计算机研究与发展
Journal of Computer Research and Development
ISSN 1000-1239/CN 11-1777/TP
43(9):1644～1650, 2006

基于 HTML 模式代数的 Web 信息提取方法

李石君[1,2]　于俊清[3]　欧伟杰[1]
[1](武汉大学计算机学院　武汉　430072)
[2](中国科学院计算机科学重点实验室　北京　100080)
[3](华中科技大学计算机科学与技术学院　武汉　430072)
(shjli@public.wh.hb.cn)

Web Information Extraction Based on HTML Pattern Algebra

Li Shijun[1,2], Yu Junqing[3], and Ou Weijie[1]
[1](*School of Computer Science, Wuhan University, Wuhan* 430072)
[2](*Laboratory of Computer Science, Chinese Academy of Sciences, Beijing* 100080)
[3](*School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan* 430072)

Abstract　Generating wrapper efficiently for extracting Web data has broad application prospect, but is also a difficult problem that is not yet solved efficiently till now. To tackle this problem, a pattern algebra for HTML documents is introduced, which includes key concepts, such as consistent pattern set, and the addition operation of pattern, and based on it a new approach to extract Web information is presented. It induces the consistent pattern set which represents identifying rules of each attribute by exploring the whole samples, and then extracts data by the consistent pattern set with multiple patterns. It can apply Web pages with tabular structure, in which there are missing attributes or attributes with multiple values or different order and hierarchical structure, and has been validated experimentally in the prototype.

Key words　Web information extraction; wrapper induction; Web mining

摘　要　高效地生成提取 Web 信息的包装器有着广阔的应用前景,同时也是至今没有得到有效解决的难题. 为此,提出了基于 HTML 文档的模式代数,该代数包括一致模式集等重要概念以及模式的加法运算. 在此基础上,提出了一种提取 Web 信息的新方法,该方法采用在整个训练例子中学习表示各属性提取规则的一致模式集,再由多个模式组成的一致模式集提取数据,适用于提取具有缺省属性、多值属性、属性具有多种不同顺序的表结构网页和层次结构网页,其有效性在原型系统中通过实验得到验证.

关键词　Web 信息提取;包装器归纳学习;Web 挖掘
中图法分类号　TP311.135

高效地生成提取 Web 信息的包装器在 Web 信息检索、集成、挖掘中有着广阔的应用前景. 尽管基于 RDF,XML 和 OWL 的语义 Web 的目的是使新一代 Web 能被机器理解,但 Web 现有的和目前每天发布的主要信息都基于 HTML,因此,目前期待整个 Web 采用语义 Web 标注还不切实际. 由于 HTML 只描述数据显示,不描述数据的内容,不便于机器理解和查询,因此,提取 Web 信息是十分困难的.

包装器是一个从特定信息源提取数据的过程. Web 包装器以网页作为输入,返回网页中用户感兴趣的数据. 返回数据通常采用关系或者 XML 表示. 手写包装器开发复杂且维护困难. 归纳学习方法通过用户标注例子 Web 网页中需提取的数据,归纳学

收稿日期:2005-05-11;**修回日期**:2006-03-14
基金项目:国家自然科学基金项目(60573095);湖北省自然科学基金项目(2005ABA238).

习提取 Web 数据的包装器. 现有的包装器归纳学习方法仅在数据的左右两边标记(token)中学习该数据的提取规则,并仅采用单一提取规则来表示每个需提取的数据. 因此,其表达能力较弱,仅能提取部分 Web 网页,不能适用于格式多样、经常变化的 Web 网页. 为此,本文扩展了现有方法,首先为形式化表示 HTML 文档中数据的识别规则和特点,提出了基于 HTML 文档的模式代数,该代数形式化地定义了 HTML 文档的模式、模式匹配、模式空间、一致模式集等重要概念;第 2,在整个例子标记中学习识别需提取数据的模式规则,而不仅仅只在属性左右两边标记中学习,因此提高了表达能力;第 3,采用由多个模式组成的一致模式集表示需提取数据的识别规则,因而适合于提取格式多样的 Web 网页.

1 Web 信息的表示和提取

本文采用类似文档对象模型 DOM 的层次树表示 Web 信息,在该层次树中页结点表示需提取的数据,内部结点表示复合结点,分为一般复合结点和 *list* 复合结点(记为 $list(x)$,表示多个 x 结点的集合).

