

HTML 表格向 XML 的智能转换

贾长云¹, 程永上²

(1. 淮海工学院计算机工程学院, 连云港 222069; 2. 河海大学计算机与信息工程学院, 南京 210000)

摘 要: XML 已经成为处理与管理信息的标准格式, 而 HTML 表格被广泛应用于 Web。为了充分利用与管理 HTML 表格信息, 需要将 HTML 表格转换成 XML。提出一种有效的处理方法, 该方法包含 2 个部分, 即表格识别与结构转换。表格识别通过检查格式、语法及语义的特征将表格提取出来并分割成值域与属性域, 使用预设的表格模板分析属性域与值域间的层次结构并将其转换成 XML 格式。通过 300 多个表格的实验表明, 所提出的方法要优于传统方法, 结果的准确率达 86.7%。

关键词: HTML 表格; 结构分析; 规范化; 信息提取; 可扩展标记语言

Intelligence Conversion of HTML Table into XML

JIA Chang-yun¹, CHENG Yong-shang²

(1. School of Computer Engineering, Huaihai Institute of Technology, Lianyungang 222069;

2. College of Computer and Information Engineering, Hohai University, Nanjing 21000)

[Abstract] While HTML tables are widely applied for Web, XML is widely accepted as a standard format to process and manage information. In order to utilize and manage XML, the HTML tables should be transformed into XML representations. This paper presents an efficient method for the process, which consists of two phases, such as area segmentation and structure analysis. The area segmentation cleans up tables and segments them into attribute and value areas by checking visual and semantic coherency. The hierarchical structure between attribute and value areas is analyzed and transformed into an XML representation using a proposed table model. Experimental results with more than 300 HTML tables show that the proposed method performs better than conventional methods, resulting in an average accuracy of 86.7%.

[Key words] HTML table; structure analysis; normalization; information extraction; XML

1 概述

表格可以方便直观地表达数据之间的关系, 因而在 HTML 网页中的使用非常广泛。但是 HTML 表格只能显示数据的内容, 并不能管理数据, 且 HTML 表格复杂多样。XML 由于其对数据的强大管理能力得到了越来越广泛的应用。因此, 提取 HTML 表格中的数据并将其转化为 XML 格式显然具有重要的实际应用意义。

近年来, 许多学者及研究机构对 HTML 表格转换为 XML 进行了不同的研究, 主要采用的方法有手工编写代码的方法、机器学习方法和启发式自动化方法^[1-3]。总体来说, 这些方法只是单纯地应用了 HTML 表格格式、布局的有限信息, 对 HTML 表格的复杂性考虑不足, 因而转换的准确率不高。

2 HTML 表格及其分类

HTML 表格通过属性-值对来描述对象, 内容为对象属性的单元格称为属性域, 内容为属性对应值的单元格称为值域。在 HTML 中, 表格由 table 元素标记, 它由若干行元素(row)构成, 每一行又由若干单元(cell)构成。这些单元可以表示表头信息的 th 元素, 也可以是表示数据信息的 td 元素。单元可以跨越多个列或多个行, 跨越的行、列数分别由 th 或 td 元素的 rowspan 与 colspan 属性指定。示例表格如图 1 所示。

Prod	2002				2003			
	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
Car	43 000	43 200	43 500	44 500	42 000	41 000	44 000	42 000
Truck	Not Available				21 000	20 000	22 000	24 000

图 1 示例表格

示例表格对应的 HTML 文档如下:

```
<table border="1">
  <tr><td rowspan="2">2" align="center">
    <b>Prod.</b></td>
    <td colspan="4" align="center">
      <b>2002</b></td>
      <td colspan="4" align="center">
        <b>2003</b></td></tr>
    <tr><td align="center">
      <b>Q1</b></td>
      <td align="center"><b>Q2</b></td>
      ...
    <tr><td><b>Car</b></td><td>4,3000</td>
    ...
    <tr><td><b>Truck</b></td><td colspan="4">Not Available
  </td><td>2,1000</td>
  ...</tr>
</table>
```

HTML 表格结构比较复杂, 从形式上可以分为简单表格与复杂表格 2 类。其中, 简单表格又可以分为 3 种^[4]:

- (1) 行标题表(属性在前 n 行);
- (2) 列标题表(属性在前 m 列);
- (3) 行列标题表(属性分别在前 n 行与前 m 列)。

作者简介: 贾长云(1960—), 男, 副教授、硕士, 主研方向: 数据库应用, 软件工程; 程永上, 副教授、博士

收稿日期: 2009-03-01 **E-mail:** lyghitjcy@vip.sina.com

3 转换原理

要将 HTML 表格转换为 XML 文档一般要经过表格识别与结构转换 2 个阶段。表格识别主要对表格进行规范化并确定表头;而结构转换则根据预定的表格转换模型提取表格的逻辑层次结构并将其转换为 XML 文档。

