一种基于未知结构网页抽取本体的方法

强 宇[1,2] 胡运发[1]

(复旦大学计算机与信息技术系 上海 200433)[1] (蚌埠坦克学院计算机室 蚌埠 233013)[2]

摘 要 在Web上数据大多是结构化的,但事先并不熟知数据的结构,因此不能有效地查询感兴趣的数据。提出了一种独立于文本抽取本体的方法,其过程包括表的理解、数据集成和本体生成,其中表理解是搜寻定位兴趣表、识别及匹配属性和值,并形成记录;数据集成是匹配源记录和目标模式;本体卷积是将源记录的数据抽取到目标模式。结果表明这种方法可以通过已知的目标模式有效地抽取未知结构的数据。

关键词 异质数据集成,语义对应,表理解,本体抽取

Method for Ontology Extraction Based on Unknown Structure Web

QIANG Yu[1,2] HU Yun-fa[1]

(Department of Computer and Information Technology, Fudan University, Shanghai 200433, China)[1]

(Research Room of Computer Science and Technology, Tank College, Bengbu 233013, China)[2]

Abstract To the user, the structure of the data in HTML tables on the Web is usually unknown, thus, the data of interest can't be queried directly. We presented a solution to this problem. The solution entails the understand of table element, data integration and wrapper creation. Table unstanding is to find interest table, recognize attribute and value in the table, pair attributes with values and form records. Data integration is to match source records with a target schema. Ontology specified wrappers is to extract the data from source records into a target schema. Results show that the data with unknown structure can be directly queried through a known target schema.

Keywords Hetero-data integration, Semantic correspondence, Table understanding, Ontology extraction

1 引言

基于易质数据集成的模式映射问题是研究的热点[1-3],其主旨是要发现一个或多个源模式与目标模式间的语义对应。语义对应的最简形式是映射元素的集合。其中一个元素完成一个源属性到一个目标属性的绑定,或者一些源属性到一些目标属性的绑定,但仅有这些正常定义的绑定,去发现兴趣数据还是不够的。源表常具有结构不完整和信息缺失的问题,因此需要做一系列处理,包括表的理解、数据集成和本体抽取。尤其是大型网页数据库的自动抽取很值得研究,在本文中,我们的研究范围局限于Web上的HTML表。

2 表理解与数据集成

2.1 源表与目的匹配

目标模式如图1所示,源表如图2、图3、图4所示。

Car	Year	Make	Model	Mileage	Price	Phone
0001	1999	Pontiac	Firebird	32,800		405-936-8666
0002	2000	Acura	RL3.5	36,600	$23,988	405-936-8666
0003	2002	Honda	Accord EX	13,800	$21,988	405-936-8666
...						
0101	1992	Acura	legend		$9500	
0102	2000	Audi	A_4		$34,600	
0103	1985	Bmw	325e		$2700,00	

Car	Feature
0001	blue
0001	
0003	white
0003	Air conditioning
0003	Driver side air bag
0101	auto
0101	AM/FM

图1 目标模式

vehicles
Search new
Pre-owned
special
Get a quote
financing
Vehicle pricing
Finance
specials
contact
service
about
Home

Year	Make And model	Price	Miles	Exterior	Photo
1999	Firebird	Contact us	32800	Blue	
2000	Acura	$23,988	36600	Silver	
2002	Hond Accord ex	$21,988	13800	White	
...	...	...	...	...	...

Show checked vehicles (注: 按钮) new search show 25 more

Bob Howard Honda
14137 Broadway Extension
Oklahoma City,Ok 73013,

Toll Free:1-877-944-2842
Phone:405-936-8666
Fax:405-936-8674

图2 含表格的网页

Preowned inventory

2002 Honda Accord EX **$21,988**

Features	***price*** *$21000*
Air conditioning	***mileage*** *13875*
Driver side airbag	
Passenger side airbag	***body type*** *car*
Anti-lock brakes	***body style*** *coupe*
Am/Fm Cassette	***exterior*** *white*
Security Features	***transmission*** *automatic*
Alloy wheel	***engine*** *3.0L 6 cyl fuel injection*
Automatic transmission;	***fuel type*** *gas*
Compact disc player	***stock number*** *350291A*
Cruise control	***vin*** *1HGCG22562A*
Front wheel drive	
Intermittent wipers	
Man light	

