汉语查询可移植接口的设计与分析

邰晓英, 童 颋, 徐洁磐
(宁波大学 上海大学(嘉定校区) 南京大学)

摘要: 本文提出在关系数据库上实现汉语查询可移植接口设计的一种方法,这一方法的主要特点是采用语义文法(SEMANTIC GRAMMAR)分类的ATN模式进行语法语义分析. 通过读取用户数据库信息建立词类知识库,从而可以支持各种应用系统对不同的SQL标准关系数据库进行汉语查询.

关键词: 汉语查询接口,可移植性,语义文法

THE DESIGN AND ANALYSIS OF CHINESE QUERY TRANSPOTABLE INTERFACE

Tai Xiaoying, TongFu, XuJiepan
(Ningbo University Shanghai University at Jiading, Nanjing University)

ABSTRACT: The paper proposed a method of realizing Chinese query transpotable interface design for the Relational Databases. The method using ATN model based on the semantic grammar for syntax and semantics.Through reading user's database information to build dictionary knowledge base,thus the Chinese query transpotable interface can be realized for relational database for all application systems.

KEYWORDS: Chinese query interface, Transportability, Semantic Grammar

1.引言

一般数据库上的自然语言接口大多是针对特定的数据库系统,特定的应用领域而设计的.所以当出现一新的数据库系统或开辟一个新领域时,几乎要重新设计一个接口系统,而自然语言接口系统本身开发周期长,这就必然造成人力物力的极大浪费.为此,我们考虑建立一个汉语查询可移植接口. 当出现一个新的数据库系统或一新的用户系统时本接口系统仍能适用.

另一方面,由于数据库语言SQL是关系数据库的语言标准,为实现可移植性,采用了将汉语查询句转换为SQL语言查询句的策略.也就是说可移植性限制在SQL语言标准关系数据库.因此,只要是SQL标准关系数据库下开发的任何领域的应用系统,都可将此接口系统连接到应用系统上,从而实现汉语查询. 这里需说明的是,当用户系统的开发者开发成功一个应用系统后,就可将本接口系统与用户的应用系统相连接.连接方法是在本系统引导下以对话方式进行,连接成功后,用户开发者将本系统作为应用系统的一个模块交付用户使用即可.

注: 本研究工作受到国家教委宁波市留学回国科研资助费资助

2.设计思想及实现方法

为了将汉语查询句转换为SQL语言,需对查询句进行词法、语法和语义分析;建立相应的分析用词类知识库;探讨便于生成SQL语言的分析结果的语义内部表示;最后将此语义表示转换为SQL语言. 因此,本系统由汉语理解,领域知识获取子系统和SQL语言转换执行子系统组成.汉语理解完成对输入的汉语查询句进行词法、语法和语义分析;领域知识获取子系统则从用户数据库中读取有关信息组成词类知识库;SQL语言转换执行子系统实现从语义内部表示到SQL语言的转换及执行.

由于词法分析在本系统中不具有特殊性所以本文不加以讨论. 问题的关键所在是对不同的应用领域来说,如何组织词类知识库,如何给出单词的语义含义,采用何种语法语义分析方法以及如何将分析结果恰当地进行语义表示.

我们首先针对汉语查询句的特点,关系数据库模式和SQL语言进行了考察,在考察的基础上提出词类知识库的建立由应用系统开发者与接口系统交互进行,分析策略采用语义语法分类的ATN (Augmented Transition Network)模式,分析结果则采用框架表示.以下分别讨论词类知识库的组织与建立方法,采用ATN模式的分析策略及所生成的框架表示.

2.1.词类知识库的组织与建立

关系数据库的基本结构是二维表.这种二维表称为关系.关系中的列称为属性.每个关系有一个关系模式,一个关系模式由一个关系名以及它的所属属性名(域名)构成.如某个数据库应用系统由三个关系组成,它们的模式是:

学生(S)(学生号(SNO),姓名(SN),系名(SD),年龄(SA))

课程(C)(课程号(CNO),课程名(CN),先修课程号(PCNO))

选读(SC)(学生号(SNO),课程号(CNO),成绩(GRADE)

设有一查询句:"检索选读课程号为C1的所有学生的姓名.",其SQL查询如下:

```
SELECT S.SN
FORM S,SN
WHERE S.SNO=SC.SNO AND SC.CNO=C1
```

由上例知,从汉语查询句到SQL语言的转换过程中,所需要的语义信息为:关系名,域名,域值,各种条件的语义表示等.而不是通常的人称,动作,工具,场所等语义信息.