3.1 表格识别

3.1.1 表格的规范化

表格的规范化方法研究已经比较成熟,本文采用文献[5]的方法进行表格规范化。图 1 的表格经过规范化如图 2 所示。

Prod.	2002	2002	2002	2002	2003	2003	2003	2003
Prod.	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
Car	43 000	43 200	43 500	44 500	42 000	41 000	44 000	42 000
Truck	Not Available	Not Available	Not Available	Not Available	21 000	20 000	22 000	24 000

图 2 规范后的表格

3.1.2 表头的确定

表格进行规范化后,要确定属性域与值域,这是提取 HTML 表格逻辑层次结构的基础。一般而言,作为属性域来说还是存在许多与值域不同的特征,如格式特征、语法特征、语义特征等。

(1)格式特征值检查

表格作者为了让用户容易识别,一般会采用一些格式化的信息来表示表头。比较重要的格式化信息有字体、字号、粗体、斜体等,而且通常情况下作为属性域的单元格具有相同的格式信息。可以采用格式特征值来区分属性和值。为此,需要计算每行每列的平均格式特征值。具体计算办法见文献[5],本文不再详述。

如图 2 的表格可确定前 2 行为属性行,前 1 列为属性列。

对于有格式设置的表格通过以上的格式特征值检查可以区分表格的属性域与值域,但有些表格根本没有设置格式(如图 3 的表格所示),所有单元格均使用相同的默认格式。在这种情况下,再通过该方法就无法确定属性域了,可以通过语法特征值来区分属性域与值域。

产品	价格
A8800V(C2Q Q6600 2G500sV(VP))	¥ 16 000 元
A6000V(PDC E2160 1G160sB(VP))	¥ 7 500 元
M2600V(ICP 420 51280sD(XP))	¥ 4 799 元

图 3 无格式特征表格

(2)语法特征值检查

表格的语法特征可从 2 个方面来检查:一是单元格的数据类型特征;二是单元格数据长度特征。因为通常情况下,表格的属性单元都是字符型的数据,而值单元的数据类型则有可能是数值型、字符型、日期型等。但对行标题表格而言,每一列的值单元肯定是一种数据类型(列标题表格则每一行的值单元取相同的数据类型)。另一方面,一般而言值单元的数据长度与对应属性单元也有所区别。因此,通过对这 2 个方面的综合检查也可以确定属性域与值域。

语法特征值的计算通常从表格右下方的单元开始,以此单元为基准块计算其与相邻单元的语法特征值,然后将该相邻单元并入基准块,再计算相邻单元的语法特征值,如此反复直到一列的所有单元。显然在属性与值的分界线上语法特征的差异最大,据此也可以确定属性域。考虑到数据类型特征(D)比长度特征(L)更为明显,使用不同的权重来综合这 2 个特征值。语法特征值(μ)的计算方法如下:

$$\mu=w1\times D+w2\times L \tag{1}$$

其中, $w1$ 为数据类型特征权重; $w2$ 为长度特征权重。

数据类型特征值的计算按下式进行:

$$D_{i,j}=n'/n \tag{2}$$

设表格为 p 行 q 列, n' 为第 j 列第 p 行到第 i 行中具有主要数据类型的单元数, n 为第 j 列第 p 行到第 i 行的总单元数。

长度特征值定义为在一定范围内单元数据长度在 α 范围之内的比值, α 由基准块中平均长度的 0.5~1.5 来确定。长度特征值按下式计算:

$$L_{i,j}=m'/m \tag{3}$$

其中, m' 为基准块中长度在 α 范围之内的单元数; m 为基准块中的单元总数。

如果表格为 $2\times N(N>2)$, 只要计算行的语法特征值来分割。如果 2 行的语法特征没有明显的差异, 则认定第 1 行为属性行, 第 2 行为值域。同样理由, 对 $N\times 2(N>2)$ 的表格, 只计算列的语法特征值。而对 2×2 表格, 则要检查语义一致性而不检查语法一致性。

(3)语义特征检查

如果发现表格不能进行格式或语法特征检查或表格为 2×2 , 那么就需要通过语义特征检查来分割表格。语义特征检查主要是从语义上来测试属性与其值之间的语义对应关系, 这是因为表头属性与其对应的值之间总存在语义上的对应。例如带有“E-mail、email、电子邮件”的属性关键字, 其值域显然是电子邮件的格式字符串; 同样, 对于“电话、联系方式”等这样的属性关键字所对应的值域显然都是由数字构成的电话号码等。所有这些统计表可以通过正则表达式匹配来进行检查。

