基于 PTCT 语义的自然语言查询系统

王华[1]

（1. 中山大学逻辑与认知研究所 510275）

内容提要：本文关注的是形式化自然语言时出现的超内涵问题，在分析一阶超内涵逻辑系统 PTCT 的基础上提出了 HIQ 系统的构想，它以 PTCT 为语义基础，以历史关系数据库模型 HRDM 为数据库概念模型，把 Squirrel 语言系统加入时态的因子，从而支持带有时间副词的自然语言查询语句。

关键词：超内涵；HIQ；PTCT

中图分类号：B81 **文献标识码：A**

本文关注的是语言逻辑中的超内涵问题，目前国内涉及此研究领域的人数不多，研究现状不尽人意。笔者希望通过此文引起更多学者对此问题的关注，并为今后探究自然语言的逻辑问题提供有益的启迪和全新的视角。

一、超内涵问题

20 世纪以来，逻辑学家、语言学家和计算机科学家一直在从事关于自然语言形式处理的研究。其中蒙太格(Montague，1974)关于英语部分语句系统的形式化方案为这个研究方向迈出了重要的一步。其引进了从句法和语义两个层面分析自然语言的强有力的方法，并发展成为形式化工具，为深刻理解自然语言的语义学提供了必要的技术背景。

然而值得注意的是，蒙太格处理命题态度句时引入了公式的内涵概念，即命题是从可能世界到真值的外延函数，这样推广的结果是在 MIL (蒙太格内涵逻辑)中任何表达式的内涵都是从可能世界到该表达式外延的函项。因此即使在内涵语境中，从自然语言到公式的对应本质上也是外延的。从而引出的问题就是：这样的函数给出的等价标准不足以刻画自然语言的精粒度。例如，必然真理(典型的比如那些数学命题)，在命题态度句中就成了相等的了。

举例如下：

(1)玛丽相信费尔玛大定理。

(2)玛丽相信 1+1=2。

在这里费尔玛大定理与“1+1=2”都是外延等价的，因此句(1)和(2)是逻辑等价的，然而这样的两个句子是内涵同一的吗？显然，答案是否定的，因为人们对于问题的认知是有很大的差异的。但我们却能从蒙太格内涵逻辑中推出其内涵也是同一的答案，因为传统内涵逻辑坚持认为：对任意语句α和β，如果α和β在所有条件下都有相同真值，那么α和β具有相同的意义。可见，传统内涵逻辑用逻辑等价对内涵的刻画是不符合自然语言语义的确切理解的。很明显，我们由$\varphi = \psi$可以得到$\varphi \leftrightarrow \psi$，但由后者我们无法得到前者。例如，尽管$(\neg\varphi \vee \psi) \leftrightarrow (\varphi \to \psi)$，但并非 believe(Mary，$(\neg\varphi \vee \psi)) \leftrightarrow$believe(Mary，$(\varphi \to \psi)$)。这样的例子就说明了自然语言所要求的精粒度内涵性系统的本质。即一个陈述中的命题态

收稿日期：2006-6-13

基金项目：国家社科基金项目；教育部哲学社会科学攻关项目。

作者简介：王华（1978-），女，汉族，山西太原人，中山大学逻辑与认知研究所硕士研究生

度，如相信，并不意味着对等价命题也有同样的态度。也就是说如果坚持使用外延原则解释表达式的内涵，是不能解释信念语境引发的替换失效问题的[1]。

当然，我们会说，如果 Mary 有初等逻辑的知识，只要她相信 $\neg\varphi\vee\psi$ 她就会相信 $\varphi\to\psi$，但这不是问题的关键：我们总可以添加一些公设到个体的信念理论中去，使得我们能从一个信念推出另一个信念来。问题关键在于，我们应该有足够精粒度的内涵，使我们不必在所有情况下都将这样的信念等同起来。我们的语义理论原则上应该允许表达式之间有最精粒度的区分，因此发展超内涵逻辑去解决超内涵问题就具有重要意义。

二、超内涵逻辑系统 PTCT

超内涵的概念由 Cresswell 在 1975 年正式提出，但一般的内涵问题及其思想来源于 Frege 和 Russell，前者关注的是内涵与外延的区分及其逻辑刻画问题，后者关注的是内涵引入逻辑系统中的悖论问题。超内涵逻辑的系统研究则始自 Carnap[2] (1947)和 Church[3] (1951)，至 1980 年代达到高潮，Bealer[4] (1982)，Barwise and Perry[5] (1983) ，Cresswell[6](1985) ，Zalta[7] (1988)，Reinhard Muskens[8] (1995)等逻辑学家在过去的二十多年里创建了一系列超内涵理论。艾塞克斯大学计算机科学系教授 Chris Fox 和伦敦大学国王学院大学哲学院计算语言系 Shalom Lappin 教授，在 2005 年出版了他们的最新研究成果《Foundations of Intensional Semantics 》专著。在专著中他们刻画了一阶超内涵逻辑系统 PTCT。

