第35卷 第21期 计 算 机 工 程 2009年11月
Vol.35 No.21 Computer Engineering November 2009
·软件技术与数据库· 文章编号：1000—3428(2009)21—0051—03 文献标识码：A 中图分类号：TP311.12

基于XML的Web数据半自动采集

蒋宏潮，王大亮，班晓娟，阮进喜
(北京科技大学信息工程学院，北京 100083)

摘 要：如何在信息量巨大的互联网上准确获取并长期跟踪用户关注的内容，是数据采集和挖掘的重要方面。探讨Web数据采集理论及其应用技术，给出一个半自动采集模型，设计基于旅游业数据的采集系统，验证数据半自动采集的可行性。
关键词：数据采集；信息采集；半结构化数据

Web Data Sime-automatic Extraction Based on XML

JIANG Hong-chao, WANG Da-liang, BAN Xiao-juan, RUAN Jin-xi
(School of Information Engineering, University of Science and Technology Beijing, Beijing 100083)

【Abstract】It is an important aspect of data extraction and mining that how to exactly gain and chronically trace the content regarded by users on Internet with huge information. This paper discusses Web data extraction theories and its application technologies, gives a sime-automatic extraction model, and designs a extraction system based on tourism industry data to prove the feasibility data sime-automatic extraction.
【Key words】data extraction; information extraction; semi-structured data

Web是一个广泛分布、高度异构、半结构化且不断进化的信息仓库，它是一个巨大的文档集合，包括超链接信息、访问和使用信息。随着互联网技术的发展，人们越来越难以准确迅速地获得自己需要的数据。虽然目前有多种查全率较高的搜索引擎，但它们的查准率普遍不高，很难进一步挖掘深度数据。因此，研究如何进一步获取互联网上某个特定范围内的数据具有重要意义。

1 相关概念和技术

相对于Web的半结构化数据而言，传统数据库中的数据结构性很强，其数据为完全结构化的数据。半结构化是相对于完全结构化的传统数据库数据而言的。从结构上看，Web上的数据基本都属于属性结构。因此，适合采用XML技术进行数据采集，如静态页面的生成、HTML到XML的转换、信息地位的获取等。Web数据采集与搜索引擎、信息提取概念和技术密切相关。

1.1 XML语言

以XML为基础的新一代WWW环境是直接面对Web数据的，可以很好地兼容原有Web应用，并能更好地实现Web中的信息共享与交换。XML可以视为一种半结构化数据模型，用户能容易地将XML的文档描述与关系数据库中的属性对应起来，实施精确查询与模型抽取。

XML是一套定义语义标记的规则，这些标记将文档分成很多部件并对这些部件加以标识[1]。它是元标记语言，定义了用于定义其他与特定领域有关、语义的、结构化的标记语言的句法语言。XML的上述特点决定了其卓越性能。

XML已经成为正式规范，开发人员能用XML格式标记和交换数据。XML在3层架构上为数据处理提供了很好的方法。它可以从存在的数据中产生，使用XML结构化数据能使数据从商业规范和表现形式中分离出来。这使XML在Web数据采集过程中起到了桥梁作用，定位采集数据和采集数据后的清洗等。

1.2 Web数据采集与搜索引擎

Web数据采集与搜索引擎有很多相似点，如利用信息检索技术。但两者侧重点不同，搜索引擎主要由网络爬虫、数据库和查询服务3个部分组成。爬虫在网上的漫游是无目的性的，只是尽量发现较多内容。查询服务尽可能多地返回结果，不关心结果是否符合用户习惯、专业背景等。而Web数据采集主要针对某个具体行业，提供面向领域、个性化的信息挖掘服务[2]。

1.3 Web数据半自动采集与信息提取

信息提取是指从一段文本中抽取指定的一类信息并将其形成结构化数据填入一个数据库中，以供用户查询使用的过程。信息提取是面向不断增长和变化的某个具体领域的文献特定的查询，这种查询是长期的或持续的。与传统搜索引擎基于关键字的查询不同，信息提取的查询即要包含关键字，也要匹配各个实体间的关系。Web数据采集很大程度依赖信息提取技术，需要实现长期的、动态的追踪。

