计算机科学 1997Vol. 24 №5

# 数据库自然语言查询技术研究*)

A Study on Natural Language Query Technique in Database

许龙飞

(暨南大学计算机系 广州 510632)

**摘 要** This paper surveys database English Natural Language query techniques and Chinese Natural Language database interface studies in China. Major technical issues are presented for implementing a practical Chinese Language database interface based on the restrictive Chinese Language.

**关键词** Database model, Natural language interface, Chinese language query, transportability

## 1 引 言

近年来,随着DB技术的日益成熟,作为与计算语言学、人工智能等技术紧密结合的产物,数据库自然语言查询界面(DBNLI)的研究受到高度重视,成为新一代计算机系统研究的重要课题,具有很高的理论价值和广泛的应用前景。对复杂问题的查询,使用自然语言描述,较传统的命令菜单或触摸屏的查询方法来得更自然,更灵活,更全面和更准确。无需对用户作任何特别的训练。国外在这方面的研究技术已渐趋成熟,拥有多个商业化应用系统,如ROBOT, PLANES, IRUS, CO-OP, TEAM, NLI+等,而国内的汉语自然语言查询技术正处于探索阶段,近年虽也取得一些进展,但由于汉语的复杂性,基于对汉语理解的计算模型尚未完全建立,对汉语分析中的语法、语义、语用研究尚不成熟,因而直接影响接口研究的实用性。本文在深入分析国外DB英文自然语言查询技术的同时,结合当前国内汉语DB接口技术的研究动向,对实现商业化的汉语数据库接口进行研究。

## 2 英文自然语言查询技术分析

70年代末,美国研制成功采集月球岩石样本系统LUNAR,伊利诺大学开发了用于军用飞机维修的PLANES[1],还有可移植性的轻便系统LADDER等。80年代,人们的研究集中在提供自然语言界面(NLI)的"习惯性"(customization)使用上,所谓的"习惯性"是指NLI接受新领域和新的数据库系统的处理能力,如Datalog系统[2],其移植性可从办公室到家庭的信息领域,PRE系统则是办公室之间的信息移植。另外研制具有一定协调性响应的NLI也是这一时期的子目标,如美国斯坦福大学研制的CO-OP系统。90年代,数据库技术的发展,除继续向DB提供"习惯性"使用外,对新一代的数据库系统(Genesis, Exdus, Starburst)功能予以扩充,使用不同的知识表示策略,扩充的数据库系统以最自然的方式表示领域知识,完成用户的新应用。

以下将分几个方面介绍自然理解技术在数据库系统上的成功运用。

### 2.1 数据库英文自然语言理解技术的要点

**2.1.1 词法分析** 是专门领域知识的主要来源,可以为未知输入串构造新词,能定义不同词义,作为来自数据库词的生成对象。它提供解释问题中回答所需的语义知识。词法分析的主要手段是匹配,输入字串与数据库词典匹配。实用性的商业系统允许单词拼错或漏字,具有较强的容错能力。

**2.1.2 语法分析** 多数商业化数据库的NLI的语法分析采用了扩充转移网络法(ATN),它由递归网络加一测试集合以及一组寄存器组成,分析句子时,测试条件用来确定是否与一弧匹配,寄存器则保存其中间结果或全局性状态。如PLANES系统,它的语法分析器由匹配输入的机内对应子网和构造概念框架(语义语句模式)集组成。匹配模式后转换成语义构成的无序集,对每一子句,用规范化的形式取代用户输入条目,并用相关代词与省略解决。运用启发式分析法去寻找子句的分界线(边界),具体地,

*)得到国家自然科学基金资助,编号69633020。许龙飞 副教授,主要研究领域为数据库、知识工程等。

分成三个动作：①匹配先前已存贮的子句模式并设置正文登录表；②匹配正文登录值与概念框架所构成的模式；③填入缺少的正文信息，以形成有意义的查询句。其中(1)和(2)构成语言理解处理的核心，主要由子网处理，每一个子网就是 ATN 短语分析器，它匹配着一特殊意义的短语。在 PLANES 系统中，即对应飞机类型、日期、故障维护类型，多数子网匹配名词或前置词短语，子网构造基于名词短语的 Winograd 分析法，量词由特别的量词子网处理，如图 1 表示"数量"短语匹配子网。