图3 带附加信息的链接页

到稿日期:2008-03-04

强 宇 博士后,研究方向为人工智能、知识工程、数据挖掘;胡运发 教授,博导,研究方向为网络、人工智能、知识工程。

Make	Model	Year	Colour	price	Auto	Air cond	AM/FM	CD
acura	legend	1992	grey	$9500	Yes	no	Yes	No
audi	A_4	2000	blue	$3,4500	Yes	Yes	Yes	Yes
bmw	325e	1985	black	$2700.00	No	No	Yes	No
Chevrolet	Cavalier z_{24}	1997	black	$1199500	No	Yes	Yes	No

图 4　表格

图 2—图 4 的源 html 表与目标模式匹配[5-8]时，存在诸多问题，分别是：

a)合并的属性/值，在目标模式中，Make 和 Model 是两个属性，但在源表 2 中是一个。

b)子集，colors 在目标模式中是一类特殊的 feature，图 2 和图 4 中属性 colors 的值集是目标模式中 feature 值的子集。

c)同义，目标模式的 mileage 和图 2 的 miles 意义相同。

d)额外信息，目标模式的表中没有属性 photographs，但在图 2 中有。

e. 链接信息，属性 Make 和 Model 的值可链接到进一步信息。

f)list，一维表格与 list 外观相似，如图 3 所示的 features 是一个 list，但容易被格式化成表格。

g)属性的位置，图 3 的链接子表属性位置在左边的列，而非顶部的行。

h)遗失信息，在目标模式中有属性 phone number，但在图 2 至图 4 的表格中没有。

i)外因数据，phone number 在图 2 的表格以下的正文、图 3 的表格以上的正文出现。

j)重复数据，在图 2 和图 3 中，Honda EX 的价格出现了 3 次。

k)未期望的多值，在目标模式中，每辆车最多有一个联系电话，在图 2、图 3 中却有多个联系电话。

l)属性作为值，在图 4 中，Auto，Aircond，Am/Fm，CD 是属性，但在目标模式中，它们是属性 feature 的值。布尔值 Yes/No 决定了它们是否在属性 Feature 列中出现。

2.2　抽取策略

信息的抽取策略包含 5 部分，分别是：

1)定位兴趣表格

基于本体描述的期望属性，识别兴趣表。

2)形成属性-值对

给予本体识别的单个字串，采用表识别技术形成属性值对。例如(year，1999)，(exterior，blue)是图 2 中的第一条记录。

3)调整属性-值对

图 4 表格中第二行的属性-值对(CD，YES)转换成串 CD，第一、三、四行属性-值对(CD，NO)转换成空串。

4)分析抽取模式

给定表的设计，子表的链接，及正文部分的定位，可以抽取特定的模式，例如本体识别图 2 中第一行的值 32833 应抽取成目标模式第一辆车的属性 mileage 的值，值 Pontiac Firebird 的第一部分应抽取成属性 Make 的值，第二部分应抽取成属性 Model 的值。

5)推导映射

给定识别出的抽取模式，系统可以推导源到目的的映射，例如图 2 中的属性 Miles 值映射到目标模式的 Mileage，Make and Model 值串的第一部分映射到目标属性 Make，第二部分映射到目标属性 Model.

