Web 表格信息抽取研究综述

赵 洪 肖 洪 薛德军 师庆辉
(中国学术期刊(光盘版)电子杂志社 北京 100084)

【摘要】介绍 Web 表格的特点与结构、Web 表格信息抽取及其过程,分析 Web 表格信息抽取的 4 个关键技术:Web 表格定位、Web 表格结构识别、Web 表格内容整合和抽取结果表示,以及 Web 表格信息抽取的应用。最后指出目前国内外该项研究的不足之处及未来发展方向。

【关键词】Web 表格 信息抽取 表格定位 表格结构识别 表格内容整合
【分类号】TP391

A Survey of the Research on Information Extraction over Web Tables

Zhao Hong Xiao Hong Xue Dejun Shi Qinghui
(China Academic Journal(CD) Publishing House, Beijing 100084, China)

【Abstract】This paper firstly introduces the characteristics and structure of Web tables and describes the process of information extraction over Web tables. Then four key technologies are analysed, including Web table detection, Web table structure recognition, Web table interpretation and presentation of table extraction. It also analyses the application of the research and points out the problems in current researches, and finally presents a prospect of its future.

【Keywords】Web tables Information extraction Web table detection Web table structure recognition Web table interpretation

1 引 言

表格(Tables)作为一种重要的信息表现形式已广泛地应用于 Web 文档中。对于表格而言,“表格中的句法和语义概念是相互混合的,表格逻辑单元格以它的相对位置信息(可认为是句法)来获得语义,但此类句法结构比自然语言更为复杂”[1]。因而,如何让机器准确地抽取表格信息,一直是个具有挑战性的难题。

传统的表格信息抽取研究主要着眼于 ASCII 文件或由光学字符识别得到的表格,主要围绕表格识别[2-6]、单元格分类[3,7,8]等展开研究。20 世纪 90 年代末期,随着 Web 信息的膨胀,逐渐提出了 Web 表格信息抽取任务[9,10]。目前,Web 上的数据绝大部分是由 HTML 语言描述的,缺乏对数据本身的描述,不含清晰的语义信息,模式也不明确[11],使得 Web 表格抽取比传统表格抽取更加困难。

目前,国内外对 Web 表格信息抽取的研究还处于探索阶段,本文旨在总结 Web 表格信息抽取的研究现状,以明确 Web 表格信息抽取概念及其过程、关键技术、应用及发展趋势。

2 Web 表格信息抽取及其过程

2.1 Web 表格概述

Web页面中表格的主要形式有HTML、PDF、图片、TXT、XML等,目前Web表格信息抽取的研究对象主要为

收稿日期:2007-12-11

HTML 表格。

表格由多个单元格组成,如图 1 所示。Web 页面中规范的表格通常包含在标记 < table > 与 </table > 之间,table 元素内有专门标记来表示表格的结构,如 < caption > 表示表格标题信息, < th > 表示表项标题信息, < tr > 表示表格行信息, < td > 表示数据项信息等。这些标记元素虽然能揭示一部分结构信息,但实际的 Web 表格更为复杂:例如表格标题和表项标题可能是用 < td > 标记显示的, < table > 元素仅仅用作页面布局使用等。

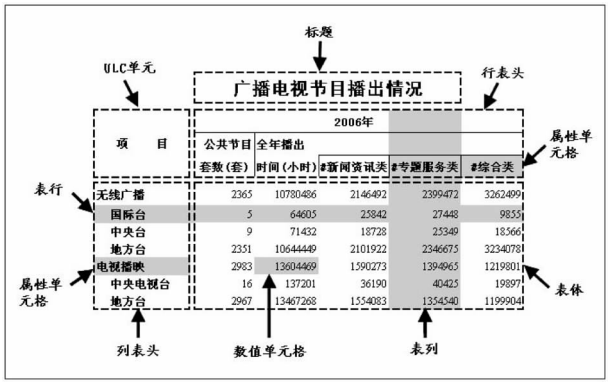

图 1　表格功能结构示例

Web 表格依据其显示特点可以分为:

(1)嵌套表格:表格以互相嵌套的方式出现,类似于网页的框架结构。

(2)“假表格”与“真表格”:如商业站点广告、导航栏或其他站点的链接,不是为了表示真正的数据,而是为了让页面美观、易读,这种表格称之为“假表格”;称真正有实际数据的表格为“真表格”。

