基于 WebMagic 垂直网络爬虫的 CSDN 博客分析

课程名称：《Java 编程技能训练》

成员名单：

姓名	学号	备注
贾世琳	1501210924	组长
董志	1501210894	
曹天元	1501210876	

2016 年 1 月

目　　录

1. 系统简介

使用 WebMagic 网络爬虫架构，从 CSDN 爬取技术类博客，从整个博客网页上分析出技术博客的标题、时间、作者、浏览次数和文章内容，并保存网页中需要的链接，将所有的数据存入数据库中，通过网站对抓取的技术类博客按照用户要求进行分类展示、搜索和统计等功能。通过将 CSDN 技术类博客进行汇总分类，方便使用者更容易找到所需要的技术博客。如按编程语言、数据库等进行分类。通过将抓取的技术博客按浏览次数进行排序，将浏览次数最高的博客展示给用户。对编程语言博客数量进行分析统计，使用户了解当前热门的编程语言以及高质量博客。

2. 相关技术

本次实验采用的相关技术有：WebMagic 垂直网络爬虫、Mysql 数据库、Hibernate 对象关系映射框架和 MVC 架构:Jsp+Servlet。

3. 系统框架

3.1 总体架构

总体分为三部分：网络数据爬取，数据存储和结果展示。

网络数据爬取:使用 WebMagic 垂直网络爬虫架构，爬取 CSDN 技术类的相关博客。

数据存储:将技术类博客存入 Mysql 数据库中，采用 Hibernate 对象关系映射框架。

分析结果展示:通过网站对抓取的技术类博客按照用户要求进行分类展示、搜索和统计等功能，采用 MVC 架构：Jsp+Servlet。

总体架构如下图所示。

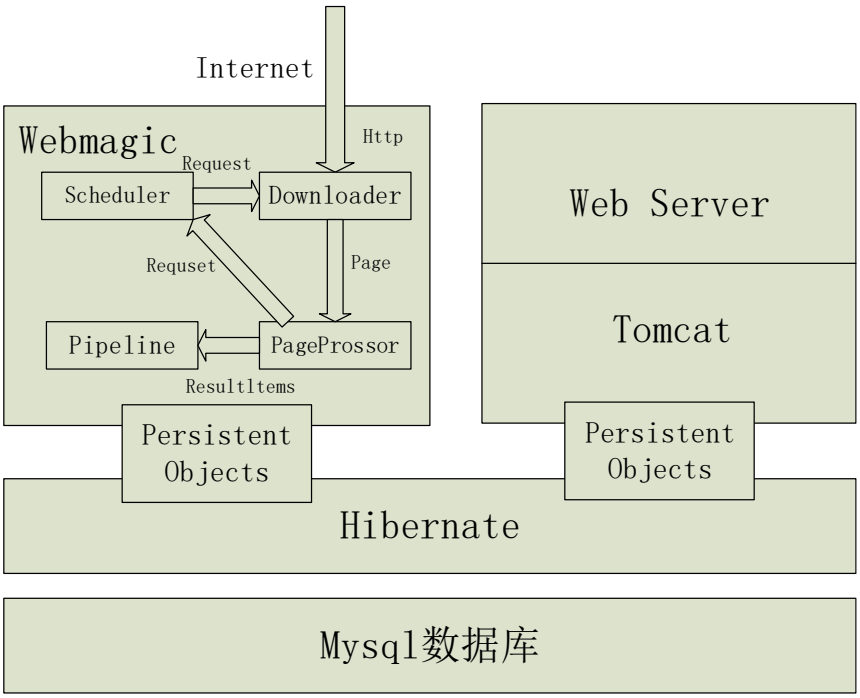

图 3.1　总体架构图

3.2　WebMagic 爬虫架构

WebMagic 的结构分为 Downloader、PageProcessor、Scheduler、Pipeline 四大组件，并由 Spider 将它们彼此组织起来。这四大组件对应爬虫生命周期中的下载、处理、管理和持久化等功能。WebMagic 的设计参考了 Scapy，但是实现方式更 Java 化一些。而 Spider 则将这几个组件组织起来，让它们可以互相交互，流程化的执行，可以认为 Spider 是一个大的容器，它也是 WebMagic 逻辑的核心。WebMagic 总体架构如图 3.2 所示。

