June. 15 2014 Vol. 25 No.06 43-46

信息采集技术研究与应用

黄继鸿*，赵新华，王强
中航工业第一飞机设计研究院， 陕西 西安 710089

摘　要：在分析网络爬虫技术与ETL技术的基础上，探讨了基于网络爬虫技术与ETL技术相融合的算法，并将该算法应用于飞机研制信息采集，实验结果表明，该算法完全满足非结构化数据采集的要求。

关键词：信息采集；网络爬虫；ETL

中图分类号：V328　　**文献标识码**：A　　**文章编号**：1007-5453（2014）06-0043-04

随着局域网和数据库技术的飞速发展，出现了巨大的信息流和与之相伴的大数据流，于是又面临着对庞大的信息、数据无所适从的局面。例如，简单地在一个页面搜索或在数据库内输入关键字，就会有成千上万的相关信息罗列出来，必须经过一系列繁琐的挑选过程，才能找到真正需要的内容，即使找到了想要的信息后，又不知该如何保存和管理这些信息。

信息抽取技术的初始研究最早开始于20世纪60年代中期，最早的研究注重从自然语言文本中获取结构化信息。但是由于信息的逐渐复杂化，多样化，早期的研究已经不再适合目前的信息格式。直到20世纪80年代，为了解决这些问题，信息抽取技术研究才蓬勃开展起来。

F- 22项目开发过程中，积极实施CALS（Computer Aided Logistic Support）计划，所有项目成员都可以在网络环境里进行数据采集、信息传递和后勤支援等活动。其中的数据和信息采集主要应用信息抽取技术来实现。

目前，国内市场上已经有一些网络信息采集产品推出并被使用，比较成熟的信息采集系统有网站万能信息采集器、通达信息采集系统等，采用基于模板的网页解析方式对网页信息进行抽取。

1 信息采集与处理技术

目前的信息采集处理技术有多种，包括爬虫技术以及ETL技术等一些优秀的技术。

1.1 ETL技术

ETL（Extraction Transformation Loading）技术就是数据挖掘中用来进行数据抽取、转换、处理与装载的一门技术，它是数据仓库的核心技术之一。ETL技术首先从数据源中取出数据，对这些数据进行清洗、转换，最后按照一定的数据仓库模型，将这些数据装载到数据仓库中。通过使用各种技术手段，把数据转换为结构化信息，才可以提高一个企业的核心竞争力，而ETL技术是一个主要的技术手段。

1.2 网络爬虫技术

网络爬虫是一种能自动从网络上收集信息的工具，可根据用户的需求定向采集特定主题信息的工具，自动在网络上获取网页源码。对于采集数量较少的工作而言，实现一个网页下载程序不会很麻烦，但是，当从网络上采集海量信息的时候，爬虫系统的实现将会变得十分复杂。

相对ETL技术，网络爬虫技术仅仅处理数据抽取这一步。网络爬虫只将网络上的网页下载下来，对下载的非结构化数据没有进行处理，而直接将其保存到数据库之中。

这样做的劣势在于：搜索引擎的后续操作首先要从数据库中获取到这些非结构化的信息，然后再进一步的进行信息的处理。这样无形之中增添了两次操作数据库的任务量，在数据量极其庞大的背景下，会大大的降低整个系统的效率。

收稿日期：2013-10-27；　退修日期：2014-03-21；　录用日期：2014-05-10
*通讯作者．Tel.：029-86832376E-mail：hjh603@qq.com
引用格式：HUANG Jihong，ZHAO Xinhua，WANG Qiang．Research and aplication of information collection[J]．Aeronautical Secience & Technology，2014，25(06)：43-46．黄继鸿，赵新华，王强．信息采集技术研究与应用[J]．航空科学技术，2014，25(06)：43-46．

1.3 融合的信息采集处理技术

基于以上分析，将ETL这种技术理念应用于网络爬虫技术之中，将信息抽取与信息处理模块结合为一个模块，在信息抽取的同时对信息进行一些必要的处理，并最终将处理好的结构化信息保存到数据库之中，那么会使得整个系统的性能有很大的提高。

2 信息采集技术的应用

将网络信息采集技术应用于飞机研制信息采集，就是基于现有工程设计和管理平台，根据用户自定义的任务配置，无需访问现有平台的后台数据库，通过合法用户账号自动登录各个平台，访问Web页面的内容，批量而精确地跟踪和提取目标网页中的半结构化和非结构化研制信息，并进行结构化等综合处理与分析，最终生成飞机的研制报告，以满足飞机研制全寿命周期管理的要求。飞机研制信息采集系统架构如图1所示。

采集系统通过用户登录Web页面自动跟踪和采集现有平台信息，包括工程设计平台、OA系统、综合保障管理平台、标准化信息平台以及各部门的设计管理平台等，实现对飞机架次信息、构型基线、更改单等信息和数据的收集和后期处理。

