数据库汉语查询句中查询目标信息的研究*)

郑逢斌　葛　强　汤赛丽　党兰学
(河南大学计算机与信息工程学院数据与知识工程研究所　开封 475001)

摘　要　在自然语言处理的研究中,最有希望的应用领域之一是自然语言接口。计算机对自然语言中的查询语言理解的正确程度是自然语言接口质量好坏的关键。通过对汉语查询句——即用于数据库自然语言查询的祈使句和特指问句、是非问句、正反问句、选择问句等共五类句型的研究,建立并研究了复合概念、逻辑概念与标准概念的内在联系,将查询目标概念分解为直接查询目标、逻辑推理目标和比较判断目标三个阶段,研究了这三个阶段的关系。

关键词　自然语言处理,人机接口,查询目标

A Study of the Query Aim Information of the Chinese Query Sentence in Database

ZHENG Feng-Bin　GE Qiang　TANG Sai-Li　DANG Lan-Xue
(Institute of Data and Knowledge Enginerring, Computer and Information Engineering College, Henan University, Kaifeng 475001)

Abstract　Natural language interface is one of the most hopeful fields in the research on Natural Language Processing. Accuracy of computer understanding query of natural language is key to quality of the natural language interface. Through the study of the Chinese query sentence, the Chinese query sentence is consist of the imperative sentence and special question, the yes-or-no question ,the positive and negative question , choose question etc. Build and study the relation of composing conception , logical conception , standard conception. The conception of the query aim is decomposed into three phases, there are direct query aim, logic discursion aim and compare judge aim, the relation of the three phases has been studied .

Keywords　Natural language processing, Man-machine interface, Query aim

1　引言

本文所说的汉语查询句是指用于数据库自然语言查询的疑问句和祈使句两种类型,疑问句分成特指问句、是非问句、正反问句、选择问句四类。计算机理解汉语查询句正确的程度是数据库自然语言接口好坏的关键。所谓自然语言接口就是允许人们用某种自然语言的子集在限制领域内同计算机进行通讯。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、数据库技术、人机接口等方面的研究[2,3]。狭义上的数据库自然语言接口仅仅指数据库自然语言查询接口。

计算机处理汉语自然语言是多学科的研究工作,但我国语言学界多数着重汉语教学方面的研究,并且研究的成果是"给人看"的,而不是"给计算机看"的。成果是"给人看"时,研究成果很难形式化,不适合直接用计算机处理,有些成果根本无法形式化,它们对计算机处理用处不大。成果"给计算机看"时是形式化的,可以用软件逐步实现[1]。下文是作者从程序实现的角度出发对汉语数据库查询句进行了深入研究后关于查询目标部分的结论。

2　查询目标分类

作为以查询数据库内容为目的的自然查询语句,主要包括查询实体、查询条件、查询目标等方面的信息以及语句表达上需要而与查询内容无关的干扰噪音。自然查询语言处理中,最主要的也是系统最关心的就是要分清查询目标与查询条件,一般的查询句有以下形式[4~6]:

(查询句)::=(〈查询实体〉||〈条件信息〉||〈目标信息〉||〈干扰噪音〉)*

(查询实体)::=〈实体名〉或间接〈实体名〉

(条件信息)::=[(域名)十][(关系符)十]域值[十环境词]

(目标信息)::=〈域名〉或间接〈域名〉

(关系符)::={=,≠,≤,≥,<,>,∈}

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)[…]表示其内容或者没有或者一次;(…)表示其内容可一次;(…)*表示其内容可重复零次或多次。

定义 1　本文中标准概念是指收录在理解自然语言所使用的各个知识库中的概念。与它同义的其它词语称为非标准概念。

定义 2　本文中的复合概念是指能分解为多个标准概念或分解为一个标准概念与常量的算术运算式的概念。如"中级职称"、"明年"。

一个复合概念对应多个标准概念,它是多个标准概念算术运算或逻辑运算的结果;当一个概念解释为唯一的一个标准概念时它就是标准概念同义词了。

*)基金项目:河南省科技攻关(No. 0424220146),河南大学重点理工科项目(No. 04ZDZR001)。郑逢斌　博士,副教授,主要研究方向为自然语言理解,软件工程。

