

基于 Web 的表格信息抽取研究

秦振海, 谭守标, 徐 超

(安徽大学 电子科学与技术学院, 安徽 合肥 230039)

摘 要:如今, Web 成为了网络信息的主要平台。根据研究发现, 表格在 Web 文本中被经常使用。正因为表格形式简洁并且含有丰富的信息, 自动理解表格在知识管理、信息检索、Web 挖掘等应用中有着广泛的用途, 所以研究 Web 表格信息抽取有着重要的现实意义。互联网上有大量信息采用 HTML 表格表示, 由于 HTML 不描述数据的内容, 机器不能理解和查询。论文首先将 HTML 文档转换为 XML 文档, 结合本体形成启发式规则, 对表格定位、表格结构识别两个关键技术进行了分析。在此基础上, 利用 HTML 表格属性, 将 HTML 表格标准化, 从而适用于复杂表格的信息抽取。

关键词: HTML 表格; 信息抽取; Web; XML

中图分类号: TP393

文献标识码: A

文章编号: 1673-629X(2010)02-0217-04

Study on Tables Information Extraction Based on Web

QIN Zhen-hai, TAN Shou-biao, XU Chao

(Department of Electronic Science and Technology, Anhui University, Hefei 230039, China)

Abstract: Nowadays, web becomes the main information resource. According to the report, tables are used frequently in web documents. Since tables are inherently concise as well as information rich, the automatic understanding of tables has many applications including knowledge management, information retrieval, web mining and so on. Study on tables information extraction based on web has an important practical significance. A large amount of information available on the web is formatted in HTML tables, which are not content-oriented, and are not suitable for understanding and query by machines. In this paper, firstly transform HTML documents to XML documents and combine ontology to discover heuristics. Then two key technologies are analysed, including web table detection, web table structure recognition. On this basis, we normalize the HTML tables according to the attributes of HTML tables and thus this approach is appropriate to extract complicated tables information.

Key words: HTML tables; information extraction; Web; XML

0 引 言

随着信息技术飞速发展, 互联网已经成为最流行的信息发布媒介。人们无论是发布信息还是阅读信息都变的极为方便。然而, 随着互联网信息爆炸性地增长, 人们想要精确获取一条所期望的资料犹如大海捞针般困难。在这种背景下, 人们希望提高有用信息获取的效率。目前 Web 信息获取主要有两种方法: 通过搜索引擎查询或者进行 Web 信息抽取。搜索引擎帮助人们通过关键词来获取相关的文档。用户必须从获得的文档中自己查找有用的信息。因为这些文档并不考虑用户的知识领域, 对用户来说并不容易定位到自

己需要的资源上。然而 Web 信息提取则自动从网络里分析和发现有用的信息, 废弃并不需要的数据, 可充分提取用户知识领域的知识。由于 Web 页面大量使用表格元素这一现象, 所以对表格进行信息抽取具有重要的现实意义。

1 XML 介绍

XML(eXtensible Markup Language, 扩展标记语言)是由万维网协会(W3C)设计的专门为 Web 服务的, 它是 SGML(Standard General Markup Language, 标准通用标记语言)的一个简化子集。XML 是一种类似于 HTML(Hypertext Markup Language, 超文本标记语言)的数据描述语言, 它以一种开放的、自我描述的方式定义数据结构, 是用来自动描述信息的一种新的标记语言。XML 文档由标记和字符数据组成, 通过 DTD(Document Type Definition, 文档类型定义)或 Schema 使文档结构化, 这样很容易验证文档数据的合

收稿日期: 2009-06-13; 修回日期: 2009-09-02

基金项目: 安徽省自然科学研究重点项目(2005KJ004ZD)

作者简介: 秦振海(1981-), 男, 安徽阜阳人, 从事网络与智能信息系统研究; 谭守标, 博士, 教授, 从事网络与智能信息系统研究; 徐超, 博士, 教授, 从事网络与智能信息系统研究。

法性,容易提取(查询)文档中的数据。在浏览器中同一 XML 文档可以利用 CSS (Cascading Style Sheets,层叠式样式表)或 XSL (eX-tensible Style Sheet Language,扩展样式表语言)实现多种显示形式。利用 XSLT 也可方便地将 XML 文档译为 HTML 文档或不同标记表示的 XML 文档。

