Eworm 系列文档之

作品说明书

开发团队：刘家鸣 虞剑飞 吕游
指导教师：王树梅 赵洋

2012 年 6 月

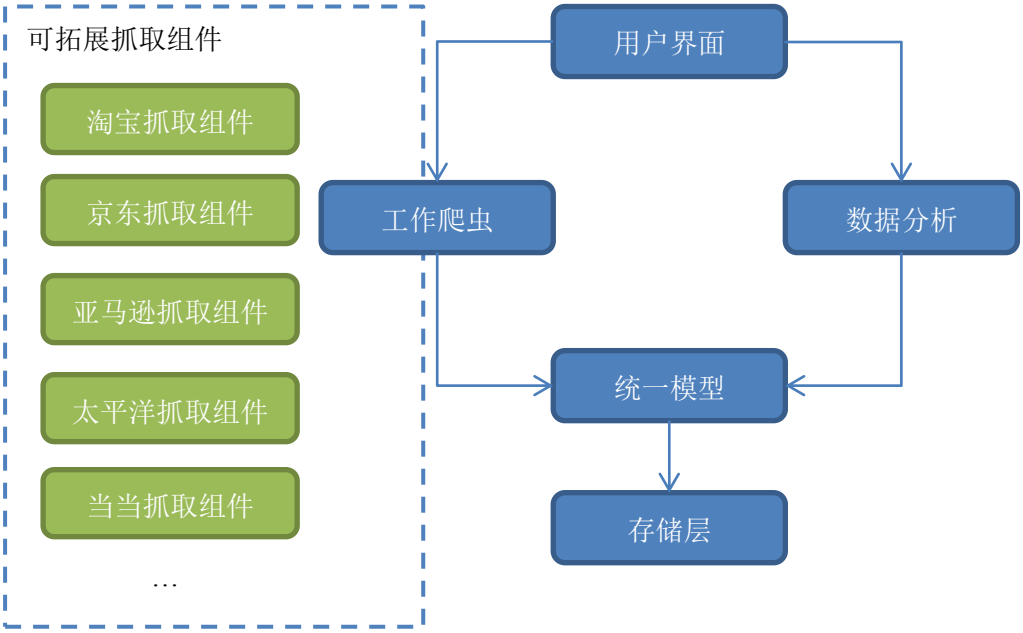
一、 开发背景

随着互联网的迅猛发展以及 B2C、B2B、C2C 网站的成长，足不出户的网络购物已经在人们的生活中扮演着越来越重要的角色，但是随之而来想要从各种网站中找到最好、最划算的商品也非常困难，所以提供一个高质量的在线商品搜索平台无疑会为人们提供更多的便利以及给网络购物带来更好的未来。但是传统搜索引擎所返回的结果包含大量用户不关心的信息，并且他们往往对这些信息含量密集且具有一定结构的数据无能为力，不能很好地发现和获取这些信息，因而人们已经不能满足于使用传统搜索引擎，他们对聚焦于网上商品的搜索引擎的需求正与日俱增。

本系统 Eworm 就是根据上述需求设计的，它能够匹配用户输入的关键字，将各大购物网站上与该关键字有关的商品呈现在用户面前。如果用户对一件商品感兴趣，用户可以通过 Eworm 及时地了解该商品在各大网站的销售以及评价详情，从而选择心仪的卖家。