例 1. 以下是一个有关产品信息的 HTML 文档节选,其层次树表示如图 1 所示.

```
〈B〉Teatime Chocolate Biscuits〈/B〉〈BR〉
  〈I〉9.20〈/I〉〈BR〉
〈B〉Specialty Biscuits, Ltd.〈/B〉
  〈B〉29 King's Way〈/B〉
    〈B〉(26) 555-4448〈/B〉〈BR〉
      〈LI〉〈I〉100-500〈/I〉〈I〉5 %
〈/I〉〈BR〉〈LI〉〈I〉501-〈/I〉〈I〉8 %〈/I〉〈BR〉…:
```

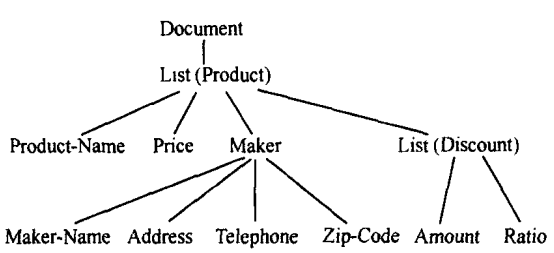

Fig. 1 The hierarchical tree of the Web page in example 1.

图 1 例 1 中 Web 页的层次树

为归纳学习 Web 信息的提取规则,需提供被标注的例子页[1]. 例如,用户可用鼠标点亮标注需提取的数据. 对 $list(x)$ 结点,可由标注的实例学习提取各 x 结点的规则. 在上一层结点(父结点)已被正确提取后,可由标注的实例学习父结点中各个子结点的提取规则. 这样 Web 层次信息可由上至下逐层被正确地提取. 因此,Web 层次信息的提取问题可归结为在父结点被提取后如何正确提取各子结点的问题.

例 2. 设需提取例 1 中子结点厂家名、地址、电话、邮编的值. 首先由用户标注的商品例子,学习 *list*(商品)中各商品的提取规则并提取各商品的值,然后由用户标注的厂家例子,学习各厂家的提取规则并提取各厂家的值,最后由用户标注的厂家名、地址、电话、邮编例子,学习厂家名、地址、电话、邮编的提取规则并提取其值. 设被正确提取的各厂家值为如下所示的 e_1, e_2, e_3, e_4: e_1: *Specialty Biscuits, Ltd.* 〈/B〉〈B〉29 *King's Way*〈/B〉〈B〉(26) 555—4448; e_2: *New England Seafood Cannery*〈/B〉〈B〉2100 *Paul Revere Blvd.*〈/B〉〈B〉(617) 555—3267〈/B〉〈B〉02134; e_3: *G'day, Mate*〈/B〉〈B〉170 *Prince Edward Hill*〈/B〉〈B〉2042〈/B〉〈B〉(02) 555—5914〈/B〉〈B〉(02) 555—5915: e_4: *Pavlova, Ltd.* 〈/B〉〈B〉 *Rose St. Moonie Ponds*〈/B〉〈B〉03-444-2343〈/B〉〈B〉3058. 其中需提取的子结点数据采用关系模式可表示为厂家(厂家名、地址、电话、邮编). 但与传统关系模型不同的是提取的元组中存在缺省属性(如 e_1 缺省邮编)、属性存在多属性值(如 e_3 包含两个电话号码)、属性的顺序不一致(如 e_2 的顺序是电话、邮编,而 e_3 是邮编、电话). 现有方法不能正确提取例 2 中的数据. 以下研究通过例子学习父结点 e_1, e_2, e_3, e_4 中各子结点厂家名、地址、电话、邮编的提取规则,并正确地提取其值.

2 基于 HTML 的模式代数

HTML 文档是由标记(token)组成的序列,这些标记包括 HTML 标记、标点符号、分隔字符(空格、"-"等)、数字、小写字符串、大写字符串等. 本文用 $Num(-)$, $Punc(-)$, $SP(-)$, $SL(-)$, $SUL(-)$, $SU(-)$ 分别表示 Web 页中数字标记、标点符号、分隔字符标记、首字母为小写字母的字、仅首字母为大写字母的字、其余字的泛化标记[2](generalization token). 在以下归纳学习中,各种类型的具体标记都可以泛化(generalizing)为相应的泛化标记. 如例 1 中的具体数字标记"29","100"等在归纳学习中都可以泛化为数字的泛化标记 $Num(-)$. 从父结点中提取各个子结点,关键是获得子结点在父结点中开始和

结束识别规则.因结束识别规则与开始识别规则类似,以下只讨论开始识别规则,识别 x 指识别 x 的开始.