3　本体抽取

一个抽取本体[4]是一个基于狭窄领域的概念模型实例。包括两部分，分别是：

1)一个对象/关系模型的实例(概念的集合)，包括多个对象集合、对象间的关系集合、对象和关系集合上的限制条件。

2)数据框架，对每个对象集，数据框架定义了对象集中的对象基于分类法的出现、及对象在文本中被提及时文本的关键词。

对网页采用本体，可以识别网页中的对象、关系，并将其与本体概念模型实例已命名的对象、关系相关联。

4　表的定位

原始表存在着诸多问题：

a)多模板，图 2 有 3 个模板(框架)。

b)表设计，标签一个在表上正文，一个在表下正文。

c)信息行不在表中，最后一行包含联系信息，次行包含按钮。

d)零碎部分的显示，如图 2 所示，每页显示 25 行，若显示剩余表行，须点击按钮“show 25 more”。

e)信息跨多页，点击图 2 中 Honda accord ex，获图 3 的页。

f)无表格标签，类似图 3 的链接页有一个单列的表格，是 HTML 列表。

若要成功定位表，对主表采用以下策略：

a)主表，须嵌入在 HTML 表格。

b)表格大小，三行三列。

c)网格设计，N 代表具有普遍单元数的表格行数，M 代表表格行数，N/M 值应大于 2/3。

d)属性，基于本体描述的关键词和对象集名，知道属性名，定位表，可做属性名的数据条应超过 60%。

e)值密度，基于不同分类法的对象集的值，用本体识别字串，如识别字串的因子与表格字串的因子比超过 10%，则定位该表。

f)折叠表，设计者通常将表折为两半。属性分两行显示。

g)因子值，是所在表行有一半单元未填充，且填充单元集中在左端区域的值。

对链表采用以下策略：

a)表大小，两行两列。

b)属性，与顶层表相同。

c)属性值-对表，对图 3，本体将左边列抽为属性，右边列抽为值。

d)单属性表格，为发现图 3 的列表，须寻找标签。

e)跨页表格，跟随顶层表格，可得到几个链接表，当表跨越多页时，通常值改变，属性不变。

5 源-目标模式[9-11]映射

源到目标的映射，包括4部分，分别是：

1)形成属性值对

采用表理解技术输入HTML表格，输出标准格式的记录。每条记录是一个属性-值对的集合。

例如图4的第一条记录是：

{(Make, Acura), (Model, legend), (Year, 1992), (Colour, grey), (Price, $9500), (Auto, yes), (Air cond, no), (Am/Fm, yes), (CD, no)}

采用μ算子可以解套循环模式，例如将图5的a表解到b表，需做以下计算：

$\mu_{(Make,Model,Price)}{}^{*}\ \mu_{(Model,Price)}{}^{*}\ T_a$

在第一个$\mu_{(Model,Price)}{}^{*}$运算后，属性Make值分配给了Model，例如ford出现在属性Make值列的空单元中。

在第二个$\mu_{(Make,Model,Price)}{}^{*}$运算后，属性Year值出现在属性Year值列的空单元中。则得图5的b表。而采用运算$\mu_{(Model,Price)}{}^{*}\ \mu_{(Make,Model,Price)}{}^{*}\ T_a$，效果相同。

Year	Make	Model	Price
1995	Ford	F150 super cab	$6,988
		ContourGL	$3988
	Acura	Integra LS	$14,500
	Honda	Civic ex	
1994	Ford	F150	$4,488
		Probe	$3,988
		Taurus	$2,988

a表

Year	Make	Model	Price
1995	Ford	F150 super cab	$6,988
1995	Ford	ContourGL	$3,988
1995	Acura	Integra LS	$1,4500
1995	Honda	Civic ex	
1994	Ford	F150	$4,488
1994	Ford	Probe	$3,988
1994	Ford	Taurus	$2,988

b表

图5 表的内在因子化

2)调整属性-值对

图4中的第一条记录，经调整变成：

{(Make, Acura), (Model, legend), (Year, 1992), (Colour, grey), (Price, $9500), (Auto), (Am/Fm)}。

采用B算子可把布尔指示符转成属性值。例如$B^{Auto}_{yes,no}$ $B^{Aircond}_{yes,no}$ $B^{Am/Fm}_{yes,no}$ $B^{CD}_{yes,no}$ T将图4转成图1的目标模式。

3)执行抽取

调整完属性值对，采用本体，对图4的第一条记录生成如下信息：

{(Car, 0101), (Year, 1992), (Make, acura), (Model, legend), (Mileage,) (Price, $9500), (Phonenr,)}

{(Car, 0101), (Feature, grey)},

{(Car, 0101), (Feature, Auto)},

{(Car, 0101), (Feature, AM/FM)}

采用e算子可抽取非结构化信息，例如$e_{phoneNr}$ t可从图2中的非结构化脚注中提取

1-877-944-2842，返回(PhoneNr, 1-877-944-2842)。

4)推导映射

如果要从Ta表映射到图1中的目标模式year列，可先应用μ算子，解嵌套，得到T_b，然后抽取来自T_b表的Year列。

$Year=\pi_{Year}\mu_{(Make,Model,Price)}{}^{*}\ \mu_{(Model,Price)}{}^{*}\ T_a$

采用d,c,U算子可分别做分裂属性、合并属性、并集运算的操作。

$d^{A}_{B_1,\cdots,B_n}r$表示属性A的值分裂成若干值$v_1,\cdots,v_n$，并赋给新属性$B_1,\cdots,B_n$。

例如$d^{MakeandModel}_{Make,Model}$T(T为图2的表)，属性Make and Model值前半部赋给Make，后半部赋给Model。