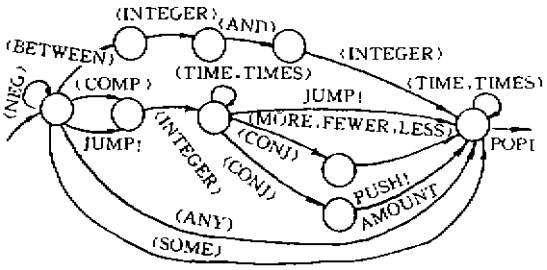


图 1　关于"数量"的语义子网

在子网与短语匹配后，短语将其信息存入与之匹配的子网中并输入下一个请求，试图以同一子网去匹配新输入串，若前置词短语后面为名词短语中的主名词，则 PLANES 会先假定前置词短语为主名词的量词。

多数子网与短语匹配后，系统会将其规范值送入正文登录表，该表主要用于相关代词和省略上，由栈实现。子网在识别短语类型时，可能会丢失某些信息，如"那些飞机"，可以理解成飞机类型或飞机组，但用户准确请求的那些飞机并不在其中，这时系统会从先前的上下文登录中选择丢失的值填入当前的上下文登录表。

语法分析中会涉及到"概念框架"，它由系统存入所理解的问题的模式(与数据库询问类型相关)，并含有动作(与动词相关)与名词短语表(涉及子网/上下文登录表)，这里的"子网＋概念框架"，构成了 PLANES 的语义语法。

一旦句子的组成成分丢失(如拼写)或由代词或有关短语取代时，系统需要了解短语的语义类，需查找所有的概念框架以完整该概念。

子网具有可扩充性。对网上的短语，可加上状态与弧与之匹配(可递归调用)，对专门的子网，可使用最少的新弧与状态数。一旦新短语加入到子网后，短语在任何句子的上下文即可被匹配。

2.1.3　语义分析　先通过句法分析产生句子的句法结构(如派生树)，再将该结构送入语义解释程序。在 DB 的 NLI 中，对复杂的句子还需要考虑某些语义信息，如修饰词(限定词)处理，子句边界的确定，含二义性(模糊)词(短语)的处理策略等。如 PLANES 系统，在名词短语中，修饰词通常出现在主要名词的后面，由相关代词引入，即"which"，"that"，如"planes which crashed in May"，"planes with poor maintenance records"，修饰限定词可从应用修饰子网对名词短语中主要名词后面的请求部分找到。当修饰短语及子句出现时会产生以下作用：

(1)启发式分析器请求证实修饰词出现且找出其边界；

(2)如修饰词被发现，则处理在当前子句中挂起，并由当前的上下文登录表值向下推；

(3)修饰词短语外的主要名词代换成有关代词，或插入作为修饰词短语元素。

(4)修饰短语或子句处理时，询问请求来自子句的主名词，并注意到动词形式所发生的变化，即词根＋词尾变化。

子句边界的确定由启发分析器进行，先是处理括号中的子句，将请求分成带非嵌入子句的简单问题。从而可知嵌入子句的左边界，直接处理外子句的主要名词，亦找到了子句的右边界(主要名词由内外子句分享)。寻找正确的子句界线的两个模式是：

【p1】{which|that}＋VP＋〈VP 外的其它成分〉＋右边界＋〈VP|语句尾部〉}。

【p2】{with}＋NP＋右边界＋{VP|语句尾部}。

TEAM 系统对修饰词的语义范围的确定采用了树结构信息，它运用语法树机制，获取所有可能的范围，由 Hendvix 所确定的基本策略是，对每个测试均带两种类型信息，即限定词的强度和关联逻辑形式的修饰词，整个算法是典型的生成-测试形式，生成的可能性范围由专家评估提供测试。

在语义转换中，对基本的语义函数调用转换器，生成与语法分析树相关的逻辑形式片段(LFF)，其中 LFF 标记含动词、形容词与前置词的谓词，以及各种来自名词性短语的类。

PLANES 系统对具有模糊语义的词(短语)采用定义特殊函数的策略，如"好"的维修记录被解释为"最低耗费(人/时)，地面时间/飞行时间的比率最小"等，TEAM 对二义性名词处理，在语法分析时系统强行分配相应词典的唯一语义形式，对二义性动词、形容词采用构造多描述谓词的策略，从而生成供

用户选择的多个解释。

通常DB的NLI具有语义解释生成器，允许用户检查系统是否理解其请求，系统反馈其对用户请求的理解，由用户选择适合的语义样板。

2.2 关于中间语言

在所有的商业化或实验性的DBNLI中，从用户的NL文本到数据库查询语言的转换间均设置了一种中介逻辑形式，称为中间询问语言(Meta-Query Language)，出自系统实现和优化方面的考虑，不同的系统所设置的MQL形式均不同。