因此,我们对单词的分类采用了语法和语义共同确定的分类法.例如:学生(S),其类标识为rname,表示为一关系名,而姓名(SN)则为fname,表示域名,域值的类标识为fval.即单词的类型是根据关系数据库的特点分类的,与何种应用领域无关.除单词类型外,为了分析的需要,必需收集所有有关数据库信息.例如,对关系名来说,还应记录所属域名表以及哪些是关键字的信息.对域名来说,则要记录所属哪个库,是否是关键字的信息.而域值应记录其所属属性的信息.这些信息可以用单词来表述;再把单词组织为词类知识结构.下面是一个词类知识结构(主要几项)及几个单词描述(称为A类单词)的一个例子:

(中文词名, 词类标识,机内表示表,域名表, 所属库表,所属域表,...)
(学生,ranme,S,(SNO,SN,SD,SA))
(学号,fname,(SNO),((S ,SNO,Y)(SC,SNO, dY)))
(李杨,fval,(S.SN))

A类单词指的是与用户数据库相关的关系名,域名,域值等单词.除A类单词外还有一类起语法功能作用的B类单词,如"的","和","为","最大","大于"等等.B类单词在查询条件的判断中起着很重要的作用.如"课程号为C1的"短语中,"为"这一词类标识是"=",由此可得到CNO=C1这样一个条件短语."的"则表示其前面的内容修饰其后的内容.下面是几个B类单词的词类标识:(的 de),(为 =),(最大 max),(大于 >)(和 he)(个 ge)....

系统主要是靠A,B类单词的类标识通过ATN模式进行语法语义分析的.A类单词及其信息由系统的领域知识获取子系统从用户数据库中读取数据,并经过推理学习得到.B类单词及其信息由系统事先给定.除A,B类单词外的单词均称为C类单词,"情况","公司","信息","字母","字典"等等.C类单词在输入词类知识库时,一律以"NUL"标识.

A,B,C类单词及其有关信息构成本系统的词类知识库.

当用户系统的开发者将本接口系统模块与用户系统连接后,系统便读取用户数据库的信息并进行推理学习组织自己的知识库.此时尚需用户数据库的开发者与本系统交互通信获取知识. 系统从库中读出关系名,域名,域值(CHAR型)的各种信息并显示其机内表示.用户数据库开发者根据提示输入中文表示.在此过程中一种中文表示有可能对应多个中文词名,也可能一个中文表示对应多种机内表示,而词类知识库是以中文词名作关键字的.在学习各个单词的过程中需要经过逻辑推理,有时还需用户给一些启发信息以判断是增加一条新的词项,还是对已有词项增加信息等等.此外,每当用户数据库更新,修改后,接口系统均需重新扫描数据库,以获取最新知识.

由上得到的单词知识是对A类单词而言的,对B类起结构作用的单词,系统预先尽可能地全部将其存入原始词类知识库中. 若在查询中遇到生词可通过生词学习得到. 对C类单词,通过对大量例句分析,得知其数量有限,所以我们将一部分能预知的作为原始库的词类信息外,余者也可以通过生词学习得到.

系统在扫描汉字串时,每当遇到B,C类生词便调用生词学习模块(每次可处理一个生字串).生词学习模块提示用户输入有关信息,用户便根据提示进行输入, 系统将其添加到词类知识库.

2.2 分析策略

对查询句如何进行语法语义分析,是理解的关键.我们通过对大量查询句的分析,得出查询句有如下特征:

1) 查询句可分为查询条件(COND),查询对象(OBJ),排序语(ORD),定义语(DEF)四部分,其语序如下:

[DEF|ORD](COND)$^{+}$(OBJ)*[ORD]

[...]表示其内容或者没有或者一次, (...)$^{+}$表示其内容可一次或多次.(...)*表示其内容可零次或多次.并且前后两个ORD不应同时出现.

2) 查询对象一般位于句尾,这样适于从右向左扫描先识别出查询对象后再判查询条件.我们先对查询对象的若干模式进行匹配,当查询对象模式匹配不成功时再进入到查询条件的判断.