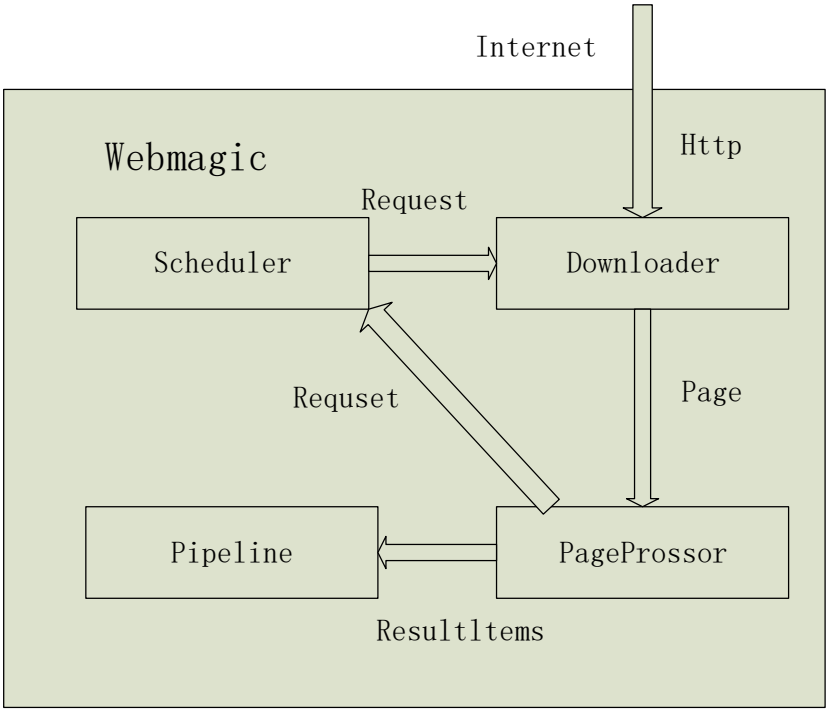

图 3.2　WebMagic 架构图

3.2.1　WebMagic 的四个组件

1) Downloader

Downloader 负责从互联网上下载页面，以便后续处理。WebMagic 默认使用了 Apache HttpClient 作为下载工具。

2) PageProcessor

PageProcessor 负责解析页面，抽取有用信息，以及发现新的链接。WebMagic 使用 Jsoup 作为 HTML 解析工具，并基于其开发了解析 XPath 的工具 Xsoup。在这四个组件中，PageProcessor 对于每个站点每个页面都不一样，是需要使用者定制的部分。

3) Scheduler

Scheduler 负责管理待抓取的 URL，以及一些去重的工作。WebMagic 默认提供了 JDK 的内存队列来管理 URL，并用集合来进行去重。也支持使用 Redis 进行分布式管理。

4) Pipeline

Pipeline 负责抽取结果的处理，包括计算、持久化到文件、数据库等。WebMagic 默认提供了“输出到控制台”和“保存到文件”两种结果处理方案。Pipeline 定义了结果保存的方式，如果你要保存到指定数据库，则需要编写对应的 Pipeline。对于一类需求一般只需编写一个 Pipeline。

3.2.2 用于数据流转的对象

1) Request

Request 是对 URL 地址的一层封装，一个 Request 对应一个 URL 地址。它是 PageProcessor 与 Downloader 交互的载体，也是 PageProcessor 控制 Downloader 唯一方式。除了 URL 本身外，它还包含一个 Key-Value 结构的字段 extra。你可以在 extra 中保存一些特殊的属性，然后在其他地方读取，以完成不同的功能。例如附加上一个页面的一些信息等。

2) Page

Page 代表了从 Downloader 下载到的一个页面——可能是 HTML，也可能是 JSON 或者其他文本格式的内容。Page 是 WebMagic 抽取过程的核心对象，它提供一些方法可供抽取、结果保存等。在第四章的例子中，我们会详细介绍它的使用。