如CO-OP的MQL是一图结构，结点表示由询问提供的实体集合，而边表示这些集合间的关系(由输入询问的词法与语法结构得到)，其主要作用是：

(1)准确反映输入的数据结构(小于一棵可修改的语法树)，并获取表面语法特点的机制。

(2)提供一种有效的语言描述级别，是对NL理解的形式化解释。

(3)MQL独立于数据库组织，即数据库设计者的观念视图的改变对询问的范围或回答内容无影响，即设计者所固有的DB视图对询问者透明。

在TEAM系统中，中间语言起着核心作用，特殊询问中的谓词和条目的语义来自词典和概念模式，故逻辑形式的选择会间接地影响系统组成的设计，并要确定DBE(数据库专家)所提供的某种信息范围。

TEAM使用先序逻辑，由某些内部和高阶操作予以扩充，且以特殊的修饰词作变量来定义WH定式(即what和which)，文[3]给出TEAM中所使用的中介逻辑形式规范的类BNF表示。TEAM的中间语言及其转换见2.5节中TEAM的查询例。

2.3 数据库自然语言查询响应的协调性

具有自然语言界面的数据库系统，其"自然"不仅体现在输入语言的"自然"，还要有系统作出响应的"自然"。如CO-OP系统，对"自然"询问的"自然"回答是：

(1a)Did Mahler complete an 11th Symphony?

(1b)Mahler began an 11th Symphony.

(2a)Which Soviet destroyers in the Mediterranean carry torpedos with a range of 50 miles or more?

(2b)Some torpedos have a range of 50 miles or more.

为此系统会经常向询问者提出重要提示(会话协调性的通讯途径)，询问器会离开选择直接响应，即以询问者观点提问题，会有多个直接回答，故它允许响应推导出提问者不知道的问题。

另外，系统会设置"预假设"，保证询问者所作问题的合理前提，当"预假设"失败时会消除对问题给予任何正确回答的可能性。这种预假设命题类具有这样的特性，它的失败需要对问题的回答至少有一个是潜在正确的，这种命题又称之为推测。

系统还设置了"协调"协议，即：响应是正确的任一推测，相信与其为假(对问题的非直接回答)倒不如给一个直接正确的回答，即使推测结论存在。根据"协调"协议，系统的策略是，询问者必须找到问题假设的所有可能性，以便能适当地提出问题，即该响应不允许选择直接的回答。如(1a)中的非直接回答"Mahler从未开始过作第11交响曲"，可以直接回答成"NO"，对(2a)，可接受一个非直接响应，"没有一个鱼雷具有50英哩以外的距离"。这个技术等价于推测。

为了生成各种协调(非直接)回答，系统的设计基于两个前提：①为了得到协调性响应，所作的推理可以直接地从输入问题的词法、语法结构中得到。②为了处理自然语言询问中有意义的一类问题所要求的领域上的专门知识应出现在数据库系统的标准路径上，且事先扩充了合适的已编码词典。

2.4 数据库NLI的可移植性

DBNLI的可移植性是NLI对面向不同知识领域的DBS的可适应性。自然语言询问的输入对NLI是大体相同的，而对不同的DBS，其输出的码可以完全不同，即输出依赖于数据库的结构特征。NLI可移植性的主要功能之一是作必要的转换，使用户界面与数据库特性独立。

为了提供这种"独立性"和沟通用户与DB数据结构之联系，要求将一般信息与特殊域相结合，尤其是系统必须具有一个应用域的专题模型，包括领域上的课题信息，必须了解该模型上实体与DB中信息间的关联，在构造可移植性系统时，重要的是提供获取专门域上信息的任何语义。

这里"可移植性"看作为"学习"问题，又称为"获取"，自然语言界面必须了解其新领域与数据库，第一步要求建立具有新域的超数据库，以增强给定NLI的可移植性。数据库系统为了增强特殊的NLI对新领域的可移植性，设置与相关数据库相独立的NLI子目标是：①NLI在域规则改变时，不需要修改。②特殊域上的NLI在DB重构时，亦不需要修改。

可移植性内容包括域(模式、数据模型、数据库系统)的可移植性。称 NLI 提供域知识的"轻易性"为域的可移植性,而模式可移植性为 NLI 能接受新数据库模式的轻易性等。

