数据库自然语言接口系统的研究

李保利 周锡令 刑光荣 （北京信息工程学院软件工程研究中心 100101）

摘要：作为人机接口的重要研究领域，数据库自然语言接口的研究具有重要的理论价值和巨大的实用价值。本文考察了国内外该领域的研究历史及现状，重点介绍了数据库自然语言接口系统的体系结构和性能评价问题，并指出今后的研究方向。

关键词：自然语言处理 自然语言接口 数据库自然语言接口

一、引言

数据库自然语言接口是人工智能与数据库技术相结合的产物，涉及到人工智能、自然语言处理、数据库系统、人机接口等方面的研究。许多年来，它作为智能接口的重要组成部分引起了广泛的兴趣，成为具有重要理论价值和巨大实用价值的研究领域。

广义上，数据库自然语言接口应当包括数据库设计、数据库定义、操纵（查询、更新）等方面，它旨在为有关数据库的各种操作提供一个自然语言界面。狭义上，数据库自然语言接口仅仅指数据库自然语言查询接口。因为对普通用户而言，查询接口是最为重要的。自然语言查询接口可以使用户直接以日常生活中使用的自然语言提出查询请求，获取数据库中的信息。

与数据库系统本身提供的形式化查询语言（如SQL）相比，用自然语言查询数据库的优势在于：①.用户只以应用领域的概念访问数据库，无需了解数据库的逻辑和存储结构，具有更强的非过程性；②.用户不需要或者只需要很少的培训就能够直接查询数据库信息，大大减轻了用户的培训负担；③.用户可以简单明了地提出查询请求，比如若要在人事信息库中查找年龄最大的人，用形式化查询语言就有些麻烦，不如“年龄最大”或“the oldest”表述简洁。

我们知道，目前所用的人机接口大多是以窗口、菜单为主的图形用户接口 GUI（Graphical User Interfaces）。这种接口简洁、直观，用户只用鼠标点击以及少量的键盘操作就能从数据库中获取所需的信息。但我们会发现有些问题是无法或难以用这种方式表达的，比如带有全称量词的查询请求“找出选修了所有课程的学生”。

二、数据库自然语言接口研究的可行性

自然语言是人类用来传递信息、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，其中存在着大量的歧义性和模糊性现象，是一个“不规则”系统。要从根本上整体上理解处理自然语言目前的理论和技术都还有差距，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。吕光楣、陈清波等人曾经总结了这种可能性[1]：

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.歧义性和上下文相关现象大量减少，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地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、国外的研究情况

国外有关数据库自然语言查询接口的研究可以追溯到本世纪六十年代。早期的代表作是格林的 BASE-

© 中国科学院软件研究所 http://www.c-s-a.org.cn

BALL 系统,这是一个专用数据库接口系统。该系统的数据库中记载着一年内美国全国棒球联赛的各种信息,系统允许用户用限定的英语进行查询。与其他早期的自然语言处理系统一样,BASEBALL 对英语的分析策略主要依赖于关键字匹配技术[2]。

美国的伍兹(W. A. Woods)设计的 LUNAR 系统是七十年代自然语言专用接口的代表。该系统利用英语对美国国家航空和航天管理局提供的一个从月球上采集的岩石标本的数据库进行查询。LUNAR 系统的一个重要特点是对英语的句法和语义做出了比较深入的分析,它是借助于扩充转移网络 ATN 来处理句法问题的第一个程序。

尽管 BASEBALL、LUNAR 和其他一些自然语言专用接口可以出色地完成指定领域的数据库查询任务,但是要把它们移植到其他应用领域非常困难。因此,人们开始探索通用接口的设计。所谓通用接口就是一种用来设计和调试各种自然语言接口的开发工具(或支撑环境),专用接口的建造者利用它可以很快地建立起特定领域的词典、句式和相应的响应式。这样做的好处主要是使新系统的设计开发可以复用已有的成果,从而缩短系统的设计、调试周期,避免大量的重复性劳动。

