

# 关系数据库汉语查询接口的设计与实现

吕光楣 陈清波

(哈尔滨船舶工程学院计算机系)

**【摘要】** 汉语接口一直是我国AI界研究的热门课题之一。本文首先对汉语接口的可行性进行了论证,然后在分析了自然查询语言功能特征的基础上,提出了以词汇为基础,以语义特征为先导的综合处理技术。为使接口能进行移植,又引入了数据库模式字典,设置移植和学习模块,使之重建专用字典,与新库连接。最后给出了实验系统的结构和流程图,并作出该系统的性能评价和测试结果。

## 一、引言

近年来开发出的数据库系统大部份采用菜单式来引导用户进行查询的,这种方式虽然简单方便,但形式比较呆板,查询内容比较狭窄,满足不了各种各样用户的要求。另一方面,随着人工智能技术不断发展,自然语言理解目前虽还未获得根本性的突破,但对一定范围内语言现象的分析处理却已打开了可行性的大门。实际上国外早已研制了若干自然语言接口系统,国内也有不少单位对汉语接口问题进行了探讨和尝试,并取得了一定效果,我们就是在这种既可能又需要的情况下提出此课题的。特别在今天,计算机的汉音识别已初步达到实用地步,人机接口的汉语理解今后若与汉音识别配套成龙,前景更为诱人,到那时,人机可以直接对话,交换信息就更自然方便了。

## 二、汉语接口的可行性分析

汉语接口实质上就是让机器对用户用自然语的汉语对数据库内容所提出的各种操作要求进行分析,然后转换成数据库内部操作语言的一个转换器,或者说它是一个应用处理自然语言系统。

计算机最终能否理解处理自然语言一直是人工智能中一个非常热门又有争议性的研究课题。我们知道,自然语言是人类数千上万年生产斗争阶级斗争的产物,是人类用来表达知识、传递信息、交流思想感情的媒介,它是一个非常庞大复杂、又在不断发展演变的开放式符号系统,其中不仅存在着大量的多义性模糊性现象,而且还带有浓厚的感情色彩,夹杂些

本文1991年3月14日收到,修改稿1991年6月2日收到

丰富的成语俚语，故也可称为是一个‘不规则’系统，因此从根本上整体上来理解处理它目前还不可能，但如果把自然语言限制在一定范围内，也就是应用于它的一个子集，特别是像某一具体的数据库接口上则应该是完全可行的，这是因为

①数据库中的内容一定是明确的有限的，而用户的提问又总是围绕着数据库进行的，因此提问中的名词必为数据库概念模式中定义的词或其同义词、或可由它们定义的词，提问中的动词一般为数据库操作命令词、或与数据库关系名属性名有关的领域性动词。

②由于是向数据库提问，不可能出现带有感情色彩的词汇，也杜绝了成语俚语的出现。

③句型有所限制，句法有所简化，例如只剩下了祈使句，疑问句及相应的省略句。

④多义性和上下文有关现象大量减少，且有一定的规则可循。

⑤更重要的一点是，由于接口的最终目的是把自然语言转换成数据库内部查询语言，所以它并不要求完全彻底的去理解语言的深层含义，只要我们从语言的功能结构和语义的某些特征上去分析处理它，达到转换的目的就行了。

### 三、自然查询汉语功能结构与语义特征

#### 1. 自然查询汉语的功能结构

自然查询汉语虽然还是比较复杂，但从功能角度上看可以说是极为简单的，其结构主要表现为如下形式。

[前置词组] + 查询条件 + 查询对象 + [后置词组]

