维普资讯 http://www.cqvip.com

面向数据的句法分析消歧*)

张玥杰[1] 张 涛[2] 朱靖波[3] 姚天顺[3]

(复旦大学计算机科学与工程系 上海市智能信息处理重点实验室 上海 200433)[1]

(上海财经大学信息管理与工程学院 上海 200433)[2]

(东北大学信息科学与工程学院 沈阳 110006)[3]

摘 要 面向数据的分析技术(Data-Oriented Parsing, DOP)是一种概率分析策略,其概率模型的主要目的在于为一个给定的句子找到最可能的分析,即分析消歧。实际上,有关算法计算复杂度的大量研究证明,该类消歧问题属于NP-完全问题。因此,为有效实现最可能的分析,国外学者提出许多近似分析算法。本文主要论述在DOP框架中,基于Monte Carlo方法找到最可能分析的近似分析算法,并说明该方法可在合理的算法时间代价范围内实现,而且在统计上受控,以确保所获得的近似解确实对应着分析消歧后的精确解。

关键词 面向数据的句法分析,随机树替换文法,消歧,Monte Carlo方法

Disambiguation for Data-Oriented Parsing

ZHANG Yue-Jie[1] ZHANG Tao[2] ZHU Jing-Bo[3] YAO Tian-Shun[3]

(Dept. of Computer Science & Engineering, Shanghai Key Laboratory of Intelligent Information Processing, Fudan University, Shanghai 200433)[1]

(School of Information Management and Engineering, Shanghai University of Finance & Economics, Shanghai 200433)[2]

(School of Information Science & Engineering, Northeastern University, Shenyang 110006)[3]

Abstract Data-Oriented Parsing(DOP) technique is a kind of probabilistic parsing strategy. The main goal of DOP model is to find the most probable parse for a given input sentence, that is, parse disambiguation. In fact, it is proved through a lot of research work about algorithm computation complexity that this kind of disambiguation problem belongs to the class of NP-Complete problem. So in order to implement the most probable parse efficiently, some researchers have proposed many approximation parsing algorithms. This paper mainly presents a kind of approximation parsing algorithm based on Monte Carlo method in DOP framework, which can be implemented at reasonable(i. e. polynomial) algorithmic cost. And at the same time, under statistical control, it is guaranteed that an obtained approximate solution indeed corresponds to an exact solution of the problem after disambiguation.

Keywords Data-oriented parsing(DOP), Stochastic tree substitution grammar(STSG), Disambiguation, Monte carlo method

1 前言

"面向数据的分析技术"(Data-Oriented Parsing, DOP)首先由Scha在1990年提出[1],之后由Bod逐步发展[2,3]。其中具体表达了这样的假设,人类对语言的领悟和创造依赖于以往具体的语言经验,而不是依赖于抽象的语法规则[4]。基于DOP的处理模型建立在包含大量语言现象的语料库基础上,把经过标注的语料库看作一种语法(Grammar)[5]。模型首先预设了一个具有带标短语结构树标注的语料库,然后从中抽取所有任意大小规模和复杂结构的片段。其次,通过对语料库中片段的组合操作来实现新输入的分析,然后考虑输入的所有派生结果的概率总和的大小来选择最具有可能性的分析结果[6]。

DOP概率模型的主要目的在于:对于一个给定的句子,为其找到最有可能的分析(Most Probable Parse, MPP),即分析消歧。假设,A_i 表示针对输入言语 i 的分析集合(可能为空集),则消歧过程可通过以下优化公式给出:

$$A^* = \mathrm{argmax}_{a \in A_i} P_{DOP}(a)$$

该优化公式说明,所优选的分析 A^* 是根据模型 P_{DOP} 具有最高概率的一种分析。由此可见,在应用DOP概率模型进行消歧中的一个重要问题为:是否能够通过一条合理的途径,将上述优化公式转换为一种有效的算法。然而,对于该类消歧问题,不可预知是否能够设计出一种确定的多项式时间求解算法。实际上,有关算法计算复杂度的大量研究证明,该类消歧问题属于NP-完全问题。对于NP-完全问题,唯一存在的确定算法具有指数时间复杂度,其相关概率模型在应用类别以及规模等诸多方面存在严格的约束条件[7]。因此,为有效解决消歧问题,实现最可能的分析,国外学者提出了许多分析算法,如Bod提出的Monte Carto近似分析算法[8],该方法应用广泛且较为有效;Sima'an提出的优化算法[9];Joshua Goodman提出的基于概率上下文无关文法的算法[10]等等。

本文主要论述在DOP框架中,基于Monte Carlo方法找

*)本文得到国家自然科学基金(编号:60203010、70501018与605333100)和上海财经大学"211工程"(2004年)资助。张玥杰 副教授,主要研究领域为机器翻译、中文信息处理相关技术;张 涛 副教授,主要研究领域为机器学习、优化建模及相关技术;朱靖波 副教授,主要研究领域为计算语言学、中文信息处理及相关技术;姚天顺 教授,博士生导师,主要研究领域为计算语言学、自然语言理解及相关技术。

维普资讯 http://www.cqvip.com

到最有可能分析的近似分析算法，并说明该方法可在合理的算法时间代价范围内实现（即算法时间复杂度为多项式时间），而且在统计上受控，以确保所获得的近似解确实对应着分析消歧之后的精确解。

2 随机树替换文法

随机树替换文法（Stochastic Tree Substitution Grammar，STSG）可利用五元组$\langle V_N, V_T, S, R, P\rangle$描述，其中$V_N$表示非终结符集合；$V_T$表示终结符集合；$S\in V_N$表示文法开始符；$R$表示基本树有限集，这些树的根结点和分支结点标注为非终结符，叶结点标注为非终结符或终结符；P为概率权值函数，对任何一棵属于集合R的以a为根结点的树t，利用函数$P(t)$表示用树t替换结点a的概率权值，因而$P(t)$满足下面两个条件：(1)$0<P(t)\leqslant 1$；(2)$\sum_{t:\mathrm{root}(t)=a}P(t)=1$。

STSG中完整的派生过程基于基本树与替换运算符进行，替换运算符$\circ$被称为最左优先替换或者简单替换。其操作原则为：如果基本树t_1的最左非终结符叶结点和基本树t_2的根结点相同，则$t_1\circ t_2$的运算结果为用t_2来替换t_1的最左非终结符叶结点。STSG G的一个最左优先派生是一组基本树$\langle t_1, \cdots, t_n\rangle$，其中树$t_1$的根结点为$S$，$t_1\circ\cdots\circ t_n$的叶结点都标注为终结符。因此，STSG G的最左优先派生集合可以描述为：

$$\mathrm{Derivations}(G)=\{\langle t_1,\cdots,t_n\rangle \mid t_1,\cdots,t_n\in R\wedge \mathrm{root}(t_1)=S\wedge \mathrm{yield}(t_1\circ\cdots\circ t_n)(V_T^+\}$$

根据给定的STSG G，如果对于一棵树T，存在一种最左派生$\langle t_1, \cdots, t_n\rangle=T$，则称之为树$T$的一种派生。如果派生$\langle t_1, \cdots, t_n\rangle$的叶结点可产生字符串$S$，则称之为串$S$的一种派生，且派生$\langle t_1, \cdots, t_n\rangle$的概率定义为$P(t_1)*\cdots*P(t_n)$。设STSG G产生一棵分析树$T$，则存在派生$\langle t_1, \cdots, t_n\rangle\in$ Derivations(G)，满足$t_1\circ\cdots\circ t_n=T$。由G产生的所有分析树的集合，即树语言，可描述为：

$$\mathrm{Parses}(G)=\{T \mid (\langle t_1,\cdots,t_n\rangle\in \mathrm{Derivations}(G): t_1\circ\cdots\circ t_n=T\}$$

