

基于受限汉语的数据库自然语言接口技术研究*

许龙飞¹, 杨晓昀¹, 唐世渭²

¹(暨南大学 计算机科学与技术系, 广东 广州 510632);

²(北京大学 信息科学中心, 北京 100871)

E-mail: txlf@jnu.edu.cn

摘要: 介绍了一种新的基于受限汉语的数据库自然语言接口 NLCQI(natural language (Chinese) query interface) 的系统模型及设计框架. 给出系统实现中具有特色的多栈结构的中间语言以及以关联路径搜索方法实现的中间语言向 SQL 转换的策略. 实验表明, 该系统采用的非过程化汉语查询句表达方式较自然, 对汉语句型的理解、处理能力有较大的改进.

关键词: 自然语言界面; 受限汉语; 自动分词; 中间语言; 数据库模型

中图法分类号: TP311 **文献标识码:** A

数据库自然语言接口是自然语言理解与数据库技术结合的产物, 近年来, 作为 AI 中自然语言理解的智能接口技术而受到重视, 尤其与汉语的手写体及语音识别的结合研究, 具有很高的理论价值与广泛的应用前景.

在该研究领域内, 纵观近年来国内所研制的多个系统, 所采用的技术主要有基于数据库的 E-R 汉语理解模型、类关系代数逻辑式的中间语言转换、以条件为中心的句型匹配以及多语句组合模板等方法. 在此基础上, 我们在文献[1,2]中提出一种新的基于受限汉语^[3]的数据库自然语言查询界面 NLCQI(natural language (Chinese) query interface), 给出这种接口系统的模型框架、基本原理与设计思想.

该模型的主要特点是:

(1) 采用了数据库技术、计算语言学与人工智能等多学科结合的新思路. 近年来的研究实践表明, 要想最终解决数据库的汉语自然语言接口问题, 靠纯语言学或纯数据库技术都是行不通的. 为此, 本系统让模型建立在受限汉语集合上, 运用汉语自动分词技术, 受限汉语语法是根据数据库汉语查询句中常用的词法与语法而建立的一系列的语法、语义规则^[1,2], 并采用数据库 E-R 模型与其指称的数据库模型语义及背景知识结合的技术. 与近年来国内同类系统相比, 在构思上有新的特色^[4~7].

(2) 所输入的汉语查询句型比较符合中国普通用户的思维习惯和表达方式, 本系统采用了完全非过程化的汉语自然语言方式, 在表达形式上较为灵活与多样性, 同一语义的查询语句可以有多种不同的表达形式.

(3) 从汉语句型到 SQL 的中间语言形式都采用类关系代数形式的语义查询树, 而在实现技术上采用了多栈结构形式, 既能准确地表达原查询句的语义, 而且在形式上也更灵活, 便于向 SQL 的自动转换^[7].

(4) 提出了以关联路径搜索方法实现 MQL 到 SQL 的转换策略, 解决了 SQL 中多层嵌套子查询的搜索难题. 这在国内同类系统中尚未见到.

东南大学的 CQI 系统^[4]是国内首次研制成功的基于 E-R 模型的 DB 中文查询接口, 与之相比, 本系统在计算语言学的结合研究、对 E-R 语义模型的理解以及系统对汉语的处理能力方面(如对汉语句型修饰段的处理

* 收稿日期: 1999-11-28; 修改日期: 2001-07-06

基金项目: 国家自然科学基金资助项目(69633020); 北京大学视听觉信息处理国家重点实验室资助项目; 暨南大学“211 工程”资金资助项目

作者简介: 许龙飞(1946 -), 男, 广东开平人, 教授, 主要研究领域为数据库系统, 知识工程; 杨晓昀(1974 -), 男, 广东湛江人, 硕士, 助理工程师, 主要研究领域为数据库应用系统开发技术; 唐世渭(1939 -), 男, 浙江镇海人, 教授, 博士生导师, 主要研究领域为数据库与信息系统, 数据仓库技术.

有更大的灵活性与适应性)已有了较大的改进.与中国人民大学和香港中文大学研制成功的著名的中文数据库查询界面 Chiqi^[6]相比,也有自己的特色,本系统采用完全非过程化的汉语自然语言方式,表达方式更加自然,用户不必理解、记忆和选择多个语句模板.同时,克服了由于采用多语句执行方式而影响查询性能等不足.