3) ReusltItems

ReusltItems 相当于一个 Map，它保存 PageProcessor 处理的结果，供 Pipeline 使用。它的 API 与 Map 很类似，值得注意的是它有一个字段 skip，若设置为 true，则不应被 Pipeline 处理。

3.2.3 控制爬虫运转的引擎--Spider

Spider 是 WebMagic 内部流程的核心。Downloader、PageProcessor、Scheduler、Pipeline 都是 Spider 的一个属性，这些属性是可以自由设置的，通过设置这个属性可以实现不同的功能。Spider 也是 WebMagic 操作的入口，它封装了爬虫的创建、启动、停止、多线程等功能。

3.3 分析框架

软件架构采用经典的 MVC 三层架构模式，即表示层（View）、控制层（Controller）、数据层（Model）。区分层次的目的即为了“高内聚，低耦合”的思想。

表现层（View）包含表示代码、用户交互 GUI、数据验证。该层用于向客户端用户提供 GUI 交互，它允许用户在显示系统中输入和编辑数据，同时 系统提供数据验证功能，用户通过表示层查询相关的数据并将想要的博客内容进行展示。

控制层（Controller）：针对具体问题的操作，也可以说是对数据层的操作，对数据业务逻辑处理，此层对博客数据进行数据分析。

数据层（Model）：该层所做事务直接操作数据库，针对数据的增添、删除、修改、更新、查找等，本实验采用 Mysql 数据对数据存储。

系统的软件架构图如图 3.3 所示。

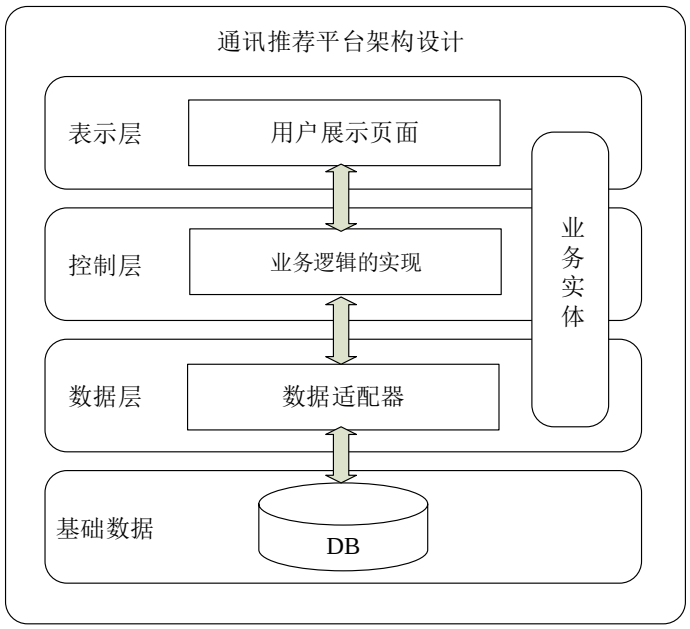

图 3.3　原型架构图

4. 实验结果

通过实现 PageProcessor 接口进行网页元素获取和链接的发现，本实验获取的页面元素有：用户的 ID、文章标题、提交时间、阅读次数、标签和正文内容。用正则表达式从 HTML 页面获得需要的网址格式，将匹配的链接加入到待抓取的队列中。网络爬虫的爬取入口地址为：http://blog.csdn.net/。将所获得的数据通过 hibernate 对象存入 Mysql 数据库中将其进行持久化。通过网站与用户进行可视化交互，用户通过网页提交相关请求，在数据库中查找有关数据在网页上展示结果。如图 4.1 和 4.2 所示，为网站首页展示结果，首页推荐五篇最近热门文章。

