第 37 卷 第 23 期
Vol.37 No.23

计 算 机 工 程
Computer Engineering

2011 年 12 月
December 2011

• 软件技术与数据库 •　　文章编号：1000—3428(2011)23—0072—03　　文献标识码：A　　中图分类号：TP311

自然语言向 SQL 代码的转化方法

杨鹤标，陈 力

(江苏大学计算机科学与通信工程学院，江苏 镇江 212013)

摘 要：为解决智能学习系统查询语言的转化问题，提出一种自然语言向 SQL 代码转化的方法。利用所建立的字典扫描单词和理解语义，采用改进后的单词提取技术扫描自然语言串，以生成语义依赖树，并将其语义关系划分为若干独立的集合块，通过对该集合块遍历生成与自然语言等价的 SQL 代码。实验结果表明，该转化方法简单有效。

关键词：自然语言处理；中文分词；语义依赖树；中文查询数据库；智能学习系统

Transforming Method for Natural Language to SQL Code

YANG He-biao, CHEN Li

(School of Computer Science and Telecommunication Engineering, Jiangsu University, Zhenjiang 212013, China)

【Abstract】To solve the problem of generate SQL code automatically in an intelligent study system, this paper proposes a method to transform natural language to SQL code. This paper builds a dictionary as the base of scanning and comprehension of natural language, uses the improved word collecting technique to distinguish the words and build a semantic dependency-tree, uses the set-block technology to separate the tree to several set blocks, and scans the set blocks and generate the SQL code. Experimental results show that this method simple and effective.

【Key words】natural language processing; Chinese word segmentation; semantics dependency tree; Chinese query database; intelligence learning system

DOI: 10.3969/j.issn.1000-3428.2011.23.024

1 概述

演练测验是自主学习过程中十分重要的一步，在自主学习 SQL 语言时，需要有针对性地演练大量的习题，以巩固学习效果。目前大量习题采用自然语言描述，并且缺少相应的参考答案，这不利于学习者自我查究。因此，希望构建一个智能学习系统来解决这一问题，该系统能够自动分析理解自然语言描述的习题，并生成对应的参考答案。要达到这一目标，需要利用中文自然语言数据库处理的相关技术。这一领域的研究主要有以下方式：基于中间描述语言的架构[1](通过中间描述语言转换成 SQL)；基于语法的框架[2-4](自然语言查询语句通过句法分析得到语法树，通过语法树转换成 SQL)；以及基于 Ontology 的设计框架[5]等。

考虑到问题域限定在已经规范描述的自然语言查询语句，这些语句的特点是，描述规范、语义歧义小、逻辑嵌套级别低等，因此将这些语句分析生成语义树较为简单。针对上述特点，使得以上研究方法相对于问题域来说过于复杂。为此，本文提出一种自然语言向 SQL 代码转化的方法。

2 自然语言代码分词

自然语言理解是本文转化流程中最基础的一部分。目前针对这一领域的研究方法大多数是在建立词典的基础上，辅之以词法、语法和语义规[6]。本文研究领域所设计的语料相对集中，不需要对语句进行语法及语义分析。因此，建立一个针对查询语句的字典即可满足对计算机自动分词的要求。

2.1 字典建立

字典建立是为了帮助扫描程序准确高效地扫描出代码中各个单词，出于对该方法实现的考虑，将字典中加入属性、量词、所属表等字段，其目的是为了在语义分析阶段减轻算法的负担。字典的结构如下：

对象：表示具有独立语义的单词结构；

属性：表示单词对象所属类型。具体如下：

$Attribute \in Type\{E,Q,F,R,J,N,U,T\}$

其中，$E \in$ {学生，教师，课程，…}，实体类型；$Q \in$ {所有，存在，…}，量词类型；$F \in$ {年龄，性别，成绩，…}，字段类型；$R \in$ {等于，小于，在…之间，…}，关系类型；$J \in$ {并且，或者，…}，连接类型；$N \in$ {70, 80, …}，数字类型；$U \in$ {分，元，…}，单位类型；T，所属表(如成绩所属课程表)。

