，这个数据库在设计的时候很好地考虑了这些问题，并且该数据库支持高并发，支持记录的顺序存储和随机存储，是一个不错的选择。

URL 去重存储库使用 Berkeley DB，压缩后的 URL 字符串作为 Key，或者直接使用压缩后的 URL 字节数组作为 Key，对于 Value 可以使用 Boolean，一个字节，或者使用字节数组，实际 Value 只是一个状态标识，减少 Value 存储占用存储空间。

第五，基于布隆过滤器（Bloom Filter）的存储。

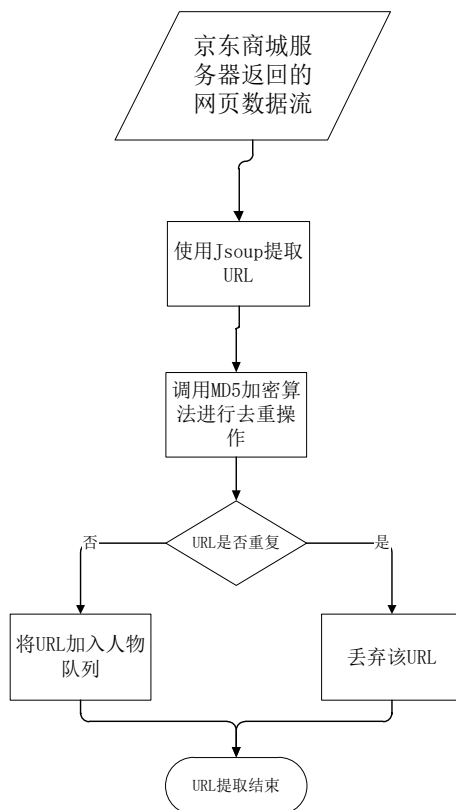
使用布隆过滤器，设计多个 Hash 函数，也就是对每个字符串进行映射是经过多个 Hash 函数进行映射，映射到一个二进制向量上，这种方式充分利用了比特位，使得存储所占空间更少。

纵观所有方案，都在减少 URL 的存储位数和提高检索效率方面做了非常多的努力。

（4）数据采集关键部分——URL 提取和去重

是一个不断读取并对新的 URL 信息进行检测存储处理的过程，包括 URL 提取和去重操作，此模块最终的产物为 URL 库。URL 提取通过 Jsoup 解析器和去重通过 MD5 加密算法

处理流程如下：



图表 8 URL 处理流程图

2.2.2.2 数据分析模块

(1) 数据分析工作原理

从数据库中读取数据，对读取的每条数据进行分割以及字段提取，得到电脑商品的属性信息如价格、品牌、cpu 类型、屏幕尺寸等。现将这些数据用键值对的方式进行存储，便于进行分析统计。根据系统需求，对每个商品的不同属性信息和由中文分词得到的评论信息进行分析统计，综合得出分析结果。对于不同的数据采取不同的分析方式，对于容易得到的数据通过关键字段提取的方法，对于较难直接获取的评论数据通过利用盘古中文分词进行分割，获取评论中出现频率最高的关键词。

(2) 数据分析所用技术总结

■ HashMap 方法

从数据库中读取的每条数据，得到商品的标题(title)、价格(price)、品牌(brand)、CPU 类型(CPU)、尺寸(Size)等信息。利用这些数据初始化定好的 computer_info 对象。

按照<brand, computer_info>, <price, computer_info>, <CPU, computer_info>, <size, computer_info>的键值方式对数据进行存储，便于分析统计。

■ 中文分词——盘古分词

通常情况下，用户评论的价值对于商家、用户还是京东商城而言都是十分重要的资源，会很大程度上左右一款产品的销量，而我们需要从大篇幅的用户评论中获取针对一款商品的

关键字是十分必要的，因此选择好的中文分词算法至关重要。

在这个问题上，选择盘古分词，盘古分词采用字典和统计结合的分词算法，分词准确率较高，并且盘古分词也是一个开源组件。盘古中文分词有两个重要部分组成：`devideywords.exe`和 `PanGu.xml`

直接通过调用命令行执行，并通过缓冲区输入流获取分词结果
代码片段如下：

```
dir_exe="C:\\Users\\guicun\\workspace\\wangguicun\\bin\\Segment\\devideywords.exe";
    arg1="
C:\\Users\\guicun\\workspace\\wangguicun\\bin\\Segment\\pangu\\PanGu.xml ";
    arg2="";
Process p;
arg2=arg2.replaceAll(" ", "");
    arg2=arg2.replaceAll("\\t", "");
    String cmd=dir_exe+arg1+arg2;
    p=java.lang.Runtime.getRuntime().exec(cmd);
```

