基于规则引擎的大规模网页信息抽取平台设计与实现

任宪臻　朱　义

(北京信息职业技术学院　北京　100018；清华大学信息技术研究院　北京　100084)

摘　要：信息抽取是数据挖掘和知识发掘的重要方法，基于规则自动化或半自动化地从互联网中提取准确有效的数据是知识挖掘的关键。本文构建了一个通用文本信息抽取平台，采用多种信息匹配技术从网络数据源中抽取数据和信息，并采用规则处理方式对网页信息进行智能化抽取。该平台采用 Eclipse RCP 开发，对其功能可进行插件式扩充，在业务逻辑上采用规则引擎。该平台具有界面友好、易于扩展、使用方便等特点，并能够从大规模网页中自动地获取有效的数据和信息。

关键词：信息抽取；规则引擎；富客户端平台；增量爬取

中图分类号：TP3　　**文献标识码：**A　　**文章编号：**1673－4513（2010）－05－067－04

1、前言

信息和知识的获取就是从大量数据中获取有效的、新颖的、准确的、最终可理解的模式的非平凡过程。准确的说，信息抽取就是从大量数据中提取符合要求数据的过程。截止到 2009 年 4 月底，全球互联网网站数量已经超过 2.3 亿，互联网已成为最大的信息库。而网络上信息格式多种多样，同一网站上的信息虽然具有一定的相关性，但又经常有新的表现形式，不同网站的信息差异性更大。当爬虫程序在网络上爬取网页的时候，需要对获取的链接进行字符串处理，得到有效链接，从而实现聚焦爬取。数据提取工具对指定的网络数据的处理时，如果需求固定，通常只需要几个简单的判断处理就可以完成，但如果需要对任意网站链接处理，且需求不断变化，就不得不重新编写代码并编译处理，整个程序就需要不断修改，而后续的处理也需要不断的修改。

因此，如何自动获取有效准确的数据，是一个需要迫切解决的问题。设计并实现一个可扩展性强、简单易用、适用范围广泛的数据抽取平台是实际急切需要的。为此，本文将给出一个采用 Eclipse RCP 开发，使用规则库和规则引擎的处理方式，利用规则引擎来处理复杂业务逻辑，可以实现聚焦爬取数据等特定功能的数据抽取实验平台。该平台的特点是易于扩展、适用范围广，而且可根据需要定制编写工作流扩展功能，能够适用于任意数据源的数据抽取工作。同时，该平台可作为信息发现和知识挖掘的重要部分。

2、相关技术与知识

2.1 RCP

RCP（Rich Client Platform）是基于 Eclipse 项目推出的一个开发富客户端应用框架，目的

收稿日期：2010 年 7 月 8 日

作者简介：任宪臻（1977－），女，硕士研究生，北京信息职业技术学院讲师，研究方向：计算机软件应用开发技术。

朱　义（1973－），男，硕士研究生，清华大学信息技术研究院 WEB 与软件研究中心工程师，研究方向：海量数字媒体管理，数字图书馆。

在于为开发人员提供一个功能强大的、快速的、可扩展的应用平台。瘦客户端应用程序很多情况下无法满足用户要求，富客户端又成为流行的开发模式。但是与早期的富客户端相比，富客户端的内涵有了变化。在需求变化异常频繁复杂的今天，用户不仅希望有丰富的图形用户界面，还希望能够具有智能更新、跨平台、可扩展等特性。而 Eclipse RCP 则可以满足上述需求。简单来说，Eclipse RCP 主要具有以下优点：

1、组件化。基于 Eclipse 的系统设计由被称为 plug－ins 的插件构成，可以通过扩展点进行配置，也可以被不同应用程序共享。

2、便利性。Eclipse RCP 对各个平台下的产品包装提供了强有力的支持，其开发的 RCP 甚至可以在嵌入式设备、掌上电脑上运行。

3、智能安装和升级。Eclipse 提供了专门的 Update 组件，可以实现通过 HTTP 、Web 站点、复制等多种方式进行安装和更新，可以解决富客户端应用部署升级的麻烦。

4、可扩展性。Eclipse 基于插件进行扩展的思想使得用户可以方便地搭建各种规模、类型和用途的应用程序。

5、开发工具支持。目前随着 Eclipse 插件开发环境的日趋成熟和不断流行，基于 Eclipse RCP 的插件越来越多，特别是有大量免费的插件可供下载，这有利于加速产品的研发进程。

