一个Web查询界面域序列模式图模型及其应用

郭文宏 范学峰
(同济大学电子与信息工程学院 上海 201804)

摘 要 针对Deep Web查询界面集成问题,定义了一种面向专门领域的域序列模式图(FSRG)模型,用于表示和发现同一领域查询界面中的所有域序列模式。该模型将领域内不同查询页面的域序列模式统一到一个有向有环图中。基于序列模式图进行研究可发现领域模式中域的结构化组织排列规律。还论述了域序列模式图的构造、域子序列模式划分和领域所有域的整体序列模式发现方法。在有限领域下封闭测试表明,该模型及其算法对结构化Web界面分析有较大应用价值。研究为实现智能化的Web数据模式处理提供了域序列分析方法,对大规模智能集成和搜索应用有一定参考价值。

关键词 域序列模式图,模式分析,Deep Web,智能信息处理
中图法分类号 TP391,TP393 **文献标识码** A

Model of Field Sequence Pattern Graph for Web Query Interface and its Application

GUO Wen-hong FAN Xue-feng
(College of Electronics and Information Engineering, Tongji University, Shanghai 201804, China)

Abstract Defined a model of field sequence pattern graph (FSRG). It was provided to analyze field sequence patterns in query Web pages of special domain on the research of Deep Web query interface integration. Directed cycle graph is used to represent and discover domain field sequential pattern in the model. The paper also discussed the methods of model construction, sub-sequence partition and global domain fields sequence discovery. We had worked on it and achieved tangible results in several domains. The research provided a field sequence pattern analysis method for semantic Web information processing. It is valuable to large scale applications of intelligent integration and information retrieval.

Keywords Field sequence pattern graph, Schema analysis, Deep Web, Intelligent information processing

为屏蔽各种Deep Web资源在来源和查询方式上的差异,消除查询平台的异构性,可将同一领域的各Deep Web查询界面集成为统一的界面。在对界面模式分析中,除了需要考虑界面接口各域的语义信息外、还需要考虑接口整体和局部的模式信息。不同的界面接口域的组织布局不尽相同,图1显示了Web上2个网站上海鲲鹏票务中心网(www.jp114.com.cn)(图1(a))和中国便民网(http://www.zgbm.com/tools/4.htm)(图1(b))的实时单程机票查询接口界面,这两个接口域相同但域序列不同。为发现同一应用领域查询接口域序列整体上在语义结构和使用惯例方面潜在的组织规律,本文定义了一种面向专门领域的域序列模式图模型FSPG,该模型将领域内不同查询页面的域序列模式统一到一个有向有环图中。利用领域数据模式及Web查询界面模式来建造特定领域的域序列模式图,用于表示和发现领域查询界面中的所有域的局部域序列模式和整体序列模式。

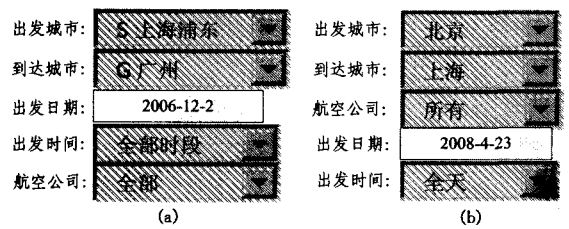

图1 同一领域中域相同域序列不同的Web查询接口示例

1 域序列模式图模型

1.1 模型定义

单个查询模式的域序列模式图(FSPG-SS Field Sequential Pattern Graph for Single Schema)定义为$Gs(V,s,t,R)$,其中节点集V中元素表示领域查询界面模式中的域,s为查询界面模式中起点域,t为终点域,$s,t\in V$,弧集R中元素表示域间的直接相邻关系。

在界面分析集成中,为发现同一领域中的所有域的模式规律,提出了领域域序列模式图模型SPG-DS Field(Sequential Pattern Graph for Domain Schemas),它是一种带权的有向有环连通图,定义为$Gd(V,S,T,R)$,其中节点集V中元素表示特定领域中所有查询界面模式中的域,S,T分别为出现在单个查询界面模式中的起点域和终点域的集合,即对应于在单个具体模式域序列模式图中的s和t的集合。对于单个

到稿日期:2008-05-05
郭文宏(1971—),男,博士研究生,主要研究领域为计算机网络信息处理、语义Web、知识工程,E-mail:guowh2003@sohu.com;范学峰(1947—),男,研究员,博士生导师,主要研究领域为计算机网络与信息处理。