PTCT 指的是带 Curry 类型的性质论(Property Theory with Curry Typing)，作者通过带柔性的 Curry 类型的逻辑系统在一阶框架上发展根本内涵语义的观点。PTCT 可以看作是 PT 的发展，主要的不同之处表现在第一：PT 只是用性质来模仿类型，而 PTCT 有一个完全成熟的类型语言；第二：PT 用全称类型来处理自然语言的多态现象，比如连接词，动名词和不定式，但是因为没有对类型进行任何约束导致有悖论产生。而 PTCT 允许受限的多态性，因此它可以用不同的函数类型来表达相应的自然语言的协调性，一致性等语言现象，并且因为受限而避免悖论；第三：PT 允许自我应用，PTCT 不允许自我应用，增加这种约束是为了防止悖论；第四：不同于作者早期对于性质论研究的工作的是，带 Curry 类型系统的 PTCT 完全具有了清晰完整的证明论[2]和模型论[3]。

PTCT 的核心语言由以下的子语言构成[9]：

项 $t::=x\,|\,c\,|\,l\,|\,T\,|\,\lambda x(t)\,|\,(t)t$ （3-3）

逻辑常量 $l::=\hat{\neg}\,|\,\hat{\wedge}\,|\,\hat{\vee}\,|\hat{\to}|\hat{\leftrightarrow}|\,\hat{\forall}\,|\,\hat{\exists}|\hat{=}_T|\hat{\cong}_T|\in T$

类型 $T::=B\,|\,\mathrm{Pr}op\,|\,T\Rightarrow T'$ （3-4）

合式公式 $\varphi::=\alpha\,|\,\neg\varphi\,|\,\varphi\wedge\psi'\,|\,\varphi\vee\psi'\,|\,\varphi\to\psi'\,|\,\varphi\leftrightarrow\psi'\,|\,(\forall x\varphi)\,|\,(\exists x\varphi)\,|$ （3-5）

α 是原子公式wff $\alpha::=(t=_T s)\,|\,t\in T\,|\,t\cong_T s\,|\,{}^{T}t$

可以看到，项的语言是无类型的 $\lambda-$ 演算，并且还带有常量c和逻辑常量 l，类型 T 也是项。类型的语言包括基本类型 B，命题类型Prop和函数转换类型 $T\Rightarrow T'$。合式公式的语言是带有类型判断 $t\in T$、类型同一 $=_T$ (内涵同一)、类型相等 $t\cong_T s$ (外延相等)以及真值判断 ${}^{T}t$ 的一阶语言。特别的是我们可以看到在PTCT中，有两种相等的概念。$t\cong_T s$ 表明项 t,s

[2]具体可参见《Foundations of Intensional Semantics》75—83 页
[3]具体可参见《Foundations of Intensional Semantics》96—102 页

是类型T且外延相等，外延相等在项的语言中表达为$t \hat{\cong}_T s$。$t =_T s$表明项t,s是内涵同一，它的规则本质上就是$\lambda\partial\beta\eta$演算，内涵同一在项的语言中表达为$t \hat{=}_T s$。对于任何一种类型，我们都有$t =_T s \to t \cong_T s$但是没有$t \cong_T s \to t =_T s$，所以PTCT能在等价命题上支持精粒度内涵区分。

三、HIQ 系统

目前，为了使用数据库，用户必须学会某种数据库操作语言(广泛使用的SQL)，然后用这种语言来表达自己的意图。这种工作方式要求用户接受专门的训练，不便于数据库技术的进一步推广应用。很自然的我们能够想到的一种较为理想的方式是:用户只需用某种自然语言来表达自己的意图，然后系统能自动理解用户的意图并执行相应的操作，最后系统用同一种自然语言来给出结果，从而大大改善人机交互的容易程度。