(3)分段表格:为提升页面的视觉效果,网页设计者将一个完整的表格分割成几个片断显示,这样的表格便是分段表格。

(4)跨页表格:同分段表格类似,不同的是将一个完整的表格在不同的页面中显示。

(5)无 < table > 标记表格:不是以 < table > 标记而是利用其他标记如 < div > 、< ul > 、< li > 、< br > 、< p > 等表示的表格。

依据表格的结构类型,即“属性 - 值”对的展开方式,可分为横向(Horizontal)和纵向(Vertical)两类,也称之为按行方向展开(Row - wise)型和按列方向展开(Column - wise)型,特殊情况下还有混合型(Mix - wise)的展开方式[12]。图 2 给出了表格结构类型的部分模板,模板(a)和模板(b)只存在行表头或列表头,称为一维表格;模板(c)和模板(d)中行表头和列表头同时存在,称为二维表格。此外,还存在一些特殊结构的表格,如图 3 所示。

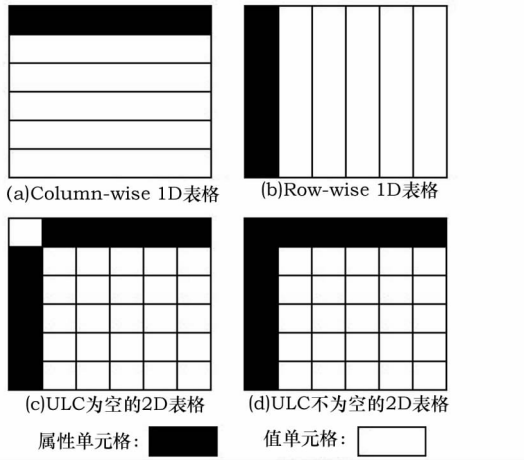

图 2　部分表格结构类型示例

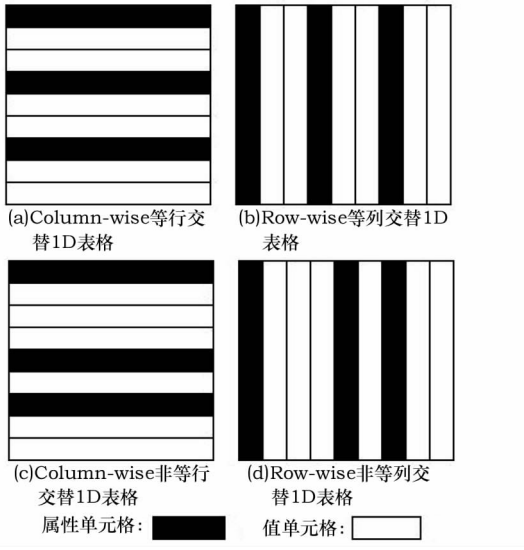

图 3　部分特殊表格结构类型示例

图 3 仅仅示例了按交替形式出现的一维表格,还有更多的一维、二维表格形式以及混合表格等[12-14]。

2.2　Web 表格信息抽取概述

信息抽取(Information Extraction)是指从各种不同的文本里定位、识别和提取出需要的信息点,表示成一种统一的、结构化的形式[15]。主要通过召回率、准确率和 F 值(F - Measure)来衡量信息抽取系统的性能。

Web 表格信息抽取是指从 Web 表格中抽取语义一致性的、结构化表示的数据和知识。目前 Web 表格信息抽取主要有 3 种方法:基于 Wrapper 学习的方法,基于表格结构分析的方法和基于本体的方法。

(1)利用归纳学习方法生成抽取规则。可以利用自动化、半自动化的手段来进行抽取器的构造工作(例如通过样例学习等),如 Lerman 等人和 Cohen 等人通过基于实例的学习算法构造包装器[16-18],学习规则的定界(Token)由 HTML 标记或关联文本组成。虽然抽取效果良好,但没有改变抽取器对页面结构的依赖,扩展性和可重用性不强。

(2)通过分析表格结构,将 HTML 表格转化为一种逻辑结构表格来抽取单元格内容。依据网页分析的描述方式,可分为基于树结构(Tree)和基于视觉线索(Visual Clues)两种抽取模式。前者利用 DOM 解析器等工具将网页解析成树状结构,抽取和分析 < table > 、< div > 或 < ul > 等特定标记对应的结点;后者利用 CSS2 Visual Box Model[19](盒状模型)等工具对 Web 文档进行解析,依据解析结果中的视觉信息(Visual Information)及空间关系(Spatial Relations)对 Web 表格的信息进行抽取。图 4 是 DOM 树和 Visual Box Model 两种 Web 页面描述方式的比较[20]。