$c_{B\leftarrow A_1+\cdots+A_n}r$，$A_i$是属性或字串，$B$是新属性。

U运算可将子集合并成集合。

例如，$\rho_{Exterior\leftarrow Feature}\pi_{Exterior}$ T $\cup$ $\rho_{Feature\leftarrow Feature}$ T/MakeandModel/Features

其中T/MakeandModel/Feature指示了图2到图3的链接。

图2的Exterior colors和图3的Features合并，可得到图1的属性Feature的子集。

6 实验及讨论

在实际中，对表的规模、属性个数和信息覆盖率均需做出限制。在汽车广告领域所做的实验表明，可以有效地抽取未知结构的信息。

首先收集来自多个不同英语网站的汽车广告，例如，将60个广告表分成两组，分别是“训练型”和“测试型”表。采用7个“训练型”表生成启发式规则，用以表定位和表理解。对训练表，可成功定位所有的顶层链表和链接页的内容。对测试表，可成功定位46个顶层链表，其中的28个链接到附加页，13个含结构化信息，15个含非结构化信息。对46个可识别表，产生映射319个，正确或部分正确296个，丢失23个，错误13个(占已发布映射309个的4%)，在正确映射296中，228个来自顶层链表，58个来自附加表，10个既来自顶层链表，又来自附加表。296中121个是直接映射，即源和目的中的属性是同一，175个是间接映射。

结束语 从HTML表抽取信息，实现目标模式，需在设计表的显示、智能元素和图形对象识别上做进一步工作。通过表定位的理解，我们知道很多表是隐藏模式的，即隐藏网页，为获取尽量多的系统可处理信息，需要处理隐藏网页。进一部的工作实现综合抽取工具。

参考文献

[1] Biskup J, Embley D W. Extracting infromation from heterogeneous information sources using ontologically specified target views. Information Systems, 2003, 28(3): 169-212

[2] Crescenzi V, Mecca G, Merialdo P. Roadrunner: Towards automatic data extraction from large Web sites// Proceedings of the 27th International Conference on Very Large Data Bases(VLDB'01). Rome, Italy, September 2001

[3] Doan A, Domingos P, Halevy A. Reconciling schemas of disparate data sources: A machine-learning approach// Proceedings of the 2001 ACM SIGMOD International Conference on Management of Data(SIGMOD 2001). Santa Barbara, California, May 2001: 509-520

[4] Embley D W, Ng Y K, Xu L. Recognizing ontology-applicable multiple-record Web documents// Proceedings of the 20th International Conference on Conceptual Modeling(ER2001). Yokohama, Japan, November 2001: 555-570

[5] Embley D W, Xu L. Record location and reconfiguration in un-

structured multiple-record Web documents// Proceedings of the Third International Workshop on the Web and Databases (WebDB2000). Dallas, Texas, May 2000: 123-128

[6] Hu J, Kashi R, Lopresti D, et al. Why table ground-ruthing is hard// Proceedings of the Sixth International Conference on Document Analysis and Recognition. seattle, Washington, September 2001:129-133

[7] Liddle S W, Embley D W, Scott D T, et al. Extracting data behind Web forms// Proceedings of the Joint workshop on Conceptual Modeling Approaches for E-business: A Web Service Perspective(eCOMO 2002). Tampere, Finland. October 2002:38-49

[8] Liddle S W, Yau S H, Embley D W. On the automatic extraction of data from the hidden Web// Proceedings of the International Workshop on Data Semantics in Web Information Systems (DASWIS-2001). Yokohama, Japan, November 2001:106-119

[9] Madhaven J, Bernstein P A, Raham E. Generic schema matching with Cupid// Proceedings of the 27th International Conference on Very Large Data Bases (VLDB'01). Rome, Italy, September 2001:49-58

[10] Miller R, Haas L, Hernandaz M A. Schema mapping as query discovery// Proceedings of the 26th International Conference on very Large Databases (VLDB'00). Cairo, Egypt, September 2000:77-88

[11] Raghavan S, Garcia-Molina H. Crawling the hidden Web// Proceedings of the 27th International Conference on Very Large Data Bases(VLDB'01). Rome, Italy, September 2001

(上接第 161 页)