方括号中的内容为可选项。查询命令通常包含在前置词组或后置词组中，有时省缺，省缺时默认为显示。

如i. 请列出数学成绩大于90分的学生名单。

ii. 请将鲁川各门课程的成绩打印出来。

iii. X-2型巡洋舰的最大航速，满载排水量是多少？

其中下划线标出的为查询条件，用虚线标出的为查询对象。查询汉语另外还有如下特征：

(1) 通常查询条件位于查询对象之前，但如果查询对象由疑问词引导，则词序会有所改变，这时查询对象又往往为单项组成。

如i. 哪些学生是赵明老师教的。

ii. 我国是在哪年生产潜水艇的。

(2) 缺了查询条件，或查询条件为代词，则为承上疑问句，此时查询条件就是上一句的查询条件。

如 i. 鲁川的数学成绩是多少分？

ii. 他的物理成绩呢？

iii. 英语成绩呢？

后两句的查询条件均为鲁川。

(3) 当缺了查询对象时，或为承上疑问句，或为一般疑问句。

如 i. 鲁川的数学成绩为多少？

ii. 王红的呢？查询对象为上句的数学成绩。

iii. 王红的英语成绩大于80分吗？此为一般疑问句。

(4) 自然语言允许嵌套，一般在查询条件中再嵌入一查询语句，但寻常嵌套深度不会超过二层。

如：请打印出数学成绩大于鲁川英语成绩的学生名单。

句中鲁川英语成绩就是嵌在查询条件中的一个子查询语句。

## 2. 查询条件的语义特征

(1) 凡提问中名词为数据库中属性字符值或其同义词者必为查询条件。

如‘请打印年龄为60岁以上的教授名单’中的教授为数据库里属性名为职称的字符值，故可断定查询条件之一为：职称 = ‘教授’

(2) 名词虽为属性名，但后面紧跟有‘为’、‘在’、‘是’、‘大于’等关系连词以及数字单位时，则亦为查询条件，如上例中‘年龄为60岁以上’亦同时为一查询条件：年龄 > = 60

(3) 在若干查询条件中出现‘，’、‘的’、‘和’等分割符，助词连词时，这些条件多半是‘与’的关系，省缺也多是‘与’的关系，只有明显出现‘或’的连词才是‘或’的关系。如上例的查询条件为

职称 = ‘教授’ · AND · 年岁 > = 60

## 3. 查询对象的语义特征

(1) 凡名词为数据库中的属性名或其同义词，其后又不跟关系连词和数值者均为查询对象。

(2) 查询对象常在其前后带有疑问词，数值型的查询对象有时还带上单位量词。

如：i. 王红的数学成绩是多少？

ii. 巡洋舰上装有几门大炮？查询对象：火炮数。

iii. 巡洋舰上装的是什么火炮？查询对象：火炮名

(3) 多项查询对象之间有时夹有分割符‘，’或连词‘和’，通常情况是省略。

自然查询汉语虽有上述的普遍功能特征（当然还需进一步挖掘）可供我们在分析处理时作为主要依据，但另一方面，查询汉语作为自然语言的一个子集，仍然还存在着多义性现象，这是一个比较复杂难处理的问题，需认真加以对待。

自然查询语言的多义性主要表现在以下几个方面：

(1) 结构功能的多义性

由于库结构的不同，查询语言功能成分的划分也会随之而变。例如当库关系框架分别为学生（学生名，课程名，分数）

学生（学生名，数学成绩，物理成绩，英语成绩）

对同一查询语句‘请告诉我王红的数学成绩’的功能分析就不一样，按前者分析认为数学为查询条件，分数为查询条件，而后者则认为数学成绩整个为查询对象。

(2) 词用法的多义性

一个词在不同句中，在不同的位置有不同的作用，例如‘的’可作多个查询条件的分割符，也可作查询条件和查询对象之间的分割符。如‘请找出赵明教的数学成绩大于80分的学生姓名’，前后两个‘的’的作用就不一样。又例如‘和’在查询条件中一般是‘与’的关系，但有时又作‘或’的关系，如‘请将年龄大于60岁的教授和付教授的名单打印出来’，其中的‘和’便

是‘或’的关系。

### (3) 属性的多义性。

一个词的属性由于前后搭配不同，属性名可以不同，如‘赵明教了多少学生’？，‘赵明教了哪些学生？’，同一学生前者指的是学生数，后者指的是学生名。

词的多义性是一个很隐晦细微的语言现象，只有多方面的从习惯用法，上下文搭配，应用环境等来加以辨认。

## 四、一种新的分析方法——语义功能特征分析法

根据对接口的目的以及对查询汉语功能特征分析，我们认为传统的自然语言分析方法已不能很确切的适应我们现在要处理的问题，为此，我们提出一种新的方法——语义功能特征分析法，它是一种以词汇为基础，以功能特征为主导，把语义、语法、语用、上下文等信息综合起来加以分析的方法，此法对词的多义性有较强的识别能力，能高效达到转换的目的。

词汇是所有语言现象产生的基础，所以我们把它看作是一种知识源，它应该包括自身的词性、语义、习惯用法、以及与应用环境相关的特别是在当前数据库结构中所处的地位等各种有用信息，接口就是靠词汇提供的信息来分析查询语言的成份，确定其结构，并排斥其多义性的，因此设计组织好词汇字典，关系到整个系统的质量和效率。

