文章编号:1007-1423(2019)22-0038-05 DOI:10.3969/j.issn.1007-1423.2019.22.008

数据库汉语自然语言查询技术研究

王代星

(贵州大学教育教学评估中心、高等教育研究所,贵阳 550025)

摘要:
从数据库汉语自然语言查询的目的和对象出发,研究利用数据库预查询技术,结合数据库语义分析技术、自然语言句子结构分析技术,来分析处理自然语言查询的方法。明确提出查询要素切分、条件值归属模糊、探测查询、超关系等概念,使自然语言查询技术研究概念清楚、目的明确、方法简单易于实现,具有很强的可移植性和通用性。

关键词:
自然语言查询;汉语自然语言处理;数据库查询;SQL 语言

0 引言

计算机人机交互界面一直在朝着简单、易用、智能、人性化方向发展,数据库自然语言查询技术也正是顺应这一趋势,研究使用自然语言查询数据库的方法,从而扩大数据库用户群体,方便广大用户使用,而无需掌握数据库专业技术知识,摆脱数据库形式化查询语言的桎梏。本文着重从数据库角度出发,探讨了数据库汉语自然语言查询技术的实现(以下简称自然语言查询)。

1 国内外研究现状

计算机自然语言处理 NLP(Natural Language Processing)早在上个世纪 60 年代,国外就展开了研究,并在机器翻译领域获得了成功。数据库自然语言查询技术的研究也伴随着自然语言处理而展开,在 80 年代进入高潮,前前后后开发了许多具有代表性的系统。如:60 年代美国 B. Green 开发的基于关键字匹配技术的 BASEBALL 系统,允许用户用限定的英语句子查询数据库内记录的美国全国棒球联赛信息;1978 年美国国际人工智能研究所(SRI)C. Hendrix 等人设计的 LIFER 系统,通过将分析程序与知识库相分离的做法,设计出了自然语言查询通用接口,在它的基础上,美国成功地建立了一批专用接口;80 年代,美国人工智能公司(AIC)推出 Intellect 英语人机接口系统,Frey Associates 公司推出 Themis 人机接口系统,加利福尼亚工学院推出 ASK 系统,日本日立公司推出 HICALTS 英日、日英翻译系统等,标志着语言产业的形成,研究的重点也向通用系统转移;在 80 年代末,90 年代初,由于受到图形用户界面技术的冲击,数据库自然语言查询研究跌入低谷,之后朝着两个方向发展,一是将前期的系统引入实用阶段,二是不断地探索新的理论和方法,引入多模式界面和人工神经网络等技术。

我国于上世纪 70 年代末 80 年代初开始汉语自然语言数据库接口系统的研究,在借鉴国外研究成果的基础上,根据汉语的特点,设计了一批专用接口和通用接口,主要采用关键词匹配、句法模式匹配、语义语法、扩充转移网络(ATN)等技术。主流实现方式有基于数据库 E-R 汉语理解模型、类关系代数逻辑式中间语言转换、以条件为中心的句型匹配以及多语句组合模板等方法,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、通用性。但从整体来说,进展缓慢,多数系统只停留在原型系统水平,未考虑向实用系统转化。

2 基本术语约定

从实体联系模型出发,数据库逻辑模式与概念模

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

式具有较为直接的对应，利用图 1 所示数据库语义词典，可以方便地进行转换，因此，在后文提到关系、属性以及查询要素时，不再严格区分逻辑模式和概念模式。为方便讨论，将文中用到的部分术语在此作简略解释。

目标属性：自然语言查询结果所涉及的数据库关系属性。即查询结果是满足查询条件的目标属性值的子集。

条件属性：自然语言查询中对查询结果的限定条件涉及的数据库关系属性。

条件值：自然语言查询条件部分的具体限定值。分字符串型、数值型和日期型三类。

查询要素：指一条自然语言查询包含的目标属性、条件属性、条件值以及排序属性、分组属性等要素。

数据库语义词典：数据库逻辑模式与概念模式映射工具。主要用于自然语言查询要素切分、SQL 语言转换。简称语义词典。

通用词典：包括标点、介词、连词、查询词、是词、有词、聚集词、比较词、数词、量词、疑问词等，涉及自然语言查询的通用词汇，它们对句子的结构分析、查询要素之间的关联分析具有特殊的意义，对不同类型的词汇需要作相应的处理。

查询要素切分：利用数据库语义词典和通用词典，采用正向最大匹配或反向最大匹配方法从自然语言查询句子中切分出查询要素，必要时预先进行数据库探测查询。要素切分的同时也完成了数据库概念模式与逻辑模式的转换。