2 研究意义

2.1 信息迷失问题的解决

随着互联网的快速发展，网上存在越来越多的对用户没有价值的冗余信息，导致人们越来越难以及时准确获取有用信息，信息利用效率和效果极大降低。互联网上的信息冗余主要体现在信息的过载性、无关性和重复性等方面。

在当今高度信息化的社会里，信息冗余、信息过载已成为互联网上一个急需解决的问题。而Web数据采集可以通过一系列方法，依据用户兴趣，搜取网上特定种类的信息，去除无关数据和垃圾数据，筛选虚假数据和迟滞数据，过滤重

基金项目：国家“863”计划基金资助项目(2007AA01Z170)
作者简介：蒋宏潮(1975－)，男，博士研究生，主研方向：知识获取与系统实现；王大亮，博士；班晓娟，副教授；阮进喜，硕士研究生
收稿日期：2009-05-12 **E-mail：**jhc@ustb.edu.cn

复数据。用户无需从包含大量无用信息的搜索结果中获取少量有用价值，而是直接获得按其要求呈现的信息。

2.2 搜索引擎智能化程度的提高

互联网上存在海量信息，但对某个特定的个人或团体而言，需要获取的相关信息或服务及其关注的范围只是很小一部分。目前，人们查找网上信息的主要渠道是搜索引擎，如 Google, Baidu, Yahoo 等。此类搜索引擎涉及面大而广，检索智能度不高，查准率和查全率的问题日益凸现，且难以针对用户需求提供个性化服务。

2.3 人力物力成本的节约

与传统人工采集数据相比，基于 XML 的 Web 数据半自动采集可以减少很多重复性工作，极大缩短采集时间，节约人力物力成本并提高效率，且不会出现人工数据采集中的遗漏、偏差和错误问题。

3 基于旅游业的数据采集系统

Web 数据采集由领域驱动或数据驱动，本文基于上述理论，设计基于旅游业的数据采集系统原型。

3.1 研究目标

旅游业是当今最活跃的行业之一，拥有众多信息供应者和需求者。网上存在大量信息提供者，但用户不可能浏览所有相关网页。通过搜索引擎获取信息，用户要准确找到有用信息较困难。针对该情况，本文设计一个旅游信息自动采集系统，实现数据采集的高效化和自动化。

3.2 系统结构

数据半自动采集系统采用 B/S 模式，用 hibernate+spring+struts 的架构进行开发，数据库采用 mysql。

(1)系统架构

运用 hibernate 将数据表对象化，很好地处理数据库的移植性问题。struts 是典型的模型视图控制器(Model View Controller, MVC)模式，将模型-视图-控制分开，更易于系统维护。由于反转控制和面向切面的编程是 spring 的 2 大特点，因此系统具有很好的扩展性。系统结构主要分成 5 大块：1)数据层；2)业务层，处理实际业务逻辑；3)页面处理层，即页面的请求程序入口；4)jsp 页面；5)XML 的配置文件，主要包括 hibernate 中与数据表对应的实体 XML，spring 的 bean 对应的配置文件和 struts 的 action 和 form 对应的配置文件等。

(2)运行方式

该系统采用多种运行方式，用户可以随时监测采集网页的最新变化情况。如果采集数据量较大、网络堵塞，则数据采集系统会在对方服务器空闲时运行。例如，让采集系统每天凌晨开始搜寻最新网页更新内容，执行数据采集工作。采集系统可以依据实际需要，选择各种灵活的运行方式，充分考虑了采集者和被采集者的情况。

4 Web 数据半自动采集模型及其实现

4.1 采集模型框架

该系统可以分为 3 大功能模块[3]：采集规则生成模块，数据采集与过滤模块，数据展现模块。

4.2 采集规则生成模块

采集规则生成模块的工作流程如图 1 所示。从动态页面链接集合中任意一个样本页面的链接，根据链接生成静态页面，通过 tidy 工具将 HTML 转换转换成 XML，用户标记出感兴趣的节点。系统读取节点上被标记过的、通过算法自动生成信息的采集规则。采集规则主要通过 XML 文档表示，对生成的采集规则，用户可以对需要采集的数据进行处理，如时间格式等。因此，采集规则是数据采集的基础和依据。