图表 9 调用分词程序的代码片段图

2.2.2.3 界面展示模块

使用 JSP 开发网页展示界面，通过接收数据分析模块的分析结果来调用 highcharts 库生成图表来进行展示，其中界面的设计与实现主要利用了 html、javascript、DOM 技术。

(1) 应当遵循的界面设计规范

遵循人机界面设计规范，追求界面个性化、简单性、易于操作的界面风格设计，让用户轻松的使用系统。

- 清楚一致的设计。所有界面的风格保持一致，所有具有相同含义的术语保持一致，且易于理解和使用。
- 拥有良好的直觉特征。以用户所熟悉的现实世界事物的抽象来给用户暗示和隐喻，来帮助用户迅速学会系统的使用。
- 较快的响应速度。从操作上为用户提供简单、方便、快捷的操作途径。

(2) 界面的关系图和工作流程图

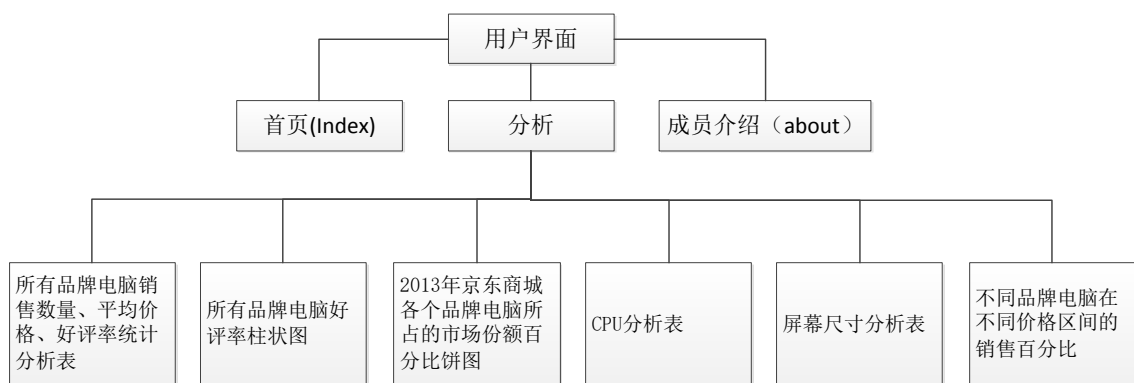
■ 给所有界面视图分配唯一的标识符

表格 2 界面视图汇总表

序号	界面名称	界面标识	功能说明
1	首页 (Index)	index. jsp	介绍项目背景和项目目的
2	成员分工 (About)	about. jsp	介绍小组成员及分工
3	所有品牌电脑销售数量、平均价格、好评率统计分析表	statistics_base_brand_combo. jsp	根据笔记本的各种品牌，分别给出不同笔记本电脑品牌的销售数量、平均价格、好评率，并对其进行比较
4	所有品牌电脑好评率柱状图	analysis_item_1. jsp	比较不同品牌电脑好评率

5	2013 年京东商城各个品牌电脑所占的市场份额百分比饼图	analysis_item_2.jsp	比较不同品牌电脑在 2013 年一年的市场份额
6	CPU 分析表	cpu_analysis_combo.jsp	比较不同品牌和型号的 cpu 被应用在电脑上的数量百分比 比较应用较少数几种应用较广泛的 cpu 的电脑的销售数量和平均价格
7	屏幕尺寸分析表	size_analysis_combo.jsp	比较不同大小的屏幕尺寸被应用在电脑上的数量百分比 比较应用较少数几种应用较广泛的屏幕尺寸的电脑的销售数量和平均价格
8	不同品牌电脑在不同价格区间的销售百分比	analysis_item_3.jsp	对不同品牌电脑统计比较其在不同价格区间的销售比例