定义 1. 模式是由 HTML 的标记、泛化标记、通配标记“ * ”或直到标记“ $until(t)$ ”组成的一个序列,其中 t 为 HTML 的标记或泛化标记,直到标记“ $until(t)$ ”只能出现在最后.

定义 1 模式中通配标记“ * ”表示通配任意标记,直到标记“ $until(t)$ ”的语义表示“直到 t 出现”.

定义 2. 设标记序列 $S = S_1 x S_2$, x 为子结点,若模式 r 的最后一个标记不为直到标记“ $until(t)$ ”,则模式 r 匹配 S_1 是指模式 r 中的非通配标记在 S_1 中依次出现. 若模式 r 的最后一个标记为直到标记“ $until(t)$ ”,则模式 r 匹配 S_1 是指模式 r 中最后一个标记前的非通配标记在 S_1 中依次出现,而 t 在 x 中出现. 若模式 r 匹配 S_1, 则称模式 r 识别 x;若模式 r 在 S 中不匹配,则称模式 r 与 S 无关.

例 3. 对例 2 中的 e_3,模式 $r_1 = (\ *\langle B\rangle)$ 识别地址结点,模式 $r_2 = (\ *\langle B\rangle\ *\langle B\rangle)$ 识别邮编结点,但对 e_2, r_2 将导致错误识别,因邮编结点的第 1 个标记为数字,而电话号码的第 1 个标记为标点符号. 模式 $r_3 = (\ *\langle B\rangle\ *\langle B\rangle until(Num(-)))$ 将识别 e_3 中的邮编结点,同时,对 e_2, r_3 将不会导致错误识别,其中 $until(Num(-))$ 的语义表示“直到数字标记出现”.

定义 3. 以下对标记序列 $A = a_1 a_2 \cdots a_k$,记 $M_A = \{(a_1' a_2' \cdots a_k') \mid a_i' = a_i \vee a_i' = * \vee a_i' = g(a_i),\ i = 1,2,\cdots,k\}$,其中, $g(a_i)$ 表示 a_i 的泛化标记. 设 $m_1 \in M_A$, $m_2 \in M_B$, $m_1 = a_1^1 a_2^1 \cdots a_k^1$, $m_2 = a_1^2 a_2^2 \cdots a_k^2$,则 m_1, m_2 之和定义为 $m = m_1 + m_2 = \{a_1' a_2' \cdots a_k' \mid a_i' = a_i^1 + a_i^2,\ i = 1,2,\cdots,k\}$,其中 $a_i' = a_i^1 + a_i^2$ 定义如下:若 a_i^1, a_i^2 中有一个为 $g(a_i)$,则和为 $g(a_i)$;否则,若有一个为 a_i,则和为 a_i;若两者都为 *时,则和为 *.

定义 4. 设标记序列 $S = BxA$, x 是需提取的结点,定义 $\Omega = \{(m_1, until(m_x)) \mid (m_1 \in M_B) \wedge (m_x(M_x))\}$ 为序列 S 中子结点 x 的模式空间.

定义 5. 设序列 $S = BxA$ 中子结点 x 的模式空间为 Ω, $r_1 = (m_1, until(n_1)) \in \Omega$, $r_2 = (m_2, until(n_2)) \in \Omega$,其中 $m_1, m_2(M_B, n_1, n_2 \in M_x$,则 r_1 与 r_2 的和定义为 $r = r_1 + r_2 = ((m_1 + m_2), until(n_1 + n_2))$.

设序列 S 中子结点 x 的模式空间为 Ω, r_1, r_2, $r_3 \in \Omega$, 由模式和的定义可得:

性质 1. ① $r_1 + r_2 = r_2 + r_1$; ② $(r_1 + r_2) + r_3 = r_1 + (r_2 + r_3)$; ③若 r_1 与序列 K 无关, r 是模式空间 Ω 中任一模式,则 $r_1 + r$ 也与序列 K 无关.

