第 36 卷 第 3 期 计 算 机 工 程 2010 年 2 月
Vol.36 No.3 Computer Engineering February 2010
·软件技术与数据库· 文章编号：1000—3428(2010)03—0064—02 文献标识码：A 中图分类号：TP391

用于 Web 查询接口集成的模式匹配算法

何玲玲，刘国华，孔令民
(燕山大学信息科学与工程学院，秦皇岛 066004)

摘 要： Web 查询接口是 Deep Web 的访问入口。通过集成内容相关的 Web 查询接口，能为用户访问提供方便。现有查询接口集成的模式匹配算法效率低，针对该问题提出一种模式匹配算法，以概念团选择定理为依据，直接形成最优概念划分，并生成最优模型。理论分析和实验结果表明，该算法具有可行性，可以减少运算量并提高匹配效率。

关键词： 查询接口；概念团；模式匹配

Mode Matching Algorithm for Web Query Interface Integration

HE Ling-ling, LIU Guo-hua, KONG Ling-min
(College of Information Science and Engineering, Yanshan University, Qinhuangdao 066004)

【Abstract】 Web query interface is the visiting access of Deep Web. By integrating content pertinent Web query interface, it can provide convenience for the users' visiting. Existing mode matching algorithm for query interface integration is inefficient. Aiming at this problem, this paper proposes a mode matching algorithm. The algorithm is based on the concept clique choice theorem, forms the optimal concept partition directly and generates the optimal model. Theoretical analysis and experimental results show that the algorithm is feasible. It can reduce the computation and improve the matching efficiency.

【Key words】 query interface; concept clique; mode matching

1 概述

网络的迅速发展为人们提供了可以利用的海量信息。整个网络按其蕴含信息的“深度”可以划分为 Surface Web 和 Deep Web。后者表示 Web 中不能被传统搜索引擎索引到的内容，又称为 Web 数据库，其蕴含的信息量是 Surface Web 的 400 倍~500 倍，超过 50%的 Web 数据库是面向特定领域的[1]。面对网络上的大量信息，如何实现对同一领域不同 Web 数据库的访问自动化、高效化，已成为目前的研究热点之一。为了给用户提供一个统一的访问途径，需要对每个 Web 数据库的查询接口进行集成并获得一个统一接口，该接口称为集成接口[1]。通过在集成接口上提交查询，可以自动实现在多个 Web 数据库的查询接口上提交查询的目的。

现有查询接口的集成算法[2-3]效率较低，文献[2]提出统计模式匹配方法，虽然其效率较高，但当模式集规模较大时，仍然不能满足要求。为解决上述问题，本文提出一种模式匹配算法。此算法在模型生成时，选择性地生成最优模型，有效避免了冗余工作，提高了执行效率。

2 基本定义

为了合理地定义模式模型及其他相关概念，给出以下 3 条假设：(1)概念团相互独立，即实例化模式时，对不同概念团的选择是相互独立、互不干扰的；(2)同义词相互排斥，即实例化模式时，对同一概念团中同义词的选择是相互排斥的；(3)概念团无交集，即不同的概念团在语义上没有重叠。

基于上述 3 条假设，相关定义描述如下：

定义 1(模式模型[2]) 设模式模型 M 为四元组(V, C, P_c, P_a)，属性集合 V 是该领域所有查询接口属性的集合，即$\{A_1, A_2,\cdots, A_n\}$。概念划分 C 是概念团的集合，即$\{C_1, C_2,\cdots, C_m\}$，$V=\cup C_i$ 且 $C_i\cap C_j=\varnothing$。P_c 是概念团概率函数，P_a 为属性概率函数。

定义 2(概念网络图[2]) 给定输入模式集 I_s，其属性集合为 V，图中节点与属性一一对应。当且仅当 2 个属性没有在任何模式 I 中共同发生($I\in I_s$)时，代表属性的节点之间由线段连接，并称这些属性为同义词或具有一致性。由任意数量具有一致性的属性形成的集合称为概念团。

定义 3(同义词最大集) 给定输入模式集 I_s，其属性集合为 V，概念划分为 $C=\{C_1,C_2,\cdots,C_m\}$，若 C 存在概念团 $C_i=\{,\cdots,A_i,\cdots,\}$，即 C_i 中包含与属性 A_i 表示同一语义的所有属性，则称该概念团 C_i 为同义词最大集。