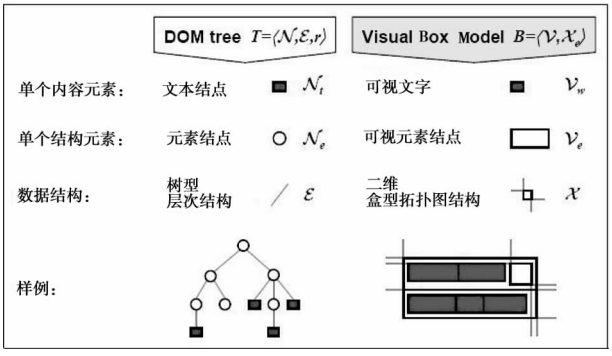

图 4 DOM 树和 Visual Box Model 两种 Web 页面描述方式的比较

(3)面向具体领域,在分析表结构的基础上,依据领域本体中对表格结构和内容的定义产生抽取规则。

2.3 Web 表格信息抽取过程

Web 表格信息抽取过程包括 Web 表格识别(Table Recognition)和 Web 表格内容抽取(Table Extraction)。Web 表格识别过程是从 Web 页中定位目标表格区域并分析表格结构;Web 表格内容抽取过程则是从表格中提取“属性—值”对并以结构化表示。

台湾学者 Chen 等人首次正式研究了 Web 表格抽取的过程,包括表格定位、表格结构识别以及“属性—值”对的提取[21]。BYU 研究小组的 Embley 等人将该过程划分为表格理解、数据整合和信息抽取几个部分,基于本体来完成对 Web 表格的定位、识别和抽取[22]。Tengli 等人的系统通过对样本表格属性内容的词汇学习及启发式规则来对表格进行定位、结构识别和“属性—值”对的提取[23]。Pivk 等人将 Web 表格抽取分析划分为 4 个层次:物理层、结构层、功能层和语义层,分别对 Web 表格进行规整(Normalization)与定位、结构识别、功能定义和语义分析[24]。Zhai 等人的方法包括两步[25]:利用标记字符的编辑距离(String Edit Distance)等视觉信息识别 Web 页面中的数据记录区域;利用基于树匹配(Tree Matching)的部分对齐技术(Partial Alingment Technique)从数据记录区域中对齐和抽取数据项。Gatterbauer 等人提出的 VENTex(Visualized Element Nodes Table EXtraction)方法则利用 Web 表格的拓扑结构、样式等视觉线索,基于 CSS2 Visual Box Model 构造启发式规则,该方法完全独立于表格所属领域,相应过程包括 Web 表格抽取和内容整合[20]。

吴扬扬等人提出了一种基于语义和数据特征的方法,包括 Web 列表识别和关系元组抽取[26,27]。林科锵、林琳在 BYU 研究小组的研究基础上,将表格处理过程分解为表格的定位、表格结构识别以及表格内容抽取 3 个步骤,并给出一个基于本体的通用 Web 表格信息抽取系统(UWTIES)模型[12,28],如图 5 所示。

3 Web 表格信息抽取关键技术分析

Web 表格信息抽取的关键技术包括:Web 表格定位(Table Detection)、Web 表格结构识别(Table Structure Recognition)、Web 表格内容整合(Table Interpretation)和抽取结果表示(Presentation of Table Extraction)等。

3.1 Web 表格定位

Web 表格定位是指从 Web 页内找到表格区域,并去除“假表格”等噪音。真假表格的判断需要构造分类器,目前有基于机器学习分类、基于人工构造规则分类及基于本体辅助分类 3 种方式。

(1)需要选取表格特征信息以及样本集来训练分类器;