■ 绘制各个界面之间的关系图



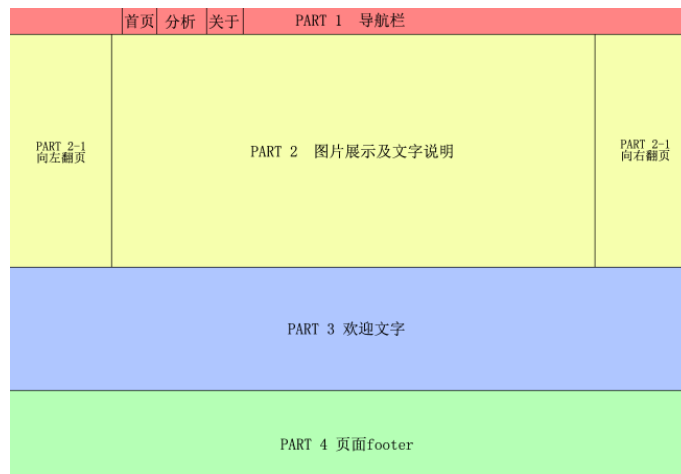
图表 10 界面关系图

■ 各个界面之间的访问流程
各个界面可以实现相互跳转。

(3) 各个界面视图布局

■ 首页界面：

(1) 绘制首页视图；



图表 11 首页视图布局

(2) 说明主界面中所有对象的功能和操作方式；

本界面主要分为 4 部分，分别为导航栏、图片展示及文字说明区、欢迎文字区、页面 footer 区

a. 导航栏: 首先在该栏最前边标注了一个本项目主要功能，其后面的三个分别为连接回首页、分析页面和用户分工的页面，其中分析会做成下拉菜单，让用户可以更加方便的选择需要的分析图表。

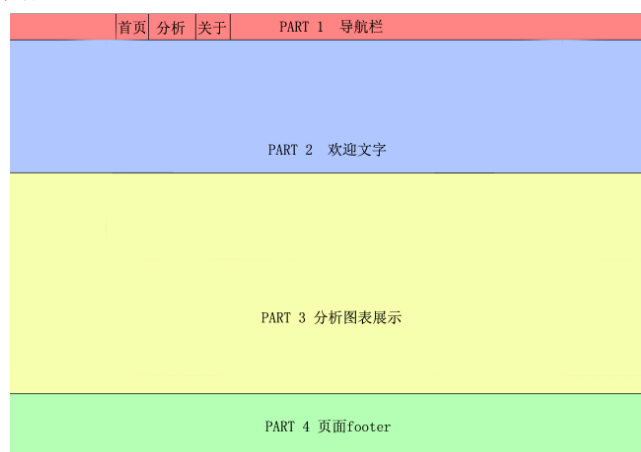
b. 图片展示及文字说明区: 通过图片滚动或是用户操作左右两边的向左向右按钮来查看项目的背景、内容和目的的图片 and 文字。

c. 欢迎文字区: 本栏目显示欢迎语。

d. 页面 footer 区: 本栏目项目版权归 java 第四小组所有。

■ 分析界面:

(1) 绘制分析视图;



图表 12 分析界面视图布局

(2) 说明子界面 A 中所有对象的功能和操作方式;

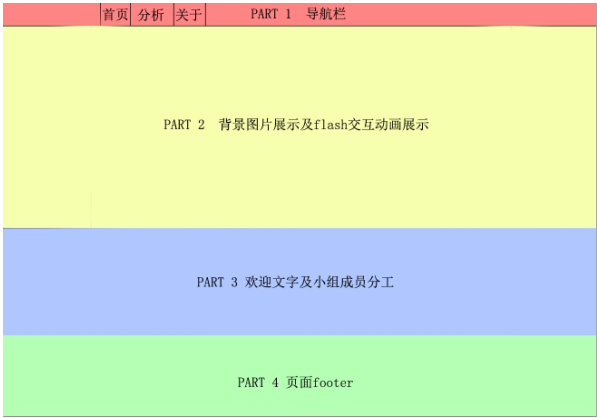
本界面主要分为 4 部分，分别为导航栏、欢迎文字区、分析图表展示区、页面 footer 区。

a. . 导航栏: 首先在该栏最前边标注了一个本项目主要功能，其后面的三个分别为连接回首页、分析页面和用户分工的页面，其中分析会做成下拉菜单，让用户可以更加方便的选择需要的分析图表。

- b. 欢迎文字区：本栏目显示欢迎语和感谢并引用 highcharts。
- c. 分析图表展示区：本栏目显示通过引用 highcharts 和分析数据得出的分析图表。
- d. 页面 footer 区：本栏目项目版权归 java 第四小组所有。

