计算机科学1999Vol. 26№. 8

数据库汉语自然语言查询模型研究*)

A Study on Natural Language(Chinese) Query Model in Database

许龙飞
(暨南大学计算机系 广州 510632)

唐世渭
(北京大学信息科学中心 北京 100871)

Abstract This paper presents a advanced Database natural Language(Chinese)query model, which based on E-R Model and keywords driving. Its formal description is given. It analyses the surface semantics and deep semantics for understanding Chinese sentences using the model, and discusses some problems of the expansion and transportability of the model.

Keywords Database model, Natural language understanding, Chinese Language inqury, Semantic rule formula

1. 引言

近年来国内数据库中文查询界面中，汉语查询模型主要有类关系代数表达式的中间语言转换模型、数据库 E-R 语义的汉语查询模型以及以条件为中心的语言理解模型等[3,4,5,8,10]，作者仅就以上模型的长处和不足提出一种新的基于数据库 E-R 语义的查询模型，该模型的主要特点是采用数据库 E-R 语义理解模型，摆脱纯语言学理论的传统框架，将汉语查询句子与其指称的数据库模型的语义以及背景知识相结合，建立类 SQL 的表格式中间语言 MQL，通过表层与深层的语义处理达到对汉语查询句子的较深入理解。较国内传统的句型匹配法有更好的理解能力。采用直接输入汉语句型由系统自动进行汉语词切分形成中间语言，经过 MQL 与 SQL 间的多语句模板的自动转换规则形成 SQL 查询句。整个过程是非过程化自动进行的。与目前国内外的研究方法不同的主要特点之一是采取了数据库技术、汉语语言学与自然语言处理等多学科相结合的新思路。

鉴于目前国内对这种基于数据库汉语查询的计算模型尚缺乏较深入的形式化描述，对模型的可扩充性与移植性尚待进一步深入研究，本文的宗旨正是围绕以上问题展开的。

2. 数据库汉语查询模型的形式描述

在自然语言处理中，由于汉语的复杂性，至今尚未有面向机器分析的汉语形式语法，给汉语自然语言理解造成极大困难。本文仅就作者所提出的数据库汉语查询模型，以 NLCQI[6]为例，给出这种数据库汉语自然语言查询模型的较为形式的描述，该模型分成五个主要部分，包括汉语查询树生成规则、汉语词的语义指称规则、汉语修饰词组词规则、深层语义转换规则和背景语义词典，是一种基于受限文法的汉语查询模型。

2.1 受限的汉语查询文法

实践表明，对于数据库，不加限制的汉语句型查询的实现是极其困难的。对文法予以适当约束而得到汉语自然语言子集，又称为受限规则汉语。本模型旨在一个汉语子集上建立在一定范围内对汉语句型的理解，文[6]从语言学的角度给出类上下文无关文法，本文进一步给出其形式描述为：

〈语句 S〉::=〈$S_{查}$〉|〈$S_{陈}$〉(暂不处理)
〈$S_{查}$〉::=〈查询动词〉〈修饰短语〉〈目标短语〉
〈查询动词〉::=[请]〈基本动词〉[出]
〈基本动词〉::=查[询|找]|列|给|统计
〈修饰短语〉::=〈条件短语〉+[V]+[N]+"的"|[P]+[V]+〈条件短语〉+[N1]+[P]+[N2]+"的"|〈条件短语〉+[P1]+N1+V+[P2]+N2+"的"|[P]+N1+[V]+〈条件短语〉+[N2]+"的"[V]+[P]+N+"的"|略
〈条件短语〉::=〈基本条件短语〉[{,〈连接词〉+〈基本条件短语〉}]|

其中 N_i 表示表名，n_i 表示属性名。P 为限定词，如"所有"，"各个"，"不"等。C 为关联词，如"在"，"为"，"等于"，"是"等，V 为关联动词(对应相关表

*)本文得到国家自然科学基金和北京大学视觉听觉信息处理国家重点实验室资助。

名)。

限于篇幅,目标短语,基本条件短语,连接词不再展开,修饰短语14个子句仅列出5个。另外,文法还作了若干语义规则限定,如:

r_1:P 限定词仅修饰最近的表。

r_2:查询表中的属性值应同时指出其相应的属性名。

r_3:限定词"各"字后面应紧跟表名(或表的属性名)。

如"列出(供应产品类型为 A 类的供应商的)产品",修饰短语为括弧部分,可表为:

(修饰短语)::=V+(条件短语)+N+"的"