本文字典的建立是以实际可能涉及的单词为基础，将学生数据库实验过程中经常遇到的单词进行汇总，并结合其语义加以分类，最后将其归纳入字典。

2.2 单词扫描与 DSE 单词提取改进

在字典建立工作完成之后，便得到扫描程的序参照源。扫描程序的工作是从左至右地扫描自然语言代码，匹配出最合理的单词并输出。已有大量针对自然语言分词的研究，有基于概率统计的分词方法[7]，也有将字典和统计结合的分词方法[8]。这些方法均采用概率统计理论来消除歧义，同时在分词之后即输出语法树。考虑到本文研究内容现定于教学实验中出现的查询语句，并且前一阶段的工作已经将单词及其属性完整地存储在字典中，即不需要处理自然语言查询要求中可能出现的歧义，因此借鉴文献[9]提出的一种基于数据库模式的提取(Database based Schema Extracting, DSE)技术，并

基金项目：国家“863”计划基金资助项目(2007AA04Z1B2)

作者简介：杨鹤标(1960－)，男，教授，主研方向：自然语言处理，数据库技术；陈 力，硕士研究生

收稿日期：2011-06-29　**E-mail：**wormchenli@126.com

结合本文研究对象对其加以改进。

文献[9]指出，在 DSE 的提取过程中，采用的提取步骤为：基于提取规则库进行自动提取；采用提取模板的方式为前一步提取的结果获取附加的信息(如同义名、特征词、动词格关系等附加信息)。

提取模板的提出是为了减轻用户对语言知识面的掌握，而所要处理的问题是，将已经描述规范的中文查询语言转化为 SQL 语言。因此可将基于数据库模式的单词提取过程中参照提取模板的步骤省略。改进后的提取流程如图 1 所示。

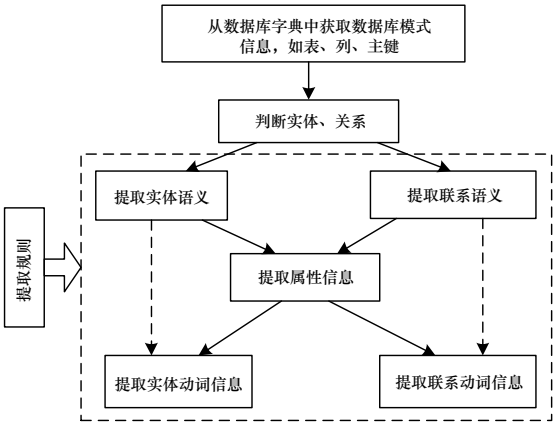

图 1　改进后的 DSE 提取流程

由图 1 可知，提取的规则库由一些启发式规则构成。例如对实体和关系的判断是基于如下规则：

(1)在数据库模式 $R(R_1,R_2,\cdots,R_n)$中，若 R_i 存在主码 $k_i(A_1, A_2,\cdots,A_m)$，$m>0$，则：

1)当 $m=1$ 时，R_i 为实体类。

2)当 $m>1$ 时，且存在 R_j 主码为 k_j，$k_i \cap k_j=\varnothing$，则 R_i 为关系类；否则 R_i 为实体类。

(2)对数据库模式中词与词的修饰关系，由于组合的复杂性，根据实体与实体、实体与联系之间的相互关系总结以下提取规则：

1)实体名能修饰该实体的所有属性名。

2)主属性的值能修饰所有其他属性名。

3)若某属性能指代实体，那么该属性的值也能修饰所有其他属性名。

3　语义关系描述

经过第 2 节的阐述，可以从自然语言查询语句中提取出单词集合，且集合内每个元素结构包含自身的属性信息。接下来将考虑将这些单词组合起来，使其能正确地表达出自然语言的语义信息。针对语言、语法和语义的分析及描述的经典理论，得出所提出的描述方式包括 0 型文法、上下文无关文法、上下文有关文法以及正则表达式，但是这些描述方式相对于本文研究的领域而言过于繁杂。由于工作领域针对性很强，并且要求算法实现便捷，因此可采用依存语法树[4,9-10]这一结构化的描述方式。

依存语法主张动词应是一个句子的中心，它支配别的成分，而它本身不受其他任何成分的支配，所有其他成分都从属于支配者。这一特性使得能将查询语言中的依赖关系完整地表示出来。

由于单词抽取完成后每一个元素都包含了各自的属性以及关系信息。因此可以方便地构造出对应的语义依赖树。如语句“查询李明的英语成绩”，其分析生成的语义依赖树如图 2 所示。