■ 成员分工展示界面：

(1) 绘制成员分工展示界面视图：



图表 13 成员分工展示界面布局

(2) 说明子界面 A 中所有对象的功能和操作方式：

本界面主要分为 4 部分，分别为导航栏、背景图片展示及 flash 交互动画展示区、欢迎文字及小组成员分工展示区、页面 footer 区

- a. . 导航栏: 首先在该栏最前边标注了一个本项目主要功能，其后面的三个分别为连接回首页、分析页面和用户分工的页面，其中分析会做成下拉菜单，让用户可以更加方便的选择需要的分析图表。
- b. 背景图片展示及 flash 交互动画展示区: 本栏目显示小组动态成员名字和照片介绍。
- c. 欢迎文字及小组成员分工展示区：本栏目与背景图片展示及 flash 交互动画展示区进行交互，当点选某个小组成员名字时，本栏目显示此成员的分工。
- d. 页面 footer 区：本栏目项目版权归 java 第四小组所有。

(4) 各个界面视图布局

■ 阐述界面的布局及理由


以大图和简要文字作为页面布局，使页面显得简洁易懂，使用户容易抓住重点。

■ 阐述界面的色彩及理由

整体界面以蓝、白色为主基调，突出商业性，给人以专业可靠的感觉。

(5) 界面资源

表格 3 图标资源表

序号	图标样式	图标标识	功能说明
1		favicon.png	用做回到顶部和网页 logo

表格 4 图像资源表

序号	图像样式	图像标识	尺寸规格	功能说明
----	------	------	------	------

1		01-1. jpg	1366x775	首页展示, 用来说明大数据时代
2		02. jpg	1366x775	首页展示, 用来说明精确分析的重要性
3		05. jpg	1366x775	首页展示, 用来说明项目目的
4		06. jpg	1024*683	成员分工页面展示, 用来说明一个团队

(6) 界面相关技术总结

相关技术包括以下几种:

■ Jsp

JSP (Java Server Pages)是由 Sun Microsystems 公司倡导、许多公司参与一起建立的一种动态网页技术标准。**JSP** 技术是在传统的网页 **HTML** 文件(*.htm,*.html)中插入 **Java** 程序段(**Scriptlet**)和 **JSP** 标记(**tag**), 从而形成 **JSP** 文件, 后缀名为(*.jsp)。由于本次的项目使用了 **MVC** 架构模式, 因此页面不再涉及复杂的控制操作即 **jsp** 标签和 **java** 程序段, 页面的作用只是用来对分析结果进行显示, 而其余的工作交由控制器 **sevlet** 来做。

■ Javascript

JavaScript 是一种基于对象和事件驱动并具有相对安全性的客户端脚本语言。同时也是一种广泛用于客户端 **Web** 开发的脚本语言, 常用来给 **HTML** 网页添加动态功能。例如成员分工界面中可以进行 **flash** 动画与 **js** 交互, 并利用 **js** 对 **html** 表单控件进行修改的操作, 代码片段如下:

过程为 **flash** 通过动作代码在响应到 **flash** 的鼠标事件后, 调用对应 **jsp** 中的 **javascript** 中的 **clickMeToSendTheValueToText** 函数, 并将参数传给它, 然后利用 **DOM**, 定位 **html** 表单控件 **ID**, 最后将值传给 **html** 表单控件, 并显示在页面上

```
//flash动作代码
import flash.net.*
import flash.external.*;
ball_3.addEventListener(MouseEvent.CLICK, goto);
function goto(e:MouseEvent):void{
    var temp_str =
```



```
String(ExternalInterface.call("clickMeToSendTheValueToText", "马笑", "
负责页面展示、文档撰写")));
    }

//html标签及表单控件
<h1 class="row" align="center">
<input type="text" id="text_1" value="" align="center"
readonly="readonly"
style="color:#000000;text-align:center;font-size:30px;
BORDER-BOTTOM: #000000 0px solid;
    BORDER-LEFT: #000000 0px solid;
    BORDER-RIGHT: #000000 0px solid;
    BORDER-TOP: #000000 0px solid;" >
</h1>

//javascript方法
<script language="javascript">
function iHaveAReturnValue1(txt){
    return txt;
}
function iHaveAReturnValue2(txt){
    return txt;
}
function clickMeToSendTheValueToText(txt1,txt2){
    document.getElementById("text_1").value =
iHaveAReturnValue1(txt1);
    document.getElementById("text_2").value =
iHaveAReturnValue2(txt2);
}
</script>
```