超关系：将目标属性所在的基本关系，以及从该关系出发通过主键和外键两两关联的所有基本关系连成一个虚拟的大关系，称之为一个超关系。

条件值归属模糊：指自然语言查询中未指明条件属性的条件值可能归属超关系的几个属性的现象。系统需要采用探测查询等技术才能消除这种模糊。

探测查询：即数据库预查询。指在查询要素切分过程中，对条件值可能归属的超关系属性进行预查确认，消除归属模糊和排除领域动词等无关词汇。

3 数据库语义分析

自然语言查询有两种实现方案：一是对数据库管理系统进行扩充；二是在数据库管理系统之上开发应用接口。两种方案都必须建立数据库逻辑模式与概念模式的映射。本文采用的是第二种方案，通过建立如图 1 所示的数据库语义词典，完成模式转换。该词典从具体数据库抽取出来而独立于数据库存在，一般与分析处理程序一起放入 Web 服务器中，以实现多服务器、多数据库的访问。考虑到自然语言词汇的丰富性和用户用词的个性，词典中需要加入大量的同义词。同时还需要包括许多辅助信息，例如：属性的类型、域、量词、单位；关系的主外键约束；超关系；数据库服务器的连接方式等。词典采用树型结构，这种结构与 XML 文档结构非常相似，用 XML 文档词典实现平台无关性。语义词典的建立过程如下：

（1）从数据库的词典中自动提取逻辑模式。

（2）从系统 ER 模型、需求分析文档中的数据词典、系统说明书等提取概念模式、同义词。需要人工参与，由数据库管理人员或系统开发人员手工添加。

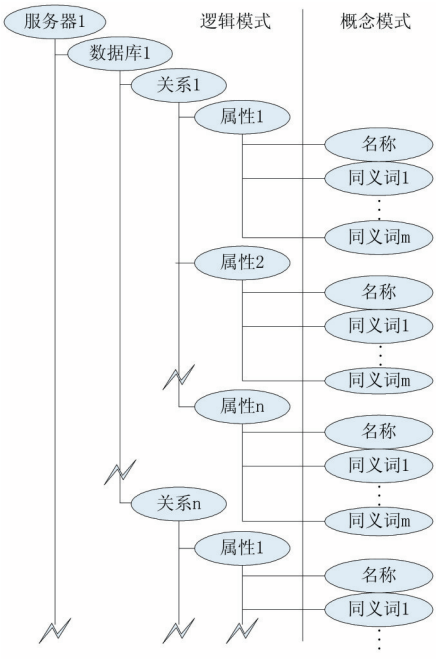

图 1 数据库语义词典

4 汉语自然语言查询的语言特征

表示查询的自然语言有祈使句和疑问句。祈使句只针对数据库的内容，而疑问句则分两种情况，一种是对数据库内容提问，另一种是基于数据库内容进行推

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

理和判断性要求提问。疑问句的后一种情形涉及人工智能领域的研究,需要知识库的支持,本文不作讨论。在现实中,人们的查询请求基本上都是比较简洁的单句,可简化为短语结构,如:

例 1 查询数据库的课程号和学分

简化:数据库的课程号和学分

例 2 张三住什么地方? 或:张三的家庭地址在哪里?

简化:张三的家庭地址

例 3 查询学号为 98001 的学生姓名、性别

简化:学号为 98001 的学生姓名、性别

可编程实现这种简化,因此本文只针对这种短语结构进行讨论。查询要素在自然语言查询短语中主要有如下规律:

(1)目标属性(组)名称前一般都有关系名修饰,或实体关系的名称属性的某个值限定,如:

例 4 学生的姓名、年龄

其中目标属性组“姓名、年龄”由其实体关系“学生”修饰。

例 5 张三的性别、年龄

其中目标属性组“性别、年龄”由实体关系的名称属性“姓名”的值“张三”修饰。

(2)当条件值前无属性名称修饰时,一般都是名称类属性的值。如例 5 中的“张三”。

(3)实体关系名经常单独出现,其后无属性跟随。取其默认属性组为目标属性,如:

例 6 张三选修的课程

其中“课程”是实体关系,包含“编号、名称、学分、先修课”等属性。可以为其指定一组默认属性。

(4)属性名后无是词、比较词等与条件值关联时,为目标属性。如例 4、例 5。

(5)属性名之后有是词、比较词等与条件值关联时,属性和条件值组合成查询条件,如:

例 7 学号为 95001 的学生姓名

由此可知,虽然自然语言很不规范,但仅就表达查询这一有限的自然语言集合来说,其用词是有限的、句子结构是有规律可循的,各查询要素之间是有一定的固定搭配的。综合运用这些结构信息,是自然语言查询处理的依据之一。