3 模式类包装器归纳学习

以下设例子集 E 分为包含子结点 x 的例子集 E_1: $E_1 = \{e_1 : S_1 = S_{11} x S_{12}, e_2 : S_2 = S_{21} x S_{22}, \cdots, e_m : S_m = S_{m1} x S_{m2}\}$ 和不包含子结点 x 的例子集 E_2,即 $E = E_1 \cup E_2$.

定义 6. 若模式 r 识别 E_1 的子集 E_j 中每一个子结点 x,且 r 与 $E_2 \cup (E_1 - E_j)$ 中的每一个无关,则称模式 r 为例子集合 E 中子结点 x 的一致模式. 对模式集 $R = \{r_1, r_2, \cdots, r_k\}$,其中 r_i 是与例子集合 E 一致的模式,令 r_i 从 r_1 到 r_k 循环,每次循环置 $E = E - E_j$, E_j 为例子集 E 中 r_i 能识别的子集,则若循环结束时 $E = E_2$,则称模式集 R 为例子集 E 中子结点 x 的一致模式集.

设需从例子集 E 提取 n 个子结点 $x_1, x_2, \cdots, x_n$. 若对任一 $x_i (i = 1,2,\cdots,n)$,存在例子集 E 中的一致模式集 L_i,则提取 n 个子结点为扫描各个例子标记序列,依次从 $L_1, L_2, \cdots, L_n$ 中查找,若存在 L_i 包含匹配该序列的模式,则匹配的位置是子结点 x_i 的开始位置. x_i 的结束位置可同样得到. 因此,提取Web 信息的关键是得到各个子结点的一致模式集.

定义 7. 设 $L_1, L_2, \cdots, L_n$ 分别为子结点 $x_1, x_2, \cdots, x_n$ 的(开始)一致模式集, $R_1, R_2, \cdots, R_n$ 分别为子结点 $x_1, x_2, \cdots, x_n$ 的结束一致模式集,则模式类包装器定义为 $\{\langle L_1, R_1\rangle, \langle L_2, R_2\rangle, \cdots, \langle L_n, R_n\rangle\}$.

穷举搜索子结点 x 的模式空间可以找到与子结点 x_i 一致的模式,但是一个 NP 完全问题. 以下采用启发式方法,称 HTML 标记、标点和分隔符、其他标记的顺序为特征顺序. 在集合 E_1:$\{S_{11}, S_{21}, \cdots, S_{m1}\}$中按特征顺序选取最长的序列. 按特征顺序的原因是 HTML 标记是决定属性的开始和结束的最重要的标志,其次为标点和分隔标记. 选择最长的原因是能够提供足够多的信息将错误匹配转变为正确匹配或无关. 设 E_1 中特征最长序列为 S_{i1}, $S_i = S_{i1} x S_{i2} = a_1 a_2 \cdots a_k x S_{i2}$,设结点 x 的首标记为 b,算法分为以下两步.

第 1 步:选取 S_i 关于 x 的一元模式 $r_1 = (\ * \cdots * a_k)$ 或 $r_2 = (\ * \cdots *\ until(b))$,用 r_1 或 r_2 去匹配 S_i,若两者都提前匹配 S_i,则按特征顺序从 a_{k-1} 到 a_1 中选取一个标记 a_j,改进模式 r_1 和 r_2 为 $r_1 = r_1 + (\ * \cdots * a_j\ * \cdots *)$, $r_2 = r_2 + (\ * \cdots * a_j\ * \cdots *)$,如此循环,直到 r_1 或 r_2 识别 x. 记改进模式 r_1 和 r_2 的过程为 $AddSelfRule$.

$AddSelfRule(S = a_1 a_2 \cdots a_k x S_2)$

$r_1 = (\ * \cdots * a_k)$, $r_2 = (\ * \cdots until(b))$ / * b 为 x 的首标记 */

While r_1 和 r_2 都提前匹配 S,

按特征顺序从 a_{k-1} 到 a_1 中选取一个元素 a_j

$r_1 \leftarrow r_1 + (\ * \cdots * a_j\ * \cdots *)$, $r_2 \leftarrow r_2 + (\ * \cdots * a_j\ * \cdots *)$,

If r_1 匹配 S Then 返回 r_1 Else 返回 r_2.