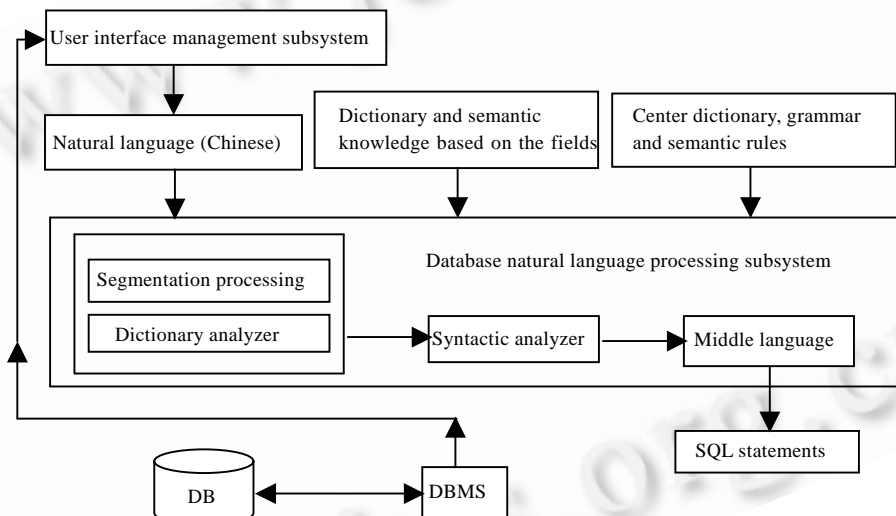
下面本文将在此基础上对系统模型的关键性技术的实现作深入研究,并对涉及系统实用性的重要技术——应用领域的可移植性作有益的探讨.

1 系统的模型及总体设计

NLCQI 所采用的模型是数据库基于 E-R 语义的汉语关键词理解模型,是一个八元组,即 $= (S, VN, VT, RS, P, S, S', Wd)$.

其中 S 为文法开始符号, VT 为汉语基本词集, VN 为汉语词类复合范畴(如短语等), P 为语义规则式集合(有限), RS 为汉语词的语义指称规则, S 为汉语修饰词的组词规则集, S' 为深层语义映射规则集, Wd 为汉语理解的背景词典(包括通用词典与专用词典),详细意义见文献[1,2].在系统的实现中,对原设计的汉语查询树生成规则集 P 与受限汉语文法规则作统一处理.

整个系统可分成用户界面管理子系统、数据库汉语自然语言处理子系统两大部分,后者又分为分词子系统、中间语言生成器和中间语言向 SQL 的转换器几大部分,如图 1 所示.



用户界面管理子系统, 自然查询语言, 基于领域的词典及语义知识, 核心词典、语法及语义规则, 数据库自然语言处理子系统, 分词处理, 词典分析, 句法分析, 中间语言, SQL 语句.

Fig.1 NLCQI system architecture

图 1 NLCQI 系统体系结构

目前,本系统已在 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,该系统已通过 20 余种不同的汉语句型进行测试,取得较满意的效果,并继续对更为复杂的汉语句型进行再测试与系统的进一步完善.

以下仅就受限的汉语文法、系统词典设计、中间语言的组织与生成以及中间语言(MQL)向 SQL 转换等关键性技术作一介绍.

2 受限的汉语文法与规则

由于汉语自然语言所具有的复杂性与开放性,要使系统完全准确地理解其语义是很困难的,为此,近年来,国内学者转向对受限汉语查询模型的研究,我们在文献[1,2]中提出有关的方法与规则,设计的原则是既能基本覆盖常用数据库的查询句型,又便于在计算机上的实现,本系统在实现中所采用的文法规则是:

<S 查> = [<查询动词>][<修饰段>]<目标段>
 <修饰段> = [<分组段>]<修饰短语>[<分组段>][<分组段>][L]
 <修饰短语>[<分组段>]][Z]
 <分组段> = D[E]P
 <修饰短语> = E+Z+F+P+O+V|E+F+P+O+V|E+Z+E+O+V|E+E+O+V(略)
 <目标段> = <目标短语>{[L]<目标短语>...}
 <目标短语> = E+Z+F+P|E+F+P|E+Z+P(略)