图表 14 js 的应用代码片段图

■ Highcharts

Highcharts 是一个用纯 JavaScript 编写的一个图表库，能够很简单便捷的在 web 网站或是 web 应用程序添加有交互性的图表，并且免费提供给个人学习、个人网站和非商业用途使用。HighCharts 支持的图表类型有曲线图、区域图、柱状图、饼状图、散状点图和综合图表。本项目主要选取这个类库来进行分析数据的展示。调用库的方法即引入外部已经写好的.js 文件，如：

```
<script type="text/javascript"
src="js/jquery-1.10.2.min.js" ></script>
<script src="js/highcharts.js"></script>
<script src="js/exporting.js"></script>
```

图表 15 引入外部 js 文件的代码片段图

依据 Highcharts 的架构，写 js 函数并向其中填入数据即可生成图表。

■ CSS

是能够真正做到将网页表现与内容分离的一种样式设计语言。相对于传统 HTML 的表现而言，CSS 能够对网页中的对象的位置排版进行像素级的精确控制，支持几乎所有的字体字号样式，拥有对网页对象和模型样式编辑的能力，并能够进行初步交互设计，是目前基于文本展示最优秀的表现设计语言。CSS 能够根据不同使用者的理解能力，简化或者优化写法，针对各类人群，有较强的易读性。

样例如下：

```
body {
  padding-bottom: 40px;
  color: #ffffff;
}
```

图表 16 CSS 代码片段图

2.3 数据库分析与设计

（1）数据库表设计

数据库主要包含两张表 `urls` 和 `items`，这两张表分别存储爬取到的 `url` 信息和分析到的商品信息。数据库的两张表如下：

表一：urls

表格 5 urls 数据库字段及说明表

UrlMD5	url 的 MD5
Url	url
Depth	url 对应的深度（列表页为 0，简介页为 1）
IsAnalyzed	是否已经分析
InsertTime	插入时间

表二：items

表格 6 items 数据库字段及说明表

itemUrl	商品 url
itemUrlMD5	url 对应 md5
detailUrl	商品详细 url
commentUrl	商品评论 url
Price	价格
Title	标题
Brand	电脑品牌
Model	型号
Color	颜色
Platform	平台
os	操作系统
CpuType	Cpu 类型
Ram	内存
HardDrive	硬盘
CdRom	光驱
Monitor	显示器

Weight	重量
Size	大小
IsDetailAnalyzed	商品详细是否已经分析（0-未分析，1-已分析）
IsCommentAnalyzed	商品评价是否已经分析（0-未分析，1-已分析）
GoodCommentNum	好评数
MediumCommentNum	中评数
BadCommentNum	差评数
commentContent	评价内容

（2）数据库管理

在数据采集部分，四个子模块都使用了数据库进行数据的存储和读取，而每个用户使用的数据库的配置都可能不一样。为了能使系统能够更加方便地修改数据库的主机，用户名，密码，数据库名等配置信息，数据搜集部分使用了统一的数据库管理模块。

这数据库管理模块对应 DBMnager 类，这个类实现了数据库的初始化，表的创建，数据库连接的获取等常用的功能。

3. 项目特色

运用 Git 对版本控制，加强了小组成员的协作关系，也对于项目更好的开展产生了巨大积极的影响。

4. 小组成员分工

表格 7 小组成员分工表

Java 第 4 组小组成员分工	
范英明	负责整合各模块，编写控制器 servlet，管理 Git 仓库；
牛童	负责文档撰写与修改；
李洪宇	负责用户界面模块的设计与实现；
马笑	负责文档撰写和用户界面模块的设计与实现；
王贵存	负责数据分析模块的设计与实现；
赵仕荣	负责数据采集模块的设计与实现；
李轶	负责系统测试的设计与执行。

5. 致谢

本学期的 JAVA 高级技术课程即将接近尾声，借着大项目检查之际向老师和助教表示由衷地感谢，在本学期的 JAVA 课程，要感谢褚伟杰老师，您对于每个知识点的讲解都非常详

细、清晰，对于同学的学习非常负责任，要求严格，并且对同学们进行无私的指导和帮助，充分照顾没有学过 JAVA 的同学的学习感受，这让我们在本学期的 JAVA 课程中学到了许多以前忽略的细节，并且将课程中学到的知识充分地应用到了项目中，使项目的进程较为顺利，这也使我们感受到了老师的良苦用心，再次衷心的感谢您和助教。