3) 查询条件的若干种匹配模式

查询条件按条件短语的不同有以下若干种匹配模式:

```
meiy+fname+de                                  ;没有fname的
bi3+fval+de+fname+(>|<)                        ;比fval的fname大或小
zai4+[COND]+he+[COND]+zhi1jian1+de ;在(条件短语1)和(条件短语2)之间的
fname+compop+fval +de                          :fname (大于|小于|等于|大于等于|...)fval的
fname+=+fval+de                                :fname(为|是|等于|...)fval的
.....
```

通过上述对查询句的考察和分析,我们采用扩充状态转移网络(ATN)作为语义语法分析模式.ATN是从有限状态转移网络进化而来,在递归状态转移网络基础上对它的每条弧又附加了测试,行动与分析结果的存储等功能,因此ATN既可看作是一种语法的形式化,又可看作一种自动机器,为自然语言理解系统所采用.但ATN也有其不足之处,即它是一种非模块结构,随着结点的增多,复杂性会急剧增长,修改一个很大的ATN会起到意想不到的副作用.而且它对语法的紧密依赖,也限制了它对不符合语法的(然而是有意义的)话语的处理能力,又因为对无意义的语法成分的大量分析,从而降低了系统效率.

而本系统却不存在ATN的这种缺陷.其原因为:

1)我们描述的是汉语查询句,由查询句构成的ATN网络的结点数很有限,因而复杂性较低.

2)我们采用的是语义语法分析,每当识别出一个语句成分时,它是符合语法而且具有一定含义的,可避免对无意义的语法成分的大量分析.

查询句的ATN如图1所示.

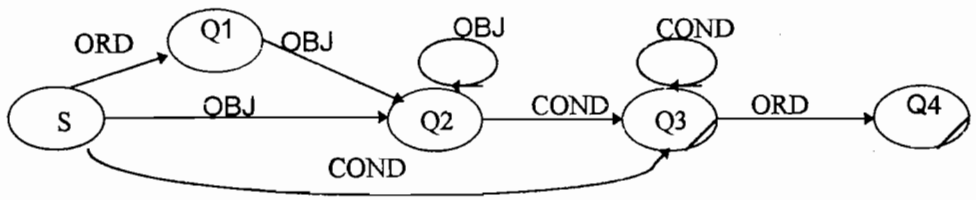

图1.查询句的ATN

这里,OBJ,ORD,COND 都是一个个子网络,系统由S状态开始分析,Q3,Q4为结束状态.系统首先测试当前单词是否在ORD的类标识集合中,若满足条件则进入ORD子网络;否则系统进一步测试当前单词是否在OBJ的类标识集合中,若在则进入OBJ子网络;否则进入COND子网络.每个子网络都有若干种短语匹配模式.在匹配的过程中需随时保留识别出的中间结果并执行为相应短语设计的动作子程序.子程序的功能一是进一步作语义判断,二是产生SELECT框架.SQL转换执行子系统根据此框架生成SQL语言的查询句.

2.3.语义框架表示

我们采用框架结构表示分析后的结果,主体框架如下:

```
( SELECT
   (OBJECTLIST  (DISTINCT[Y/N]
      ( FNAME1 :FNAMEVAL1  FUNTYPE : FUNOPSET  OPERptr : OPERSET
          (FUNTYPE1 :FUNOPSET ) )...
      (FNAMEn :FNAMEVALn  FUNTYPE : FUNOPSET  OPERptr: OPERSET
          (FUNTYPEn  :FUNOPSET) ))
   ( RELATIONLIST  (RNAME1,RNAME2,...,RNAMEm) )
   (BETWEEN V1:FNAMEptr NOT:[Y/N]  V2:FVALptr)) ...
   (CONDITION
      (COMPARISON1  V1:FNAMEptr  COMPOP:COMPOPVAL V2:FVALptr
        AND:COMPARISONptr  OR:COMPARISONptr)...
      (COMPARISONP V1:FNAMEptr COMPOP:COMPOPVAL  V2:FVALptr
        AND:COMPARISONptr  OR:COMPARISONptr))
   (ORDER  (FNAMEptr1  ASC:[Y/N]  DESC:[Y/N]) ...
           (FNAMEptrt  ASC:[Y/N]  DESC:[Y/N])))
```

这里,OBJECTLIST表示查询对象表,RELATIONLIST表示关系表,CONDITION表示各种条件框,ORDER表示排序框.$FNAME_1$,...,$FNAME_n$为各个查询对象.FUNTYPE为聚合函数类型.OPERptr为算术操作运算及运算对象的联结.$COMPARISON_1$,...,$COMPARISON_p$为各个条件短语框架.这种框架表示既适于存储分析结果又适于生成SQL语言.

