通用数据库自然语言查询接口
General-purpose Interface of Natural Language for Data Retrieving

石油大学(华东) 计算机系　　王新民　叶延滨

摘　要

本文介绍了一种适应于数据库查询的汉语语言接口。该系统可以接收相当广泛的一类以汉语方式提供的对数据库的查询；它经过抽取输入语句中的查询信息，生成有关的查询操作。本系统不仅可以针对某一特定环境完成数据查询功能，而且还具有根据语用环境进行扩展和裁减的机制，从而使得该系统可以适用不同的应用。经实验表明，该系统具有适应面广，查询功能强等优点，是一种非常受欢迎的产品。

1．引言

随着计算机日益普及以及数据库技术日趋完善，现在已经有越来越多的人利用数据库进行有关日常事物的管理，其中有许多人没有经过计算机训练，实验表明，对于这些人，无论是交互式语言式查询，还是窗口菜单式查询都有不尽人意之处，而最可能受到各阶层人士欢迎的查询工具就是自然语言式查询，本文就是对其中有关问题进行探讨，并介绍了一个通用的数据库自然语言查询接口系统。

2．数据库查询的自然语言一般形式

人类自然语言是文字的奇妙组合，它除了有上下文关系外，还具有各种各样的语义、语气修饰，因而要想全面地了解人类的自然语言目前还有困难，然而就与数据库查询有关的自然语言而言，由于它受到较高的语用环境限制，且有严格的目标语言标准，因此，对数据库的自然语言查询语句进行理解是完全可行的。

就任何一个有意识操纵数据库的人而言，数据库查询的自然语言必然由一系列的条件所构成，其中可以包含一定的连接词，修饰词等，也可以包含一定的无关词语。在所有的条件中，已知的条件属于查询的基础，而未知的条件是查询目标，为便于用近似形式的方法表述自然语言查询语句，我们定义如下符号：S：查询语句；CC：完全条件，即包括域名，域值以及相关符的条件，如“年龄=25岁”等；NC：非完全条件，它或只有域名，或只有域值，分别称之为名条件和值条件，如“年龄多大?”，“45岁的厂长”，等；FN：数据库的域名：FV：数据库的域值。近似的，一个数据库自然语言查询问句可表述为入下形式：S：-(CC|NC)*，CC：-FN op FV，NC：-FN|FV。其中*是条件之间连接符，CC和NC中可能含有若干种噪声，而域名、域值条件可能以隐含形式出现，但其骨架形式必为上述形式。

3．数据库自然语言查询语句理解及实现

根据以上分析，将输入语句滤除噪音后，就成为若干个条件，其中已知条件为查询基础，未知条件是查询目标，根据这个原则，在语法分析中，我们首先根据连接词性质将查询语句分成若干查询词，然后在各查询子句获得查询条件和查询目标，如果在一个句子中没有域各条件，则认为关键字域或所有域为隐含查询目标。　例如：对于“年龄小于45岁的女厂长”可形式化为：(年龄<45) and (性别=女) and (职务=厂长)，在该句中没有查询目标，则可认为关键字姓名域为查询目标。

根据以上分析，我们设计并实现了一种数据库自然语言查询理解系统，它经过分词及滤波后，按上述原则进行条件抽取，然后将抽取的条件表转换成数据库操作，实验表明，该系统能够较好地理解各种自然语言查询。由于在设计中，条件抽取与数据库无关，因而，该系统可以适用各种数据库系统，另外，更改或添加专业知识库可以使该系统适用于各种语用环境。

4．参考文献

1. ALAN F. SMEATON，'Progress in the Application of Natural Language Processing to Information Retrieval Task'.
The Computer Journal, Vol. 35, NO. 3, 1992.