XML 解决了 HTML 不能解决的两个 Web 问题:一是 Internet 发展速度快而接入速度慢的问题;二是可利用的信息多,但难以找到用户所需要的信息的问题。XML 能增加结构和语义信息,可使计算机和服务器即时处理多种形式的信息。XML 将网络信息标准化,使开发者和计算机易于辨认信息,能创建不依赖于平台、语言或格式的开放数据。

XML 作为一种标记语言,有以下几个特点:

(1) 简单 XML 能创建一种任何人都能读出和写入的世界语。整个规范简单明了。它由若干规则组成,这些规则可用于创建标记语言,并能用一种常常称作分析程序的简明程序处理所有新创建的标记语言。

(2) 开放 XML 可以用许多成熟的软件来帮助编写、管理等,开放式标准 XML 的基础是经过验证的标准技术,并针对网络做最佳化。XML 解释器可以使用编程的方法来载入一个 XML 文档,当这个文档被载入以后,用户就可以通过 XML 文件对象模型来获取和操纵整个文档的信息,加快了整个文档的运行速度。

(3) 可扩展性 XML 在两个意义上是可扩展的,首先它允许开发者创建他们自己的 DTD,其次,使用几个附加的标准,您可以对 XML 进行扩展,可以向核心的 XML 功能集增加样式、链接和参照能力。

(4) 互操作性 XML 可以在多种平台上使用,而且可以用多种工具进行解释。因为文档的结构是相容的,所以解释它们的语法分析器就可以以较低的费用建立,XML 支持用于字符编码的许多主要标准,允许它在许多不同的计算机环境中使用。

2 系统概述

笔者在现有 Web 表格信息抽取技术的基础上,提出了一个使用预定义领域本体知识库^[1-4]的 Web 表格的信息抽取方法,其系统结构如图 1 所示。

2.1 获取 Web 页

即从互联网上获取包含 HTML 表格的 Web 页。

2.2 WEB 页清洗

获取的 Web 页包含很多无用信息,这些无用信息会导致后面模块中将 Web 页转化得到的 XML^[5,6]文档无法正常装载,所以通过清洗器,把这些多余的内容

去掉。

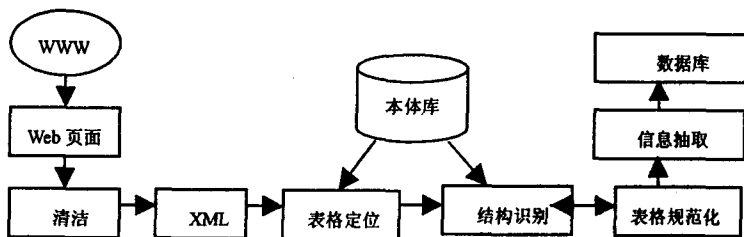


图 1 系统结构

2.3 HTML 文档转变成 XML 文档

因为 HTML 文档对其格式完整性没做严格的要求,所以首先要对格式非良好的 HTML 文档进行整理,把 HTML 文档转变成格式良好的 XML 文档。然后通过分析 XML 文档得到用户需要的信息。

2.4 Web 表格定位

在此模块中滤除非数据表格和非用户感兴趣的数据表格等额外的信息,识别出满足要求的数据表格。

2.5 Web 表格结构识别

在该模块中识别出表格的展开方式和表格属性行(列)、数据单元格所在位置。

2.6 Web 表格规范化

网上表格由若干行组成,每一行又有许多单元组成,单元可能跨越多个行与列。这种结构比传统的关系数据库的表格结构复杂得多。这给表格数据抽取带来了不少困难。文中采用拆分与合并单元格的方法,使表格标准化,即使表格的每一行(列)都具有相同数目对齐的单元格。

2.7 信息抽取

网上表格被规范化后,表格的各个数据单元和其相应的表头属性被一一确定,因此可以根据表头和单元格数据来理解表格的属性值对。将表格数据存入关系数据库中。

3 Web 表格信息抽取各模块关键技术分析

3.1 清洗器

当把包含要抽取信息的 Web 页面输入清洗器时,Web 页面不仅仅包含待抽取的表格,而且还包含脚本、样式表、注释、广告、图片等等其它信息。在这个模块中,清洗器采用以下步骤,剔除多余信息。