2.1 信息采集

信息采集模块主要用来对网络上的专题信息进行收集下载，包括三个核心部分：网络爬虫程序、网页的搜集策略及网页的源码获取。

(1) 网络爬虫程序

网络爬虫程序主要是用来自动提取网页。遍历的起始网页称为种子URL，从起始网页开始，所有指向外部的链接被保存在URL队列中，然后按照顺序访问队列中的网页，这些网页包含的链接也保存在队列中。

(2) 网页的搜集策略

广度优先搜集算法，一般需要存储产生的所有结点，占用的存储空间会比深度优先搜集大；而广度优先策略一般无回溯操作，故运行速度比深度优先策略要快些。而飞机研制信息采集要求对研制过程的信息进行搜集，要求全面，要求时效，故选择广度优先搜集策略研制信息的网页搜集策略。

(3)网页的源码获取

根据浏览器的原理，通过模拟浏览器向WEB服务器发送URL请求，然后获取WEB服务器的响应信息，即网页源码。

2.2 信息处理

该模块将网页这种非结构化信息转换为结构化信息，主要包括：网页内容提取、噪音页面的处理、URL去重和页面语义去重。

工作是对一个网页的标题、正文、URL等进行结构化提取，这个工作需要对网页的具体结构有一个详细的分析，才能进行相关信息的提取；其次进行去除噪音页面的工作，对一些无用网页进行过滤工作；再次需要对收集到的URL地址进行去重工作，即防止程序对同一URL进行一次以上的爬取而陷入死循环；最后需要进行网页语义去重工作，即将多个

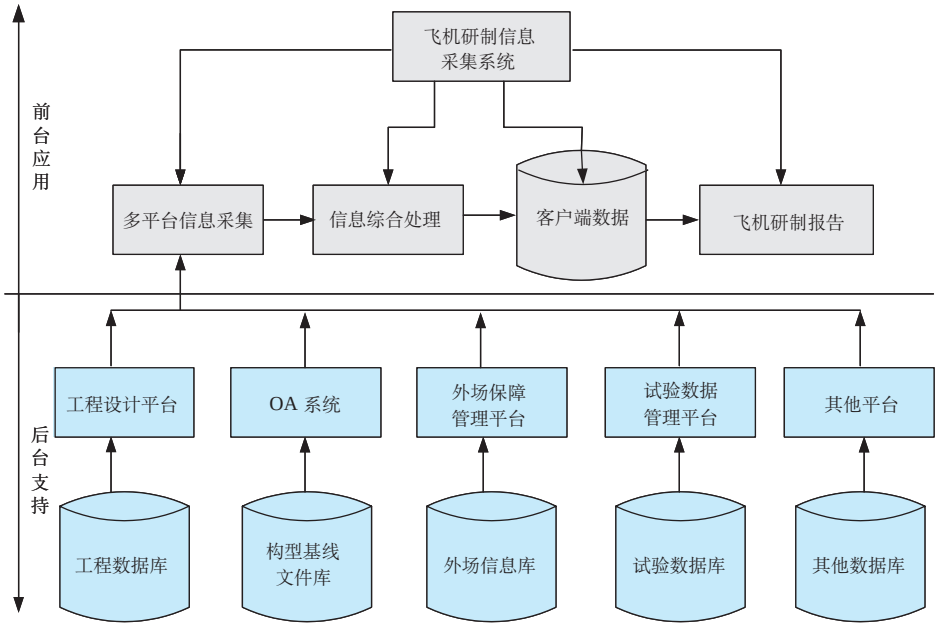

图1 飞机研制信息采集系统架构

Fig. 1 System architecture of information collection in aircraft development

同质的网页找出来，只保留其中一个。

(1) 网页内容提取

网页内容抽取是将一个网页中的标题，正文，URL提取的一个过程。采用基于模板的网页信息提取方法：首先观察同一个类型网页的一些特点，形成规则，然后根据这些规则设计封装器，最后进行结构化抽取。

半结构化文本在网页中非常普遍，比如表格、框架等，并且同一站点的半结构化文本格式通常都是比较固定的，如图2所示。

名称	值
型号	xx飞机
更改名称	总图更改
编号	ECP-xxx-xxxx
要求更改内容	固定后缘内襟翼舱
请求优先级	高
创建日期	2013/08/26
工程更改类别	二类
更改方式	正式更改
更改属性	完善设计
更改原因	扭力管更改引起内襟翼舱零、部件更改
受影响专业	结构
有效性	101-105
附件	ECP-xxx-xxxx.pdf