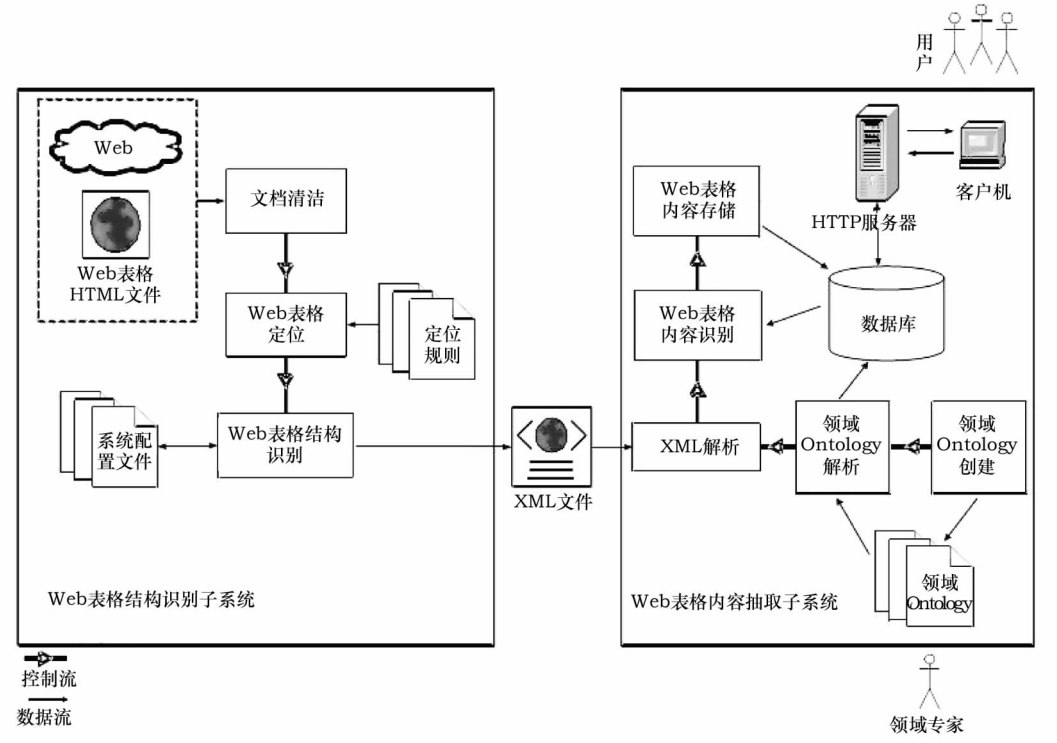

图 5　UWTIES 模型框架图

(2)需要构造表格特征的启发式规则;

(3)在前两种方式的基础上,在限定领域内利用本体判断真假表格。

国外关于 Web 表格定位的研究中,Hurst 归纳了 Web 表格的两类特征:DOM 特征(5 个)和几何模型特征(3 个),并利用两种训练算法:贝叶斯(Naive Bayes)和甄别(Winnow),对表格进行特征训练,两种算法实验结果 F 值最高分别为 94.2% 和 95.9%[29]。Wang 和 Hu 提出了 Web 表格定位时需考虑的 3 类特征:布局特征、内容类型特征和词组特征,在基于决策树学习方法和基于 SVM 学习方法的分类算法实验结果中,F 值分别为95.88% 和 95.89%[30]。BYU 研究小组的 Cui Tao 将 Web 表格分为顶层表格(Top - Level Tables)和链接页面表格(Linked - Page Tables),基于页面训练集分别提取相应的表格特征,构造启发式规则,并引入领域本体对真假表格进行判断[11]。Kim 和 Lee 将表格定位过程分为预处理和关系抽取两个阶段,从区域分割、语法、语义 3 个层次识别真假表格[31]。Liu 和 Zhai 提出 MDR(Mining Data Records) 算法,通过计算 DOM 子树结点的编辑距离,将连续相似的数据记录当作目标表格区域[32]。Gatterbauer 等人利用 CSS2 Visual Box Model 对 Web 页面进行描述,对页面中任意两个可视元素结点(Visual Element Nodes,VENs)的空间关系用"对齐方式"(5 种类型)和邻近间距这两个参数来表示,定义目标表格 (Tables of Interest,TOI)所必须具有的空间特征和语义特征[1,18]。

在国内,台湾学者 Chen 等人提出识别 Web 表格的两条规则[21]:至少含有两个单元格以表示属性和值;内容中包括许多超链接、表单和图像的将视为非表格区域。在此基础上,利用表格单元格的 3 种相似度(字符串相似度、命名实体相似度和数值类型相似度)与阈值的比较来进一步过滤非目标表格,实验结果显示 F 值为 86.5%。王放等人利用 Web 表格的 HTML 标记特点构造 Web 表格的启发式识别准则[33]。吴扬扬等人认为,Web 列表的 HTML 代码中有一系列结构相似的嵌套代码段重复出现[25,26]。林科锵构造了一系列启发式规则,并提出 Web 表格定位框架(见图 6)及各模块的主要算法,实验结果显示 F 值达 89.70%[12]。