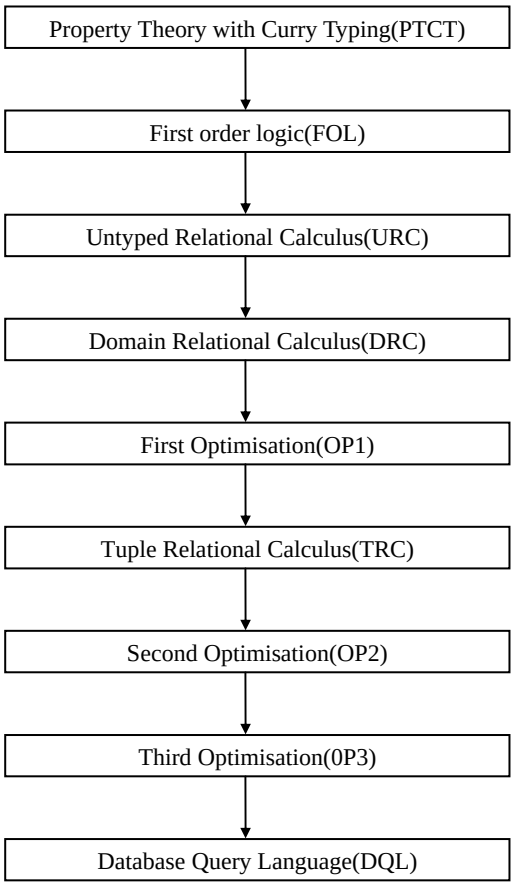
一般来说，用户对数据库的操作包括数据定义，数据查询，数据更新(增加 删除 修改)等。但是对普通用户而言， 查询操作是使用最为频繁和最为重要的，因此对数据库关注的问题主要集中在自然语言查询处理的过程中. 当我们使用自然语言进行数据库查询的时候，因为自然语言是丰富的，所以我们对数据库提出的问题形式描述也是多样的。

现有的比较成熟的基于自然语言理解的语言系统有：中文数据库自然语言查询系统Nchiql，WTCDIS系统，Rchiql系统，IVR自动语音应答系统等。Chris Fox创立了Squirrel语言系统[10]，这是一种基于关系型数据库的自然语言前端，可以对受限的英语子语言集以自然语言形式提问，然后将问句形式化，最终由系统翻译为SOL语句。Squirrel语言本身比较简洁，翻译自然语言到SQL时比较高效，并且可以消除摸棱两可的语义或着处理多语义的词汇项。但是系统也有不足之处，一是它对数据库的自然语言查询有诸多的限制，比如要求查询中的名词、动词等实词必然与数据库内容相关，还要求查询句中涉及到的概念只限于特定数据库的概念模式等。二是它只能处理静态数据的快照数据库。快照数据库仅仅能体现现实世界中数据的当前状态，只反应了一个对象在某一个时刻的状态(快照)，不联系其过去和未来。但是如前所述，现代的信息含量是巨大和复杂的，日益广泛的数据库应用要求管理被处理事件的时态信息，包括时刻信息(Instant Information)， 时间区间信息(Interval Information)和相对时间信息(之前、之后、重叠)等等[11]。HIQ(Historical Information Query)系统HIQ将Squirrel语言系统加入时态的因子，使之有更强大的功能，从而可以处理具有时间副词的自然语言查询语句。

我认为要实现对数据库的自然语言查询，核心的问题就是实现从自然语言向数据库语言的转换，即自然语言的语义与数据库语义一致的一个过程。关键的几个步骤如下：首先对查询语句进行分词处理，然后通过结合词典信息，对自然语句作初步的语法语义分析以及歧义排除处理，从而得到了系统对句子的理解的中间形式[12]，这样就具备了向数据库语句转换的基础(典型的如SQL语句)，最后通过元组关系运算和近一步的优化得到符合数据库语法的有效语句。

实现 HIQ 这样的系统首先要选择一个时态数据库模型，例如 Squirrel 语言系统是以关系型数据库为模型的。针对当前已有的比较成熟的 13 种时态数据库模型，我选择 HRDM 模型。HRDM 模型[13]是 1982 年 Jmes. Clifford 在他的博士论文“A logical frame work for the Temporal Semantics and Natural Language Querying of Historical Database ”中提出的，文章详细阐述了历史关系数据库模型 HRDM（Historical Relational Data Model）和相应的自然语言查询系统 QE-III。选择这样的模型有两个原因，一是因为 HRDM 对时态数据库的理论发展具有划时代的意义，二是因为在 HRDM 中只处理了过去时态，没有涉及将来时态等更多的时态算子，这样为扩展 HIQ 系统减小了复杂程度。