第 2 步:用第 1 步得到的模式 r 去匹配例子集 $E = E_1 \cup E_2$,设 r 对 E 产生错误匹配的子集为 E',对 E 的每个例子 S_j,设匹配的标记序列为 S_j',按特征顺序从 a_{k-1} 到 a_1 中选取一个不在 S_j' 的逆置标记序列中的标记,若不存在这种标记,则例子集 E 不存在子结点 x_i 的一致模式集. 否则,设该标记为 a_j,则将模式 r 改进为 $r = r + (\ * \cdots * a_j\ * \cdots *)$,改进的模式 r 必将对例子 S_j 的匹配后移,直到 r 要么匹配 S_j,要么与 S_j 无关. 记改进模式 r 的过程为 $AddNewRule$. 若子结点 x_i 存在一致模式,循环调用 $AddNewRule$,得到的改进模式 r 对 E 的每一个例子要么匹配,要么无关,即 r 是 E 中 x_i 的一致模式. 记产生一致的模式的过程为 $GenModeRul$. 利用 $GenModeRul$,由一致模式集的定义可产生 E 中 x_i 的一致模式集,算法记为 $GenModeRuleSet$.

$AddNewRule$(模式 $r, S = a_1 a_2 \cdots a_k x S_2$) / *改进的模式 r */

While r 对 S 产生错误匹配(设匹配的元素序列为 S'),

If 按特征顺序从 a_{k-1} 到 a_1 中选取一个不在 S 的逆置元素序列中的元素 a_j

Then $r \leftarrow r + (\ * \cdots * a_j\ * \cdots *)$,

返回 r,

$GenModeRule(E = E_1 \cup E_2, E_1 :\{ e_i : S_1 = S_{11} x S_{12}, e_2 : S_2 = S_{21} x S_{22}, \cdots, e_m : S_m = S_{m1} x S_{m2}\})$ / *产生一致的模式 */

$i \leftarrow$从例子集 E_1 中选择特征最长序列 S_{i1}

$r \leftarrow AddSelfRule(S_{i1} = a_1 a_2 \cdots a_k)$

/ *设 $S_i = S_{i1} x S_{i2} = a_1 a_2 \cdots a_k x S_{i2}$ */

For $E = E_1 \cup E_2$ 的每个例子 $S_j = S_{j1} x S_{j2}$

If r 对 S_j 中 x 错误识别 Then

$r \leftarrow AddNewRule(r, S_j = S_{j1} x S_{j2})$

/ *改进错误匹配的模式 r 为正确匹配或无关 */

While E 中存在滞后匹配或虚假匹配例子 $S_j = S_{j1} x S_{j2}$,

$r \leftarrow AddNewRule(r, S_j = S_{j1} x S_{j2})$

$E = E - \{ S_j \}$

返回 r

$GenModeRuleSet(E = E_1 \cup E_2)$

/ *产生一致的模式集 */

While$(E_1 \neq \varnothing)$

$r \leftarrow GenModeRule(E = E_1 \cup E_2)$

/ *由例子集 E 产生一个与 x 一致的模式 */

For E_1 中每一个例子 e_i,

/ *从 E 中删除能由 r 识别的例子 */

If r 识别 e_i Then $E_1 \leftarrow E_1 - \{ e_i \}$

$R \leftarrow R \cup \{ r \}$ / * R 初值为空 */

返回 R.

由 $GenModeRule$ 算法可知,由该算法产生的模式是与各子结点一致的模式;因此,$GenModeRuleSet$ 算法是有效的. 至于算法的完备性问题是一个有待进一步研究的问题.