3.2　Web 表格结构识别

Web 表格结构识别是指通过识别 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、表格展开方式识别以及表头和表体识别(属

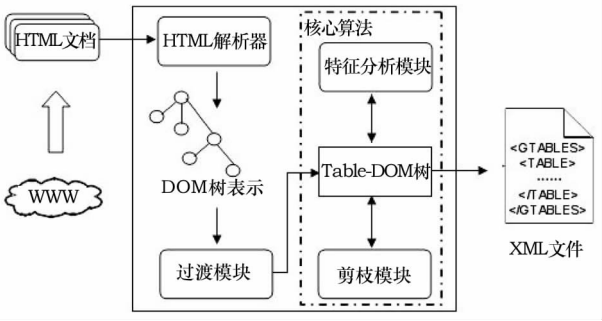

图 6　Web 表格定位框架图

性、值区域识别)。

国外关于 Web 表格结构识别研究中,Hurst 提出将 Web 表格转化成逻辑结构模型需考虑 6 个因素:单元格的内在结构、单元格拆分、单元格跨越错误(Rowspan、Colspan 值计算错误)、单元格冗长、表格对象及关联元素约束、HTML 标记规整重构[23]。Tengli 等人设计的抽取系统中,通过样本表格学习属性的词汇信息,利用向量空间模型对表格的单元格进行模糊匹配,从而识别属性单元格所在位置来定位属性和值[23]。Yoshida 等人使用特定领域的词汇信息,通过期望最大化算法(Expectation Maximization)匹配预定义的 9 种表格结构模型[13]。Pivk 等人比较表格行之间相似度和列之间相似度的数值大小来确定表格的展开方式,依据匹配表格区域的内容模式来识别属性 - 值区域并规整表格逻辑单元[24]。Yang 和 Luk 识别表格标题行和内容行的基本规则为[35]:标题行中的单元格相对较少;标题行中不包含"属性—值"对;标题行的视觉特性(Visual Characteristics)明显不同于内容行。采用模板匹配的方式对一维和二维表格进行结构识别,实验结果显示 F 值分别为 94.29% 和 94.45%。BYU 研究小组的 Cui Tao 利用表格属性的位置信息和其他结构化信息定位表格标题和内容行,确定顶层表格和链接页面表格中的属性单元格,并利用本体过滤表格属性中的无关信息[11]。

在国内,Chen 等人利用表格单元格之间的 3 种相似度计算表格相似行、列数目,确定表格展开方式及"属性—值"区域边界[21]。王放等人利用本体的学习和积累,对表格的结构识别提供指导,识别过程不依赖于所抽取的 Web 页面的格式,当应用领域发生改变时,只需要修改应用本体即可[33]。林科锵从单元格类型特征、单元格几何特征和单元格字体特征识别表格的展开方式,依据表头结构特征计算表格的类型、表头位置以及表头深度等参数,并对某些特殊表格的结构识别进行了探索[12]。

3.3　Web 表格内容整合

Web 表格内容整合是指在 Web 表格的逻辑结构上识别并规整表格存放的内容。Web 表格内容整合技术有基于规则的内容整合、基于决策树的内容整合和基于本体的内容整合。

Yoshida 等人的研究分为表格聚类和表格合并两个过程[13]。前者通过计算表格集合中属性单元格的特异度(Degree of Peculiarity)来完成,令 Uniq(a)表示属性 a 的特异度,U(a)表示出现属性 a 的表格集合,V(a)表示 U(a)中出现的属性集合,Freq(b,a)表示属性 b 在 U(a)的表格中出现的频率。

$$Uniq(a)=\left(\frac{1}{|V(a)|}\sum_{b\in V(a)}Freq(b,a)\right)\times\left(\frac{1}{|V(a)|}\sum_{b\in V(a)}\frac{Freq(b,a)}{|U(b)|}\right)$$