复合概念的含义在本系统中是固定的，如果在不同的条件下含义不同属于逻辑概念。

定义 3 本文中的逻辑概念是指在不同的条件下有不同的含义(或值)的概念。如："退休年龄"。

在本文中逻辑概念用产生式表示，存储分为静态存储和动态存储两种形式，静态存储可用二维关系表表示，动态存储用二级链表表示。所谓静态存储是指系统处于未运行时的存储状态，动态存储是指系统启动运行中的存储状态。系统在启动时自动将静态存储状态的知识转化为动态存储状态，系统使用逻辑概念只对处于动态存储状态的知识进行操作。求逻辑概念的值时要进行逻辑推演才可以得知。

概念等价变换包括概念分解等价变换和概念合成等价变换。

定义 4 概念分解等价变换是指用复合概念知识库将用户输入自然语言句子中复合概念逐步变换为只剩下域名和逻辑概念组成的复合式的过程。

在复合直接查询目标转换为简单直接查询目标、间接逻辑推理目标转换为直接逻辑推理目标、带复合概念的查询条件处理过程中都涉及概念分解等价变换。

定义 5 概念合成等价变换是概念分解等价变换的逆变换，指利用复合概念分解等价变换链表将域名和逻辑概念的值逐步求出复合概念的值的过程。

定义 6 查询目标是指查询句要查询的并需要推演运算的结果。

查询目标分为直接查询目标、逻辑推理目标、比较判断目标三个阶段目标。要得到这些目标有时需要对概念进行等价变换和逻辑推理。

定义 7 直接查询目标是指通过知识库的计算机操作命令直接对知识库进行操作得到的数据，或这些数据经过概念合成等价变换的结果。

直接查询目标分简单直接查询目标和复合直接查询目标，简单直接查询目标是指只包含域名和聚集函数的目标；复合直接查询目标是指由简单直接查询目标经过概念合成等价变换的目标。

用自然语言描述直接查询目标可以有多种表达方式，用户不仅可以直接指定所要查询的目标，也可以用疑问代词指代所要查询的目标，另外，根据用户的需求，在查询目标中还可以出现聚集函数。据此，本文又可以将简单直接查询目标分为显性目标、疑问目标和聚集目标三类。

定义 8 显性目标是指用域名直接给出的目标。

定义 9 疑问目标是指用疑问词给出的目标。

定义 10 聚集目标是指用聚集函数给出的目标。

在表达上，当将自然查询语言中的每一个词转换为知识库内部表示后，显性目标和疑问目标都是与知识库某些域名相对应，而聚集目标则与一个聚集函数(如：COUNT、SUM、AVG、MAX、MIN)相对应。

定义 11 数据提取目标是指简单直接查询目标中能转换为同一个计算机命令语句的那些目标。计算机实现数据提取目标的过程称为数据提取过程。

每个数据提取目标最终转换为一个 SQL 语句，简单直接查询目标包括一个或多个数据提取目标，因此简单直接查询目标应转换成一个或多个 SQL 语句。这几个数据提取目标由数据提取过程来实现，它们是并列的。如：张三与李四谁先退休？通过一系列变换得到简单直接查询目标为：张三的出生日期、性别、职业、职称和李四的出生日期、性别、职业、职称，它对应如下两个数据提取目标：

数据提取目标 1：张三的出生日期，性别，职业，职称。

数据提取目标 2：李四的出生日期，性别，职业，职称。

定义 12 数据提取子目标是指数据提取目标中每一个相互独立的最小的数据目标；如："出生日期"、"性别"。

数据提取目标＝提取子目标 1＋提取子目标 2＋……＋提取子目标 $n(n\geqslant 1)$。数据提取目标在分解为数据提取子目标时有如下几种情况：

(1)提取子目标 1、提取子目标 2、…、提取子目标 n 为同一个知识库的域名。

(2)提取子目标 1、提取子目标 2、…、提取子目标 n 为多个不同知识库的域名。

(3)提取目标 $i(i=1,2,\cdots,n)$ 为知识库的域名与常量的算术运算符(＋，－，＊，/)组成的简单的算术运算式。

(4)$n=1$，即只有一个提取子目标，且为集函数的函数值。

定义 13 逻辑推理目标是指自然查询句中包含的在应用领域逻辑概念推理知识库的产生式结论中出现的知识目标，或者经过概念分解等价变换最终转换为逻辑概念推理知识库的产生式结论中出现的知识目标的目标。