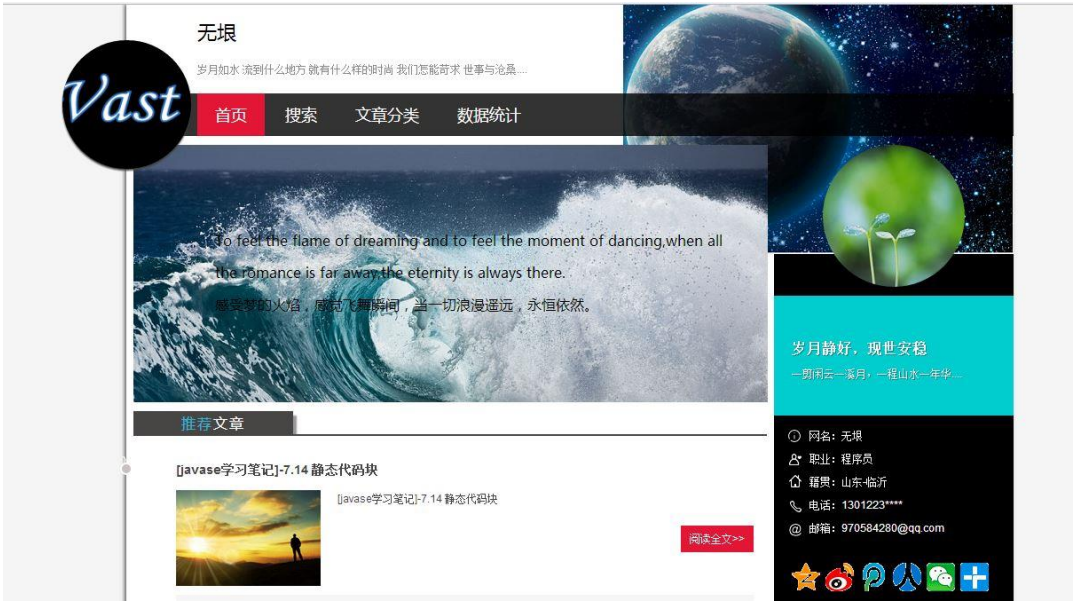

图 4.1　首页展示 1

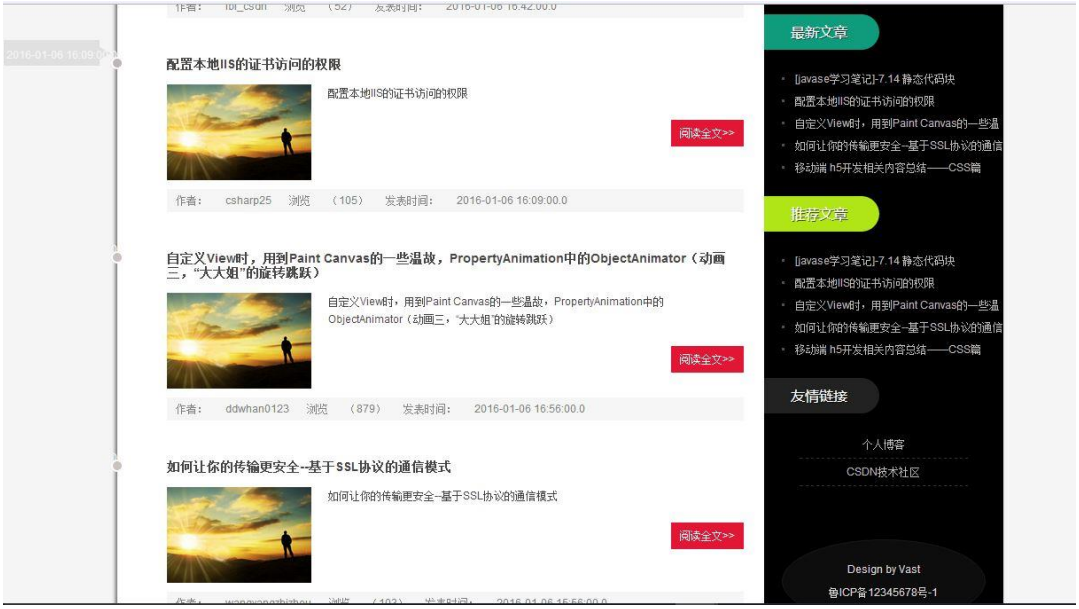

图 4.2　首页展示 2

图 4.3 所示，为搜索页面，通过填入所要搜索的关键词，在数据库中通过与文章标题、标签和内容进行字符串匹配，返回相关的技术文档，搜索后的返回结果如图 4.4 所示。