Uniq(a)值越大,则 U(a)中表格的主题为相似对象的可能性就越大。后者则是利用表格聚类的结果,将同类表格中相似属性合并到同一表格。Pivk 等人利用 WordNet 词典和 GoogleSets 服务识别属性单元格间的语义关系,将表格逻辑结构模型转化为 F - 逻辑框架[24]。Cui Tao 通过表格中"关系"源属性到本体中目标属性的推理映射实现内容整合,包括:预生成和调整"属性—值"对,进行映射识别;利用已识别的"属性—值"模式进行推理映射[11]。Zhai 等人提出一种基于简单树匹配(Simple Tree Matching)的部分树对齐(Partial Tree Alignment)技术,实现多重数据记录的准确对齐[25]。Gatterbauer 等人通过对 Web 表格内容整合时全局模式的统计和对应视图的概率计算,返回与具体查询内容相关的内容整合结果及其概率值,而不是传统的返回某条确定性结果[20]。

国内 Chen 等人利用 Web 表格单元格的一些标记规则对属性单元格进行内容规整,如 Rowspan、Colspan 属性等[21]。王放等人从标题识别和数据项内容识别两个角度来研究,首先利用学习模块对表格标题进行规则匹配,匹配成功的表格由应用本体识别其数据项内容,否则返回由人工处理不匹配标题与本体对象集的映射关系[33]。林琳采用本体技术实现 Web 表格的

内容识别,识别成功的条件包括表头属性列的内容层次与本体层次相对应和表头的属性列单元格内容与本体内容相匹配[28]。

3.4　抽取结果表示

抽取结果表示是指从内容整合后的 Web 表格中提取规整化内容,并以结构化的形式表示。Web 表格内容整合的结果一般为“属性—值”对的序列,“属性—值”对的提取与结果表示是抽取结果表示的主要功能。结果表示可以是 < 层级属性,值 > 的二元组、XML、关系数据表等形式。

Chen 等人认为,在“属性—值”序列中存在两种情况:某一单元格可能是多个“属性”单元格的“值”;某一单元格既是另一个“值”单元格的“属性”,也是另一个“属性”单元格的“值”。依据以上分析将最终结果表示成“属性$_1$ - 属性$_2$ - ... - 属性$_n$ - 值”结构形式[21]。Tengli 等人、Wohlberg 和 Gatterbauer 等人的研究结果中将“值”单元格与行、列标签信息相对应表示成 XML 的形式,行、列标签由“属性”及其层级构成[20,23,36]。Cui Tao、Embley 等人、Yoshida 等人、林琳和王放等人依据“属性—值”序列中的关系信息,将结果表示成关系数据表的形式,对表格内容进行结构化存储[11,13,22,28,33]。

4　Web 表格信息抽取技术的应用

目前,Web 表格信息抽取技术应用的范围主要包括:

(1)搜索引擎:与普通文本不同,Web 表格的内容文字有特定的空间关系,利用关键词搜索时难以实现准确匹配。Web 表格信息抽取技术可实现对 Web 页中表格的理解,改善搜索效果。如 CNKI 数字搜索项目[37]中,利用 Web 表格信息抽取技术对中国各地区政府统计网站中的表格进行信息抽取,抽取结果以数值知识元形式表示,满足用户数字搜索的特定需求。

(2)本体学习:手工构建本体费时费力,因而有研究者对本体的自动构建进行研究。Web 表格语义信息丰富且结构完整,有利于本体的自动学习。BYU 研究小组的 TANGO(Table Analysis for Generating Ontologies)系统就是一个从 Web 表格信息中生成领域本体的应用项目,基本过程包括:

①理解 Web 表格的结构和概念内容;

②发现概念内容间的相互约束关系,生成小型本体(Mini - Ontology);

③利用已构建的应用本体对小型本体进行概念匹配,发现本体内部的映射;

④将小型本体合并到应用本体[38]。

(3)Web 文档聚类和分类:Web 表格信息抽取技术应用于 Web 文档聚类和分类,不仅考虑到文档内容的文字特征,还顾及到结构特征和表示形式,使得结果更为准确,如 TRS InfoRadar[39]。

(4)Web 数据挖掘:Web 表格信息抽取技术最初的研究便是在 Web 表格数据挖掘的研究中出现的[21]。许多垂直搜索引擎如酷讯生活搜索[40]、Google 生活搜索[41]就是从大量中文网页表格中抽取住房、工作职位、火车票、机票等分类信息,提供专项搜索。

此外,Web 表格信息抽取技术在知识导航、机器翻译、自动文摘及数据库深加工等领域均有重大的应用价值。

5　结　语

目前,国内外关于 Web 表格信息抽取研究仍处于起步阶段,还无法达到通用领域内大规模工程项目应用的要求,主要表现在:

(1)可定位的表格范围狭窄:目前 Web 表格信息抽取技术的处理对象仍以 < table > 标记的表格对象为主,非 < table > 标记的 Web 表格定位技术可扩展性不强(如 Wrapper 学习方法等)或效果不佳(如基于视觉线索的方法等)。此外,目前的技术较难对 Web 页面中以 PDF、图片等形式出现的表格进行准确定位。

(2)可处理的表格结构不够:目前很多研究的对象仅仅是一维表格或简单的二维表格,很少有研究处理列表头单元格之间复杂的语义关系。

(3)Web 表格内容整合技术受限:一般而言,独立于具体领域的 Web 表格内容整合技术都是基于规则或基于决策树学习的,但样本量的规模和代表性制约抽取结果的准确度。面向具体领域时虽然可以利用本体解决内容整合时的语义难题,但大规模领域本体的构造较为困难。

此外,如何完成对分段表格和跨页表格的整合和抽取,以及对本文 2.1 节中提到的某些特殊类型的 Web 表格的结构识别等,都是 Web 表格信息抽取技术未来需要解决的问题。

参考文献:

[1] Gatterbauer W, Bohunsky P. Table Extraction Using Spatial Reasoning on the CSS2 Visual Box Model[C]. In: *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI 2006)*, Washington: AAAI Press, 2006: 1313 – 1318.

[2] Douglas S, Hurst M. Layout and Language: List and Tables in Technical Documents[C]. In: *Proceedings of ACL SIGPARSE Workshop on Punctuation in Computational Linguistics*, New Jersey: Association for Computational Linguistics, 1996: 19 – 24.

[3] Hu J, Kashi R S, Lopresti D, et al. Evaluating the Performance of Table Processing Algorithms[J]. *International Journal on Document Analysis and Recognition*, 2002, 4(3): 140 – 153.

[4] Ng H T, Kim C Y, Koo J L T. Learning to Recognize Tables in Free Texts[C]. In: *Proceedings of the 37th Annual Meeting of the Association for Computional Linguistics*, New Jersey: Association for Computational Linguistics, 1999: 443 – 450.

[5] Wang Y, Haralick R, Phillips I. Document Zone Content Classification and Its Performance Evaluation[J]. *Pattern Recognition*, 2006, 39(1): 57 – 73.

[6] Wang Y, Phillips I T, Robert R M, et al. Table Structure Understanding and Its Performance Evaluation[J]. *Pattern Recognition*, 2004, 37(7): 1479 – 1497.

[7] McCallum A, Freitag D, Pereira F. Maximun Entropy Markov Modals for Information Extraction and Segmentation[C]. In: *Proceeding of the 17th International Conference on Machine Learning*, 2002: 591 – 598.

[8] Pinto D, McCallum A, Wei X, et al. Table Extraction Using Conditional Random Fields[C]. In: *Proceedings of the ACM SIGIR*, 2003: 235 – 242.

[9] Hammer J, Garcia M H, Cho J, et al. Extracting Semi – structured Information From the Web[C]. In: *Proceedings of the Workshop on Management of Semistructured Data*, 1997: 18 – 25.

[10] Lim S, Ng Y. An Automated Approach for Retrieving Heirarchical Data from HTML Tables[C]. In: *Proceedings of the 8th International Conference on Informaiton and Knowledge Management (CIKM'99)*, 1999: 466 – 474.

[11] Cui Tao. Schema Matching and Data Extraction over HTML Tables[D]. Brigham Young University, USA, 2003.

[12] 林科锵. Web 页中表格结构识别的研究与实现[D]. 成都:电子科技大学, 2006.

[13] Yoshida M, Torisaw K a, Tsujii J. A Method to Integrate Tables of the World Wide Web[C]. In: *Proceedings of the First International Workshop on Web Document Analysis (WDA)*, 2001: 31 – 34.

[14] Embley D W, Lopresti D P, Nagy G. Notes on Contemporary Table Recognition[C]. In: *Proc. 7th Int. Workshopon Document Analysis Systems (DAS)*, 2006: 164 – 175.

[15] 李保利, 陈玉忠, 俞士汶. 信息抽取研究综述[J]. *计算机工程与应用*, 2003, 39(10): 1 – 5, 66.

[16] Lerman K, Getoor L, Minton S, et al. Using the Structure of Web Sites for Automatic Segmentation of Tables[C]. In: *Proc. of SIGMOD*, 2004: 119 – 130.