5 自然语言查询要素切分

查询要素切分不同于分词概念。分词技术必须尽可能准确地、彻底地把句子切分成语言的最小组成单位“词”,而自然语言查询分析只需要切分出查询要素即可。例如“家庭地址”,在数据库中它是一个独立的概念,而不用细分为“家庭”和“地址”。对条件值的切分区别更为突出,例如公司名称“联华科技责任有限公司”、书名“高级数据库技术与应用”等,分词结果则显得画蛇添足。因此,查询要素切分的概念更适合于自然语言查询处理。

查询要素切分使用的两种汉字串切分方法:正向最大匹配法和反向最大匹配法,同时也是自然语言三种常用分词技术中的两种方法[1]。查询要素切分交替使用正向和反向最大匹配方法,有利于效率的提高。每一轮匹配,当语义词典匹配、通用词典匹配、探测查询匹配都失败时,才考虑舍弃一个字,再进行剩余字串的匹配。

算法 5.1 查询要素切分

输入:自然语言查询字串、语义词典、通用词典

输出:目标属性、查询条件(条件属性 = 条件值)

1. 采用反向最大匹配法或正向最大匹配法或交替使用这两种方法,查询数据库语义词典,切分出关系、属性,同时完成模式转换,确定所属超关系。

2. 采用同样的方法,查询通用词典,切分出常用词汇,结合第 1 步的结果,判断目标属性、条件属性、以及与条件属性关联的条件值。

3. 采用同样的方法,在超关系中进行探测查询,消除条件值归属模糊。

4. 若剩余的字串不空,重复上述过程,直到空串。

5. 输出目标属性、查询条件。

6 探测查询

目前对条件值归属模糊或整个自然语言查询的处理主要有以下几种方法:

(1)规范查询用语。对查询用的自然语言进行一定的限制,要求用户使用规范的句型格式。优点是简单、易于实现,缺点是限制太多,要求用户熟悉数据库的概念模式。

(2)人机交互确认方式。对未登录词、专有名词等系统无法解释的词汇,作出几种可能的推测,由用户进

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

一步选择确认。优点是增强了人机互动，提高了分析处理的精确度，具有自学习功能。缺点是用户必须熟悉数据库的概念模式，必须清楚地知道他要查询的内容属于哪一个实体或联系的哪一个属性。

(3)句型模式匹配方式。分析、统计数据库中关系与关系之间、关系与属性之间、属性与属性之间、值与值之间、值与属性之间可能存在的修饰关系，与自然语言句子结构结合起来，归纳出若干特定的句型，然后将实际的查询句子与这些句型模式匹配，取相似度最高的句型作为实际问题的解。这种方法查询效率高，但实现起来麻烦，要归纳出一个庞大的数据库的所有句型模式是非常困难的，普通的数据库技术人员很难胜任这项工作，系统缺乏可移植性。

实际上数据库内既然包含了我们想要知道的内容，就完全可以利用这些内容来帮助我们分析查询。探测查询正是基于这样的思想。首先，数据库内的数据是有组织、有结构的，作为一个整体，反映现实世界某个领域的客观对象的信息，查询也会紧紧围绕这些信息进行，这种紧密相关性正好映射一个超关系的概念。其次，数据库内的数据类型分成三类：数值型、日期型、字符串型，前两种数据类型一般与“年月日”或量词同时出现，而字符串型数据，在隐含条件属性的情况下，都是现实客观对象的名称。第三，并行计算技术可以同时实现多个属性的探测，提高查询速度。最后，探测查询不向用户返回查询结果，不占用网络带宽。总之，对条件值归属模糊，在超关系内按属性分类进行探测，是行之有效的。例如：

例 8 张三的家庭住址

例 9 工程项目管理的学分

例 10 张三的工程项目管理的考试成绩

分析例 8：假设数据库内有学生、教师、课程三个实体关系，选课、授课两个联系关系，学生实体通过选课与课程实体关联，教师实体通过授课与课程实体关联，这五个关系构成一个超关系 R。首先，通过反向最大匹配，查找语义词典，得知“家庭住址”是属性 student.address，同时由 student 关系确定超关系 R；其次，查找通用词典，匹配出“的”字，得知前面的“张三”修饰 student.address 属性；第三，查找语义词典和通用词典，无法匹配“张三”，转而进行探测查询；第四，“张三”属字符串型数据，是客观对象的名称，而超关系 R 中含有三个实体名称型属性：student.sname, teacher.tname, course.cname，且句中没有明确指出“张三”到底是学生、教师还是课程名称，因此产生条件值归属模糊，需要分别对这三个属性进行探测查询，依次或并行地执行下面三条 SQL 查询语句：

(1)select * from student where sname='张三'

(2)select * from teacher where tname='张三'