以上给出的受限汉语语法规则基本上覆盖了常用的数据库查询句型,目前模型所接受的汉语子集受到文法规则和应用领域的限制。

2.2 模型的形式定义

定义2.1 数据库基于 E-R 语义模型的汉语关键词理解模型 $\sum$ 是一个八元组,即

$$\sum=(S,V_N,V_T,R,P,\delta,\delta',W_d)$$

其中 S 是文法开始符号,V_T 为汉语基本词集,V_N 为汉语词类复合范畴(如短语等),P:语义规则式集合(有限),即:

$$p\Leftrightarrow\{p\to\alpha\mid(p\in V_N)\wedge\alpha\in(V_T\cup V_N)^*\wedge(\exists i)(S\in P_i\wedge(P_i\to\alpha))\wedge i\in Nat\}$$

如 $P1:S\to Q_k\oplus E_k\mid Q_k\theta\mid\theta Q_k\mid Q_{k1}\theta Q_{k2}$

其中 Q 为修饰短语,k 为修饰短语析出序号,E 为目标短语,亦为相应实体指称,$\oplus$ 为目标短语关于相应修饰短语的连结符,θ 为疑问助词,Nat 表示自然数集。

$$P2:Q_k\to Q_{k-1}\oplus V\oplus E_{k-1}\mid Q_{k-1}\oplus V\oplus Q_{k-2}\mid V\oplus Q_{k-1}\oplus E_{k-1}\mid Q_{k-1}\oplus E_{k-1}\mid Q_{k-1}\oplus V\mid V\oplus Q_{k-1}\mid\sim Q_{k-1}\mid P_{k-1}\odot Q_{k-1}\mid Q_{k-1}\oplus P_{k-1}\odot E_{k-1}\mid P_{k-1}\odot Q_{k-1}\oplus E_{k-1}\mid E_{k-1}$$

这里的 V 为关联指称(含动词与非动词性关联),P 是限定词,$\odot$ 是谓词与相应实体(属性)指称的析出符,或属性与其相应实体指称的析出符,$\sim$ 为否定符。

本模型的语义生成规则有六条。

R_3 为汉语词的语义指称规则,汉语字串与数据库语义模型之间存在着相应的语义变换。即汉语词 W 对应于 E-R 模型中某一实体(联系)或其属性,将 W 称为实体(联系)或其属性的指称。如词 W 不显式对应于 E-R 模型上的元素,需通过语义分析后才得其指称者为其蕴涵指称(见后面定义3.8)。

δ 为汉语修饰词的组词规则集[1]。

δ' 为深层语义映射规则集(见后面定义3.9)。

W_d 为汉语理解的背景词典,包括自动分词的专用词典和数据库词典。含汉语理解必需的词汇和有关应用领域的实体、联系、属性信息。

从以上模型的定义中不难看出,模型中的语义规则集 P 源于上下文无关文法 CFG,但又与之完全不同,因为汉语语句经 CFG 分析后只能得到其层次结构,而不可能得到更深层次的语义信息,这也正是该模型的主要特色。

3. 汉语查询模型的表层语义和深层语义

计算语言学认为,计算机对自然语言的理解是分层次的,各层次间在理解上存在着单向依赖关系,本模型对汉语的理解主要分成表层语义和深层语义的理解。

(1)汉语句型理解的表层语义

定义3.1 语言单位(词,短语)的表层语义表示语言单位所代表的基本语义内涵,本模型表示为一种类 SQL 的中间语言形式表,见表1。

表1 汉语句型的中间语言表

表名 属性名	关系符	属性值	条件分组字段	条件关系符	目标分组字段	排序字段
部门-楼层	=	二楼			部门名	
产品-类型	=	A 类		AND		
SUM(销售-销售量)						

定义3.2 语义函数 f 是汉语词 W 到它指称语义对象的映射。

语言单位的表层语义结构表是由其语义结构文法生成的,即:

定义3.3 语义结构文法是一个六元组 $SSG=(S,W,T,\rho(o),F,P)$,其中 S 为文法开始符,W 为汉语词集合,T 为语义结构表,$\rho(o)$ 为词 W 所指称对象集的幂集,F 为语义函数集,而 P 为语义规则式集。

该文法是一个汉语句型分析器,输入是汉语词集 W',分析器利用语法、语义规则式,以双向扫描的方式分析汉语词串得到汉语句子的表层语义结构表,存于中间语言栈中,其算法大意是:

算法3.1 (表层语义结构分析算法)

输入:汉语字串 $\alpha:w_1w_2\cdots w_n$,初始状态令指针 P_0 与 P_1 分别指向 w_1 与 w_n,中间语言栈为空,栈指针指向

栈成单元。

输出：识别后的汉语词置中间语言区。

算法要点：

1 取语义规则式 $S\rightarrow V_E\oplus Q_i\oplus E_i$，采用双向扫描，如 w_1 与 V_E 匹配，w_n 与目标短语 E_i 匹配，将目标表及属性存入中间语言区（自上而下匹配）。