二、 系统结构

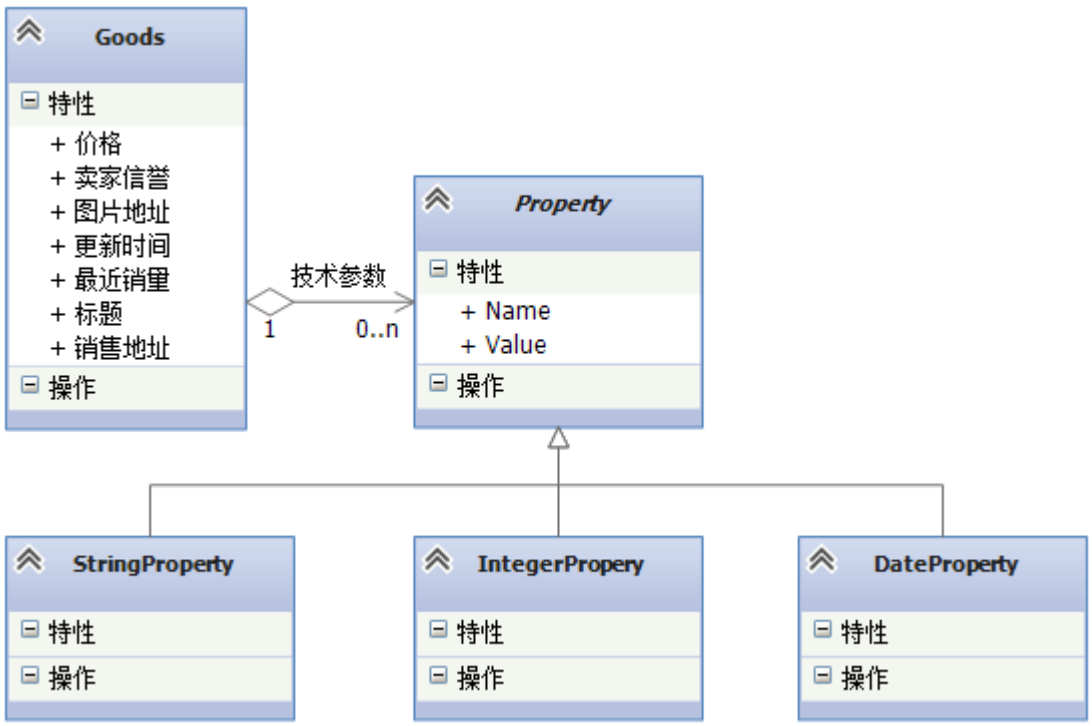
系统的整体结构如图表 二-1 所示，组件分为五大块，分别是用户界面、工作爬虫、数据分析、统一模型、存储层。



图表 二-1 系统结构示意图

三、 模型与存储

本着 KISS（Keep It Simple And Stupid）的原则，设计了一个简单而又能存储所有商品信息的模型。模型定义如图表 三-1 所示：



图表 三-1 统一模型示意图

商品本身只维护业务相关的字段（如价格、销量、卖家信誉需要用来计算优先程度），而技术参数用户关心系统不关心，所以只需呈现。

如此一来，存储变得非常简单，仅需要两张表：t_goods 和 t_property。前者保存商品的信息，后者保存商品的技术参数信息。

四、 抓取组件发现机制

各大商业网站的网页抓取组件是通过 MEF 框架发现载入的，这样有利于后续的拓展 – 无需更改任何代码即可实现网站支持的增加。

抓取组件需要包含一定的元数据和实现一定的接口，元数据是用 Attribute 打在类上的。而接口则是程序集给出，指明了一个抓取组件需要做的事情。

组件元数据

组件元数据描述抓取组件自身的属性。包括以下属性：

1. Guid，用来唯一标识该组件
2. Name，支持的网站的名称（如 Dangdang 或 Taobao）
3. Url，支持的网站的地址（如 http://www.taobao.com）
4. Disabled，用于表示组件是否暂时被停用，停用的组件在发现时被忽略

下面是淘宝抓取组件的元数据表示的一个实例：

```
[GoodsFetcher(
    guid: "525E1313-1E04-47C2-A05A-D93079865079",
    name: "Taobao",
    url: "http://www.taobao.com",
```

```
        disabled: false)]
    public class TaobaoItemFetcher : IGoodsFetcher {…}
```

组件接口

组件接口描述组件要实现的功能，接口定义也很简单：

1. 给定的搜索关键字和限制数量，给出需要抓取的页面的 Uri 列表
2. 给定属于上述列表的 Uri，给出商品抓取的结果

下面是接口的原始定义：

```
    public interface IGoodsFetcher
    {
        IEnumerable<Uri> GetGoodsUriByKeyowrd(string keyword, int count);
        Goods FetchGoods(Uri goodsUri);
    }
```