3 模式匹配算法

在通常情况下，查询接口被认为是接口属性的集合，则接口模式匹配问题转化为接口属性匹配问题，即发现同义词的对应关系。本文提出的用于 Web 查询接口集成的模式匹配算法借助概念网络图，可以同时匹配大量数据源的查询接口并得到所有同义词的对应关系。

3.1 模型结构

依据定义 1，模型 $M=(V, C, P_c, P_a)$可以表达为 3 层树形结构：模型 M 为根节点，概念划分 C 为第 2 层节点，属性集合 V 为叶子节点。例如，给出模式模型 M_B，属性集合 $V_B=$ {author, title, ISBN, subject, category}，概念划分 $C=$ {(author), (title), (ISBN), (subject, category)}，则其树形结构如图 1 所示。该模型也可以表示为：$M_B=$ {(author), (title), (ISBN), (subject,

基金项目： 国家自然科学基金资助项目(60773100)；国家“十一五”科技支撑计划基金资助项目(2006BAK05B02)

作者简介： 何玲玲(1984－)，女，硕士研究生，主研方向：空间数据库；刘国华，教授、博士生导师；孔令民，工程师

收稿日期： 2009-08-09 **E-mail：** helingling20051984@163.com

category)}。

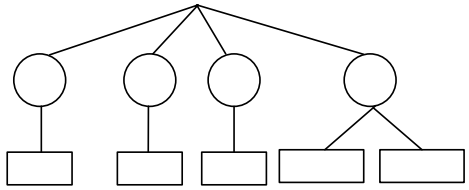

图 1　M_B树形结构

3.2　模型生成

本节将介绍如何确定模型中的各个参数及其参数值，即属性集合 V、概念划分 C、概念团概率函数 P_c和属性概率函数 P_a。

3.2.1　概率值计算

给定模型 M，为了量化模型实例化模式的概率值，相关公式描述如下：

(1)如果概念团 C_i被选中，则概率为 $\Pr(C_i|M)=\alpha_i$，否则为 $\Pr(\neg C_i|M)=1-\alpha_i$

(2)选中任意属性 A_j的概率为

$$\Pr(A_j|M)=\begin{cases}\alpha_i\times\beta_j & \exists i: A_j\in C_i\\ 0 & \text{otherwise}\end{cases} \tag{2}$$

(3)选中一组属性的概率为

$$\Pr(A_1,A_2,\cdots,A_m|M)=\begin{cases}0 & \exists j\neq k,\exists i: A_j\in C_i\wedge A_k\in C_i\\ \prod\Pr(A_j|M) & \text{otherwise}\end{cases} \tag{3}$$

(4)实例化模式 I_i的概率为

$$\Pr(I_i|M)=\Pr(A_1,A_2,\cdots,A_n|M)\times\prod_{\forall A_j,A_j\notin C_i}\Pr(\neg C_i|M) \tag{4}$$

若 $\Pr(I_i|M)>0$，则模型 M 能够实例化模式 I_i。

若 $\Pr(I_s|M)>0$，则模型 M 与输入模式集 I_s一致。根据以上公式可以计算出模型实例化任何模式的概率值。

3.2.2　概念划分

为确定概念划分 C，由定义 2 给出的概念网络图获得所有一致性概念团，然后依据概念团选择定理形成概念划分 C，并证明得出由此概念划分 C 所构建的模型 M，其实例化输入模式集 I_s的概率值最高。

定理(概念团选择定理) 给定输入模式集 I_s，其属性集合为 V，概念划分为 $C=\{C_1, C_2,\cdots, C_m\}$且 C 中所有概念团均为同义词最大集，如果由此概念划分 C 构建模式模型 M，那么该模型 M 实例化输入模式集 I_s的概率值最高。

证明：若给定输入模式集 I_s，其属性集合 $V=\{A_a, A_b,\cdots, A_n\}$，各属性发生的模式数目分别为 $n_a, n_b,\cdots, n_n$。初始概念划分为∅。先根据概念网络图获得所有一致性概念团，然后从中选择同义词最大集$\{A_b\},\{A_c, A_f\},\{A_a,\cdots, A_i, A_k\}\cdots$，设共得到 x 个同义词最大集，由此形成概念划分 $C^1=\{C_1, C_2,\cdots,C_x\}$。由 C^1 构建的模型 M_1，其实例化任意模式 $I=\{A_a, A_b,\cdots, A_f\}$的概率是 $\Pr^1(I|M_1)=\alpha_a\times\beta_a\times\alpha_b\times\beta_b\times,\cdots,\times\alpha_f\times\beta_f$。