其中 E 代表实体(表名);P 代表属性;V 代表属性值;F 代表谓词函数;O 代表关系符;Z 代表助词;L 代表逻辑符;D 代表分组词。

“分组词”是指“各”、“各个”等,为判断转换成 Group by 而设置的词,其他词意,限于篇幅,不再展开。

受限汉语的规则是:

r₁: 查询句型仅限于祈使句;

r₂: 暂不使用“至少”、“当且仅当”、“除了...之外”以及相关代词“其”、“该”、“它(们)”等;

r₃: 查询实体中的属性值应同时指出其相应的属性名;

r₄: 查询目标实体仅限一个等(其余规则略)。

以上文法规则基本上覆盖了常用的数据库查询句型。随着系统的完善,可逐步放松对受限文法的约束。目前本系统对同一语义的查询句可有不同的表达方式,如:

- (a) (查)部门 为 玩具部 的 员工
 E O V Z E
- (b) (查)部门 名称 为 玩具部 的 员工 姓名
 E P O V Z E P

表示同样的语义,即对应相同的 SQL 语句。

3 关于系统的词典设计

系统将查询句中的词类分成 8 种:实体词(E)、属性词(P)、助词(Z)、关系符(O)、逻辑符(L)、函数词(F)、限定词(M)与分组词(G)等。其中实体词包括数据库中的表名以及数据表间的关联助词(对应相关表名)。

NLCQI 将词典分成与领域相关与无关的两大类,即专用词典与通用词典。

3.1 通用词典

通用词典与应用无关,包括助词(Z)、关系符(O)、逻辑符(L)、函数词(F)、限定词(M)与分组词(G),如“的”、“是”、“大于”、“最大”、“而且”、“各(个)”等。如文法中的查询动词“查(列、找、给)出”也归入通用词典。

3.2 专用词典

专用词典中存放与应用领域相关的词的语义及描述,如关系表名、属性名等,为了提高查询速度以及减少存储空间,专用词典又分为:

(1) 标准词与非标准词典

标准词主要是指全部实体词(E)和属性词(P),设置的目的在于将查询句型与数据库模式语义紧密关联。

非标准词典中存放的仍然是与应用领域相关的词,但它们不是数据库概念模式中定义的词,而仅是它的同义词,或它所定义的词,即非规范表达形式,为了提高查询速度,将非标准词与相应的标准词对应地存放在一起。其结构是:

| | | | |
|------|----|--------|-------|
| 非标准词 | 词类 | 对应的标准词 | 对应的实体 |
|------|----|--------|-------|

建立非标准词词典的目的在于减轻对用户查询文法的约束,用户无须务必使用数据库的标准词查询,使汉语的查询更加自然。

(2) 数据库关联词典

数据库关联词典主要描述关系表之间的逻辑关联,又细分为直接关联与间接关联词典,如间接关联词典说

明一个关系表通过属性组与其他关系表的连接,从而可以得到间接关联矩阵.

4 中间语言的构造与生成

中间语言的选取应遵循两个原则:一方面能够准确地反映原汉语句型的语义,另一方面,又能方便地转换成 SQL 语句.目前国内大部分系统均选用关系代数表达式.本系统所设计的中间语言是经自动分词后的汉语词串,由汉语查询文法自顶向下搜索,由左向右形成汉语查询的词组结构树,每一个叶结点对应于实体(关联)或属性的语义指称,存于句子栈中^[1],较以往系统的 MQL,形式上更灵活,更便于优化处理.

4.1 中间语言的实现结构

为了便于中间语言(MQL)向 SQL 的转换,在实现时,其结构分为句子栈、实体栈、查询目标位置栈与查询条件栈等.如句子栈存放查询句经自动切分后与查询目标相关的词((非)标准词或词的形式描述等),即剔除了查询中的若干干扰信息,其数据结构为

```
Sentence Stack=Record
    Elem:Array 1..Arrmax of string;
    Top:integer
End;
```

而实体栈中存放查询句中所涉及的实体或联系;查询目标位置栈存放查询目标的个数及每个目标在句子栈中的起始位置;查询条件栈中存放查询句中的查询条件个数、各条件在句子栈中的位置等信息.

例 1:如“查出楼层为“二楼”销售类型为“A”的产品的总销量”.经汉语自动分词后所得到的句型字串为“POVEZPOVZEZP”,其句子栈和实体栈的内容分别为:

```
总销量|的|产品|的|“A”|为|类型|销售|“二楼”|为|楼层,以及产品|部门|销售
      (Top)                                     (Top)
```

其中“部门”是通过分析属性“楼层”所属实体而得,查询目标位置栈和查询条件栈等内容略.