6、本地观感及使用体验。Eclipse 为各种操作系统提供了本地图形接口包。

7、脱机操作。由于 RCP 在本机运行，不需要网络连接，可以充分利用本机硬件的处理能力高速进行大量数据的处理。

2.2 规则引擎

规则引擎（Rule Engine）起源于基于规则的专家系统，其核心思想是实现判断分支逻辑与程序代码的分离，实现程序流程的可配置化。规则引擎实现了将业务逻辑从应用程序代码中分离出来，并使用预定义的语义模块编写业务逻辑。规则引擎接受数据输入，解释业务规则，并根据业务规则做出业务决策。

规则引擎是一种根据规则中包含的指定过滤条件，判断其能否匹配运行时刻的实时条件来执行规则中所规定的动作的引擎。它主要包含以下几个部分：规则集（Rule set）、模式匹配器（Pattern Matcher）、执行列表（Agenda）和执行引擎（Execution Engine）。

1、规则集就是许多规则的集合。每条规则包含一个条件过滤器和多个动作。一个条件过滤器可以包含多个过滤条件。条件过滤器是多个布尔表达式的组合，其组合结果仍然是布尔类型。在程序运行时，动作将会在条件过滤器值为真的情况下执行。

2、模式匹配器决定选择执行哪个规则，何时执行规则。

3、执行列表从管理模式匹配器挑选出来的规则的执行次序。

4、执行引擎负责执行规则和其他动作。

3、通用文本信息抽取平台的体系架构设计

信息抽取系统通过网络爬虫、格式转换和规则引擎的结合实现了富客户端应用框架，并将抽取的数据和信息保存到数据库中。整个系统架构如下图 1 所示。从该图中可以看出，该结构层次清晰，功能模块之间的耦合度低，这更符合 RCP 开发的规范。

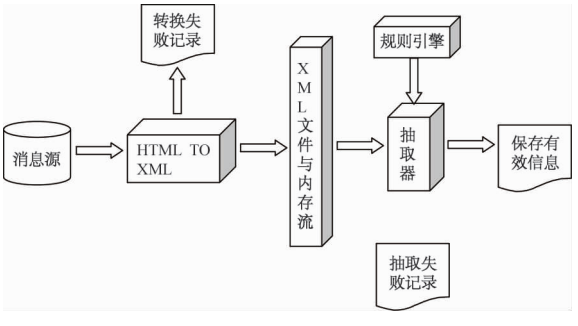

图 1　信息抽取系统架构图

从图 1 可以看出，该系统主要由信息爬取、信息转换、规则引擎、信息抽取、信息保存五部分组成。爬取 web 文本页面通常是 html 格式，通过转换成 xml，然后对 xml 进行按规则进行抽取，可以抽取具体正文内容，也可以抽取包含标签的节点信息。最后需要把抽取成功和

失败的数据分别保存在不同的数据库表中，为数据挖掘和知识发现提供更准确的信息源。

4、功能模块设计

4.1 规则配置功能

该功能模块主要是采用规则引擎配置数据抽取方式，调用规则，实现复杂业务逻辑，如：链接生成、网页爬取、信息抽取、持久化模块、链接生成结果的保存、中间结果的保存、增量爬取的初始化、链接过滤。链接过滤可根据当前处理任务的要求，过滤符合指定规则的链接，对链接进行相关处理，对符合条件的输出，对不符合条件的做保存，供后续分析使用等。

4.2 聚焦爬取功能

该功能主要用来获取网站的历史页面。

大型网站一般都有目录页提供按日期排序的所有文章目录，该功能主要针对这一特点，生成目录页链接的规则集，再获取文章链接和数据。在爬取过程中，应用聚焦爬取；通过设置规则获取链接集合，通过链接判断过滤，对链接做有效性处理，然后判断链接是否下载过，对新链接下载到本地，供多次挖掘使用；该功能是按数据流的方式处理，支持对文本、图片、多媒体等网络数据源。

4.3 增量爬取功能

网络信息每天都在实时更新，门户网站通常都把最新更新的页面放在首页中实时显示。增量爬取功能就是按照规则设定对多个门户网站定期访问，对指定门户网站首页分析，根据链接情况判断是否有新的链接，然后爬取和抽取有效信息等操作。其流程如图 2 所示。