例 4. 以下对例 2 中的例子集 E 求关于邮编子结点的一致模式集. 此时 $E_1 = \{ e_2, e_3, e_4 \}$, $E_2 = \{ e_1 \}$,选择特征最长序列为 e_2 的 S_{21}(即"*New England Seafood Cannery* 〈/B〉 … 555 —3267 〈/B〉〈B〉"),调用 $AddSelfRule$ 可得到识别 e_2 的模式:$r_1 = (\ *\langle B\rangle\ *\langle /B\rangle\langle B\rangle\ until(Num(-)))$,该模式识别 e_2, e_3,与 e_1 无关,对 e_4 错误识别. 对 e_4 调用 $AddNewRule$ 得到改进的模式:$r_1' = (\ *\langle B\rangle\ *\ -\ *\langle /B\rangle\langle B\rangle\ until(Num(-)))$, r_1' 识别 e_4, e_2,但将导致以前识别的 e_3 变为无关,由性质 1 可得 r_1' 仍然与 e_1 无关. 置 $E_1 = E_1 - \{ e_4, e_2 \} = \{ e_3 \}$,重复以上步骤选择特征最长序列为 e_3 的 S_{31},调用 $AddSelfRule$ 可得识别 e_3 的模式:$r_2 = (\ *\langle B\rangle\ *\langle /B\rangle\langle B\rangle)$. 此时 $E_1 = E_1 - \{ e_3 \} = \varnothing$. 因此,例子集 E 中邮编子结点的一致模式集为 $L_4 = \{ r_1', r_2 \}$ 即 $L_4 = \{ (\ *\langle B\rangle\ *\ -\ *\langle /B\rangle\langle B\rangle\ until(Num(-))), (\ *\langle B\rangle\ *\langle /B\rangle\langle B\rangle) \}$. 同样可得到厂家名的一致模式集为 $L_1 = \{\ * until(SUL(-))\}$,地址为 $L_2 = \{\ *\langle B\rangle\}$,电话为 $L_3 = \{\ * until((), *(\ * until((), *\ until(Num$

(-)-)}. 注意,在电话的一致模式集中,第 1 个模式识别 e_1, e_2、第 2 个模式识别 e_3 中的多个电话、第 3 个模式识别 e_4. 类似可得到识别各子结点的结尾的一致模式集 R_1, R_2, R_3, R_4. 故例 2 中的例子集 E 的模式类包装器为{〈L_1, R_1〉,〈L_2, R_2〉, …,〈L_4, R_4〉}. 由该包装器可提取各个厂家 S_1, S_2, …S_n 中厂家名、地址、邮编和电话.

4 实验结果

本文实现了模式类包装器原型系统,对实际 Web 站点进行了实验,部分站点的实验结果情况如表 1 和表 2 所示. 表 1 主要包括层次结构的 Web 站点,表 2 主要包括具有缺省属性、多属性值、属性顺序不惟一的表结构 Web 站点. 表 1 各列分别为站点 URL、Web 数据树结构的层数、由查准率,查全率,训练例子数表示的实验结果. 表 2 各列分别为站点 URL、属性个数、由 Yes 和 No 表示表结构网页信息是否具有缺省属性、多属性值、属性顺序不惟一. 例如,表 2 第 2 行中包含书目信息的网址中"作者"存在多属性值;大部分书记录具有"BibTex"属性值,但少部分缺省该属性值,另外属性"书号 ISBN"和"年份"都存在缺省;属性"书号 ISBN"和属性"年份"具有两种顺序,一是"年份"在前"书号 ISBN"在后,二是"书号 ISBN"在前"年份"在后. 第 4 列为平均查准率、平均查全率、训练例子数.

Table 1 Experiment Result of Web Sites with Hierarchical Structure

表 1 具有层次结构数据的 Web 站点实验结果

Web Site URL	Number of Layers	Mean Precision, Recall and Number of Train Examples
www.jobsjobsjobs.com	3	Type: 100 %, 100 %, 2; Date: 100 %, 100 %, 2; List of Company Name and Their URL: 100 %, 100 %, 2.
www.travlang.com/languages	2	Language: 100 %, 100 %, 1; URL: 100 %, 100 %, 2; Image: 100 %, 100 %, 6; Translation: 88 %, 100 %, 10.
www.informatik.uni-trier.de/~ley/db//groups/index.html	3	State: 100 %, 100 %, 2; Nation: 100 %, 100 %, 2; University: 100 %, 100 %, 1; Group: 70 %, 83 %, 8; List of Group Member and URL: 60 %, 78 %, 10.
www.odci.gov/cia/publications/factbook/	3	Nation Name: 100 %, 100 %, 1; List of State, Capital, Area, Population: 100 %, 100 %, 1; Total Area: 100 %, 100 %, 1; Total Population: 100 %, 100 %, 1.
us.imdb.com/top-250-films	2	Film Name: 100 %, 100 %, 1; Directed by: 100 %, 100 %, 1; Screenplay 70 %, 86 %, 3; Cast : 100 %, 100 %, 5.