查询模式中的域节点 v，$I(v)$表示为 Gd 中节点 v 的入度，$O(v)$为 v 的出度，如 $I(v)=0$ 则 $v\in S$，如 $O(v)=0$ 则 $v\in T$。弧集 R 中元素表示域间的接续关系，模式中的各域是独立但又关联的，每条弧附带一个权值 W 表示其所关联的节点间的关联度，弧(u,v)的权值 $W(u,v)$的计算公式为

$$W(u,v)=\frac{2*N(u,v)}{O(u)+I(v)} \tag{1}$$

其中 $N(u,v)$代表(u,v)出现的模式数目。与其它图相比，领域域模式图 Gd 具有的特性如下：(1) 起点集和终点集中的元素与其它图不同，Gd 中对于开始节点 s，有 $I(s)\geqslant 0, O(s)\geqslant 1$；对于终点节点 t，有 $I(t)\geqslant 1, O(v)\geqslant 0$；在 Gd 中，$S\subset V$，$T\subset V$ 且有 $|S|\geqslant 1$，$|T|\geqslant 1$，$S\cap T$ 可以不为 Φ。(2) 其它节点的入度与出度均$\geqslant 1$；(3)不存在闭环。

除了用连接弧显式地表示相邻节点的关联度外，该模型还基于模式样本集统计领域中所有域间的关联度。假设领域域模式 $S=(w_1,w_2,\cdots,w_n)$中有 n 个域，$w_i(i\in 1,2,\cdots,n-1)$与 w_n 间的关联度定义为：

$$Re(w_i,w_n)=\sum_{l=0}^{n-1} r_l\times\frac{cocur(w_i,w_n,l)}{n}$$

其中 $r_l=\log(n)-\log(l+1)$ (2)

这里 l 是同一模式中 w_i 与 w_n 间的距离，$cocur(w_i,w_n,l)$为 w_i,w_n 在相距 l 时的同现次数，n 为模式库中的域个数，r_l 反映 w_i 与 w_n 间相距 l 时的加权因子，是距离 l 的单调函数，l 越小 r_l 越大。为了获取领域模式中相邻域间的关联度，利用语料库语言学方法，对预处理后的模式样本进行统计分析，从中获取各域的词频向量、域的二元、三元接续关系。

1.2 模型的建立

对特定应用领域的 n 个域序列模式样本的样本集 SS，首先基于第一个模式样本生成初始域序列模式图 Gs，然后依次提取各模式样本域序列信息，不断对 Gs 中节点、弧进行更新。整个构造过程中随着样本个数的增加，序列模式图的拓扑结构会渐变并趋向稳定，最终得到模式图 Gd 的拓扑结构，文献[4]也通过实验证明了这样的结论：随着同一领域中可利用 Deep Web 资源的增加，查询模式的域数目趋向稳定。最后根据图中所有节点以及弧的统计数据计算每条弧的附带权值。模型建立示意图如图 2(a)和图 2(b)所示。在模式样本质量较好的前提下，所建的 Gd 拥有该领域的比较完整的域信息及域间关系信息。

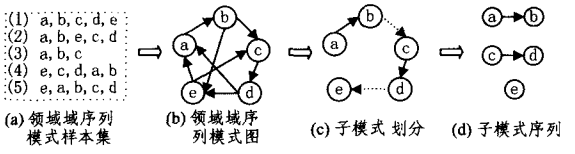

图 2 领域域序列模式图及其划分的子模式序列图

1.3 领域域子序列模式划分

领域域序列模式发现主要研究领域域序列模式划分，得到多个小的域子序列模式，来描述领域模式局部域特征。各子模式图中的域及其序列可作为模式知识单元。在一个域序列样本集中，如果序列 S 包含于一个样本序列中，则称该样本支持序列 S。一个序列的支持度定义为支持该序列的样本总数。给定一个域序列模式样本集，在划分中考虑了序列支持度因素。通过预先定义的规则集对 Gd 中 V 和 R 进行模式划分。一个领域域模式图及其划分示例过程如图 2(b)、(c)和(d)所示。设“*”表示多(单)个域统配符，“?”表示单个域统配符；PW 表示所有弧权值的均值；阈值 $\theta=0.7$。域子序列模式的边界确定规则如下：

(1) IF: $I(v)=1\ \wedge\ O(v)\neq 1$ THEN: $(*,v)$ RB:0.6; // v 称为子模式$(*,v)$的右边界节点。

(2) IF: $O(v)=1\ \wedge\ I(v)\neq 1$ THEN: $(v,*)$ RB:0.6; // v 称为子模式$(v,*)$的左边界节点。