3.实例分析

例如:"检索选读课程号为C1的所有学生的姓名并按学号升序打印输出."经词法分析后得到分析结果如下:

(检索 select) (选读 rname (SC) (SNO,CNO,GRADE)) (课程号 fname (CNO) (C.CNO Y) (SC.CNO dY)) (为 =)(C1 fval (C1)(C.CNO,SC.CNO))(的 de) (所有 suoy) (学生 rname (S) (SNO,SN,SD,SA))(的 de)(姓名 fname (SN)(S.SN N))(并 bing) (按 an4) (学号 rname (SNO) (S.SNO Y)(SC .SNO dY))(打印 ord) (输出 ord) (/ /)

语法语义分析程序按如下的ATN进行分析:

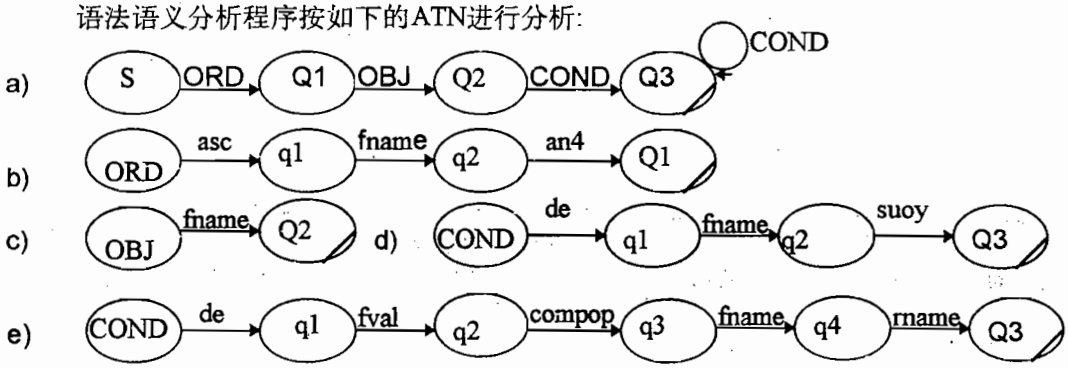

图2."检索选读课程号为C1的所有学生的姓名并按学号升序打印输出"的ATN分析图

经分析后得到的框架表示如下:

```
(SELECT
   (OBJECTLIST
        (DISTINCT:N)  (FNAME1:S.SN  FUNTYPE  : _  OPERptr : _ (FUNTYPE  : _ ))
   (ORDER
        (FNAMEptr1:S.SNO  ASC :Y  DESC:N))
   (RELATIONLIST (S.*,SC.*) )
   (CONDITION                                  (COMPARISON2
            (COMPARISON1                              V1:S.SNO
                 V1 : SC.CNO                        COMPOP: =
             COMPOP :  =                                V2:SC.SNO
                 V2: C1                              AND:_
               AND :COMPARISON2                        OR:_
                OR : _ )                          ) ))
```

4.实验结果

本文提出的汉语查询可移植接口设计,采用C语言编程,通过在DOS(其它操作系统的实验尚在进行中)下对FOXPRO和ORACLE两种关系数据库所作的实验模型验证是可行的[7].既可方便地连接到这两种关系数据库用户系统上.因关系数据库系统种类繁多,格式各异,适应的操作系统也不同,其它类型的数据库系统格式尚在实验中.文中所采用的ATN语义语法分析方法,经过几百多例查询句的分析,80%以上得到满意的结果.今后将在分组条件的判断方面作进一步的研究.

另外,本接口系统研制成功可为汉语语音识别系统所采用,组成汉语语音查询可移植接口.

参考文献

[1] Barbara J.Grosz,Douglas E. Appelt,Paul A. Matin & Fernand C.N.Perira,TEAM:An Experiment in the Design of Transportable Natural Language Interface,Artificial Intelligence,1987.5

[2] 计算机与信息处理标准化,全国计算机与信息处理标准化技术委员会会刊,1990,1.

[3] 徐洁磐,王银根 ,《数据库系统导论》,科学文献出版社,1991.

[4] 童 兆页, 沈一栋,《知识工程》,科学出版社,1992.

[5] 邰晓英,童 兆页, 限制汉语语法分析中歧义性的启发式方法,中文信息学报,1993

[6] 石纯一,黄昌宁等《人工智能原理》,清华大学出版社1993.10

[7] 王勤波 , 关系数据库的汉语查询接口---知识获取子系统,宁波大学毕业论文,1995

[8] 堂下修司等, 《音声.言语.概念の统合的処理による对话の理解と生成に関する研究》,研究成果报告書 ,1996