逻辑推理目标分为直接逻辑推理目标和间接逻辑推理目标。直接逻辑推理目标是指知识库逻辑概念推理知识库的产生式结论中出现的知识目标，如"张三的退休年龄是多少？"例句中"退休年龄"；间接逻辑推理目标是指经过概念分解等价变换最终变换为逻辑概念推理知识库的产生式结论中出现的知识目标的目标，如"张三的退休日期是何时？"例句中"退休日期"(退休日期＝出生日期＋退休年龄)。

定义 14 逻辑推理目标化解推演变换是指把直接逻辑推理目标经过应用领域逻辑概念分解知识库中的产生式逆向(即有结论到前提)推理转换为直接查询目标的过程。

定义 15 逻辑推理目标求值推演变换是指把直接查询目标所得的知识经过应用领域逻辑概念分解知识库中的产生式推理转换为直接逻辑推理目标的过程。

逻辑推理目标的结果本文用直接逻辑推理目标队列表示，间接逻辑推理目标到直接逻辑推理目标的转换用概念等价变换链表表示。

定义 16 比较判断目标是指根据自然查询句的要求，对直接查询目标或逻辑推理目标进行比较，根据比较结果产生回答的内容，这个结果就是比较判断目标。

比较的类型要考虑的因素有：句型、比较元素个数及比较关系、比较元素值的来源、比较结果类型等[7,8]。

从句型来看，特殊问句和祈使句没有比较判断目标；是非、选择和正反问句的比较判断目标为直接查询目标或逻辑推理目标的比较值。

从比较元素个数及比较关系来看又分为：两个元素比较(比较关系有：＝，≠，≤，≥，＜，＞，—)；多个元素(比较关系有：MAX，MIN，ORDER)；一个元素与一个集合比较(比较关系有：∈)等。

从元素值的来源来看又分为：查询所得数据和查询句原带数据。如：张三是北京人还是上海人？张三和李四是不是一年出生的？

从比较数据的结果类型来看又分为：逻辑值(真，假)，数据值。如：张三是南阳人吗？张三比李四大多少？

比较结果用比较判断目标(一对多)链表来表示。

(下转第 103 页)

for XML. In: Proceedings of the 8th International World Wide Web Conference, May 1999. 77～91
5 Bruno N, Koudas N, Srivastava D. Holistic Twig Joins: Optimal XML Pattern Matching. In: ACM SIGMOD, June 2002
6 Wang H X, Park S, Fan W, et al. ViST: A Dynamic Index Method for Querying XML Data by Tree Structures. In: ACM SIGMOD, June 2003
7 McCreight E M. A space-economical suffix tree construction algorithm. Journal of the ACM, 1976, 23(2): 262～272
8 Shasha D, Wang J T L, Giugno R. Algorithmics and Applications of Tree and Graph Searching. In: ACM Symposium on Principles of Database Systems(PODS), May 2002. 39～52
9 van Leeuwen J. Algorithms for finding patterns in strings. In: Handbook of Theoretical Computer Science. Vol A, Algorithms and complexity. Chapter 5. Elsevier, Amsterdam, 1990. 255～300
10 Sleepycat Software. http://www.sleepycat.com. The Berkeley Database (Berkeley DB)
11 Ley M. DBLP database web site. http://www.informatik.uni-trier.de/ ley/db, 2004
12 XMARK: The XML-benchmark project. http://monetdb.cwi.nl/ xml, 2004

(上接第 91 页)

