自然语言理解及在数据查询中的应用

徐九韵　叶延滨　王新民

石油大学（华东）计算机系
山东省东营市（257062）

摘　要： 自然语言理解作为人工智能的一个重要分支，改善了计算机和用户之间的人机界面，为计算机的普及和进一步推广起着推动作用。本文旨在探讨自然语言理解及在数据查询中的应用，提出了自然语言理解必须和语用环境及理解目的紧密相连的环境与目的驱动原理，建立以条件为中心的语言模型，并以此为基础设计了相应的理解算法。

关键词： 人机界面，自然语言理解，匹配，知识，概念。

Natural Language Understanding and its Application in Data Inquiring

Xu Jiuyun　Ye Yanbin　Wang Xinmin

Department of Computer Science ,
University of Petroleum , china 257062

Abstract: Natrual Language Understanding is an important in Artificial Intelligence, which makes the interface between computer and user more friendly. It will play a great role in computer application. The paper presented an algorithm of Natural Language Understanding. The methods of understanding must connect to pragmatic environment and purposes from this point, we proposed a condition based language mode and relative understanding algorithms.

Keywords: Natural Language Understanding ; human-machining interface ; match ; knowledge; concept.

0　引　言

随着数据库技术在生产、生活等各个领域的广泛应用，人们非常希望有一种更为方便、实用的数据库界面，这种界面不仅可以方便那些缺乏计算机知识的用户，而且还可以大大拓宽数据库的应用领域。纵观计算机几种通用的人机界面，无论是命令语言界面还是图形界面，用户都要经过必要的操作训练，才能正确操作计算机。这不利于计算机的进一步普及。而自然语言界面无疑是最理想的人机界面，用户无须训练，只要用自然

语言就可操作计算机。国内外学者投入大量精力对这一领域进行研究，但由于自然语言的复杂性和随意性，广义的自然语言人机界面的研究收效不大。但对于数据库查询这一特定的应用领域，由于所使用的自然语言句式简单且理解目的单一，即对数据库进行查询，所以实现这种意义的理解是完全可能的。本文旨在探讨一种广义的通用数据库查询界面，提出了基于环境知识的以提取关键词为目的，以模式匹配为手段的理解算法，充分利用环境和理解目的的特点，非常方便地实现对各种库的查询，达到较好的效果。

1 查询语言的结构分析

要实现对自然语言命令的真正理解，就应在考虑到用户的语用目的和具体语言环境的基础上提取相应的查询条件和查询目标，从而形成查询表达式。其中条件信息是用户命令中显式或隐含数据库的信息，目标信息是用户希望查询的域的内容。

一般的查询句有以下形式:

〈查询句〉：：＝{〈条件信息〉||〈目标信息〉||〈干扰噪音〉}*

〈条件信息〉：：＝[〈域名〉+][〈操作符〉+]域值[+环境词]

||[〈域名〉+][〈操作符〉+]相关概念词[+环境词]

〈目标信息〉：：＝〈域名〉

〈操作符〉：：＝〈 | 〉| ＝ | 在。。之上 | 在。。以下 | 比。。高|。。。

说明:

（1）查询句可由若干个条件信息或目标信息组成，其中可能有若干干扰噪声。

（2）条件信息中域名和操作符有时可以缺省，这主要因为域值本身是特定域名的值。

（3）操作符是域值和域名信息的连接符，表示它们之间的关系，从某种意义上讲操作符代表了表达条件信息的句型。

（4）条件信息可以有以下几种特殊形式:

A、仅域值条件: 这是可以由域值直接找到相关域名的条件，如在人事库中，“女”对应的域名就是“性别”。

B、隐含域名条件: 这种条件也是一种仅域值的条件，但其域值可能对应若干个域名，如“1975年出生的职工”，显然“1975年”不会只是出生时间，还可能是工作时间或定级时间，所以对于这类条件，我们通过寻找可以表示其域名的环境词来确定其域名，如“出生”在上句中就是环境词。