(3)select * from course where cname='张三'

最后，根据探测查询的结果，确定“张三”到底归属于哪一个属性，并组合成查询条件。探测查询的输出结果可能有 4 种：①student.sname='张三'；②teacher.tname='张三'；③course.cname='张三'；④FALSE(无满足条件的记录)。例 9、例 10 的分析类似。

探测查询法立足于数据库本身的内容，解决条件值归属模糊，从而简化了自然语言查询的分析处理。缺点就是在采用最大匹配方法从自然语言句子中切分出条件值时，在匹配过程中，可能需要进行多次探测查询，从而占用过多的数据库资源。

7 数据库自然语言查询系统体系结构

综上所述，得出如图 2 所示自然语言查询系统体系结构。自然语言字串经过要素切分、探测查询后，已经由自然概念转换成了数据库逻辑模式，确定了超关系、目标属性、查询条件，再经 SQL 转换模块组合成完整的 SQL 语句，最后交底层数据库管理系统执行，并以 XML 文档格式向用户返回查询结果。

SQL 转换模块主要有两个功能：首先是超关系的简化。超关系中存在很多冗余的基本关系，需要根据查询要素，筛选出实体关系，再考察各实体关系之间是否需要联系关系连接，从而确定 FROM 子句和连接条件；其次将所有的查询要素分别装配成 SQL 的子句，即 SELECT 子句、WHERE 子句(可能还有 GROUP BY、ORDER BY 子句)，然后将它们组合成完整的 SQL 语句。

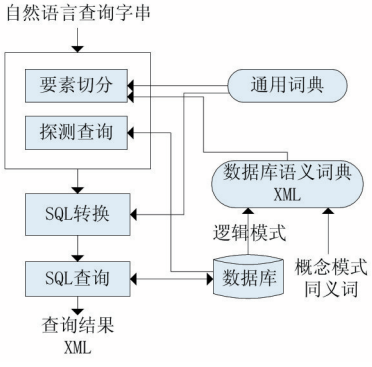

图2 自然语言查询系统体系结构

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

系统将通用词典、语义词典与分析处理模块分离，只要语义词典不同，就可实现对不同数据库的访问，从而提高系统的可移植性。探测查询既是简化系统设计、提高可移植性的关键，也是影响系统性能的瓶颈，因为对数据库反复地进行探测查询，将浪费一定的系统资源、增加客户查询的等待时间。

8 结语

自然语言查询技术具有广阔的应用前景。首先，扩大了数据库的使用群体，用户可以避免学习数据库形式化查询语言，甚至不需要了解数据库知识；其次，屏蔽了数据库模式细节，提高了系统的安全性；第三，可以基于自然语言查询技术联成松散的多数据库网络系统，从而避免模式集成、数据转换等繁琐的工作，降低数据共享的成本；第四，可以基于自然语言查询技术，通过 Internet 建立网上虚拟数据库，并与网上搜索引擎集成起来，提供完美的网上搜索查询服务。

参考文献：

[1]王晓龙，关毅，等. 计算机自然语言处理[M]. 北京:清华大学出版社，2005:35-36.

[2]孟小峰. 中文数据库自然语言查询处理研究[D]. 中国优秀博士论文全文数据库，1999.

[3]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.

[4]孟小峰，王珊. 数据库自然语言查询系统 Nchiql 中语义依存树向 SQL 的转换. 中文信息学报[J]，2001，15(5):40-45.

[5]郑逢斌. 关于计算机理解自然查询语言的研究[D]. 中国优秀博士论文全文数据库，2004.

[6]胡明耀. 数据库汉语自然语言查询接口的设计与实现[D]. 中国优秀硕士论文全文数据库，2006.

[7]萨师煊，王珊. 数据库系统概论(第三版)[M]. 北京:清华大学出版社，2003.

作者简介：

王代星(1969-)，男，贵州贵阳人，硕士，工程师，研究方向为数据库、软件工程

收稿日期:2019-07-25　　修稿日期:2019-07-28

Research on Chinese Natural Language Query Technique in Database

WANG Dai-xing

(Higher Education Evaluation Center & Institute of Higher Education, Guizhou University, Guiyang 550025)

Abstract:

Based on the purpose and object of query, this paper discussed the technique of pre-inquiry database, combined with techniques of database semantics analysis and the query sentence structure analysis, to handle the Chinese natural language query in database. Puts forward a serials of concepts such as query elements segmentation, fuzzy attributes of query condition, pre-inquiry database and supper-relation, which leads to a clear concept of query, a clarity purpose of query, a simplified and easy method to implement a portable and usable query system.

Keywords:

Natural Language Query; Chinese Natural Language Processing; Database Query; SQL

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net