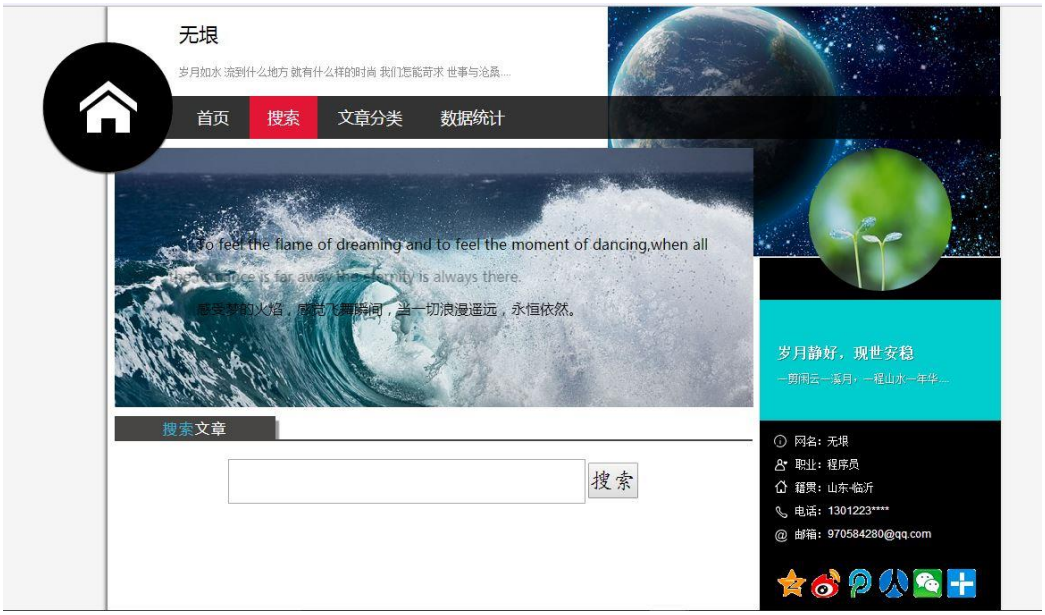

图 4.3　搜索页

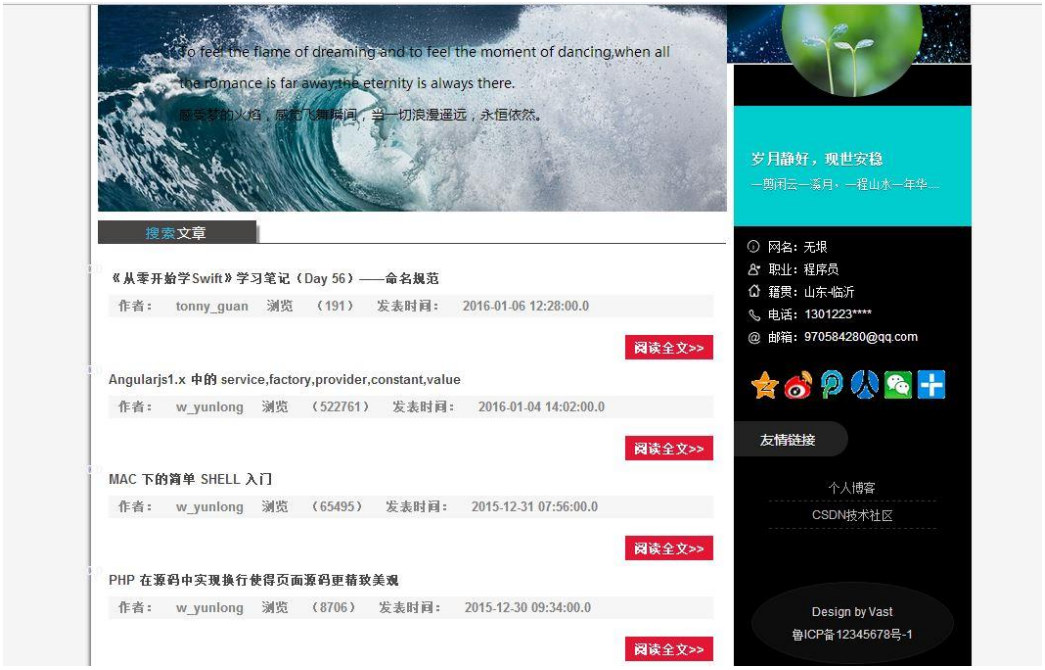

图 4.4　搜索结果

图 4.5 为文章分类页，将数据库中的文章分为编程语言、数据库、Web 前端、云计算、移动开发和其他六类，每一类又有子类，通过点击相应的标签，返回文章分类结果如图 4.6 所示。