C、简单嵌套关系条件: 就是在查询中隐含有相关子查询的条件。如“比王老师年龄大的职工”，这里要达到查询目的首先要查出王老师的年龄。这是一种比较特殊的条件。

我们根据理解的目的分析了许多数据库查询句，建立以条件和目标为中心的句型库。这个句型库可以适用于绝大多数数据库查询语句。

2 数据结构设计

词汇分析及词库设计: 要理解查询句，首先要进行分词处理，即把查询句划分为若干个有意义的词条。我们认为: 脱离语用环境的分词系统不仅效率低下而且往往会造成

错误分词结果，因而提出了基于语用环境知识的分词系统。为了便于实现通用、可靠的分词系统，我们把要提取的词条分为五大类:

1）特殊词汇:即和特定的语用环境有关的术语，这种词或者十分生僻，或者具有特定的含义，在理解过程中往往起着十分重要的地位。如在人事信息管理库中，“编号”，“工资”等词汇都属于特殊词汇，对应着库的字段，是条件信息的重要组成部分。具体地说，特殊词汇是特定数据库的域名域值及其同义词的集合。

2）结构词汇: 即构成句子语法结构的助词、介词或形容词等，如“是”，“以上”，“左右”等等。

3）专有词汇: 是一种不可列的特殊词汇，它们是一种不可列的特殊词汇，如“王刚”，“李鹏”一类的姓名，或是“A—16井”。

4）数字、标点和英文: 在技术工程数据库中，经常出现英文词汇及数字、标点等。

5）相关概念词汇: 这部分词汇是说明隐含信息的词汇。即这些词汇所表示的数据未在数据库中显式表示，但其表示的意义可通过数据库中相关数据经过一定运算获得，它通常也是由术语构成，但不同于特殊词汇，一般不直接对应某个域名的域值，这类数据的查询反映了数据库中数据的综合利用程度，是多个域名或域值共同作用的结果。目前，这方面的工作开展的较少，我们对此进行了深入的讨论，并提出相应的处理算法。

这类词汇一般可分为: 特值词汇（或称定值词汇），这部分词汇通常表示某个域名的值及其在指定条件下的不同含义。例如，在人事库中，人们往往涉及“退休”这一词汇，它不对应人事库中的某个字段或域值，但经过理解之后，它对应“年龄”这一字段中的域值，根据形成退休这一概念的具体条件不同，其对应的域值也不同，通常，男性职工的退休年龄为60岁，女性职工则为55岁，而博士生导师的退休年龄为65岁。这样，就形成了“年龄”这域名的多个含义值。与此相类似的词汇还有: “初级职称”、“处级干部”、“高产井”、“高纬度地区”等等;相关操作词汇: 这部分词汇通常表示一个域名中多个域值或者几个域名中的域值之间的运算结果。它往往含有一定的操作符连接几个不同的域值。如在人事库中，某个职工的“工资总额”这一概念，不直接对应库的字段，而是人事库中有关工资的总和，象“基本工资”、“浮动工资”、“工龄工资”等各项的累加。这样形成多个域名中域值之间的和操作。象这种词汇还有“总和”、“距离”.“厚度”、“差额”、“深度”等等。

以上五种词汇中，结构词汇的数量有限而且词义相对不变，所以可以组成一个结构词库，而特殊词汇由于对不同的库不仅含义不同而且词条也不一样，所以我们把这样的特殊词汇根据不同的库的应用建立不同的特殊库，这样一方面减少了普通词库的容量而且大大增强了该系统的通用性，并为库维护提供了便利，相关概念词汇根据分类的情况分别建立不同的词库，由于这部分词汇数量未知，所以我们将对应的词库设计成开放型的，同时还具有学习功能，以便随着知识的增加，词汇更为丰富。