2 目标短语 E_i 与修饰短语的定界原则（以反向扫描计算）：(1)第一个"表名"前的"的"为界；(2)如目标部分有多属性相连接则以第一个属性为目标属性（与(1)结合使用）；(3)目标属性所在表的确定，可根据该属性前的表名，或通过查实体表字典，或通过动宾（动补）结构，从修饰短语中的关联动词的相应实体表获得。

3 取语义规则式 $Q_i\rightarrow a_i$（修饰短语分解规则），与剩下的汉语词串匹配，若 a_i 满足，则对 a_i 中每一个成分 a_{ij}（含 V,E,Q）的相应规则式与汉语串匹配，直至完全匹配（自下而上匹配）。

[处理策略]：(1)"分词"阶段与"分析"算法结合进行，其中自动分词采用每次读一个字符与带索引的词典匹配，采用逆向最大匹配法处理；(2)所涉及的是同一实体时，条件语句带来知值时需要将同一实体作为两个实体表处理（同表嵌套）；(3)对不同实体处理时，每一实体处于中间语言栈的不同层次，等。

4 当 a_{ij} 匹配成功，则由语义词典得到每个属性或条件短语所在表，并将有关汉语词条压入栈内：(1)如各基本条件短语处于不同的关系表，则条件关系符置空，否则置 AND；(2)如汉语词串中含"各个"（按…）、"排序"等时分别在栈层相应的位置上置目标分组字段以及排序字段上置相应标记；(3)如词串中含有"不"等否定词时，在基本条件短语中作相应处理。

汉语句型的表层语义理解有如下性质：

定义3.4 汉语句型词串与相应的语义结构表的汉语词条不1-1对应的特性称为语义结构表的不完全性，该性质反映了模型在作表层语义分析时，略去一些对查询结果无影响的句子成分，仅保存其最主要的语义信息。

定义3.5 由汉语句型的歧义性而导致的同一汉语句型可生成两个以上的语义结构表中的语义表达式的性质称为语义结构表的不确定性。

该特点是由于汉语句型自身的歧义性而产生，需要给予必要的语义约束加以确定。

定义3.6 汉语词串经模型的表层语义分析后在汉语句型栈中所存放的语义表达式又称作汉语句型至 SQL 的中间语言（MQL）。

本模型的 MQL 采用类 SQL 的表格形式，如表1，其中条件关系符表示各基本条件子句的连接符（AND,OR），条件分组字段与目标分组字段分别为 having 与 group by 子句的标志。

(2)汉语句型理解的深层语义

定义3.7 语言单位（词、短语）的深层语义，表示为由其表层语义形式所作的深层语义处理，包括根据语言单位的指称语义，操作语义，根据语义映射规则库所作的语义转换而得到的新的语义表达式。

定义3.8 词语 W 的指称语义

①语义域

$e\in E$（数据库 DB 中全部实体集）
$a\in A$（DB 中全体属性集）
$r\in R$（DB 中全体关联集）
$v\in Value$（DB 中全体数据值集）
$DB=\{E,A,R,Value\}$
$w\in Input=$"$Value$（原输入的汉语词串）
$w'\in Output=Value$"（输出查询结果汉语词串）$\subseteq Value$

②语义函数

·查询结果：$result=Value\rightarrow$"$Value\rightarrow Value$；$result=\{error,stop,null\}+Value\times result$

·实体（关联）的指称函数 $E(R)Df:E(R)Df:Value\rightarrow Value$"；或 $E(R)Df(W)=\{e(r)\mid e\in E_i(r\in R_j),1\leqslant i,j\leqslant n\}$

·属性指称函数 $ADf:ADf:Value\rightarrow A\subset DB$；或 $ADf(W)=\{a\mid(\exists\ i,j,k)(a\in A_i\wedge A_i\in(R_j\cup E_k)\wedge 1\leqslant i,j,k\leqslant n)\}$

·蕴涵属性语义指称函数 $IDf:IDf:Value\rightarrow A\subset DB$；或 $IDf(W)=\{\uparrow w_i=w_i\mid(\exists\ i,j,k)((\uparrow w_i=a\in A_i)\wedge A_i\in(R_j\cup E_k)\wedge 1\leqslant i,j,k\leqslant n)\}$

其中 W 为汉语词串集，$\uparrow w_i$ 为 w_i 所蕴涵指称的属性。

这里汉语词条的指称语义的理解主要在系统的汉语表层语义的处理中进行的，关于深层语义的理解主要体现在由句子表层语义结构表向 SQL 语句的转换之中，根据语义结构表所提供的 E-R 语义信息以及类关系代数的基本语句模板转换成 SQL 子句，其转换规则有：

定义3.9 深层语义映射规则集 δ'