对于复合表格可以通过属性域的数量来确定。如果表格中属性域的数量多于一个, 则可认定为复合表格, 根据属性域的位置可以将其拆分为多个简单表格。

3.2 结构转换

表格识别完成以后, 就可以进行表格逻辑层次结构的提取。本文采用了基于表格模型的提取方法, 逻辑结构提取后还要将其转换为 XML 文档。

如图 4 所示, 表格模型将表格分为 5 类: $2\times N$ 的行标题表, 其他行标题表, $N\times 2$ 的列标题表, 其他列标题表和行列标题表。例如对图 2 的表格, 其层次结构如图 5 所示。

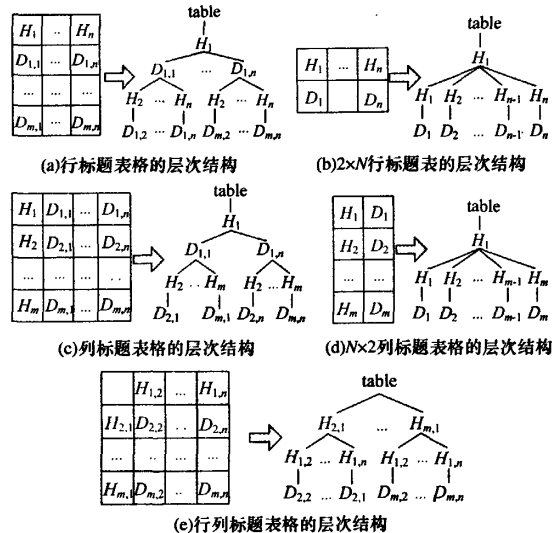


图 4 表格转换模型

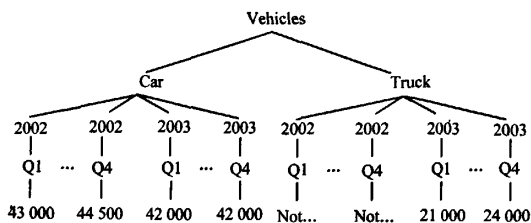


图5 示例表格的层次结构

表格的层次结构提取完成后，就可以将该层次结构转换为 XML 文档。转换的基本方法是使用深度优先策略来生成合法的标记，将层次结构转换成 XML 文档。转换后 XML 文档应该是格式良好的。为了去除 XML 文档生成后的冗余的标记，使用文献[6]提出的合并规则。

将图5所示的层次结构转换成XML，通过使用合并规则消除冗余元素最终生成了如下所示的结果，该XML文档满足了格式良好与简单化的要求。

```

<?xml version="1.0" encoding="UTF-8"?>
<Vehicles>
  <Car>
    <y2002>
      <Q1>43000</Q1>
      ...
    </y2002>
    <y2003>
      <Q1>42000</Q1>
      ...
    </y2003>
  </Car>
  <Truck>
    <y2002>
      <Q1>Not Available</Q1>
      ...
    </y2002>
    <y2003>
      <Q1>21000</Q1>
      ...
    </y2003>
  </Truck>
</Vehicles>
  
```

4 实验结果与分析

本文采用.NET平台实现了前述方法的原形系统，并对实际站点如 gaokao.chsi.com.cn, www.autoseek.co.uk, www.carssearch.com 等站点的近300个表格进行实际测试，所涉及

(上接第27页)

提出的频偏估计方法，结合2种比较算法的优势，在整个信噪比区间都有满意的表现，同时其加窗函数所要求估计的PDP，在实际系统中可以和信道估计等步骤复用，并不额外增加算法复杂度。综上，本文提出的精细同步方法，性能优良，具有很好的实用性。

参考文献

- [1] Qiao Yantao, Yu Songu, Su Pengcheng, et al. Research on an

的表格包括行标题表格、列标题表格、行列标题表格及复杂表格等4类，转换的准确率(准确率=转换正确的表格数目/表格总数)达86.7%。表1显示了各种类型表格的转换准确率。

表1 实验结果

分类	表总数	成功数	失败数	准确率/(%)
行标题表	187	179	8	95.6
列标题表	67	55	12	81.8
行列标题表	32	32	0	100.0
复杂表	26	18	8	69.7
合计	312	284	28	86.7

本文方法的性能要优于以前其他的方法，因为该方法更为系统，考虑的问题更为精确。首先，在预处理阶段就对表格进行规范化；然后不仅使用格式特征而且使用语义特征来对表格进行精确的分割。例如，对于2×2的表格，以前大多数方法都没有办法来分割，本方法却能通过识别属性与值之间的语义一致性来进行分割，这样结果要精确得多。

5 结束语

本文提出了一种智能的有效方法，从HTML表格中提取逻辑结构并将其转换成XML文档。通过规范化表格、使用格式特征及语法特征计算分割成属性域与值域，再通过表格模型提取表格层次结构将其转换为XML文档。实验结果显示，该方法可以应用于不同的表格，而且性能优于那些基于启发式规则的方法。