1978 年美国国际人工智能研究所(SRI)的汉德雷斯(C. Hendrix)等人设计的 LIFER 系统就是一个自然语言通用接口。该系统包括两个主要部分:一组交互式的语言说明函数,用来定义一种面向应用领域的自然语言子集;一个分析程序,对输入的自然语言作出解释,即把输入句子翻译成为可以对特定数据库直接进行查询的命令。这种通过将分析程序与知识库相分离来扩展系统的做法成为建造通用接口的基本思路。值得指出的是,汉德雷斯在描述语言时采用了"语义语法"。这种方法提高了自然语言的处理速度,所以后来被许多实时处理的自然语言系统所采用。在美国利用 LIFER 通用接口已经建立了一批自然语言的专用接口,如美国海军使用的 LADDER 系统。

1983 年首批自然语言接口系统打进了国际市场,标志着一个具有广阔前景的语言产业的崛起。如美国人工智能公司(AIC)率先推出的 Intellect 系统,美国 Frey Associates 的 Themis 系统,美国加利福尼亚工学院的 ASK 系统等。

在轰轰烈烈的研究中,一些学者过分乐观地认为自然语言接口的应用会持续增加,广泛应用的时代即将到来。但不幸的是,80 年代末、90 年代初图形用户接口技术的巨大进步极大地冲击了自然语言接口的研究。因为,图形用户接口解决了许多人们原来期望自然语言接口才能解决的问题;自然语言处理的诸多困难又使 NLI 与 GUI 相比没有优势可言。因此,自那以后,自然语言接口的研究开始受到冷落。这一点从这些年有关 NLI 方面的论文数量上就可以明显看到。

进入九十年代后,尽管自然语言接口方面的研究没有八十年代中期那样轰轰烈烈,但是依然有一大批学者在从事这方面的研究,也有一些试验性或商用的自然语言接口系统出现,如 BBN 公司的 PARLANCE、BIM 公司的 LOQUI、SRI 的 CLARE、微软公司在 SQL Server 6.5/7.0 中提供的 English Query、加拿大 Simon Fraser 大学开发的 SystemX 等。它们除了在系统可用性和可移植性方面有所发展外,在自然语言接口评价、领域知识的自动获取、系统的体系结构以及探索使用新的理论(如 HPSG、人工神经网络、统计与规则相结合)等方面取得了新的进展。

四、数据库汉语接口的研究

汉语数据库接口系统的研究起于本世纪七十年代末期。1980 年中国社会科学院语言研究所的范继淹、徐志敏设计实现的 RJD-80 汉语人机对话系统,成为国内第一个汉语接口实验系统。该系统的处理技术以转换生成语法和扩充转移网络语法为基础。

八十年代初,我国人工智能界的学者开始对汉语人-机接口技术予以重视,不仅设计了一批专用的汉语接口系统,如清华大学陈群秀和赵琦为该校的汽车调度专家系统设计的汉语专用接口 CNLIES 等;而且在 1986 年研制出了第一批汉语通用接口,如清华大学的 SPS 和 ZPS 系统、华中理工大学的 E-RTV 系统、上海工业大学的 LIGC 系统等。此外,汉语的人-机接口系统还被列为国家"七五"科研攻关项目[2]。

综观十几年来数据库汉语查询接口的研究,这些实验系统采用的技术主要有:关键词匹配、句法模式匹配、语义语法、扩充转移网络(ATN)等,主流技术基本上以词汇驱动、句法语义处理一体化为特征,将通用知识库与领域专用知识库相分离,利用学习模块获取领域专用知识,以此达到一定程度的可移植性。总的说来,这方面研

© 中国科学院软件研究所 http://www.c-s-a.org.cn

究的进展缓慢，多数系统只停留在原型系统的水平，未考虑向实用系统转化。

最近几年，越来越多的学者重新认识到汉语查询接口研究的理论意义和应用价值。中国人民大学、香港中文大学和北京大学在国家自然科学基金重点项目支持下，开展了“中文数据库系统及其语言和界面研究”，从查询语言分析、受限处理、界面管理和领域知识自动提取等方面进行了多方位深入探索，已经取得了一定的研究成果【3】。