DS. rule j：
if $T,C,y(Y)=$*create view* ⟨*output*⟩
as (*Select* ⟨$y(Y)$⟩
from ⟨T⟩
Where ⟨C⟩)
DS. rule k：

if (*T*,*y*1)∧(*y*2(*Y*),*C*1)∧选择条件 *C*2∧*y*3(*Y*)⇒
create view (*output*)
as (*Select* (*y*3(*Y*))
from (*T*)
Where (*C*2)
group by (*y*1)
having (*y*2,*c*1))

其中 *T* 为关系(实体或关联)名表,*y*(*Y*)为关系名表中的某些列,*C* 为查询条件,选择条件 *C* 为满足某些列 *y* 上的条件。

以上的语义转换规则集是深层语义转换器对句型表层语义结构表(中间语言 MQL)所作的深入语义理解的主要依据。目前该模型共提供11条转换规则。

算法3.2 (深层语义转换器算法策略)

初始输入状态为汉语句型经表层语义分析后所得的语义结构表,由系统的分词及数据库词典的语义信息,汉语组词规则库 *B* 以及深层语义转换规则集,逐层处理汉语句型栈的语义表达式而生成 SQL 语句,算法的主要框架见文[6]。

限于篇幅,以下仅简述转换算法中所使用的部分关键技术要点:

(1)子句嵌套的关联路径寻找策略。设关联路径 L 由关联结点集 Node(表名、属性名)组成,即形成链:$L:node_1\rightarrow node_2\rightarrow\cdots node_k$,显然 node 中的属性名为关联嵌套的关键字。具体作法是,从目标表的关联属性出发,通过查找实体(关系)表的关联结构,找出与之相关的中间表的关联属性。然后再以此关联属性出发,重复上述查找,直找到条件表,即目标表,中间表(可能多个)与条件表间有相应的共同属性,作为结点间的关联属性。如果目标表与条件表间无共同的关联属性时,需查找关系表的关联结构,增加一临时关联表,使两者相关,此时使 SQL 子句相应地增加一嵌套层。

(2)中间语言(MQL)含内部函数时的处理策略:①如 MQL 中某一条件条目(项)中含内部函数时对应的 SQL 的条件子句不用 *Where* 子句而代之以"*Group by* 条件分组属性 *having*(条件)"表达;②如 MQL 的目标(表名.属性名)含内部函数时,相应的 SQL 语句含有 *Group by* 子句,此时只需在语句中能识别出"目标分组属性",并将该函数置于紧随 *Select* 子句之后。

4. 汉语查询模型的实现

该模型已在 Windows NT 环境下,利用 Delphi 4.0的 Object pascal 开发成功,通过 ODBC 实现对 Oracle 7.3的访问。整个系统由汉语理解模块以及 SQL 码生成器两大模块组成,包括自动分词、语法与语义相结合的汉语句型分析,形成类 SQL 的 MQL 的中间语言表格形式,并实现 MQL 到 SQL 的自动转换,详细实现细节见文[6](具体实现时已对原设计的硬,软件环境作了适当调整)。目前该系统已对24种不同的汉语句型通过测试,取得满意的效果,并继续对更为复杂的汉语句型进行调试和系统的进一步完善。

5. 汉语查询模型的可移植性

汉语查询模型的实用性很大程度上依赖于该模型的扩充性和可移植性,目前国内外对这方面的研究较少,可移植性是 NLCQI 对面向不同的运行环境和不同的知识领域的可适应性,即

汉语查询模型的可移植性是指汉语自然语言查询界面对不同的硬;软件环境的可适应性,包括界面在不同的计算机平台(如在单机或分布式网络环境),不同质的 DBMS,如 Oracle,Informix,Sybase 等的可适应性;由于网络与 DBMS 接口技术的发展,ODBC,以及 JDBC,CGI 等技术的成熟,使查询界面在不同环境上的可适应性在技术上已得到解决。

领域的可移植性指的是查询界面的语法,语义解释器对不同专业领域应用的轻易性和可适应性。

通常由应用程序测试原语法成分可否在新的专业环境下修改运行,并克服语义语法过份地依赖"专门域"知识的不足。

NLCQI 可移植性的内容主要包括背景词典、语义规则库扩充、原数据库应用字典转换等。对于背景词典与规则库的扩充,在系统的实际运行中,系统会根据用户查询语料,动态增补分词词典缺少的词以及汉语组词规则,具有一定的学习功能。对于数据库应用字典的自动转换,由于每个与 NLCQI 接口的应用数据库均有相应的数据库字典,模型可移植性转换机制应能根据不同的应用环境将该系统的数据字典转化成本系统模型所需的语义词典。包括关系名词典,关系结构词典,关系表关联词典等。其中要在相应的词典上增补上相应的"实体"、"属性"等语义信息。

鸣谢 衷心感谢暨南大学计算机系硕士研究生杨晓昀在系统实现中的协助。(参考文献共10篇,略)