HIQ 系统整体设计为：



其中 PTCT 是第三章讨论过的超内涵逻辑语义，确定基本词汇的类型，分析语法树中的变量；FOL 是一阶逻辑，但是为了方便处理对数据库的查询，增加了<,>,=的初始项；URC 是无类型关系运算，将逻辑表达式映射为当前的数据库的关系运算，通常关系运算的变量是有类型的；DRC 是域关系运算，将 URC 中的有类型的关系运算的变量的域进行约束，并为变量找到合适的类型；OP1 是第一次优化，把 DRC 形式的表达式写为分句的形式并消除冗余的分句； TRC 是元组关系运算，将性质的量化转化为数据库中表的元组的量化；OP2 是第二次优化，将变量的范围从域转换到数据库中的表的关系上；OP3 是第三次优化，是为了符合数据库中关键字段的唯一性，将原来地位相等的元组字段中的关键字段选择出来，重构新的元组关系运算表达式，从而最终转换为符合 DQL 语义的数据库查询语句。

下面我们来对上述进行形式化定义：

- PTCT 如前面所定义的

项 $t::=x\,|\,c\,|\,l\,|\,T\,|\,\lambda x(t)\,|\,(t)t$

逻辑常量 $l::=\hat{\neg}\,|\,\hat{\wedge}\,|\,\hat{\vee}\,|\hat{\rightarrow}|\hat{\leftrightarrow}|\,\hat{\forall}\,|\,\hat{\exists}\,|\hat{=}_T|\hat{\cong}_T|\in T$

类型 $T::=B\,|\,\mathrm{Pr}op\,|\,T\Rightarrow T'$

合式公式 $\varphi::=\alpha\,|\,\neg\varphi\,|\,\varphi\wedge\psi'\,|\,\varphi\vee\psi'\,|\,\varphi\rightarrow\psi'\,|\,\varphi\leftrightarrow\psi'\,|\,(\forall x\varphi)\,|\,(\exists x\varphi)\,|$

α 是原子公式wff $\alpha::=(t=_T s)\,|\,t\in T\,|\,t\cong_T s\,|\,{}^{T}t$

- FOL

$\varphi ::= \alpha \mid \neg\varphi \mid \varphi \wedge \psi' \mid \varphi \vee \psi' \mid \varphi \to \psi' \mid \varphi \leftrightarrow \psi' \mid (\forall x\varphi) \mid (\exists x\varphi) \mid \alpha = \beta \mid \alpha < \beta \mid \alpha > \beta$

α 是原子公式wff $\alpha ::= (t =_T s) \mid t \in T \mid t \cong_T s \mid {}^T t$

- URC

$\Phi ::= \{\nu, \varsigma, ... \mid \varphi\}$

$\varphi ::= \alpha \mid \neg\varphi \mid \varphi \wedge \psi' \mid \varphi \vee \psi' \mid \varphi \to \psi' \mid \varphi \leftrightarrow \psi' \mid (\forall \nu\varphi) \mid (\exists \nu\varphi)$
$\mid < \alpha, \beta, ... \in \tau > \mid \alpha = \beta \mid \alpha < \beta \mid \alpha > \beta$

$\alpha ::= \kappa \mid \nu \mid *$

查询Φ由语句中的变量ν, ς抽象形式化，φ, ψ是通常的连接和量化，只是带有相等，比较，和表的成员结构，τ是表的元变量，项α, β与常量κ或者变量做比较，*不是实体，只是隐匿地提供存在量化，比如$< *, * > \in loc \leftrightarrow \exists x \exists y < x, y > \in loc$。

- DRC[14]

$\Phi ::= \{\nu \in \delta, \varsigma \in \varepsilon, ... \mid \varphi\}$
$\varphi ::= \neg\varphi \mid \varphi \wedge \psi' \mid \varphi \vee \psi' \mid \varphi \to \psi' \mid \varphi \leftrightarrow \psi' \mid (\forall \nu \in \delta\varphi) \mid (\exists \nu \in \delta\varphi)$
$\mid < \alpha, \beta, ... \in \tau > \mid \alpha = \beta \mid \alpha < \beta \mid \alpha > \beta$
$\alpha ::= \kappa \mid \nu \mid *$

δ, ε是域的元变量。

- TRC[15]

$\Phi ::= \{\alpha, \beta ... \mid \nu \in \tau, \varsigma \in \tau', ... \mid \varphi\}$
$\varphi ::= \neg\varphi \mid \varphi \wedge \psi' \mid \varphi \vee \psi' \mid \varphi \to \psi' \mid \varphi \leftrightarrow \psi' \mid (\forall \nu \in \delta\varphi) \mid (\exists \nu \in \delta\varphi) \mid$
$< \alpha, \beta, ... \in \tau > \mid \alpha = \beta \mid \alpha < \beta \mid \alpha > \beta$
$\alpha ::= \kappa \mid \nu.\alpha$