Table 2 Experiment Result of Web Sites of Table Structure with Missing Attributes or Multiple Values Attributes or in Different Order

表 2 具有缺省属性、多属性值、属性顺序不惟一的表结构 Web 站点实验结果

Web Site URL	Number of Attributes	Missing Attributes/Multiple Values/Different Order	Mean Precision, Recall and Number of Train Examples
www.altavista.digital.com	5	Yes/Yes/No	85 %,83 %,5
www.informatik.uni-trier.de/~ley/db/books/collections	6	Yes/Yes/Yes	76 %,71 %,9
www.cs.washington.edu/people/faculty.html	2	Yes/No/No	100 %,100 %,3
www2.technofind.com.sg/tf	4	Yes/No/No	100 %,100 %,5
Shareware.cnet.com	5	Yes/No/No	100 %,100 %,4

不失普遍性,本文实验网站的选取没有采用随机选择方法,而是有针对性地选择上述具有缺省属性、多属性值、属性顺序不惟一的不规范表结构网站和层次结构网站,以验证本方法的有效性.

实验采用平均查准率、平均查全率、训练例子作为评价标准. 从实验结果来看,本文提出的方法既能提取具有缺省属性、多属性值、属性顺序不惟一的表结构网页信息又能提取层次结构网页信息,其平均查准率、查全率较高,在某些网站上接近 100 %,能够较好地完成 Web 数据的提取,且其训练例子在可接受的范围内. 由实验结果可见本文提出的方法是有效的.

5 相关研究

怎样提取 Web 信息并将其转换为关系数据或机器能理解和查询处理的 XML[3]是一个困难的前沿研究课题,主要采用手工编码方法[4-5]、归纳学习方法[1-2,6-9]、启发式方法[10]和本体方法[11]. 手工编写代码方法表达能力强但需要编写复杂的代码,且由于网页经常变化而维护困难. WIEN[1]提出了采用归纳学习生成包装器的方法. 主要针对规范的表式结构信息,数据的识别和提取依赖于数据左右两边的公共标记. 其表达能力弱. SOFTMEALY[2]采用有限状态自动机归纳学习提取 Web 信息的包装器,其中数据的识别依赖于数据的左右两边标记特征,较 WIEN 表达能力强,但只针对表结构数据,不适合层次结构 Web 信息. STALKE[6]提出了采用归纳学习方法提取层次结构信息源的方法,通过例子学习由 landmark(数据的左右两边连续若干个标记)表示的提取规则. 由于 Web 信息的复杂性和多样性,很难获得 landmark 来惟一地定位要提取的信息. 本文提出的基于模式代数的包装器归纳学习方法与现有方法相比,具有以下特点:①提出了基于 HTML 文档的模式代数,其核心是识别需提取数据的一致模式集;②在整个训练例子中搜索识别需提取数据的由多个模式组成的一致模式集. 而存在的相关方法仅仅只在需提取数据相邻的标记中学习识别该数据的特征标记,且采用单一的识别规则,因而不能完全刻画识别该数据的特征. 因此,模式类包装器增强了表达能力,适合提取具有缺省属性、多属性值、属性具有多种不同顺序的 Web 网页信息;既支持提取表结构数据网页信息又支持提取层次结构网页信息.

6 结束语

本文提出了基于 HTML 文档的模式代数,在此基础上,提出了一种提取 Web 信息的新方法. 该方法定义了模式类包装器,并给出了其归纳学习生成算法. 该方法采用在整个训练例子中学习由多个模式组成的一致模式集,适用于提取具有缺省属性、多值属性、属性具有多种不同顺序的表结构网页和层次结构网页,其有效性在原型系统中通过实验得到验证. 本文进一步的研究工作包括在模式类包装器中结合对象代理模型[12],以及对模式代数的完备性等有关理论问题进行深入的研究.