以功能特征为主导，就是说我们的分析首先是从语义的角度来展开的，即在通过与字典的扫描匹配中，取得查询语句中各个词的必要信息后，先根据语义特征，自下而上的找到词与词的搭配关系，然后逐步形成查询条件和查询对象短词，以及所要访问库之间的连接关系。这种以功能为主的分析法只有在语义出现多义性时，才配以词法、句法、上下文等方法来加以解决，这里我们采用规则的办法，因此效率是比较高的，同时也便于维护扩充修改。

当然这种方法有时也会带来一些失误，由于语义特征的划分和摄取有一定主观片面性，语法又不作全面检查，多义性的细微区别考虑不周到，结果会使一些明显不合法的查询语句得出错误结果。不过随着对查询汉语功能特征规律的深入研究，字典结构的不断完善，分析处理方法的逐步修正，理解失误的现象会逐渐消灭。这点我们在调试程序中深有体会。

## 五、‘通用’接口的设计思想

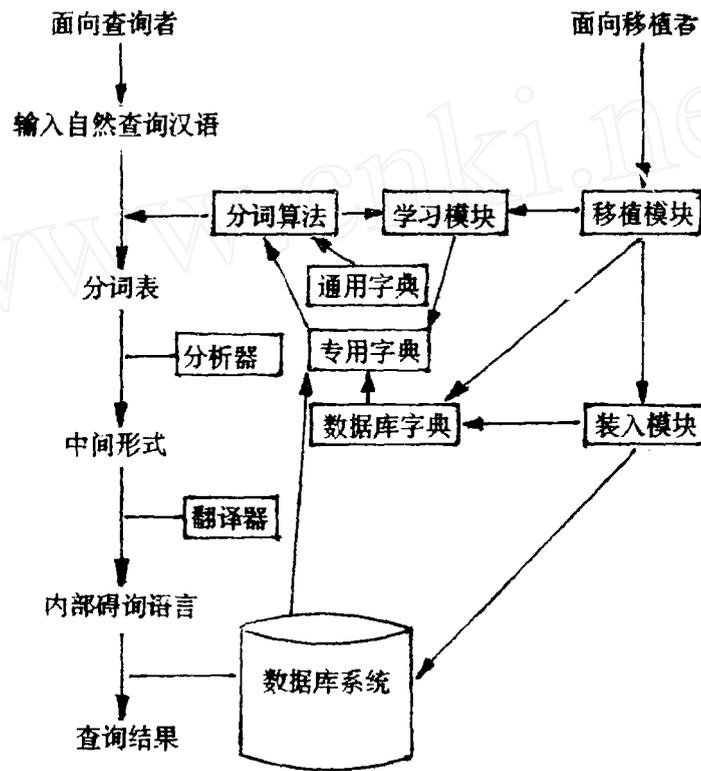
一个接口如果只能与一个具体数据库相连接，不管设计得如何精致，实际意义总是有限的。那么能否设计出一个‘通用’接口呢？我们认为这是可能的。因为在数据库模式不变的条件下（例如都是关系数据库模式），程序能够设计得与具体的库结构保持相对的独立性，也就是说这时的接口犹如一个框架工具，改换数据库，只要根据新的库结构与库内容改换相应字典的内容，就能在新数据库上进行询问了，这样也就达到了‘通用’的目的了。

具体做法是：（1）把作为知识源的字典分成二部份：通用字典和专用字典。把对各种库都通用的词，如‘大于’、‘等于’等关系词，‘列出’、‘显示’、‘打印’等操作命令词，‘最大’、‘平均’、‘总和’等运算词，‘是’、‘有’、‘哪些’、‘什么’、‘的’等连词助词，‘呢’、‘吗’等疑问词组织到通用字典中，而把与库内容密切相关的词，如主体词、关系名、属性名、属性字符值和它们大量的同义词，函数词，以及领域动词存贮到专用字典中。于是移植到一个新库

系统，只需改建专用字典即可，这样又节省了不少时间；(2)为了让接口系统能自动改建专用字典，这里我们都引入一个叫数据库结构描述字典，简称数据库字典。数据库字典也是一种知识源，它把欲相连接的各数据库的库名、主关键字、库与库之间的连接关系以及库中属性的名称、类型、特征等信息进行规范描述，这个字典规模不大，对于三五个小库组成的系统约百条左右，接口就是根据这些描述将数据库中有关内容改建成专用字典的主体部份的。(3)再设计一学习模块，使之可通过系统学习或随机学习不断扩展专用字典的延体，开阔完善接口的识别能力和范围。