发现过程

组件的发现的过程是这样的：

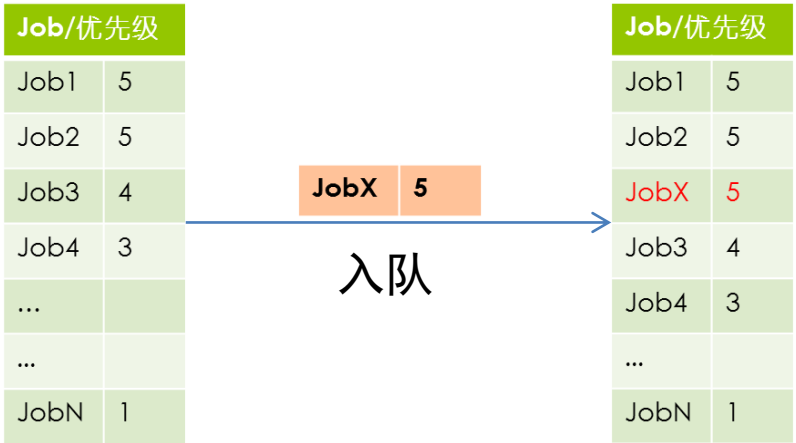
1. 爬虫的组件发现器（FetcherManager）静态初始化会调用 Init 方法
2. Init 方法会枚举当前程序集目录底下的所有 dll 文件
3. 加载这些 dll 文件的类型
4. 将满足契约的类型组合到组件发现器中，可以开始服务

五、 爬虫工作原理

任务队列

任务队列是一个有限队列，队列存放的是任务。每个任务对应一个优先级，如果新的任务入队，则应该加入到优先级刚大于新任务的任务之后。如图表 五-1 所示。

对于队列长度的增长有两个限制：队列长度超过 1000 时，扔出最后的任务；优先级小于 X 的，不予入队。此处 X 是个阈值，目前该值为 6。



图表 五-1 有限任务队列示意图

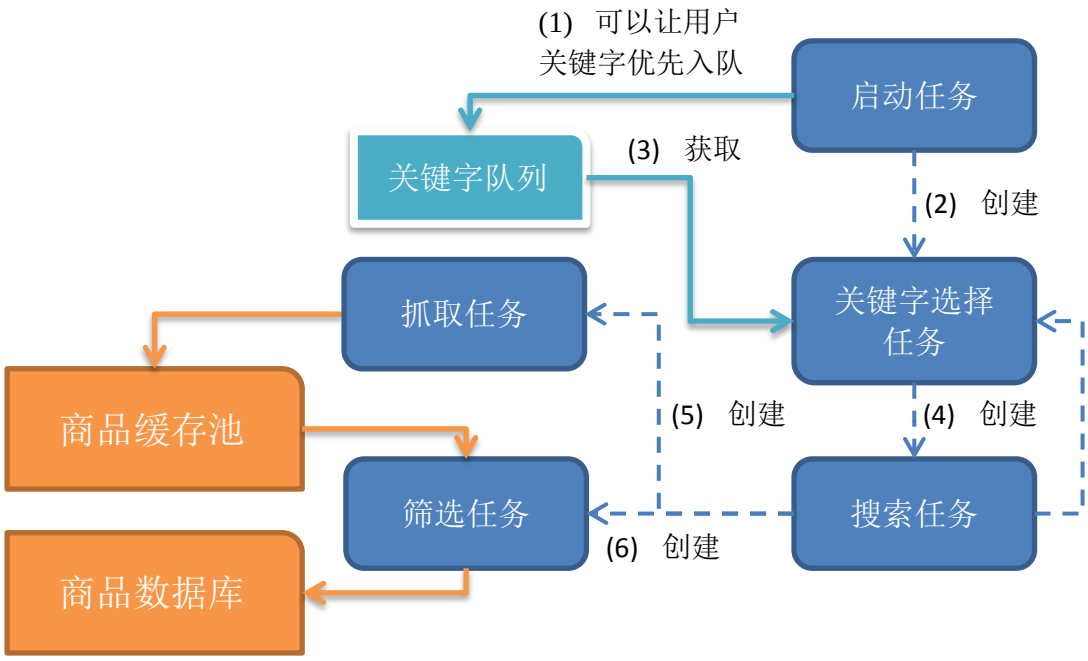
任务类型

任务分为四类：

1. 启动任务。负责爬虫的启动，并且响应用户输入的关键字。
2. 搜素任务。负责搜索指定的关键字，然后将商品的 Url 记录下来，创建抓取任务进行抓取。
3. 抓取任务。抓取商品的详细信息到缓存池中。
4. 筛选任务。筛选缓存池中的商品持久化到数据库中。
5. 关键字选择任务。负责选择下一个关键字，创建新的搜索任务。