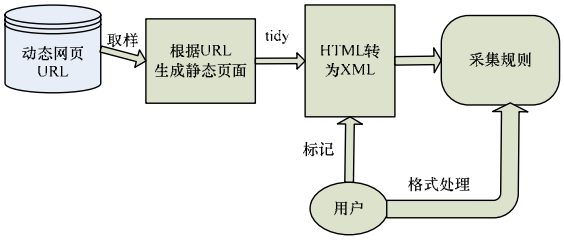

图 1 采集规则生成模块的工作流程

许多 HTML 文档中的标记不符合 HTML 语法要求，如缺乏结束标记等，此类错误会对 HTML 文档的正确解析产生影响。为便于解析，需要对 HTML 文档进行整理，将其转换成 XML 文档，可以利用各种 XML 标准技术进行分析[4]。

对 HTML 文档的整理主要包括以下 3 个方面：

(1)为不成对的标记加上结束符“ / ”，如
加上结束符为
；

(2)为所有属性值加上引号，如<a href=http://www.w3c.org>加上引号变为<a href=“http: //www.w3c.org”>；

(3)将 URL 中所有的“ \ ”换成“ / ”。

如何根据用户标记生成采集规则是本系统的重点和难点。假设待抽取的资源集合是 ε，抽取得到的信息集合是 n，那么采集规则 XML 文档就是一个从 ε 到 n 的映射或函数。采集规则的构造函数 $generateXml(s)$ 针对不同 ε 集合有不同的对象。其中，s 是集合 ε 的一个子集或元素，即样本页面。采集规则对象 g 用来处理资源集合 ε 并得到抽取信息 n。生成采集规则核心内容即如何面对不同待抽取资源集合 ε，生成相应的采集规则的 XML 文档。

在该系统中，通过对 ε 进行取样 p，将 p 页面生成 XML 供用户标记。下文简述如何记录用户标记。文档对象模型(Document Object Model, DOM)树反映样本页面 p 生成的 XML 的结构信息。一个 XML 文档如下，其中包含具有相同含义信息块：

```
<HTML>
<HEAD><TITLE>海东景区黄页，海东景</TITLE></HEAD>
<TABLE>
 <TR>
<TD>电话</TD>
<TD>010-82898217</TD>
<TD>地址</TD>
<TD/>
 </TR>
 <TR>
<TD colspan=“4”>麓景寺，位于青海省乐都县城南 21 公里处的马圈沟口。背依罗汉山，面临麓景河，北傍松花……</TD>
 </TR>
</TABLE>
</HTML>
```

对上述含有多个信息块的 Web 页面运用 XML 文档对应的页面 DOM 树，如图 2 所示。由于文档全部包含在标记<HTML>和</HTML>之间，因此选取 HTML 为标记树的根节点。标记<HEAD>和<BODY>位于<HTML>和</HTML>之间且体现为并列关系，因此，HEAD 和 BODY 是兄弟，也是 HTML 的子节点，HEAD 和 BODY 的子孙节点如图 2 所示。当用户标记 DOM 树中的某个节点时，XML 生成算法根据该节点所在位置生成 xpath 表达式。

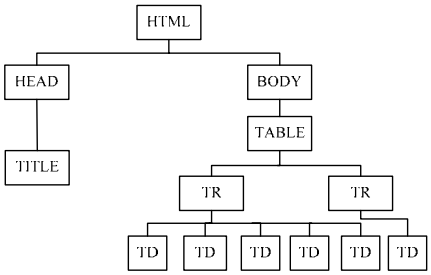

图 2　XML 文档对应的页面 DOM 树

根据生成 XML 的结构和需要查询的信息，对需要采集的数据进行精确定位，并生成 XML 文档。

采集规则的部分 XML 文档描述如下：

```
采集网页的网址
<url>www.peyed.com/groups</url>
字符集
<charset>utf8</charset>
采集的页面数
   <size>1000</size>
   <pager>
采集方法
       <method>regex</method>
数据定位 xpath 的表达式
<xpath>/html/body/table/tr/td/form/table[@class='black12']/tr/td[2]/b</xpath>
正则表达式
      <regex>[\d]*/(\d+)</regex>
</pager>
```