NLI 的语法可移植性是一种以其语法成分能采用新域的轻易性,由于域的独立性,可利用传统语法实现。语义语法的主要缺点在于过份依赖域的专门性,为了使 NLI 可移植,语法需重写,Datalog 系统提供了"一般语义"与"域语义"的分离策略。其中"一般语义"含语义产生式汇集,提供了分析的解释程序;"域语义"包括库词典、应用词典以及特殊域信息网。为了使系统采用新域,必须重构语义网,修改应用词典,使之含专门域的单词,而库词典亦予修改,以反映词的非标准解释。为此,设置词典入口指针,为其解释器指向语义产生式。

数据模型的可移植性,被认为是数据库系统中数据模型的扩充,文[9]中建议采取一种中介模型,即 $NF^2$ 模型作为 NLI 概念模式与相应的关系模式间的规范关系,它能表达更加丰富的语义,建造在 $NF^2$ 上的数据库要求系统要习惯于 NLI 知识表示策略。

TEAM 的"获取"[3]是以菜单形式由用户提供数据库的关系、字段等信息,回答系统所提出的有关问题。对每一个问题的回答,会直接影响词典、概念模式与数据库模式,DBE 无须具备任何形式语言或 NL 的处理知识。如动词的获取,TEAM 通过检测动词,将信息交给词典、概念模式、数据库模式。较好的 NLI 一般均提供两个一般性动词"have"和"be"。如"What country has an area great than 3 million square mile?"以及"What peak are more than 10,000 feet high?"。英语的动词行为很复杂,体现在获取新动词比较困难。对每个动词,TEAM 必须发现它具有多少变量以及这些变量是否可选,这些变量如何映射到各种主题或对象之中。

### 2.5 数据库自然语言查询界面系统结构

下面仅简介较为著名的 TEAM 系统。该系统分成两个主要部分:从自然语言表达式映射到其含标准中间逻辑表达式的 DIALOGIC 系统和模式转换器,后者将逻辑表达式转换成数据库询问语言的句子。图 2 表示从输入的英文句子到生成数据库询问的数据流,矩形框为构成,椭圆为各种知识源。

DIALOGIC 将语义解释器的操作分成两个主要类型:①转换器:定义短语结构的解释器。②基本语义函数:由转换器调用以构成短语解释的特殊表示。

DIALOGIC 包括 DIAMOND 语法分析器、DIAGRAM 语法表、词典、语义解释函数。基本实用函数可产生且确定修饰词的范围。该部分的主要数据结构是词典和概念模式。

模式转换器主要解决的问题是:①将逻辑形式的修饰词映射到数据库询问的操作;②确定如何将逻辑形式的谓词变量映射到数据库的关系字段和值中;③确定什么信息包含在询问回答中;④从逻辑表达式中删去冗余约束,产生有效询问。

如用户给出一请求后,系统经 DIALOGIC 处理后得到中间逻辑形式,并由模式转换器将其变换成子句询问后,经优化处理,最后形成可直接对数据库查询的 SODA 语言,详细的查询实例见文[3]。

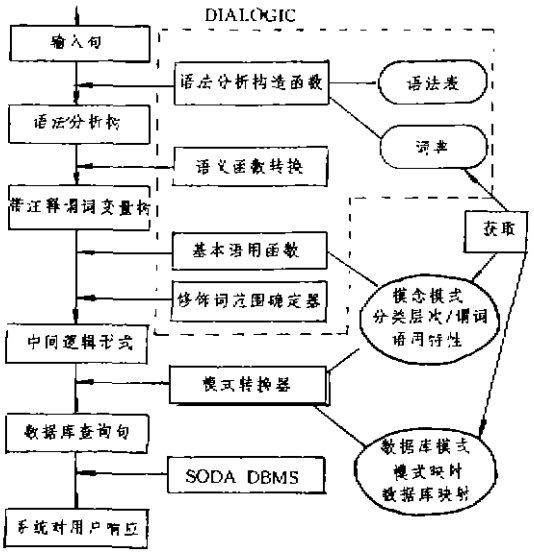


图 2 TEAM 系统结构图

## 3. 数据库汉语自然语言查询技术进展

汉语查询技术进展比较缓慢,主要原因在于尚缺乏汉语语言理解技术的深入研究,近年来国内学者在这方面作了很多有益的探讨,主要工作如下。

### 3.1 受限汉语概念的启示

文[13]提出 DB 上受限汉语的查询概念。即对汉语强加一定的规则而得到的该语言的子集,称之为受限汉语。作者试图在一汉语子集上建立查询模型,达到在一定范围内对汉语语法、语义的理解,提出一种类上下文无关文法,该文法基本上覆盖了数据库常用的查询句型,其意义在于降低汉语理解的难度,同时又为更大范围的汉语语言集的理解以及专家系统等的汉语接口技术的研究打下基础。