(3) IF: $O(u)=1\ \wedge\ \exists(u,v)\ \wedge\ I(v)=1\ \wedge\ W(u,v)>\theta$ THEN: $(*,u,v,*)$ RB:0.9; //u 和 v 可作为子模式成分。

(4) IF: $W(u,v)\approx W(v,u)\ \wedge\ W(u,v)>>PW$ THEN: $(*,u,v,*)$ RB:0.3; //u 和 v 可作为子模式成分。

(5) IF: $W(u,v)>\theta$ THEN: $(*,u,v,*)$ RB:0.3; //u 和 v 可作为子模式成分。

1.4 领域域序列模式集成

领域域序列模式集成是指发现专门领域所有域的整体序列，即在领域域序列模式图中确定最大域序列模式，并要求该模式在语义和使用惯例上容易被接受。基于域子模式发现的领域域序列模式集成在实现上抽象为构造一条 Gd 中的最佳路径算法。算法主要遍历领域域序列模式图并实现部分弧的消解。在优先保留域子序列模式的前提下，弧线排除规则如下：

(1) IF: $w(a,b)=w(b,a)\ \wedge\ W(a,b)\approx 0$ THEN: $DEL((a,b),(b,a))$，即去掉有相互关系且微弱的边：假设 2 者无关。例如价格-位置(1,1)，星级-位置(1,2)。

(2) IF: $w(a,b)\approx w(b,a)$ THEN: $DEL((a,b),(b,a))$，即去掉有相互关系且关系权值相当的边：假设 2 者序列前后无关。例如星级-店名(2,3)。

(3) IF: $w(a,b)>>w(b,a)$ THEN: $DEL((b,a))$，即去掉有相互关系且关系权值小的一条边：假设 2 者序列前后相关。例如价格-店名(8,3)，星级-价格(9,5)。

(4) IF: $I(a)=1\wedge O(a)\neq 1$ THEN: $DEL((a,?))$ 去掉右边界节点的所有出弧。

(5) IF: $O(a)=1\wedge I(a)\neq 1$ THEN: $DEL((?,a))$ 去掉左边界节点的所有入弧。

图 3 为领域域序列模式图及其界面集成后的领域序列模式示意图。

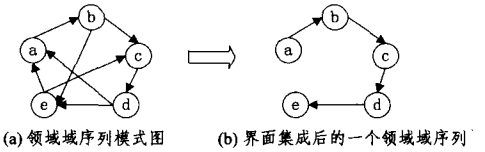

图 3

2 实验测试及分析

主要针对 4 种查询接口进行试验：酒店查询、单程机票查询、往返机票查询和楼盘新房查询。分别收集这 4 个领域查询接口数据，构造领域域序列模式图，然后依据上述规则进行域子序列分析并确定最终集成领域模式域序列。

收集 Web 领域数据库模式和 Web 查询界面模式的真实数据作为分析样本。查询界面模式数据收集原则主要考虑以下 Web 域模式：$l1+v+l2, l1+l2+v, l1+v, l1+(l1+v, l2+$

$v,\cdots,ln+v)$，$l1$，$l2$ 分别表示域提示信息和域补充说明信息，v 表示待输入或选择的域值对应的 Web 控件信息。对收集到的模式样本进行格式转换、自动分词等预处理，然后进行样本统计分析，获取构造模式图所需的各种数据。统计证实常用领域查询界面使用概念及其词汇是相对有限的。

对收集的样本集基于本文第 2 部分所述方法构造领域域序列模式图。模式图的构造是先建立拓扑结构，最后根据领域内所有域及其接续关系计算每条弧的权重。试验中的权重通过式(1)计算得到。图 4 是酒店查询界面域序列模式示意图，图中带阴影节点表示起点域，终点域表示为黑色节点。表 1 是图 4 中弧的附带权值表。

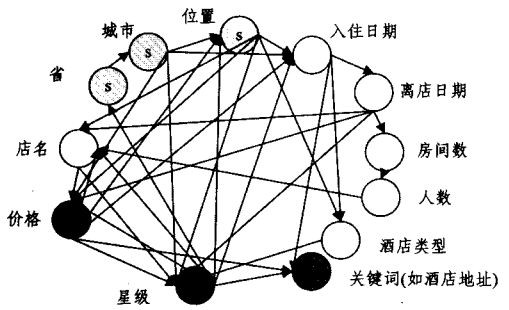

图 4　酒店查询界面域序列模式图示意图

获取模式中的域关系等模式结构特征。遍历模式图并对其进行划分发现域序列子模式。例如在酒店查询界面域序列中有省-城市，入住日期-离店日期，房间数-人数等子模式。领域模式集成序列确定时，遍历模式图生成最大领域模式序列路径，遍历过程中，节点的选择按域序列子模式优先、只有一条入弧的节点优先、弧最大权重优先原则。如果在没有完全遍历所有节点前遇到无出弧节点，则需要回溯到上一个节点重新选择路径。由于问题本身的特点决定了查询界面整体集成序列结果不唯一，因此我们设置模型算法生成的多个候选领域域序列。