相关概念词库，包括两个文件: 特值文件: 用来存放特值词汇及相应的域名和域值。

结构: 词条; 对应的域名和域值。其中词条为特值词汇;相关概念文件: 用来存放相关概念词汇和相应的域名和操作符;结构: 词条; 对应的域名，域名之间的操作符，其中词条为相关操作词汇，域名之间的操作符由操作符库来填充。

特殊词库分为四个文件：关键域名文件；关键域值文件；同义域名文件；同义域值文件。其中同义域名文件和同义域值文件存储关键域名和域值的同义词。

四个文件的结构如下：关键域名文件：词条名.该域字段类型；关键域值文件：词条名.对应关键域名的序号；同义域名文件:词条名.对应关键域名的序号；同义域值文件:词条名.对应关键域值的序号；以上四个文件中通过序号相互对应，不仅便于实现词库中词条的一致性，而且大大减少了存储开销。同时同义词的引入大大提高了理解的能力和系统容错的能力。

分词结果链表的数据结构：分词的结果以链表的形式存储起来的，我们称之为词链。词链中结点的结构如下：

```
Struct WTYPE{
    char   * word;                          /*      词条          */
    enum part_of_speech  sw;                /*      词性          */
    char * field_name;                      /*  对应的域名        */
    char * field_value;                     /*  对应的域值        */
    struct WTYPE * next; }                  /*   下一结点指针     */
```

分词结果以词链存储可以方便分词中的多次扫描及结点的修改。

在结点中，当词性是ISVALUE（是域值）或ISFIELD（是域名）时，才在field_name或field_value中填入相应的值；当词性是相关概念词汇时，只作标记；否则为空。

当词条不可识别时，词性为UNKNOWN。

每次扫描实际是对词链中的词性为UNKNOWN的结点进行处理。

句型结构:针对数据库查询命令的特点，我们将语句分为若干种骨干句型，从而将对命令语言的语法分析简化为对其句型的。句型库是用来存放句型及相关信息的知识库。其结构如下：例：序号：F+在+N+L+S，其中，F代表域名，N代表域值，L代表量词，S代表比较集。比较集中包括如“以上”，“中间”，“之外”等词汇，表示大于，小于及等于等量的概念，从而减少了句型库的容量，提高了理解的灵活性。

理解链表的结构:理解过程就是对词链进行多次扫描提取条件信息和目标信息，为便于在多次扫描时对节点进行调整，我们将这些结果以链表的形式存储，并称之为理解链表。其结构如下：

```
struct  Field_Node{
    char  *fieldname1;                  /*       域名          */
    char  interoperate;                 /*  域名间的操作符     */
    struct  Field_Node  *nodelist; }    /*       域值          */
    int     word_order;/*条件信息或目标信息在词链中的对应位置*/
    char    *relation;                  /*      关系符         */
    struct  Known_Node *intersect;      /*    嵌套条件结点     */
    int     OR_AND;                     /*      与或标志       */
    struct  Known_Node  *next; }        /*      下一个结点     */
```