3.2 基于 E-R 模型的查询策略

近年来，国内不少学者提出的基于数据库实体-联系的汉语理解模型[12,14,15]，该模型摆脱纯语言学理论的传统框架，将汉语查询句与数据库模型的指称语义及背景知识相结合，通过表层与深层的语义理解达到对汉语查询句的确切理解。文[12]对该模型作了较深入的形式化描述，并设计和实现数据库自然语言查询界面 NLCQI[13]。

我们的汉语关键词理解模型定义成七元组，即

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

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

$P\Leftrightarrow\{P\rightarrow\alpha|(p\in V_N)\wedge\alpha\in(V_T\cup V_N)^*\wedge(\exists i)(S\in P_i\wedge(P_i\rightarrow\alpha))\wedge i\in Nat\}$；

$\delta$ 为汉语组词规则集；$\delta'$ 为深层语义映射规则集；$W_d$ 为汉语背景词典。

文[12]对汉语分成表层和深层语义的理解，表层理解是源汉语句子经汉语句型分析器得到语言单位的表层语义结构树，并将其结果存入汉语句型栈中，称之为中介逻辑形式。再经汉语句型的深层语义理解，结合自动分词后目标区上的语义信息、汉语组词规则库以及深层语义转换集，按原语义结构树深度优先原则，逐层处理汉语句型栈中的词条 W，而生成 SQL 语句。

3.3 类关系代数模板的中间语言模型

鉴于自然语言理解的复杂性，无论对西文还是中文的理解，NLI 均设置一种从源语言到数据库查询的中间逻辑形式，又称中间语言。如文[17]所提出的中文界面设计思想是：

投影——在〈某表〉中，写出〈某些列〉，将答案送入〈某表〉

选择——在〈某表〉中，如果〈某(些)条件〉，写出满足这(些)条件的行，将答案送入〈某表〉

连接——如果〈某(些)列值〉等于〈某表〉的某(些)列值，写出满足这(些)条件的行，将答案送入〈某表〉。

类似地，可定义其它有关的运算，作者将关系代数的运算分解为更基本的本原结构，如

(1){在/从}〈某表〉中

(2)写出〈某些列〉

(3)如果〈某条件〉[{或/与}〈某条件〉]等。

文章认为，可"自然地"用这些中介结构去构造更复杂的查询，并将这些查询的中间形式转换成数据库查询语言。

文[16]所提出的类关系代数表达式概念，将汉语查询句化为一类关系代数的中间逻辑式。如查北京学生的姓名、性别、年龄，形如：

((＄(%1.&1＝北京)(%1)(%2.&2＝%3.&3)(学生))((%3.&3＝%4.&4)姓名(%4.&4＝%5.&5)(性别)(%5.&5＝%6.&6)(年龄)

再将该子式填充化简后得：

＃(姓名，性别，年龄)＄(学生，籍贯＝北京)(学生)

以＃，＄，R 表示选择、投影等所确定的关系，然后将此类中间语言转为 SQL 子句。

采用何种中介逻辑形式更易于设计、实现和高效，有待进一步研究和实践。

3.4 基于理解的智能汉语自然语言查询策略

目前在 DBNLI 的研究中，所应用到的 AI 技术主要有：

(1)知识表达机制。采用规则，如文[13]中汉语查询树生成规则，汉语组词规则库，中间语言 MQL 到 SQL 的映射规则等。

(2)智能自动分词系统的设置

(3)具有静态和动态的学习机制，利用"获取"技术，增强 DBNLI 的可移植性

(4)模糊技术的引入，对模糊词(名，短语)的识别和理解等。

结束语 汉语的 DB 查询技术离实用性的商业系统距离尚远，加强汉语理解技术的研究是研制新一代数据库系统理解的人机界面的关键性课题。为此，需要进一步解决的主要问题有：①完善智能分词系统；②建立有效、实用的汉语理解计算模型；③研究受限汉语的语法、语义表示与分析技术；④可移植性的大规模词典研制；⑤中间逻辑形式的规范与优化研究；⑥"获取"技术、增强学习能力的机制；⑦模糊语义的词、句型的识别、分析、处理等。

致谢 本文是作者在北京大学计算机系访问期间所作的部分工作。在此仅向该系唐世渭、俞士汶、杨冬青等教授的支持和帮助表示感谢。(参考文献共 18 篇略)