4.2 中间语言的生成

这是受限汉语理解的关键性步骤,共分 4 个阶段,即词法分析与句子栈生成;修饰段与目标段的划分;查询目标的分析与生成以及查询条件与条件栈的生成.

4.2.1 自动分词与句子栈的生成

自动分词采用逆向最大匹配法(BMM),以得到单词序列及相应词类及形式描述,根据文法自顶向下搜索,由左向右形成汉语词组结构树,算法的要点是:

算法 1.

Step 1.取切分所得到的词类(word type);

Step 2.根据词类性质作相应处理;

(1) word type=‘实体’

- 寻找 word 标准词,并将 word 标准词入栈
- 句型字串 句型字串 + ‘E’(E 为实体),

(2) 当 word type 分别为‘属性’、‘关系符’、(‘逻辑符’或函数词)、助词、量词、分组词时,亦作相应处理.

4.2.2 修饰段与目标段的划分

由受限汉语文法知,查询句的结构中对查询动词的剔除较为简单,下面我们来讨论目标段与修饰段的划分.

通常,目标短语较简单,故划分的关键在于析出目标段,其算法的要点是将文法中的目标短语根据 Length 由大到小排序,依次与句型字符串中最后长度为 Length 的字符串匹配,若匹配成功,则将此字串在句型字符串中的起始位置压入查询目标栈,并修改句型字符串,截去长度为 Length 的串,在进行下一轮匹配前判断修改后的句型串最后一个符号是否为逻辑符(逗号或顿号),以判断多个查询目标的存在.

4.2.3 查询目标的分析及生成

由上面得到的查询目标位置栈可知,查询目标的个数以及每个目标在句型字串中(或句子栈)的位置,即可知每个目标词类的表示.根据目标短语词性作相应处理.

算法 2.

Step 1. 目标短语 = P(属性)

- (1) 由目标短语位置栈可知句子栈中与 P 位置相应的词 wordp;
- (2) 从非标准词典中找出属性 wordp 对应的标准词与实体;
- (3) 若 wordp 对应的标准词其属性与实体仅有 1 个,则返回实体、属性;同时判断实体栈中是否含有分析所得到的‘实体’,若无,则将实体压入实体栈,分析结束;
- (4) 若 wordp 对应的标准词(属性或实体)不是 1 个,则 wordp 存在“二义性”,下面算法说明如何处理这种二义性.
- (5) 令 strp = wordp 在标准词典中对应的标准词;
- (6) 依次在实体栈中,从栈顶至栈底取出实体 E_i ;
- (7) 在关系结构词典中找出 E_i 的所有属性 Eallp;
- (8) 令 standard P = Eallp - strp;
- (9) 若 standard P $\neq \emptyset$,则将“ $E_i \cdot \text{standard P}$ ”压入临时目标栈(TempSta),结束分析,否则转(10);
- (10) 若 $i > 1$,则转(6),否则转(11);
- (11) 给出出错信息.

Step k. 若目标短语为 E(实体),EP,EZP,EFP,EZFP,亦对不同目标分别进行处理.

最后可得查询目标 Select Aim.

4.2.4 查询条件的分析

这一步的目的在于析出查询句中的条件,并以“实体·属性+关系符+值”或“谓词函数(实体·属性)+关系符+值”的形式表示.算法的基本思想是,分析中使用查询目标的实体(aim-entity),以确定条件中的属性是否属于目标实体,如果是,则以目标实体替换条件中的实体.因为当条件中的实体与目标实体间存在有共同属性时,则该属性即成为其嵌套连接属性,这样,不但优化了 SQL 语句,而且也降低了搜索目标实体与条件实体之间“关联路径”的难度.

5 中间语言到 SQL 的转换策略

从中间语言到 SQL 的转换原理是由多栈结构的 MQL 根据基本语句模板,通过转换规则,将 MQL 的数据及语义信息直接转换成 SQL^[1],整个转换的关键性技术在于寻找目标实体与条件实体间的关联路径.

定义 1. 如果两个关系通过某个共同属性直接连接,则称这两个关系是直接关联的,如果两个关系是通过第 3 个关系的某一属性进行关联的,则称为间接关联.