参 考 文 献

[1] Nicholas Kushmerick, D Weld, R Doorenbos. Wrapper induction for information extraction [C]. Int'l Joint Conf on Artificial Intelligence, Hyderabad, India, 1997

[2] Hsu, C M Dung. Generating finite-state transducers for semi-structured data extraction from the Web [J]. Journal of Information Systems, 1998, 23(8): 521-538

[3] Lü Jianhua, Wang Guoren, Yu Ge. Optimizing path expression queries of XML data [J]. Journal of Software, 2003, 14(9): 1615-1620 (in Chinese)
(吕建华, 王国仁, 于戈. XML 数据的路径表达式查询优化技术[J]. 软件学报, 2003, 14(9): 1615-1620)

[4] J Hammer, H Garcia-Molina, S Nestorov, *et al*. Template-based wrappers in the TSIMMIS system [C]. Int'l Conf on Management of Data, Tucson, Arizona, 1997

[5] R Baumgartner, S Flesca, G Gottlob. Visual Web information extraction with Lixto [C]. Vary Large Data Bases, Roma, Italy, 2001

[6] I Muslea, S Minton, Knoblock. A hierarchical approach to wrapper induction [C]. Third Conf on Autonomous Agents, Seattle, WA, 1999

[7] Xiaofeng Meng, Hongjun Lu, *et al*. Data extraction from the Web based on pre-defined schema [J]. Journal of Computer Science and Technology, 2002, 17(4): 377-388

[8] Li Xiaodong, Gu Yuqing. DOM-based information extraction for the Web source [J]. Chinese Journal of Computers, 2002, 25(5): 526-533 (in Chinese)
(李效东, 顾毓清. 基于 DOM 的 Web 信息提取[J]. 计算机学报, 2002, 25(5): 526-533)

[9] Hu Dongdong, Meng Xiaofeng. Automatically extracting Web data using tree structure [J]. Journal of Computer Research and Development, 2004, 41(10): 1607-1613 (in Chinese)
(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)

[10] Huang Yuqing, Qi Guangzhi, Zhang Fuyan. Extracting semi-structured information from the Web [J]. Journal of Software, 2000, 11(1): 73-78 (in Chinese)
(黄豫清, 戚广志, 张福炎. 从 Web 文档中构造半结构化信息的提取器[J]. 软件学报, 2000, 11(1): 73-78)

[11] Gao Jun, Wang Tengjiao, Yang Dongqing, *et al*. Ontology-based two-phase semi-automatic Web extraction [J]. Chinese Journal of Computers, 2004, 27(3): 310-318 (in Chinese)
(高军, 王腾蛟, 杨冬青, 等. 基于 Ontology 的 Web 内容二阶段半自动提取方法[J]. 计算机学报, 2004, 27(3): 310-318)

[12] Peng Zhiyong, Luo Yi, Shan Zhe, *et al*. Realization of workflow views based on object deputy model [J]. Chinese Journal of Computers, 2005, 28(4): 651-660 (in Chinese)
(彭智勇, 罗义, 单哲, 等. 基于对象代理模型的工作流视图实现[J]. 计算机学报, 2005, 28(4): 651-660)

Li Shijun, born in 1964. He has been professor of Wuhan University since 2006. His main research interests include database technology, Web information technology, and secure data management.

李石君,1964 年生,教授,主要研究方向为数据库技术、Web 信息技术、数据库安全.

Yu Junqing, born in 1975. He has been associate professor of Huazhong University of Science and Technology since 2005, senior member of China Computer Federation. His main research interests include retrieving of video information and artificial intelligence.

于俊清,1975 年生,副教授,中国计算机学会高级会员,主要研究方向为视觉信息检索、人工智能.

Ou Weijie, born in 1981. He has been Ph D candidate in computing science from Wuhan University. His main research interests include Web information and data mining.

欧伟杰,1981 年生,博士研究生,主要研究方向为 Web 信息技术、数据挖掘.

Research Background

Generating wrapper efficiently for extracting Web data has broad application prospect, including Web information retrieval and integration, and Web mining. In this paper, we introduces a pattern algebra for HTML documents and a new approach to extract Web information, which induces the consistent pattern set with multiple patterns by exploring the whole samples, and can apply Web pages with table structure, in which there are missing attributes or attributes with multiple values or different order, and hierarchical structure. Our work is supported by the Hubei Provincial Science Foundation of (2005ABA238) and the National Natural Science Foundation of China (60573095).