五、数据库自然语言接口系统的体系结构

软件体系结构是软件工程领域一门新兴的学科。好的体系结构便于系统的维护、扩展和复用。因此，自然语言接口系统的体系结构设计就显得相当重要。

从总体结构上看，现有系统的体系结构分为两类：单层结构和双层结构。单层结构将输入的自然语言直接映射为特定的数据库查询语言。在双层结构的系统中，先对输入进行语言分析，得到中间表示结果，再将其转化为特定数据库的查询语言。显然分层处理的思想使双层结构的系统更具有良好的模块化，便于移植和扩充。

在数据库自然语言接口系统中，最重要的处理部分是语言分析模块。根据语言分析模块中句法分析与语义处理的关系，可以将现有的系统分为三类：紧耦合型系统（tightly coupled systems）、松耦合型系统（loosely coupled systems）和混合型系统（systems with hybrid architectures）。

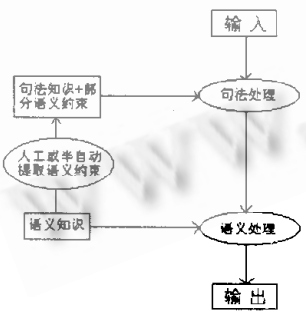

图 1　混合型处理结构

在紧耦合型系统中，句法分析与语义处理过程熔为一体，我们难以将它们分开。SystemX 系统就属于这类系统。这种结构的优点是在分析处理查询语言的过程中可以充分利用句法和语义知识消解输入句中的大量歧义，因而具有较高的性能。但另一方面，由于模块化程度较低，必然增加了系统维护的困难，也使移植到另一个领域时难以直接复用原系统的某些部分。

在松耦合型系统中，句法处理与语义处理串行进行。LUNAR 系统就属于这类系统。这种体系结构的明显优点在于系统的模块化。在将接口移植到另一个领域时，我们有可能直接复用原系统的句法知识及句法分析模块。较高程度的模块化也降低了系统维护的困难，使得不同的人能独立地研究开发句法信息和语义信息及相应的处理模块，加速系统开发过程。这种系统的最大缺点在于它的性能。由于句法处理阶段无法利用语义知识及早进行剪枝处理（Pruning），在某些领域由此引起的组合爆炸现象会严重影响系统的处理效率。

混合型系统体系结构将松耦合型系统的模块化与紧耦合型系统的高性能结合起来，其简化结构如图 1。CLARE 系统就是采用这样的系统结构的。混合型系统体系结构的缺点是需要提取应用领域的语义限制条件（选择性限制条件），而这种约束通常隐含在领域的语义知识中，提取过程需要人的参与，既耗时又会因为人的主观因素产生许多错误。当系统的语义知识发生变化的时候这些特殊的语义限制也必须更新。

六、数据库自然语言接口系统的性能评价

关于软件质量的评价问题在软件工程领域有深入全面的研究，在这里只想谈谈影响数据库自然语言接口系统广泛应用的几个主要因素。

首先是语言的覆盖面问题。尽管数据库查询语言只是自然语言的一个子集，但系统使用的句法规则要想覆盖领域内所有的查询语句也是不可能的。这就像数轴上的一个区间，有限中蕴涵着无限。我们每个人的语言习惯都不尽相同，同一个查询意图往往有多种不同的表达方式。如果一个系统只允许用户使用非常有限的句型来表达查询请求，使用其他句型系统就得不到正确结果，那么这样的系统就很难发挥自然语言接口的优势，也就难以得到用户的普遍接受。这正是目前自然语言接口没有得到广泛应用的一个主要原因。这方面需要解决的语言问题有：全称及存在量词的约束范围、与或歧义、复合名词短语、代词指代、成分省略、否定、短语修饰歧义等。