间接关联路径的搜索目的在于寻找两个间接关联关系,共同属性,在分析中需要处理关联路径的多选问题.

定义 2. 如果两个间接关联关系之间有多个共同属性,则使它们之间存在多条路径可选.

如调试例中的“销售”到“供应”有共同属性为“部门名”与“产品名”,导致两者间的关联路径链有两条,即:

(销售·部门名) (部门·部门名) (供应·部门名);
(销售·产品名) (产品·产品名) (供应·产品名).

解决多选路径的方法是,利用间接关联词典寻找目标关系与条件关系的连接属性与连接关系,并利用实体栈判断连接关系在实体栈中的情况,同时设置计数器,以确定多关联路径与相应关系嵌套属性以及正确路径的选取.

定义 3. 如果 3 个以上关系,相互间通过不同属性关联,则形成复杂关联路径,如图 2 所示.

处理复杂关联路径的算法要点是:

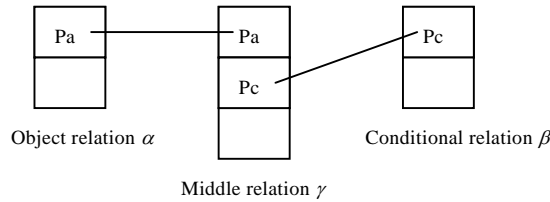


Fig.2
图 2

算法 3.

Step 1. 在实体栈中,找出与目标关系直接关联和间接关联的关系 A_d, A_u 以及与条件关系直接关联与间接关联的关系 C_d 与 C_u , 令 $A_{du} = A_d \cup A_u, C_{du} = C_d \cup C_u$.

Step 2. 若 $A_{du} = \emptyset \wedge C_{du} = \emptyset$, 表示实体栈中不含 A 与 C 有关的信息, 则从直接(间接)关联词典中找出 A, C 有直接或间接关系的 $R_{du}, R_{cu}(R_{du}, R_{cu})$ 作相应处理, 否则:

Step 3. 若 $A_{du} = \emptyset \wedge C_{du} \neq \emptyset$, 找出目标关系 A 的直接关联或间接关联关系(A 的直接(间接)关联关系不可能在 C_{du} 中)作相应处理. 显然 $A_{du} \neq \emptyset \wedge C_{du} = \emptyset$ 时与本情况相似.

Step 4. 若 $A_{du} \neq \emptyset \wedge C_{du} \neq \emptyset$, 令 $r = A_{du} \cap C_{du}$, 若 $r = \emptyset$, 则转 Step 5, 否则 r 中关系均为连接 A 与 C 的中间关系, 下面对 $N(r)(N(r)$ 表示 r 中元素个数)作判断.

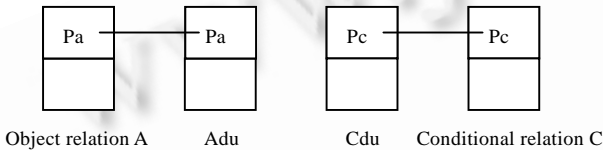


Fig.3
图 3

- (1) $N(r) = 1$, 则 r 为 A 与 C 的中间关系;
- (2) $N(r) > 1$, 其情况表示两个实体间的间接关联路径是多选问题, 按“多选路径”策略处理.

Step 5. 当 $r = \emptyset$ 时, 如图 3 所示.

从图 3 可知, 只要 A_{du} 与 C_{du} 可关联,

则 A 与 C 亦可关联, 具体算法是:

(a) 设 $A_{du_i} \in A_{du}, D_i \in A_{du_i}$ 的直接关联表, 若 $D_i \in C_{du}$, 则将路径 A A_{du_i} D_i C 作为 A 到 C 的关联路径, 否则:

(b) 令 $u_i \in A_{du_i}$ 的间接关联表, 若 $u_i \in C_{du}$, 则将路径 A A_{du_i} u_i C 作为 A 到 C 的关联路径, 否则出错.

Step 6. 若得到 A 到 C 的中间关系 r , 就可以通过直接或间接关联词典找出 A r 以及 C r 的连接属性, 而得到 A 到 C 的语句表达式.

至于分组属性的判断与处理(查询目标或查询条件中可能含有聚集函数, 使 SQL 中含有 Group by 子句)、查询条件间的关系判断以及量词处理策略等, 因篇幅所限不再展开.