## 六、系统总体结构

系统的总体结构和流程图如下图所示。



系统设置了两种工作方式：查询和移植。查询方式面对一般用户，用户只要知道数据库存贮的是关于那一方面的信息，围绕着这方面内容提出问题即可，不必知道库的结构和设置等情况。移植方式是用于将接口改换至新数据库上的，因此它需要移植者对所移植的数据库设置和结构要有比较详细的了解，这样才能在系统菜单提示下帮助系统建立新的数据库字典。装入模块是完全自动化的，不需要用户加以干预。

分词是把一个句子按词为单位将其分割开，这是汉语的特有现象。我们在分词算法中采

用了逐词遍历匹配法，即首先扫描词典，把词典中凡与句子中相匹配的词和特征提取出来，然后再处理句子中剩余的词，若是汉字通过学习赋给其特征，若是数字自动转换成数字串，最后按原句子的顺序排列得分词表。

例：请列出数学成绩大于等于90分的学生名单。

在经过逐词遍历匹配及数字处理后得分词表如下：

词 本 身	词主要特征	所在库号
请	无用词	
列出	操作词：显示	
数学成绩	属性名	2(成绩库)
大于等于	关系词 > =	
90	属性数字值	
分	单位词	
的	分割词	
学生名单	属性名：学生名	1(学生库)

分析器是自然查询语言接口的核心部份，它就是根据语义功能特征法的思想来设计的。里面除了考虑一般句子的分析外，也还考虑了省略句、指代句、反问句的分析与处理，特别是对多义词的分辨考虑得比较细。例如上述例句经分析后得到如下的中间结果：操作类型：显示；查询库：由学生库与成绩库通过学号相等条件连接成的库；查询条件：数学成绩 > = 90；查询对象：学生名。

有了查询操作类型、查询条件、查询对象和查询库这些中间结果后，翻译器就能很容易将其装配成数据库内部查询语言形式。到此，接口的转换工作结束，下面接着执行内部查询语句就能得出查询结果，这是属于数据库管理系统本身的功能了。

## 七、系统性能评述

根据上述的设计思想和流程图，我们用 Dbase plus 语言编制了一个演示系统，称为 CQID (Chinese Query Interface to Database) 系统，并在 IBM PC286 上调试通过。该程序本身约占 5K 字节，通用字典约二百余条词款，另外附带两个小型关系数据库，一为学生库，共分五个小库，共二百余条记录；一为海军装配库，共四个小库，约百余记录，专业字典主体部份分别有三百余条词款，经学习扩充目前不上四百条词款，基本满足表演所需。

经多次表演结果表明，本系统的性能可归结如下：

(1) 本系统只能用在比较简单的关系数据库系统上，即库不能过多，以几个为宜，但库结构要规范化。

(2) 本系统可以很方便地移植到一个完全崭新的数据库系统上，不过新的数据库也必须是个较简单的关系数据库。移植大约一小时左右，加上必要的系统学习也只需要数小时足矣。

(3) 本系统对输入的自然查询汉语虽在句型语法词序等方面不作任何限制和要求, 但用词不能太偏僻, 造句不能太生硬, 内容不能太离奇。测试证明, 只要提问句子通顺合情, 95%以上能给出满意回答。

(4) 本系统现在只有对库进行查询和进行简单运算统计功能, 还没有对库进行删除、插入、修改等功能。

(5) 本系统的输出形式只是数据库管理系统本身所提供的输出形式, 还未达到自然语言化地步, 不过这一步打算在一年内将其实现。

### 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] H. Ishikawa, Y. Izumida, A Knowledge-based Approach to Design a Portable Natural Language Interface to Database System. Software Laboratory Fujitsu Laboratory Ltd. Japan IEEE 1986
- [3] 黄晶宁, 汉语人机接口技术的现状与展望, 清华大学, 中国计算机用户, 1986.5
- [4] 管纪文, 黄祥善, 自然语言接口的设计方法评述, 吉林大学, 计算机科学, 1988
- [5] 杨员一, 一个数据库系统自然语言接口的设计与实现, 北京大学的计算机技术, 1988.1

## The Design and Implementation of Chinese Query Interface to Database System

LÜ Guang-Mei    Chen Qing-Bo

### ABSTRACT

The design of Chinese interface in one of the most important topics of AI in China. In this paper, first, the feasibility of natural language interface to database system in discussed. Then, based on the analysis of natural query language in function structure and semantic feature, We presented a Comprehension approach to processing Chinese query based on words, and leading by Semantic. To design a 'general' interface, a database model dictionary is used. A transplantation and a study component are set to rebuilding domain-specific dictionary. Finally, a demonstrative system is also given, it is proved that the system is practical.