[17] Lerman K, Knoblock C A, Minton S. Automatic Data Extraction From Lists and Tables in Web Sources[C]. In: *Proceedings of the workshop on Advances in Text Extraction and Mining (IJCAI – 2001)*.

[18] Cohen W, Hurst M, Jensen L. A Flexible Learning System for Wrapping Tables and Lists in HTML Documents[C]. In: *Proceedings of WWW2002*, 2002: 232 – 241.

[19] Box Model[EB/OL]. [2007 – 11 – 11]. http://www.w3.org/TR/REC – CSS2/box.html.

[20] Gatterbauer W, Bohunsky P, Herzog M, et al. Towards Domain Independent Information Extraction from Web Tables[C]. In: *Proceedings of the 16th International World Wide Web Conference (WWW 2007)*, 2007: 71 – 80.

[21] Chen H, Tsai S, Tsai J. Mining Tables from Large Scale HTML Texts[C]. In: *Proceedings of the 18th International Conference on Computational Linguistics*, New Jersey: Association for Computational Linguistics, 2000: 166 – 172.

[22] Embley D W, Cui Tao, Liddle S W. Automatically Extracting Ontologically Specified Data from HTML Tables With Unknown Structure[C]. In: *Proceedings of the 21st International Conference on Conceptual Modeling (ER2002)*, 2002: 322 – 337.

[23] Tengli A, Yang Y, Li N. Machine Learning Table Extraction from Examples[C]. In: *Proceedingds of the 20th International Conference on Computational Linguistics (COLING)*, New Jersey: Association for Computational Linguistics, 2004: 987 – 993.

[24] Pivk A, Cimiano P, Sure Y. From Tables to Frames[J]. *Journal of Web Semantics*, 2005, 3(2 – 3): 132 – 146.

[25] Zhai Y, Liu B. Web Data Extraction Based on Partial tree Alignment[C]. In: *Proceedings of the 14th International World Wide Web Conference (WWW 2005)*, 2005: 76 – 85.

[26] Wu Yangyang, Yokota H. A Method of Recognizing Tables and Lists on the Web[C]. In: *Proc. of Int. Conf. on Communication, Internet, and Information Technology (CIIT 2002)*, 2002: 479 – 485.

[27] 吴扬扬, 陈锻生. 识别和抽取 Web 列表中的关系信息[J]. *计算机科学*, 2003, 31(6): 86 – 88.

[28] 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.

[29] Hurst M. Classifying TABLE Elements in HTML[C]. In: *Proceedingds of the 11th International World Wide Web Conference (WWW*

2002).

[30] Wang Y, Hu J. A Machine Learning Based Approach for Table Detection on the Web[C]. In: *Proceedings of the 11th International Conference on World Wide Web*, 2002:242 - 250.

[31] Kim Y, Lee K. Detecting Tables in Web Documents[J]. *Engineering Appliations of Artificial*,2005(18):745 - 757.

[32] Liu B, Zhai Y. NET - A System for Extracting Web Data from Flat and Nested Data Records[C]. In: *Proceedings of the 6th International Conference on Web Information Systems Engineering*(*WISE - 05*), Washington: IEEE Computer Society Press,2005:487 - 495.

[33] 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.

[34] Hurst M. Layout and language: Challenges for Table Understanding on the Web[C]. In: *Proc. 1st International Workshop on Web Document Analysis*, *CA*: *Prima Communications*,2001:27 - 30.

[35] Yang Y, Luk W. A Framework for Web Table Mining[C]. In: *Proceedings of the 4th International Workshop on Web Information and Data Management*, 2002:36 - 42.

[36] Wohlberg T. Hypertables: Development of a Structure Description Language for Tables in XML[D]. University of Hamburg, Germany,1999.

[37] CNKI 数字搜索[EB/OL]. [2007 - 11 - 15]. http:// number.cnki.net/.

[38] Tijerino Y A, Embley D W, Deryle L, et al. Towards Ontology Generation from Tables[J]. *WorldWide Web Journal*, 2005(8): 261 - 285.

[39] TRS InfoRadar[EB/OL]. [2007 - 11 - 15]. http://www.trs.com.cn/products/wse/radar/.

[40] 酷讯生活搜索[EB/OL]. [2007 - 11 - 15]. http://www.kooxoo.com/.

[41] Google 生活搜索[EB/OL]. [2007 - 11 - 15]. http://www.google.cn/shenghuo/.

(作者 E - mail: zhaohong860112@163.com)