图 4.5　文章分类

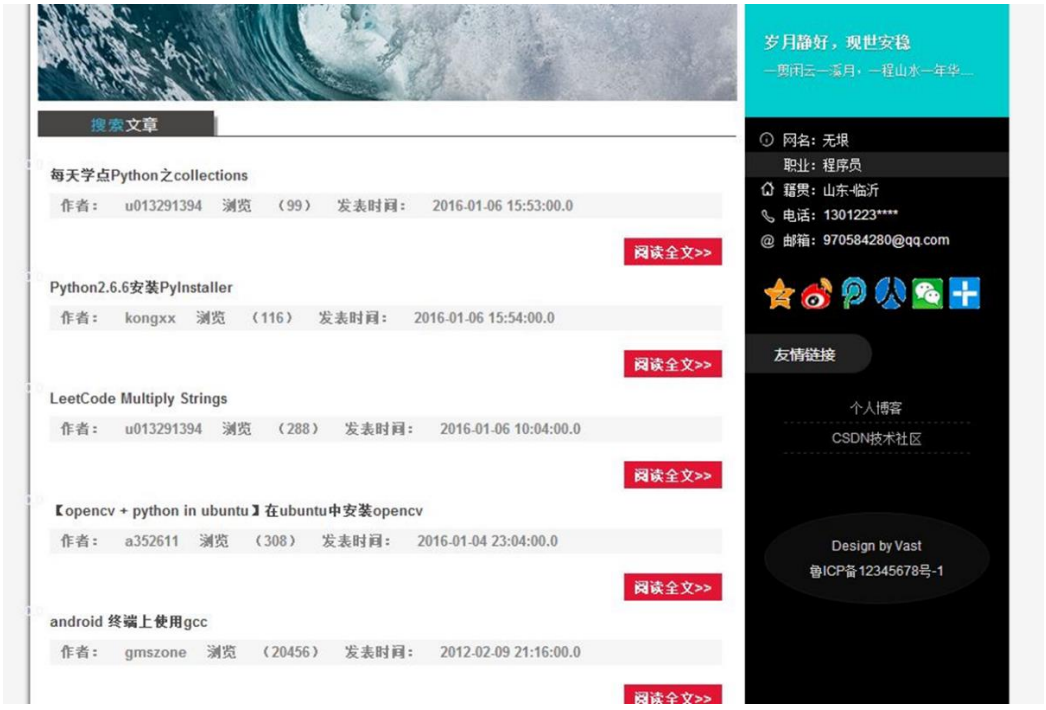

图 4.6　文章分类结果

图 4.7 为数据统计展示页，分别显示本次数据库中的文章数目，本次共爬取了 67733 条数据，浏览次数最多的文章以及根据博客数据统计出最热门的编程语言，最热门的语言是 java，总共有 5143 条数据。

图 4.7　数据统计

5. 总结

Java 编程技能训练是一门理论和实践性都很强的专业基础课，编程技能只能通过实际验证才能加深理解并真正掌握，实验就是使学生加深理解所学基础知识，对各门知识得到融会贯通的认识和掌握，通过本次实验的学习，学会了使用 WebMagic 垂直网络架构爬虫和 Mysql，并掌握了可视化展示相关技术，培养在实际综合应用中研究问题，分析问题和解决问题的能力。实现了利用爬虫爬取 CSDN 的技术博客，将所爬取的技术博客利用 Mysql 进行持久化存储，通过网站与用户进行交互，按照用户要求进行搜索、分类展示和统计等功能，实现根据数据库中的数据进行数据分析，使用户了解当前热门的计算机技术。