任务拓扑

任务拓扑如所图表 五-2 任务拓扑说明示。



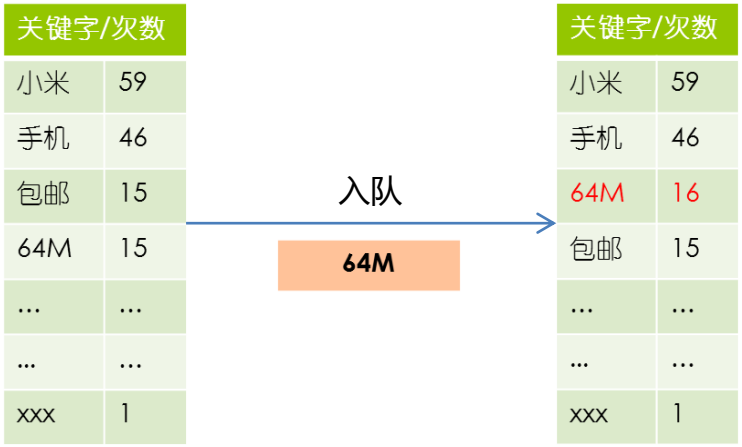
图表 五-2 任务拓扑说明

任务优先级确定

启动任务具有最高的优先级，为 10
关键字选择任务的优先级是创建者优先级减一
其他任务的优先级同创建者的优先级

关键字队列

用户搜索输入和抓取过程中，都会产生新的关键字。这些关键字将入队到关键字队列。关键字队列同样是一个优先队列，队列的优先级是关键字被入队的次数。如果一个关键字入队次数刚好大于上一个高优先级关键字，则与之交换。如所示。



图表 五-3 关键字队列

结果的筛选

对于缓存池中存放着已经被抓取的商品，筛选任务负责从缓存池中筛选出按一定条件排序的 X%个商品。主要根据商品的两个属性：卖家信誉（C）和商品销量（S）来计算这个商品的受欢迎度，公式如下：

$$R_i = \begin{cases} \frac{50C_i}{\sum C_i} + \frac{50S_i}{\sum S_i}, & C_i > 0 \text{ 且 } S_i > 0 \\ \frac{100C_i}{\sum C_i}, & C_i > 0 \text{ 且 } S_i = 0 \\ \frac{100S_i}{\sum S_i}, & S_i > 0 \text{ 且 } C_i = 0 \end{cases}$$

商品池的商品根据 Ri 的值降序排列。

六、 数据分析

这一部分的功能尚未实现。

七、 用户界面

用户界面主要包含两大部分：搜索和监视。

搜索

搜索是接受用户输入的关键字，进行两个操作：

1. 从数据库中取出结果回显给用户
2. 通知爬虫有新的用户关键字

搜索界面如图表 七-1 所示。

图表 七-1 搜索界面

监视

监视界面中，可以实时监控爬虫的工作状态（适合个人安装使用）。
监视的界面如所示。

图表 七-2 监视界面

八、 系统特点

EWorm 能够对时下最流行的六大购物网站进行商品整合，使用户高效便捷地了解各大网站商品信息的对比。除此之外，实现上还有很多架构的优势：

高拓展性和可维护性：由于采用组件发现机制，所以可以很方便的拓展新的网站支持。模块化的设计也使得系统具有高度可维护性。

人性化的操作界面：系统采用 B/S 设计模式，提供了简洁易操作的网页界面。

附录：部署说明

在线版本

如果部署有难度，或者部署失败，可以浏览线上版本：http://icpc.njust.edu.cn/Eworm

由于服务器资源有限，只给持有本说明书的人访问。如果访问要求输入口令，请输入（区分大小写）：

Csc-eWrom-2012

系统要求

操作系统：Windows Server 2003 / Windows Server 2008 / Windows 7/ Windows XP

服务器软件：IIS6.0、IIS6.5 或 IIS7.0

.Net 组件：.Net Framework 4.0，ASP.NET MVC3

数据库软件：MySql5.5.16 或以上，MySql Connector Net 6.5.4

部署步骤

1. 保证上述软件安装完成，没有达到系统要求的，需要安装相应的软件。
2. 解压“EWormDeploy.zip”到合适的位置
3. 将 IIS 默认网站位置配置为解压目录内的“Website”目录或新建虚拟目录指向该目录
4. 用 MySql 执行解压目录内的“Database.sql”（或者使用 Workbench 导入）
5. 进入 Website 目录，打开 Web.config，配置连接字符串。根据数据库的位置，填写主机、用户名和密码。
6. 用浏览器定位 IIS 网站位置便可以浏览

浏览器兼容性说明

由于时间仓促，CSS 文件没有兼容 IE9 以下的浏览器，外观可能会不如意。用 Chrome 浏览器能获得最佳浏览效果。