(1) 剔除“body”标签以外的内容。

(2) 剔除脚本内容、样式单内容、注释内容、图片、广告。

(3) 将多个连续的空格用一个空格取代。

3.2 HTML 到 XML 的转换

在这个模块中,可以把 HTML 页面转换为 XML

格式的文件。虽然可以直接从 HTML 文档中提取数据信息,但是现阶段访问 HTML 文档内容的方法并不灵活。加之存在 HTML 编码的不规范,使得直接访问 HTML 文档内容相对麻烦。可以采取把 HTML 文档先转换为 XML 文档^[7,8],而对于 XML 文档内容的提取则已经存在很成熟的技术了。

目前有一些对 HTML 页面设计进行规范化组织的工具,Tidy 就是一种过滤 HTML 文本中错误的免费产品,可以用于修正 HTML 文档中的常见错误,并生成格式编排良好的等价文档,即 XHTML(XML 的子集)。Tidy 主要做以下两件事情:

- (1)将不成对的标签加上结束符“/”,例如
转换为
,转换为。
- (2)给所有属性值加引号。例如<div class=header>转换为<div class=“header”>。

3.3 Web 表格定位

表格定位^[9]是指在网页中找到感兴趣的表格位置。HTML 页面中表格是由<table>元素标识的。在很多 Web 站点尤其是商业站点,用<table>元素标识的除了称之为其表格的数据表格之外还包含导航栏,或其它站点的链接等其他用来进行页面布局的非数据表格,非数据表格又被称为假表格^[9]。另一种情况是并不是所有的 Web 页面中的数据表格都是感兴趣的,有时即便是一个领域相关的 Web 页面也可能包含几个内容跟研究无关的数据表格。已经知道要定位的表格信息存在于<table></table>结点之间的内容中,所以只关心 table 结点。通过对 XML 文档中 table 结点的依次遍历滤除非数据表格。在该过程中使用的启发式规则如下:

- 规则 1: 表格大小至少是 3 行 3 列。
- 规则 2: 如果表格中包含<caption>或<th>标记,则该表格是数据表格。
- 规则 3: 如果表格中包含大量的超级链接、表单,则该表格为非数据表格。
- 规则 4: 基于应用本体中的对象集,可以找到所有本体可识别的字符串,如果在表格中某行(列)可识别字符串数量超过 50%,可以确信该表格是感兴趣的。

3.4 Web 表格结构识别

Web 表格结构识别^[9]是指通过识别 Web 表格的结构,生成表格的逻辑结构模型。它包括标题行和内容行识别、表格展开方式以及表头和表体识别。如何确定表格标题是本文的关键问题。根据对标题特征的分析,采用下列启发式方法:

- 规则 1: 检查<tr>及<td>属性,如果存在...项,并且属性取值不通,对于大字

体的行,其为标题行。
规则 2: 包含在<th>...</th>中的行,这些行为表格标题行。

规则 3: 从第一行(列)开始,把行(列)中的字符串分别与应用本体中的同义词词典比较,如果某行(列)中的字符串有 50% 以上匹配成功,则该行(列)为标题。

3.5 Web 表格标准化

在对表格进行抽取前,必须先进行标准化处理。把每一行(列)都具有相同数目对齐单元格的表格,叫标准格式的表格,如表 1。当出现非标准格式的表格,如表 2,要对表格进行标准化转换。

表 1 标准格式的表格

车次	硬座价格	软座价格
a	b	c
d	e	f

表 2 非标准格式的表格

车次	价格	
	硬座	软座
a	b	c
d	e	f
	g	h

在 HTML 文档中,表由 table 元素表示,一般由一个标题和许多行组成,每一行又由许多单元组成。表的单元一般用 th 元素表示头部信息,包括行头和列头,用 td 元素表示数据。表的单元可能占用多行或多列。如果一个 th 或 td 元素包含属性 colspan = n,则表示该 th 或 td 元素的单元多占了后面 n - 1 列单元的位置;如果一个 th 或 td 元素中包含属性 rowspan = n,则表示该 th 或 td 元素的单元多占了后面 n - 1 行单元的位置。