表 1　酒店查询接口域序列模式示意图附带权值表

	省	城市	位置	入住日期	离店日期	房间数	人数	酒店类型	星级	价格	店名	关键词
省		1										
城市			0.24	0.4					0.33	0.21		
位置				0.32				0.2	0.08	0.07	0.15	
入住日期					0.9			0.15				0.14
离店日期						0.2			0.08	0.14	0.37	
房间数							1					
人数								0			0.11	
酒店类型									0.11			
星级	0.12		0.18	0.07						0.51	0.18	0.11
价格			0.09	0.07					0.3		0.47	0.11
店名									0.26	0.24		

在利用模型对领域域子序列分割的测试中，首先对领域域序列模式图中的域子序列进行人工判定，设 A 为人工确定的领域子模式集，基于模型的算法得到的子模式集为 B，子模式提取的召回率定义为 $|A\cap B|/|A|$。表 2 是实验测试数据统计表。封闭测试表明，自动子模式提取具有较好的实验结果，所提取子模式的召回率平均达到了 90%以上。在利用模型对领域整体域序列集成的测试中，因此测试中同样依赖人的参与，人工事先确定领域集成域候选序列集，令 C 表示人工集成候选序列个数，D 表示机器实现的集成候选序列个数，测试中设置 $D=C$，E 为人机集成结果一致的序列个数。经过测试，集成序列平均召回率为 65%，如表 2 所列。实验显示，查询接口域个数越少的领域，整体域序列集成效果越好。对于域个数较多的领域，该模型及其算法有待进一步研究和完善。另外，如何更科学地进行该类试验的性能测试，也是值得我们今后研究的课题之一。

表 2　实验测试数据

应用领域	领域查询接口样本数	领域域个数 n	人工分析子模式集 A 元素数	系统分析子模式集 B 元素数	子模式相符个数 A∩B 元素数	子模式提取召回率 |A∩B|/|A|	人工集成候选序列数 C	机器集成候选序列数 D	人机集成结果一致序列数 E	集成序列召回率 E/C
酒店查询	42	12	4	5	3	75%	3	3	1	33.3%
单程机票查询	30	6	3	3	3	100%	2	2	2	100%
往返机票查询	30	7	3	3	3	100%	2	2	2	100%
楼盘新房查询	30	17	3	6	3	100%	4	4	1	25%
合计						93%				65%

结束语　本文利用领域域序列模式图模型发现领域整体模式结构中域的分布排列规律，确定域的子模式结构位置、同现依赖关系等，这些规律通过进一步加工整理可作为自动集成检索处理的背景知识。由于各查询接口域模式的复杂性和域布局的多样性，使得自动发现领域整体域序列具有很大的挑战性，本文的领域域序列模式集成方法对域个数多的领域处理有待加强。随着研究的深入，我们正在进行模型扩展和算法改进，扩展后的域序列模式图定义为 $G(V,C,R)$，增加了控制节点集 C，节点集 V 中元素仍然表示领域模式中的域，控制节点有开始节点、终止节点、路由节点、循环节点、或分支、或结合、与分支和与结合。其中，路由节点用来连接两个控制节点，以表达复杂的逻辑关系，如先“与”再“或”的情况。另外，弧集 R 表示域之间的顺序及路径选择关系。通过控制节点及有向弧的组合来表示任意复杂的域关系。由于篇幅所限，相关的研究将另文介绍。

参 考 文 献

[1] Liu B, Grossman R, Zhai Y. Mining Data Records in Web Pages. SIGKDD, USA, August 2003: 601-606

[2] Chang KCC. Statistical schema matching across Web query interfaces// Proc. of the SIGMOD Conf. 2003. San Diego: ACM Press, 2003: 217-228

[3] Chang KCC. Automatic complex schema matching across Web query interfaces: A correlation mining approach. ACM Trans on Database Systems, 2006, 13(1): 11-45

[4] He Bin, Tao Tao, Chang K C. Organizing structured web sources by query schemas: a clustering approach[R]. Computer Science Department: CIKM, 2004

[5] 邹宇，刘毅，陈佩文. 基于图归约法的工作流模型验证[J]. 计算机应用，2003，23(4)：1-3