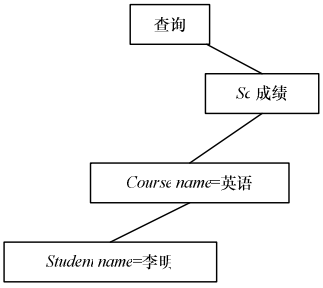

图 2　语义依赖树示例

由图 2 可知，在语义依赖树构造完成之后，通过将对其进行分析，可将其最终转化为等价的 SQL 代码。

3.1　语义依赖树分析

对中文语义依赖树的分析主要有中心词三元模型法[4]、自底向上迭代方法[9]、最大相关集合块方法[3]等。经过对其研究后发现，前 2 种方法应用于复杂逻辑结构下的语义依赖树分析。由于本文涉及的自然语言本身就是语法结构单一的对象，因此只需要保持依赖树所携带的语义信息，而不需要做更多的分析工作。因此，本文选择集合块的方法对依赖树加以分析。

3.2　集合块

在分析语义依赖树时，按照语义将它划分为若干个语义单位。通过分析其局部语义，从而得到其整体语义。首先引入集合块的概念[2,10]。

定义 1 在语义依赖树中，集合块是一个语义群，并且是一个确定了值集的数据库对象。

在语义依赖树中，将集合块形式化地描述为：

定义 2 设有集合块 $SetBlock=(Obj,Cond,T)$。

其中：

(1)$Obj=\{O|o \in TN \vee o \in AN\}$。

(2)$Cond=Cond_1 \wedge Cond_2 \wedge \cdots \wedge Cond_n$，取 $Cond_i=an_{i1}\ Op\ Value$ 或 $an_{i1}\ Op\ an_{i2}$ 或{SQL 表达式}，$1 \leqslant i \leqslant n$；$an_{i1}, an_{i2} \in AN$。$Value \in$ {数字，字符}，$Op \in \{>,<,\leqslant,\geqslant,=,\neq\}$。

(3)$AN=$ {所有属性集合}，$TN=$ {所有表集合}，$T=\{t|t=an_{ij}$ 的表名，$1 \leqslant i \leqslant n,\ 1 \leqslant j \leqslant 2\}$。其中，$Obj$ 是某表的表名或属性名；$Cond$ 为复合条件表达式；T 为条件表达式中所有表名的集合。

3.3　集合块划分

查询目标可能位于依赖树任意一个节点中，文献[2]提出先划分出查询目标所在的集合块，然后以该集合块为中心逐层向外划分临近的集合块，直至将整个依赖树划分完毕。由于在字典构造以及依赖树生成阶段已经将常用的修饰关系保存在依赖树节点中，因此只需要直接遍历语义依赖树，并将不同属性的集合块采取“分而治之”的策略即可。具体的划分策略如下：

(1)自底向上扫描依赖树。

(2)操作符节点与其操作数节点划分为不同的块。

(3)若某节点与其父节点属于同一张表，且存在自参照关系，则它们划分为不同的块。

(4)若某一动词节点的子节点依然是动词短语结构，则将其子节点划分为独立块。

(5)若节点为全称量词，则应当将其划分为该层最大的集合块。

(6)将划分好的集合块作为一个节点，返回步骤(1)直到全部划分完毕才终止。

(7)在集合块划分过程中，若某节块的子节点块是其修饰成分，则改节点块为查询主体。

考虑查询语句“查询至少选修李明选修的课程和英语的学生学号”，将其依赖树按照以上步骤划分后集合块实例如图 3 所示，其中，R 由集合块 $R1$ 和 $R2$ 构成；R 表示对查询主体 *Student.sno* 的限定；$R1$ 包含对连接词“和”的限定条件；$R2$ 包含对选择条件的限定。

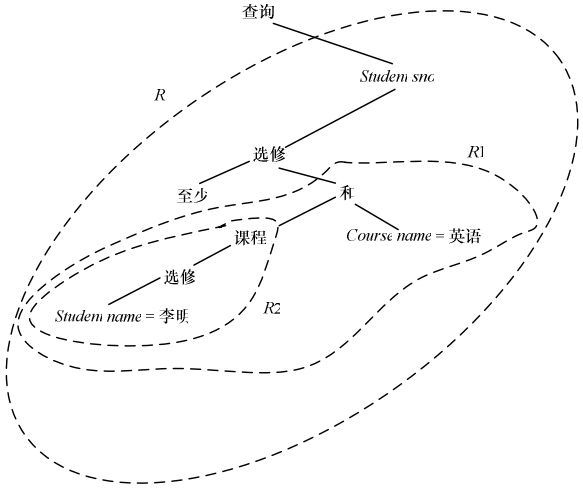

图 3 集合块划分实例

在集合块划分完毕之后，将说明如何将划分好的集合块树转化为等价的 SQL 语句。

4 SQL 代码生成

语义依赖树经过集合块划分之后，形成一颗集合块树，其包括嵌套查询在内的所有语义逻辑，如操作数、比较符等信息已经包含在每一个节点中，因此不需要担心嵌套查询或者操作数的拼接问题，只需要按照一定的算法，将集合块树中保存的信息转述出来即可。