$\nu.\alpha$ 表示变量的范围必须是对表的关系而言。

下面的例子是在 HIQ 系统中实现对具有时间副词的自然语言查询语句的处理结果。

ENG: who earned more than 2000 last year ?
SYN: (s(np(ipron(who)))(vp(v(earned))(adj(more))
(pp(prep(than))(np(propn(2000))) (adj(last))(n(n(year))))))
PTCT: ‘than: ‘2000:(‘more: ‘earn): ‘last:‘year: fva1
FOL: than(fva1, ‘more: ‘earn, ‘2000, ‘(last year))
URC: {fva1;exists(rr，[fva1,*,rr,*]：emp&rr>‘2000, tim&rr=‘(last year))}
DRC: {fva1:emp![name];exists(rr:emp![sal]，[fva1,*,rr,*]: emp&rr>‘2000, tim&rr=‘(last year))}
OP1: {va1：emp![name];exists(skf1：emp！[sal]，
[fva1,*,skf1,*]：emp&skf1>‘2000, tim& skf1=‘(last year))}
PAR: which employee earned a salary that was greater than 2000 last year?
TRC: {fva1! name;fva1:emp; exists (skf1:emp,exists(tuple1:emp,tuple1!name
=fva1!name&tuple1!sal=skf1!sal)&skf1!sal>2000&skf1!tim=(last year))}

OP2: {fva1!name;fva1:emp; exists(sk2:emp，
sk2!name=fva1!name&sk2!sal>2000&skf2!tim=(last year))}
OP3: {fva1!name;fva1:emp;fva1!sal>2000& fva1!tim=(last year)}
DQL: SELECT DISTINCT fva1.name FROM emp fva1 WHERE fva1.sal>2000 and when fva1.tim=(last year)

这里做几点说明：

- 这是一个人事管理的数据库，对其中的工资表emp进行查询，为方便表示对数据的操作沿用了SQL的一些惯用表达；
- SYN是以树形结构对自然语句的结构组成进行分析，用（）来表示层次的区分；
- PAR是相等处理，将第一次优化过的分句整理为和原英语文句同样内涵的标准句子；
- 为简明起见，将时间变量的值直接表达为 tim=(last year)，当然在实际的数据库当中这是不符合数据库语法的，但是我们完全可以通过设置时间函数去处理 last year 这样的表述，比如利用时间函数 YEAR()和 DATE ，last year 就可以表达为：YEAR(DATE — YEAR(fva1.tim))=1 这样的形式了，这样的技术细节暂且不做讨论。

参考文献

[1] 荣立武.论内涵逻辑与内涵语境下的替换失效问题. 自然辩证法研究. 2006，22 卷第 1 期: 44.

[2] Carnap.Meaning and Necessity. Chicago:University of Chicago Press . 1947.

[3] Church.A formulation of the logic of sense and denotation . Structure， Method， and Meaning.The Liberal Art Press，1951.

[4] Bealer，G.Quality and Concept.Oxford :Clarendon Press ，1982.

[5]]Barwise and Perry.Situation and Attitudes.Cambridge: MIT Press，1983.

[6] Cresswell. Structured Meaning.Cambridge:MIT Press，1985.

[7] Zalta，E.N.Intensional Logic and Metaphysics of Intensionality.Cambridge:MIT Press，1988.

[8] Reinhard Muskens.Meaning and Partiality.Stanford :CSLI and FOLLI，1995.

[9]Chris Fox，Shalom Lappin.Foundations of Intensional Semantics.Oxford: Blackwell Publishing，2005：74.

[10]， [14]， [15] Chris Fox.Squirrel Documentation. http://cswww.essex.ac.uk/SNAP/Squirrel/doc_html.1995.

[11]张师超，严小卫，聂文龙.时态数据库中的几个问题.广西师范大学学报.1995，第13卷第4期：10-11. **]**

[12]孟小峰，王珊.中文数据库自然语言查询系统Nchiql 设计与实现.计算机研究与发展.2001 ，38:1080-1086.

[13] Jmes Clifford .Natural Language Querying of Historical Database. Computational Linguistics.1988，14:10-31.

System of Natural Language Querying Based on Semantics of PTCT

Wang-hua[1]

(1. Institute of Logic and Cognition , Zhongshan University , Guangzhou 510275,China)

Abstract: This paper concerns with the hyperintensional problems that appear when natural languages are formalized. The author proposes the HIQ system on the basis of analysis of the fist-order hyperintensional logic PTCT . HIQ bases on the semantics of PTCT and uses the HRDM as the conceptual model for the database and incorporates the tense factor in the Squirrel System of language. Thus it is able to support the inquiring sentences in natural languages with adverbs of time.

Key words: hyperintensional;HIQ ;PTCT