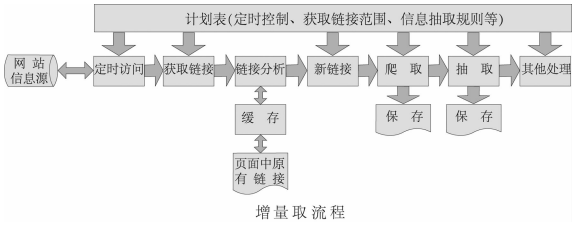

图 2　增量爬取流程

4.4 同文合并功能

网络文章或小说经常有一文多页的情况，例如某篇文章，分多页显示，在抽取信息时，设置相应的规则抽取“下一页”或“下一章”等同文页面信息，根据此信息，将多页合并成同一文章。其流程如图 3 所示。

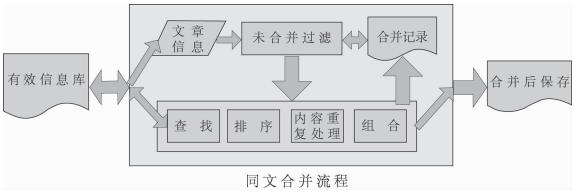

图 3　同文合并流程

5、系统的实现和应用

5.1 系统的实现

本系统使用 Java 语言，采用 Eclipse 、Mysql 、开源工具 HtmlCleaner 和 RBES 等工具和插件进行开发。该系统采用 Eclipse RCP 富客户端的架构。

使用本系统平台不仅可以用来获取网站的历史页面，也可以实时对大型网站进行增量更新爬取。

5.2 系统的应用

本系统应用的主要流程如下图 4 所示：

图 4　信息抽取流程

首先填写要抓取的网页相关信息，如 URL、待抓取信息的规则。然后，设置规则，根据规则，系统将会自动提取出符合这一规则的信息。列表页规则根据实际情况进行填写，如按日期递增，按数字由大到小顺序递增等。如图 4 所示，将新浪体育频道 2008 年 1 月 1 日开始的 100 天的体育新闻信息进行抽取保存。

提取出有用的信息后，程序将分两种方式存储提取出的信息，一种方式可以存入数据库，

另一种方式可以存为 XML 形式，而原始页面则在转换为符合 XML 规范的网页文件后，存入本地磁盘。

6、总结和展望

互联网的迅速发展对不同种类的信息分类、查找、挖掘和知识发现提出了巨大的挑战。对于大多数用户提出的与主题或领域相关的数据获取需求，传统的获取方式往往不能达到令人满意的结果。为了克服传统的获取方式的不足，本文开发了基于规则引擎的数据和信息抽取平台。此平台不仅能够获取文本信息，而且能够获取图片、音频等。研究并开发自适应的网络信息抽取器是快速发展的互联网应用所急切需要的，它不仅具有重要的经济与社会价值，而且具有较高的学术价值。此外，系统中也存在一些有待于进一步改进或增加的功能，本文作者正逐步完善这些功能。

参考文献

[1] 费尔德曼、桑格．文本挖掘［M］．北京：人民邮电出版社，2009：101－143.

[2] 王丽坤、王宏、陆玉昌．文本挖掘及其关键技术与方法［J］．计算机科学．2002，29（12）：12～20

[3] 乔智勇．Web 数据挖掘系统的设计及关键技术研究［D］．西安：西安电子科技大学；2002 年

[4] CHAKRABARTIS，VAN DEN BERGM，DOM B. Focused crawling. A new approach to topic－specific web resource discovery［A］. Proceedings of the Eighth International World－Wide Web Conference［C］，1999.

[5] 贺智平．Web 信息自动抽取技术研究［D］．西安：西安电子科技大学，2006.

[6] 陈红叶．Web 信息提取及知识发现方法研究［D］；合肥：合肥工业大学，2002 年

Design and Implementation of Web Information Extraction Platform Based on Rule Engine

REN Xianzhen，ZHU Yi

Abstract：Information extraction is an important approach of data mining and knowledge discovery，accurate and valid Internet data extraction based upon rule engine as well as automation of the action are the key to knowledge discovery. This paper develops a general text information retrieval platform，using several kinds of information matching techniques to extract data from network data source and adopt processing rules to automatically and intelligently handle information. The platform is implemented using Eclipse RCP；features are implemented as Plug－ins and business logic is embodied as rules. The advantages of the platform are user－friendly，easy expansion，and can automatically retrieve accurate and valid data from large scale web pages.

Key words：information extraction；rule engine；RCP；incremental crawling

（责任编辑：莫修明）