4.3　数据采集和过滤模块

根据采集规则生成模块生成的 XML 文档，进行数据采集和过滤，并将数据存入数据库。数据过滤方法主要分为 2 种[5]：(1)通过正则表达式进行数据过滤；(2)根据本地数据库和数据格式的要求进行过滤。第(1)种方法根据正则法规对采集的数据进行取舍，第(2)种方法根据用户的特殊要求对采集的数据进行过滤。

数据采集的部分代码如下：

```
//创建 xpath 工厂
XPathFactory factory =
XPathFactory.newInstance();
//实例化 xpath
XPath xpath = factory.newXPath();
//xpath  的表达式
XPathExpression expr;
//xpathStr 对应数据定位 xml 中的 xpath
expr = xpath.compile(xpathStr);
//获取数据
Object result =
expr.evaluate(doc, XPathConstants.NODESET);
NodeList nodes = (NodeList) result;
```

对信息进行过滤，例如去掉由生僻字变成的问号，只获取数字、标准化电话号码等。去掉生僻字变成的问号的部分代码如下：

```
String c = "?";
int qtemp = 160;//?对应的 ASCII 值
String qStr = (char)qtemp+"";
content = content.replace(qStr, "").replace(c, "");
```

4.4　数据展现模块

数据展现模块将目标数据库中的数据加工处理后呈现给用户。该模块完成数据抓取的后续工作，模块的负责程度可以根据用户需求而定。其基本功能是将数据以结构化方式呈现给用户。在该模块中可以添加报表图标等统计功能。当数据量达到一定程度后，可以对数据建模，进行时间序列分析、相关性分析，以发现各个概念规则之间的模式和关系，从而在最大程度上利用数据。

5　局限性

Web 数据半自动采集主要完成采集功能，不是完全智能化的过程。它不能理解用户业务或数据意义，必须通过用户进行标识，不会自动创造出采集规则。

6　结束语

数据采集与数据挖掘、信息检索、搜索引擎技术互为补充，各有侧重。随着数据挖掘技术的发展，出现了智能搜索引擎，使上述技术互相促进，有进一步融合的趋势。

在实际应用中，Web 数据采集针对信息过多的问题，增强信息利用效果，提高人们的工作效率并减轻其工作负担。本文模型利用 Web 数据采集的上述优点，通过一系列技术手段，使人们可以更有效、更深入地获取所需数据。

参考文献

[1] W3C. XML Path Language(XPath), W3C Recommendation[DB/OL]. (1999-11-16). http://www.w3.org/TR/xpath.html.

[2] 杨建林, 孙明军. 竞争情报收集的自动化[J]. 情报技术, 2005, (1): 40-43.

[3] 林建勤. 基于 Web 的数据挖掘应用模式研究[J]. 贵州师范大学学报: 自然科学版, 2004, 8(3): 92-96.

[4] 吴永辉. 消除结构冗余的 XML 数据库模式规范化设计[J]. 计算机研究与发展, 2004, 41(10): 30-35.

[5] 周自力, 王仁武. Web 数据自动采集及其应用研究[J]. 电子商务, 2006, (4): 120-125.

编辑　陈　晖

（上接第 50 页）

参考文献

[1] Kaim W E L, Kordon F. An Integrated Framework for Rapid System Prototyping and Automatic Code Distribution[C]//Proc. of the 5th IEEE International Workshop on Rapid System Prototyping. Grenoble, France: IEEE Press, 1994: 52-61.

[2] Girault C, Valk R. 系统工程 Petri 网——建模、验证与应用指南[M]. 王生原, 译. 北京: 电子工业出版社, 2005.

[3] 吴哲辉. Petri 网导论[M]. 北京: 机械工业出版社, 2006: 144-155.

[4] Colom J M, Silva M. Convex Geometry and Semiflows in P/T Nets: A Comparative Study of Algorithms for Computation of Minimal P-semiflows[M]. New York, NY, USA: Springer-Verlag, 1991: 79-112.

[5] 杨　涛, 郭义喜, 张　弘. 有色 Petri 网在渗透测试中的应用[J]. 计算机工程, 2009, 35(1):156-158.

[6] 钟　诚, 陈国良. PRAM 和 LARPBS 模型上的近似串匹配并行算法[J]. 软件学报, 2004, 15(2): 159-169.

编辑　张正兴