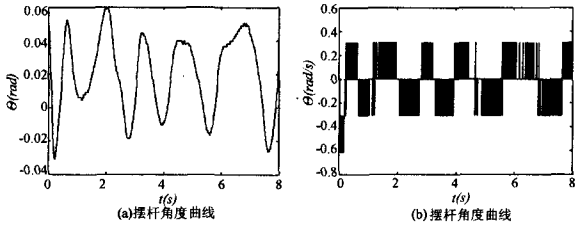

图 7 系统稳定时摆杆角度和角速度曲线图

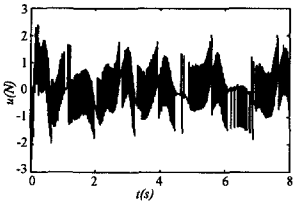

图 8 系统稳定时控制量曲线图

结束语 本文从常规模糊控制器的输入变量 E 和 EC 的基本物理意义出发，分析它们之间内在关系的本质特征，发现这种内在关系的本质是一种泛组合关系，并引入泛逻辑学中"关系柔性"的思想，将控制运算模型定义为由命题间相关性控制的算子簇，据此提出了一种柔性逻辑控制方法，可实现对复杂系统的精确控制。最后，一级倒立摆实物实验结果证明了该方法的可行性和有效性。

参 考 文 献

[1] Lee C C. Fuzzy Logic in Control System: Fuzzy Logic Controller, Part I, II[J]. IEEE Transaction on Systems, Man and Cybernetics, 1990, 20(2): 404-435

[2] Zadeh L A, Zimmermann H J. On Computation of the Compositional Rule of Inference under Triangular Norms[J]. Fuzzy Sets and Systems, 1992, 51: 267-275

[3] 李士勇. 模糊控制、神经控制与智能控制论[M]. 哈尔滨：哈尔滨工业大学出版社，1996

[4] 张思勤，施颂椒，高卫华，等. 模糊控制系统近年来的研究与发展[J]. 控制理论与应用，2001，18(1)：7-11

[5] 李洪兴. 从模糊控制的数学本质看模糊逻辑的成功[J]. 模糊系统与数学，1995，9(4)：1-14

[6] 李洪兴. Fuzzy 控制的本质与一类高精度 Fuzzy 控制器的设计[J]. 控制理论与应用，1997，14(6)：868-872

[7] 李洪兴. 模糊控制的插值机理[J]. 中国科学(E 辑)，1998，28(3)：259-267

[8] 付利华. 复杂系统的柔性逻辑控制理论及应用研究[D]. 西安：西北工业大学，2005

[9] 戴忠达，张曾科，汤俭. 一种改进的模糊控制器及其应用[J]. 自动化学报，1990，16(3)：258-261

[10] Raju G V S, Zhou Jun, Kisner R A. Hierarchical Fuzzy Control [J]. International Journal of Control, 1991, 54(5): 1201-1216

[11] Raju G V S, Zhou Jun. Adaptive Hierarchical Fuzzy Controller [J]. IEEE Transaction on Systems, Man and Cybernetics, 1993, 23(4): 973-980

[12] 胡绳荪，候文考，秦宝忠. 焊缝跟踪系统中的自调整比例因子 Fuzzy-P 控制器的研究[J]. 天津大学学报，1999，32(2)：181-185

[13] Cheng Fuyan, Zhong Guomin, Li Youshan. Fuzzy Control of a Double-inverted Pendulum[J]. Fuzzy Sets and Systems, 1996, 79(3): 315-321

[14] 甄敏，袁艳，张泰山. 三维控制规则自修正模糊算法的研究[J]. 计算技术与自动化，2000，19(1)：16-18

[15] 何华灿，王华，刘永怀，等. 泛逻辑学原理[M]. 北京：科学出版社，2001

[16] 陈志成. 复杂系统中分形混沌与逻辑的相关性推理研究[D]. 西安：西北工业大学，2004

[17] 肖军. 模糊控制在多变量非线性系统中的应用[D]. 沈阳：东北大学，2001

[18] 龙升照，汪培庄. 模糊控制规则的自调整问题[J]. 模糊数学，1982，3(3)：105-112

[19] 付利华，何华灿. 模糊推理中相异因子的研究[J]. 计算机科学，2004，31(2)：98-100

[20] 付利华，何华灿. 模糊推理中零级泛蕴涵的信息度约束研究. 计算机科学，2005，32(1)：162-164