事实上，在扫描集合块树后可以得到一个集合 $SB=\{Qo,Qc,Qt\}$，其中，Qo 为查询主体；Qc 为查询条件(包括嵌套条件)；Qt 为查询涉及到的表。对于一个语义完整地集合块来说，等价的 SQL 语句为：

```
SELECT SB.Qo
FROM SB.Qc
WHERE SB.Qt
```

可以先扫描再转化为 SQL 语句，也可以在遍历过程中直接生成 SQL 语句，下面给出遍历过程中生成 SQL 语句的具体步骤：

```
SBToSQL:
遍历集合块树，对每一个节点 node
{
扫描 SB;
if (node 为查询主体){
  SB.Qo.Add(node.name);
        在字典中查找 node 对应的表 table;
SB.Qt.Add(table);
}
if (node 为限定关系)
  SB.Qc.Add(node.name)
}
输出 SQL=SELECT SB.Qo FROM SB.Qc WHERE SB.Qt
```

该算法最后的输出即为自然语言所对应的 SQL 语句。

5 结束语

本文提出一种将自然语言转化为 SQL 代码的方法。借鉴中文查询数据库的研究方法，结合应用领域实际，构造符合实际应用需求的字典，改进 DSE 单词提取流程和相关算法。通过字典构造，基于集合块的语义依赖树分析方法将自然语言描述的查询语句转化为 SQL 语句。下一步研究工作是如何构建扩充完备的字典结构，以提高该方法的识别以及转化效率。

参考文献

[1] Androutsopoulos I. Interfacing a Natural Language Front-end to a Relational Database in Department of Artificial Intelligence[D]. Edinburgh, UK: University of Edinburgh, 1993.

[2] 郝 亮, 张文东, 袁春风. 一种数据库汉语查询接口的设计与实现[J]. 计算机技术与发展, 2010, 20(6): 13-17.

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

[4] 李明琴, 李涓子, 王作英. 语义分析和结构化语言模型[J]. 软件学报, 2005, 16(9): 1523-1533.

[5] 李 虎, 田金文, 王缓缓, 等. 基于 Ontology 的数据库自然语言查询接口的研究[J]. 计算机科学, 2010, 37(6): 200-205.

[6] 龚汉明, 周长胜. 汉语分词技术综述[J]. 北京机械工业学院学报, 2004, 19(3): 52-55.

[7] Kaplan S J. Designing a Portable Natural Language Database Query System[J]. ACM Transactions on Database Systems, 1984, 9(1): 1-19.

[8] 曹卫峰. 中文分词关键技术研究[D]. 南京: 南京理工大学, 2009.

[9] 孟小峰. 中文数据库自然语言查询研究[D]. 北京: 中国科学院计算技术研究所, 1999.

[10] 王秋月, 王 珊. NChiql 中基于语义依赖树的语言转述[J]. 计算机科学, 2006, 28(12): 5-8.

编辑 刘 冰

(上接第 71 页)

参考文献

[1] Pawlak Z. Rough Sets: Theoretical Aspects of Reasoning About Data[M]. Boston, USA: Kluwer Academic Publishers, 1991.

[2] 张文修, 梁 怡, 吴伟志. 信息系统与知识发现[M]. 北京: 科学出版社, 2003.

[3] 袁修久, 张文修. 模糊目标信息系统的属性约简[J]. 系统工程理论与实践, 2004, 24(5): 116-125.

[4] 徐伟华, 张文修. 基于优势关系下不协调目标信息系统的知识约简[J]. 计算机科学, 2006, 33(2): 182-184.

[5] 徐伟华, 张晓燕, 张文修. 优势关系下不协调目标信息系统的上近似约简[J]. 计算机工程, 2009, 35(18): 191-193.

编辑 顾逸斐