例 1 以下是 HTML 表格片段:

```
<table border="1">
<tr align="center">
<td rowspan="2">单位</td>
<td colspan="2">岗位需求数</td>
<td rowspan="2">专业</td>
</tr>
<tr align="center">
<td>教学</td>
<td>教辅</td>
</tr>
<tr align="center">
<td rowspan="2">中文系</td>
<td>1</td>
```

```

<td>1</td>
<td>现代汉语</td>
</tr>
<tr align="center">
<td>2</td>
<td>1</td>
<td>写作学</td>
</tr>
</table>

```

例 1 的 HTML 表格显示如表 3 所示。

表 3 例 1 的 HTML 表格

单位	岗位需求数		专业
	教学	教辅	
中文系	1	1	现代汉语
	2	1	写作学

表 3 显示了嵌套的标题和占用多行多列的单元的复杂结构。为了获取属性值对,需要对它进行标准化。对表 3 中占有多行的表头采用单元格重组与拆分的方法处理,对表 3 中占有多行的数据单元格采用单元格拆分的方法处理。例 1 中定义的 HTML 表格片段被标准化为如下的 HTML 片段,其对应的 HTML 表格如表 4。

表 4 规范化后的 HTML 表格

单位	教学岗位需求数	教辅岗位需求数	专业
中文系	1	1	现代汉语
中文系	2	1	写作学

```

<table border="1">
<tr align="center">
<td>单位</td>
<td>教学岗位需求数</td>
<td>教辅岗位需求数</td>
<td>专业</td>
</tr>
<tr align="center">
<td>中文系</td>
<td>1</td>
<td>1</td>
<td>现代汉语</td>
</tr>
<tr align="center">
<td>中文系</td>
<td>2</td>
<td>1</td>
<td>写作学</td>

```

```

</tr>
</table>

```

例 1 中,假定表头处于行中。对表头处于列中或表格同时具有行标题与列标题时,同样可以设计出表格标准化算法,将复杂表格标准化。表格转化为标准化形式,这为后面表格信息抽取提供了方便。

4 结束语

提出了一种转换 HTML 文档为 XML 文档的方法,该方法将本体与信息抽取相结合,利用 HTML 属性对表格单元格拆分与重组,使 HTML 表格标准化。应用此方法不依赖于所抽取的 Web 页面的设计格式,大大地提高了系统的可重用性。该方法适用于没有标出表头信息的 HTML 表格和复杂表格。

从整个项目来看,文中的研究仅仅是一个起点,有很多工作还有待进一步的开展,如对 Web 页面中无 <table> 标记表格的信息抽取,这可以作为下一个研究方向。

参考文献:

- [1] 王放,顾宁,吴国文.基于本体的 Web 表格信息抽取[J].小型微型计算机系统,2003,24(12):2142-2146.
- [2] 林琳.基于 Ontology 的 Web 表格内容抽取的研究与实现[D].成都:电子科技大学,2006.
- [3] Tan Shoubiao, Xu Chao, Jiang Yuan. Web Data Extraction System Based on Label Library[C]//In: Proceeding of the 6th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD'09). Tianjin, China: IEEE Computer Society, 2009.
- [4] Tan Shoubiao, Fan Jin, Jiang Yuan. Web Data Extraction Based on Label Library[C]//In: Proceeding of 2009 World Congress on Computer Science and Information Engineering (CSIE 2009). Los Angeles/Anaheim, USA: IEEE Computer Society, 2009.
- [5] Liu Ling, Pu Calton, Han Wei. XWRAP: An XML - enabled Wrapper Construction System for WEB Information Source [C]//Data Engineering, 2000. Proceedings. 16th International Conference. [s. l.]: [s. n.], 2000: 611-621.
- [6] 陈玉芳,葛燧和.一个基于 XML 的 WEB 数据收集模型的研究[J].计算机工程与应用,2004(10):150-152.
- [7] 李健,谭守标,徐超.一种 Web 数据挖掘系统的设计和研究[J].计算机技术与发展,2009,19(2):70-73.
- [8] 鲍仕壮,徐超,谭守标,等. Web 页面表格内容的提取方法研究[J].软件导刊,2008,9:65-67.
- [9] 赵洪,肖洪,薛德军,等. Web 表格信息抽取研究综述[J].现代图书情报技术,2008(3):24-31.