图2 工程更改网页图

Fig. 2 Web page of engineer change

对于上图页面，其对应的部分源代码如下：

<tr><td>更改原因</td><td><p>扭力管更改引起内襟翼舱零、部件更改</td></tr>……<tr><td><p>附件</td><td><p>ECP-xxx-xxxx.pdf</td></tr>

对于上面获取的源代码：首先，将页面中的注释、脚本、样式表等信息去除，结果如下：

a. 更改原因

b. 扭力管更改引起内襟翼舱零、部件更改

c. 附件

d. href=http://pdm.nic/WChill/Servlet?oid=OR:wt.content.ApplicationData:759011902

e. ECP-xxx-xxxx.pdf

然后将页面划分为若干块，具体包括文本块、图像块、链接块等，上述页面有两块：文字块和链接块；最后对信息进行分块提取，保存到相应数据库中，经过多个页面提取，发现文字块中“更改内容”、“附件”等是固定不变的，为属性名称，在数据库中对应的表中创建相应字段，紧随其后的文字块是其具体内容即“值”，存入相应字段对应的值中。对于超链接，除存储其链接外，还要存储附件的文件名称。

(2) 噪音页面的处理

对于弹出式窗口等提示页面，将其过滤掉，以此来降低系统的无用工作，提高系统的效率。在完成正文提取后，对得到的正文进行分析，然后进行判断处理。

(3)页面去重

URL去重主要是为了防止程序对同一URL进行一次以上的爬取而陷入死循环。

利用URL的哈希值来判断是否重复。通过对哈希算法的比较，并通过实例分析，选择32位Hflp与32位Strhash相结合的哈希组合算法来对URL地址进行编码，以防止对URL进行重复爬取。

(4)页面语义去重

语义去重是指通过一定的策略对于同质的网页只进行一次存储。对网页语句进行中文分词处理，使用编辑距离公式来计算出网页正文间的相似度，根据设定的相似度阀值，来判断两个网页内容是否具有相同的语义。

2.3 测试与分析

测试了整个系统在一天中的8点至18点的信息采集情况，将8点至18点分为4个时段，每个时段2个小时，记录每个时段单纯采用网络爬虫算法模式与采用网络爬虫技术与ETL结合模式的爬取网页数量及爬取网页的错误量。

通过分析测试结果，采用网络爬虫技术与ETL技术结合的融合算法，比单纯采用网络爬虫算法的执行速度有了明显提高，系统的时效性得到保障。

由错误率分析对比，网络爬虫技术与ETL技术结合的融合算法大大降低整个系统的错误率，更加高效精准的获取到网页信息，大大提高整个采集系统的性能，能真正挖掘到精准的飞机研制信息，能得到更加科学的分析结果，帮助飞机设计师能做出更加科学的决策。

3 结束语

在网络信息采集工作中，分析了网络爬虫和ETL两种技术相结合的可行性及优势，利用基于网络爬虫技术和ETL融合算法，对飞机研制信息进行自动化采集和处理。测试结果表明，该方法完全满足对半结构和非结构化信息的采集要求。

参考文献

[1] 吕剑, 谢志航, 陈明. 飞机研制信息资源共享问题初探[J].飞机设计,2003(2):75-80.
Lǚ Jian, XIE Zhihang, CHEN Ming. Research on information sharing question of aircraft development[J]. Aircraft Design, 2003(2):75-80.(in Chinese)

[2] 王建民. 工作流管理-建模、方法和系统[M]. 北京:清华大学出版社,2004:30-33.
WANG Jianmin. Workflow management - Modeling,Method and System[M]. Beijing: Tsinghua University Press, 2004:30-33. (in Chinese)

[3] 胡元军,彭四伟,许耀. 分布式专业信息采集系统[J]. 计算机工程与设计,2007,28(17):4243-4245.
HU Yuanjun, PENG Siwei, XU Yao. Acquisition system of distributed professional information[J]. Computer Engineering and Design, 2007,28(17):4243-4245. (in Chinese)

作者简介

黄继鸿(1969—) 男,硕士,高级工程师。主要研究方向:飞机技术状态管理。
Tel:029-86832376
E-mail:hjh603@qq.com

Research and Application of Information Collection

HUANG Jihong*, ZHAO Xinhua, WANG Qiang

AVIC The First Aircraft Institute,Xi'an 710089,China

Abstract: This paper analyzes the principle of web crawler, discusses the Fusion Algorithm of web crawler and ETL. Extraction Transformation Loading, technology.The Algorithm is applied to Information Collection during airplane development.The experimental results show that the performance of the Fusion Algorithm has met the Unstructured Information Collection.

Key Words: information collection; web crawler; ETL

Received: 2013-10-27; **Revised:** 2014-03-21; **Accepted:** 2014-05-10

*** Corresponding author. Tel. :** 029-86832376 **E-mail:** hjh603@qq.com