若概念划分 C^1 中存在某概念团 C_i不是同义词最大集，以拆分 $C_i=\{A_a,\cdots, A_i, A_k\}$为例，得到$\{A_a,\cdots, A_i\},\{A_k\}$2 个概念团 C_i^1, C_i^2，则概念划分由 C^1变为 $C^2=\{C_1, C_2,\cdots, C_x, C_{x+1}\}$，概念团数目由 x 变为 $x+1$。由此构建模型 M_2，其实例化该模式 $I=\{A_a, A_b,\cdots, A_f\}$的概率是 $\Pr^2(I|M_2)=\alpha_a\times\beta_a\times(1-\alpha_k)\times\alpha_b\times\beta_b\times\cdots\times\alpha_f\times\beta_f$，其中，$(1-\alpha_k)$为属性 A_a的同义词 A_k所在概念团未选中的概率。通过比较可知，概率值 $\Pr^1>\Pr^2$。

因此，当且仅当概念划分 C 中的所有概念团均取同义词最大集时，由此构建的模型实例化输入模式集的概率值最高。证毕。

3.2.3　参数值确定

在确定属性集合 V 和概念划分 C 后，通过概念团概率函数 P_c和属性概率函数 P_a获得各概念团和属性的概率值 α_i和 β_i。

$$\alpha_i^*=\frac{\sum_{A_j\in C_i}n_j}{n},\ \beta_j^*=\frac{n_j}{\sum_{A_j\in C_i}n_j} \tag{5}$$

其中，n 为输入模式集 I_s中模式的数目；n_j为属性 A_j发生模式的数目。

由上述过程得到所有参数及其参数值，即属性集合 $V=\{A_1, A_2,\cdots, A_n\}$，概念划分 $C=\{C_1, C_2,\cdots, C_m\}$，属性 A_j被选中的概率值 β_j、概念团 C_i被选中的概率值 α_i。最后用所得参数及参数值构建模型 $M=(V, C, P_c, P_a)$并输出该模型 M。

3.3　算法描述

用于 Web 查询接口集成的模式匹配算法描述如下：

```
Begin
1   V= V'=∅; C= C*=∅;
2   αi=βj=0;
3   n=ni=nj=k=0;
4   V=∪Ii;
5   if V≠∅ then (用概念网络图划分属性集合 V，形成所有的一致性概念团);
6       else goto end ;
7   while V'≠V do
8   {选择同义词最大集 Ci，V'=V'∪Ci 的属性;
9       C*=C*∪Ci;
10      k=k+1;}
11  for (i=1;i++;k)
12      αi=ni/n;
13  for(j=1;j++;m)
14      βj=nj/n;
15  用 V,C*,αi,βj 构建模式模型 M;
16  输出该模式模型;
End;
```

4　实验验证

实验环境设置为 Pentium IV 2.4 GHz 的 CPU，2 GB 内存，Windows 2003 操作系统，针对用于 Web 查询接口集成的模式匹配算法进行实验。为了体现本文算法的执行效率，将其与文献[2]中的算法 MGS_{sd}相比较。

本次实验抽取电影和汽车 2 个领域各 30 个 Web 数据库的查询接口进行操作。先抽取网络查询接口的属性，然后对属性进行预处理，如合并字面表达略有不同的属性等，得到各领域的属性集合如表 1 所示。

表 1　电影和汽车领域的属性集合

领域	属性集合
电影	title(ti), director(dr),actor(ac),genre(gn), format(fm),category(cg),keyword(kw),rating(rt), price(pr),studio(sd),star(st),artist(at)
汽车	make(mk), model(md),price(pr),year(yr), type(tp),zipcode(zc),mileage(ml),style(sy), color(cl),state(st),category(cg)

运用本文算法输出这 2 个领域的模式模型为

M_{movies}={(ti),(dr),(fm),(rt),(pr),(sd),(kw),(ac,st,at),(gn,cg)}

M_{auto}={ (mk), (md),(pr),(yr),(sy,tp,cg),(zc,cl), (st,ml)}

图 2 通过比较本文算法和 MGS_{sd}算法生成的 M_{movies}模型，测试本文算法的执行效率。从图 2 可以看出，MGS_{sd}算法的用时比本文算法长。

（下转第 68 页）