参考文献

- [1] 胡东东, 孟小峰. 一种基于树结构的Web数据自动提取方法[J]. 计算机研究与发展, 2004, 41(10): 1607-1613.
- [2] Lim S J, Ng Y K, Yang X. Integrating HTML Tables Using Semantic Hierarchies and Meta-data Sets[C]//Proc. of International Symposium on Database Engineering and Applications. [S. l.]: IEEE Press, 2002: 160-169.
- [3] Jung S W, Kwon H C. A Scalable Hybrid Approach for Extracting Head Components from Web Tables[J]. IEEE Transactions on Knowledge and Data Engineering, 2006, 18(2): 174-187.
- [4] 范莉娅, 肖田元. 自动获取HTML表格语义层次结构方法[J]. 清华大学学报: 自然科学版, 2007, 47(10): 1586-1590.
- [5] 张瑞, 李石君. 网上表格数据到XML的自动转换[J]. 计算机工程与应用, 2007, 43(2): 190-192.
- [6] Li Shijun, Liu Mengchi, Peng Zhiyong. Wrapping HTML Tables into XML[C]//Proc. of the 5th International Conference on Web Information Systems Engineering. Brisbane, Australia: Springer, 2004: 147-152.

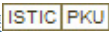
编辑 顾逸斐

Iterative Algorithm of LS Channel Estimation in MIMO OFDM Systems[J]. IEEE Trans. on Broadcasting, 2005, 51(1): 149-153.

- [2] Magnus S. Estimation of Time and Frequency Offset in OFDM Systems[J]. IEEE Trans. on Signal Processing, 1997, 45(7): 1800-1805.

- [3] 周恩, 陈茅茅, 王文博. 一种适用于多径衰落信道地OFDM系统同步算法[J]. 电子与信息学报, 2005, 27(10): 1625-1628.

编辑 陈文

作者: 贾长云, 程永上, JIA Chang-yun, CHENG Yong-shang
作者单位: 贾长云, JIA Chang-yun(淮海工学院计算机工程学院, 连云港, 222069), 程永上, CHENG Yong-shang(河海大学计算机与信息工程学院, 南京, 210000)
刊名: 计算机工程 
英文刊名: COMPUTER ENGINEERING
年, 卷(期): 2009, 35(14)
被引用次数: 2次

参考文献(6条)

1. 胡东东;孟小峰 一种基于树结构的Web数据自动提取方法[期刊论文]-计算机研究与发展 2004(10)
2. Lim S J;Ng Y K;Yang X Integrating HTML Tables Using Semantic Hierarchies and Meta-data Sets 2002
3. Jung S W;Kwon H C A Scalable Hybrid Approach for Extracting Head Components from Web Tables[外文期刊] 2006(02)
4. 范莉娅;肖田元 自动获取HTML表格语义层次结构方法[期刊论文]-清华大学学报(自然科学版) 2007(10)
5. 张瑞;李石君 网上表格数据到XML的自动转换[期刊论文]-计算机工程与应用 2007(02)
6. Li Shijun;Liu Mengchi;Peng Zhiyong Wrapping HTML Tables into XML 2004

本文读者也读过(6条)

1. 魏志华. 黄孝伦. 刘亮. 史林霞. WEI Zhihua. HUANG Xiaolun. LIU Liang. SHI Linxia 基于对称性的HTML到XML的转换方法[期刊论文]-武汉理工大学学报(信息与管理工程版) 2007, 29(7)
2. 曹风华. CAO Feng-hua XSLT在XML向HTML转换中的作用[期刊论文]-现代计算机(专业版) 2010(3)
3. 黄晓斌 HTML向XML转换的研究[期刊论文]-现代图书情报技术2003(1)
4. 黄伟. 刘娟. HUANG Wei. LIU Juan 一种基于DOM树的HTML转换为XML的方法[期刊论文]-电脑知识与技术(学术交流) 2006(7)
5. 秦振海. 谭守标. 徐超. QIN Zhen-hai. TAN Shou-biao. XU Chao 基于Web的表格信息抽取研究[期刊论文]-计算机技术与发展2010, 20(2)
6. 陈艳梅. 张斌 HTML到XML转换技术的研究与实现[期刊论文]-现代图书情报技术2003(5)

引证文献(2条)

1. 杜茂康. 李韶华. 刘苗 基于MEDL模型的HTML向XML的转换方法[期刊论文]-重庆邮电大学学报(自然科学版) 2012(6)
2. 曾广朴. 陶维安 基于信息量的Web表格信息抽取方法[期刊论文]-西南师范大学学报(自然科学版) 2010(4)

本文链接: http://d.wanfangdata.com.cn/Periodical_jsjgc200914012.aspx