其中，1）当结点是目标条件时，域值和条件关系符和嵌套结点等为空；

2）嵌套指针指向存储嵌套条件的结点；

3）word_order是提取该条件时，该词在词链中对应的起始位置。这主要是因为条件之间本身具有一定的先后顺序关系。

4）与或标志表示条件间的逻辑关系。我们缺省认为条件之间是“与”的关系，即OR_AND=1，当OR_AND=0时表示本结点条件与下一结点条件是“或”的关系。

5）当条件信息中含有相关概念词汇时，查相关概念操作的子程序库，在Field_Node结点中的各个域填相应的值；否则，只填field_name1的内容，其它为空。

理解链表是理解过程的结果，然后系统根据不同的RDBMS系统把链表的内容转化为相应的查询命令表达式。

3　算法设计

分词算法：分词的任务就是根据所定义的四类词汇分别扫描查询命令串提取各种词汇。由于四种词汇各自的特点不同，所以采用了特定的提取顺序和提取方法。

分词算法如下：

1）输入查询命令串；

2）第一次扫描词链，匹配提取特殊词汇；

3）第二次扫描词链，匹配提取结构词汇；

4）第三次扫描词链，根据专有词汇形成规则，判别提取专有词汇；

5）第四次扫描词链，匹配提取相关概念词汇；

6）第五次扫描词链，根据数字、西文的特点，提取数字、西文和标点；

7）分词结束。

［说明］

1。分词的重点是提取特殊词汇，这是因为特殊词汇中往往包含结构词汇，所以同时提取会产生歧义，所以特殊词汇必须首先提取。

2。特殊词汇和结构词汇以及相关概念词汇都是采用最大匹配法，根据词库进行匹配。

3。专有词汇的提取比较特殊，首先要确定语用环境，判断有几种专有词汇，然后根据专有词汇的不同特点设定规则进行判断。

4。由于以上四种词汇中往往有含有数字或西文，所以最后提取数字、标点和西文可以大大减少歧义分割。

专有词汇的提取算法：我们通过总结专有词汇的特征建立相应的判定规则来提取这类词汇。在专有词汇的构成知识库中，一般有专有词汇用的基本字以及组成词汇的启发式规则。我们根据其特征建立了相应的识别规则和特征库（如姓氏库），从而识别绝大部分专有词汇。

理解算法：理解过程在查询语言中实质就是一个形成布尔查询表达式的过程。我们根据建立的条件为中心的查询语言模型，设计了以搜索关键词为主要手段，以提取条件信息和目标信息为目的的理解方法。具体算法如下：

1）读入句型库；

2）按句型框架对词链进行扫描匹配；

3）若句型匹配成功，则根据该句型中的句式信息，依次搜索该句型的F（域名）N（域值）S（比较符）等；

4）如果理解形成的域名！=域值对应的域名，则对应的是嵌套条件，形成嵌套条件结点，否则形成一般查询条件结点；

5）第二次扫描词链，搜索句型匹配不成功的词条；

6）如果该词条词性=ISVALUE，则形成仅域值条件；

如果词性=ISFIELD，则是目标信息；

如果词性=ISDIGITAL，则判断是否是隐含条件，搜索相应的环境词；

如果词性=ISCONCEPT，则形成相关概念条件，调用相关处理程序；

7）理解结束。

4 系统的实现

本系统大体分为三大部分，即分词部分，理解部分和词库维护部分。分词部分功能是把查询命令串分割成有意义的词条；理解部分对以上词条进行扫描，匹配形成查询目标和查询的布尔表达式，并进而生成相应RDBMS的查询命令；词库维护主要是对以上两部分而特殊词库、普通词库以及句型库进行添加、删除及维护。

本系统在微机上运行，使用C语言实现，对人事库、地质勘探库和测井等专业领域库进行试用，理解效果很好。

5 结束语

自然语言界面是最理想的人机界面，本文所描述的数据库自然语言理解界面是该项工作中的成功尝试。与传统命令界面、图形界面相似，该界面更加方便直观，因而它对普及计算机的应用将起到不可估量的作用，不仅如此数据库自然语言理解界面的完善将直接改善数据库的性能，它既是对演绎查询、分布式查询提出的更高的要求，也是解决它的途径。

致谢 该项任务得到石油大学（华东）计算机系许多同志的支持和帮助，在此表示感谢。

参 考 文 献

[1] Peter C.Anick, "Digital Equipument Corporation Integrating Natural Language Processing and Information Retrieval in a Troubleshooting Help Desk",IEEE Expert, Vol 8, No.6 ,DEC,1993

[2]Mary Dee Harris, "Introduction to Natural Language Processing ", Reston Publishing Company,Inc. 1985

[3]刘开瑛，郭炳炎，《自然语言处理》，科学出版社，1991。

[4]顾国良，王能斌，“数据库汉语查询接口的设计与实现”，《计算机学报》，1990，12

[5]张亚南，徐洁磐，“数据库汉语查询接口的EAAD模型”，《计算机学报》，Vol.16,No.12,1993