定义 17 回答用户目标是指自然查询句要求系统回答的内容。回答用户目标可能是比较判断目标，也可能是逻辑推理目标或者直接查询目标。

回答用户目标用回答用户目标队列来表示。

3 查询目标关系

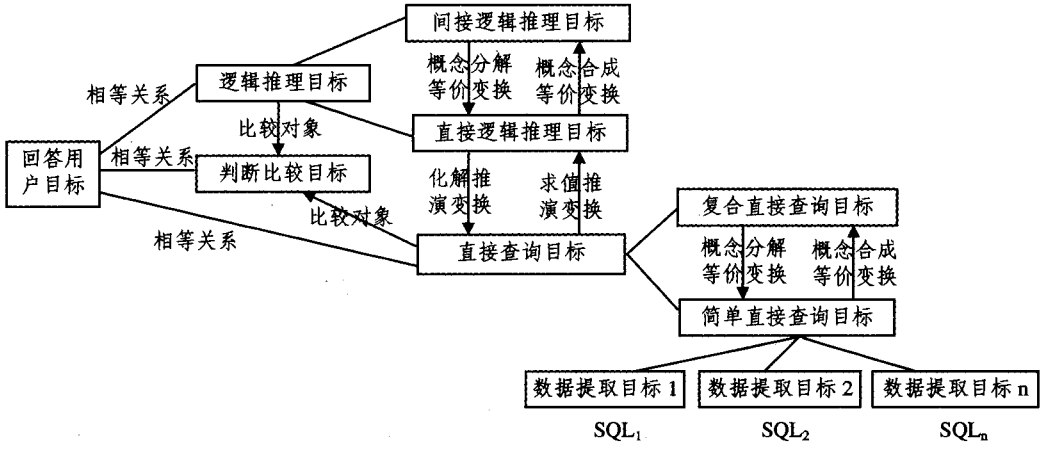

图 1 查询目标关系图

查询目标中各阶段目标的关系如图 1 所示。在实际中，许多查询句的查询目标只包括三个阶段中的部分阶段，分如下几种情况：

(1)当回答用户目标＝直接查询目标时，直接查询目标的内容就是回答用户目标的值。如：请说出张三的籍贯(祈使句)，哪里是张三的老家？(特殊疑问句)。

(2)当回答用户目标＝逻辑推理目标时，逻辑推理目标的内容就是回答用户目标的值。如：请说出张三的退休日期(祈使句)，张三哪年退休？(特殊疑问句)。逻辑概念“退休日期”的值就是回答用户目标的值。

(3)当回答用户目标＝判断比较目、只有一个比较者和一个被比较者(是非问句或正反问句)、比较者是直接查询目标或逻辑推理目标、被比较者为常量或直接查询目标或逻辑推理目标时，则回答用户目标的值是二者比较结果的逻辑值。如：张三是副教授吗？张三与李四的职称一样吗？张三与李四的退休年龄一样吗？

(4)当回答用户目标＝判断比较目、只有一个比较者和多个被比较者(选择问句)、比较者是直接查询目标或逻辑推理目标、被比较者均为常量时，则回答用户目标的值是比较结果为真的对应的被比较者。如：张三是讲师还是副教授？

(5)当回答用户目标＝判断比较目、只有一个比较者和多个被比较者(选择问句)、比较者是直接查询目标或逻辑推理目标、被比较者均为直接查询目标或逻辑推理目标时，则回答用户目标的值是比较结果为真的对应的被比较者概念对应的查询条件块。如：张三的职称是与李四一样还是与王五一样？

结论 本文深入研究了汉语查询句中查询目标信息，用祈使句或特殊疑问句查询时，回答用户目标一般等于直接查询目标或逻辑推理目标；用是非问句、正反问句、选择问句等方式查询时，回答用户目标一般等于判断比较目标。这些信息作者都设计了存储表示结构，可已用计算机软件来识别和转换，识别和转换的算法将于另文讨论。

参 考 文 献

1 郑逢斌. 计算机理解自然查询语言的研究与实现[D]:[西南交通大学博士研究生学位论文]. 2004
2 孟小峰，王珊. 中文数据库自然语言查询系统 Nchiql 设计与实现[J]. 计算机研究与发展，2001，38(9)：1080～1086
3 王英姿，宗成庆，陈肇雄，黄河燕. ITS 系统中自然语言人机接口的设计与实现[J]. 计算机研究与发展，1998，35(9)：814～818
4 许龙飞，杨晓昀，唐世渭. 基于受限汉语的数据库自然语言接口技术研究[J]. 软件学报，2002，13(4)：537～544
5 许龙飞，唐世渭. 数据库汉语自然语言查询模型研究[J]. 计算机科学，1999，26(8)：43～46
6 许龙飞. 数据库自然语言查询技术研究[J]. 计算机科学，1997，24(5)：50～54
7 卞世力，姚天顺，金鸿. 一个中间语言生成目标语言的原理和方法[J]. 软件学报，1994，5(9)：1～8
8 李保利，周锡令. 数据库自然语言接口系统的研究[J]. 计算机系统应用，1999(12)：31～34