例 2: 在例 1 查询句中的中间语言经由以上 SQL 转换策略处理后所得到的 SQL 语句为

```
select SUM(销售 销售量)
from 销售
where 部门名 in
(select 部门名
from 部门
where 部门 楼层="二楼")
and 产品名 in
(select 产品名
from 产品
where 产品 类型="A")
GROUP BY 部门名
```

目前, 本系统的主要不足是, 由于存在受限汉语语法的约束, 尚缺乏对更复杂的汉语句型的理解能力. 另外,

系统的实用性仍受到应用领域的限制等.

6 结束语

上面我们介绍了数据库自然语言接口的系统模型与总体设计的思想,给出受限汉语语法与规则、中间语言的生成以及中间语言到 SQL 的转换算法的详细描述.

进一步的工作是解决系统对应用领域的可移植性问题,应用领域的可移植性是系统实用性的关键,要求系统具有获取新领域知识能力以及系统的语法、语义分析算法不依赖于任何应用领域,其中包括:

- (1) 受限汉语语法、规则的扩充,背景词典与规则库的动态扩充.
- (2) 应用领域数据词典的知识获取,即系统从新领域中所进行的语义获取和提升方法,包括新关系名、新关系结构、新关系关联词典中“关系”或“属性名”等语义信息的增补.
- (3) 通过必要的人机交互以及文法上的适当限制,以消除汉语查询句型中的歧义性等.

References:

- [1] Xu, Long-fei, Tang, Shi-wei. Design and implementation of database natural language query interface based on the restrictive Chinese NLCQI. *Mini-Micro Systems*, 1998,19(7):26~33 (in Chinese).
- [2] Xu, Long-fei, Tang, Shi-wei. A study on natural language (Chinese) query model in database. *Computer Science*, 1999,26(8):43~46 (in Chinese).
- [3] Yu Shi-wen. *Chinese Literature Modernization*. Ji'nan: Shangdong Province Education Press, 1995 (in Chinese).
- [4] Gu, Guo-liang, Wang, Neng-bin. Design and implementation of Chinese language database query interface CQI. *Chinese Journal of Computers*, 1990,13(12):950~953 (in Chinese).
- [5] Zhang, Ya-nan, Xu, Jie-pan. An EAAD model for Chinese query on database interface. *Chinese Journal of Computers*, 1993,16(12):881~888 (in Chinese).
- [6] Meng, Xiao-feng, Wang, Shan, Lum, V.Y., *et al.* The multi-statement features and optimization in Chiqi. *Journal of Software*, 1997,8(7):549~554 (in Chinese).
- [7] Wang, Xin-min, Ye, Yan-bin. A Chinese language interface of database. *Mini-Micro Systems*, 1997,18(3):62~68 (in Chinese).

附中文参考文献:

- [1] 许龙飞,唐世渭.数据库汉语自然语言查询界面 NLCQI 的设计与实现. *小型微型计算机系统*,1998,19(7):26~33.
- [2] 许龙飞,唐世渭.数据库汉语自然语言查询模型研究. *计算机科学*,1999,26(8):43~46.
- [3] 俞士汶.关于受限的规则汉语的设想. *语文现代化论丛*.济南:山东教育出版社,1995.
- [4] 顾国良,王能斌.数据库汉语查询接口 CQI 的设计与实现. *计算机学报*,1990,13(12):950~953.
- [5] 张亚南,徐洁磐.数据库 NL 界面上汉语查询的 EAAD 模型. *计算机学报*,1993,16(12):881~888.
- [6] 孟晓峰,王珊,Lun, V.Y. CHIQL 的多语句查询特征及优化处理. *软件学报*,1997,8(7):549~554.
- [7] 王新民,叶延滨.数据库自然语言查询界面. *小型微型计算机系统*,1997,18(3):62~68.

Study on a Database Natural Language Interface Technique Based on Restrictive Chinese*

XU Long-fei¹, YANG Xiao-yun¹, TANG Shi-wei²

¹(Department of Computer Science, Ji'nan University, Guangzhou 510632, China);

²(Center for Information Science, Beijing University, Beijing 100871, China)

E-mail: txlf@jnu.edu.cn

Abstract: In this paper, model and design framework of a new database natural language (Chinese) interface technique based on the restrictive Chinese are presented. A middle language with the multi-track structure and a