© 中国科学院软件研究所 http://www.c-s-a.org.cn

其次是系统的可用性问题。这直接决定着用户是否会接受系统,在某些方面还可以弥补系统表达能力的不足。语言的覆盖面直接影响系统的可用性,除此之外,我们主要考虑人机界面的自然化、人性化,需要解决的问题有:必要的提示及引导、及时响应的对话能力、查询结果的再加工(如代码解释、空值 NULL 的处理等)、错误恢复、支持一般性查询问题(如"从系统中可以获取什么信息?"等)、输入句文本检查与更正以及与 GUI 等接口技术的有机融合等。

最后是系统的可移植性问题。这里可移植性主要包括四方面内容,即应用领域可移植性、DBMS 可移植性、自然语言可移植性、硬件和编程语言的可移植性。目前,自然语言的移植尚难以实现,因此我们所关心的是其他三类可移植性。

1. 领域可移植性 即要求系统具有获取新领域知识的能力。因为领域知识千差万别,事先不可能包罗万象。因此只能赋予系统某种适应新环境的能力,以不变应万变。解决领域可移植性问题要从整个系统范围着手,包括系统的体系结构、词库的组织、语义模型、知识获取等[3]。

2. DBMS 可移植性 即要求系统可访问不同的数据库管理系统(DBMS)。这是一个纯数据库问题,解决起来比较容易。由于当前的数据库技术已经很成熟,出现了标准的数据库接口(如 ODBC)和语言(如 SQL)。只要在系统中采纳这些标准,DBMS 上的移植是不难实现的[3]。

3. 硬件和编程语言的可移植性即要求系统能够方便地从一种软硬件平台移植到另一种平台上。早期的自然语言接口系统一般使用特殊的编程语言在大型机(mainframe)上实现,移植到其他平台时非常困难。后来微型机的出现,以及通用的人工智能语言 Prolog、Lisp 和跨平台的 Java 语言的出现,为这种可移植性的实现提供了可能。

七、进一步的研究方向

歧义现象在自然语言中普遍存在,它给自然语言处理带来很多困难。但是与其他系统相比,在人机接口中,我们有一个比较自然的消解歧义的办法,那就是人机交互。我们知道,在人与人会话过程中也时常会有歧义(或误解)发生,它们往往可以通过进一步交流来消除。如果自然语言接口系统也能够不懂就问,那么歧义现象的处理就不再成为难题。所以,从只能处理单句的自然语言接口系统向具有推理能力、篇章理解能力的对话系统转移,是目前自然语言接口研究的一个重要方向。

前面说过,自然语言接口(NLI)与图形用户接口(GUI)以及形式化查询接口各有优势。对某些问题用自然语言进行查询简单明了,而对另外一些问题使用图形用户接口则更为方便、直观。另外,不同的用户使用计算机的习惯以及熟练程度都不一样。因此,数据库接口系统应当提供多种交互方式,允许用户使用自然语言(文本及语音)、图形菜单以及形式化查询语言进行查询,以满足不同用户的需要,提高系统的可用性,成为融合文本(NLI)、图形(GUI)、语音等多种模式、多种媒体的人机接口系统。

三十多年来,数据库自然语言接口(NLIDB)方面的研究取得了很大进展,但迄今为止它还处于研究试验阶段,还没有能够广泛地推广应用,其中还有许多技术问题需要进一步研究解决。但是我们有理由相信,随着自然语言处理技术的不断进步以及自然语言手写和语音输入技术的日益成熟,数据库自然语言接口系统必然会得到广泛应用,极大地改变人们的生活。

参考文献

[1] 吕光楣 陈清波,关系数据库汉语查询接口的设计与实现,中文信息学报,1991,4

[2] 黄昌宁,汉语人机接口的现状与展望,中国计算机用户,1988,5

[3] 孟小峰 王珊,中文数据库自然语言界面研究,计算机世界报,1998 年第 34 期 D8

[4] 许龙飞 唐世渭,数据库汉语自然语言查询界面 NLCQI 的设计和实现,小型微型计算机系统,第 19 卷第 7 期,1998

(来稿时间:1999 年 7 月)

© 中国科学院软件研究所 http://www.c-s-a.org.cn