如果一棵分析树产生字符串S，则称之为S的一种分析，且该分析的概率等于其所有派生的概率之和。一个字符串的概率可以定义为其所有分析的概率之和，同样等于其所有派生的概率之和。由G产生的所有字符串集合，即串语言，可以描述为：

$$\mathrm{String}(G)=\{S \mid \exists T: T\in \mathrm{Parses}(G)\wedge S=\mathrm{yield}(T)\}$$

3 面向数据的句法分析

显而易见，人们在处理新语言现象时，寻求最有可能分析的过程是基于对曾经见过的语言现象的成功分析的基础上，一个人的语言经验不知不觉地决定了其语言分析过程的结果。面向数据分析技术的基本观点为，采用言语及其分析的语料库来表达一个人的过去语言经验，分析新言语的过程相当于出现在语料库中分析片段的构造过程。实际上，该分析过程是根据分析语料库，为一个输入句子生成一个句法分析森林的过程。

3.1 基本框架

对于DOP概率模型，可将其视为STSG形式化描述方法的一种具体应用，其中所包含的基本树与替换运算符组合，以产生整棵句法分析树的派生。每一棵基本树t被分配一个概率$P(t)$，与其在参考树库中出现的次数成比例，则派生$d=\langle t_1, \cdots, t_q\rangle$的概率$P(d)$定义如下：

$$P(d)=\prod_{t\in d}P(t)=\prod_{i=1}^{q}P(t_i)$$

其中，$t\in d$表示在派生d中出现的所有基本树。因此，一棵句法分析树T的DOP-概率定义为其所有派生的概率之和：

$$P_{\mathrm{DOP}}(T)=\sum_{d\Rightarrow T}P(d)=\sum_{d\Rightarrow T}\prod_{t\in d}P(t)$$

其中，$d=>T$表示针对句法分析树T的所有派生。

对于一个给定的句子$w_1^n=w_1\cdots w_m$，其DOP-概率定义为：

$$P_{\mathrm{DOP}}(w_1^n)=\sum_{T\Rightarrow^* w_1^n}P_{\mathrm{DOP}}(T)=\sum_{d\in(S\Rightarrow^* w_1^n)}P(d)$$

其中，$T\Rightarrow^* w^{n1}$表示T是w_1^n的一棵可能的句法分析树，$d\in(S\Rightarrow^* w_1^n)$表示对应于$w_1^n$的开始于$S$的所有可能的派生。

3.2 分析过程

对于为输入句子创建句法分析森林的算法，是从针对上下文无关文法的算法中获得的。该算法能够在多项式时间（通常是立方的量级）内，对一个包含n个单词的输入句子进行句法分析。整个句法分析过程一般由以下两个不同阶段组成：

· 分析阶段

构建描述输入句子的派生集合的形式化表示，即利用一个图或者子串表，采用上下文无关重写规则集与待分析句子的单词序列作为输入，构造与输入句子相对应的派生结果——带标短语结构图（带标短语是利用表示该短语句法类型的符号进行标注的单词序列）作为输出；

· 抽取阶段

从输入句子的派生集合中抽取句法分析结果，创建句法分析森林，即通过图中由一种句法类型指向另一种句法类型的弧，来获取与输入句子相对应的句法分析结果，从而创建类似图的句法分析森林。

在基于STSG的DOP模型中，每一棵基本树都可被看作为一条上下文无关重写规则root(t)->yield(t)。为获取与输入句子相对应的类似图的句法分析森林，不仅需要利用句法类型来标注所包含的短语，而且需要利用与其对应的基本树进行标注。同时，在通过STSG生成句法分析森林的过程中，生成相同分析树的不同派生并未发生相互冲突的问题。

下面，以一个简单的STSG G为例来说明句法分析森林的获取过程，假设其基本树集合构成如图1所示。

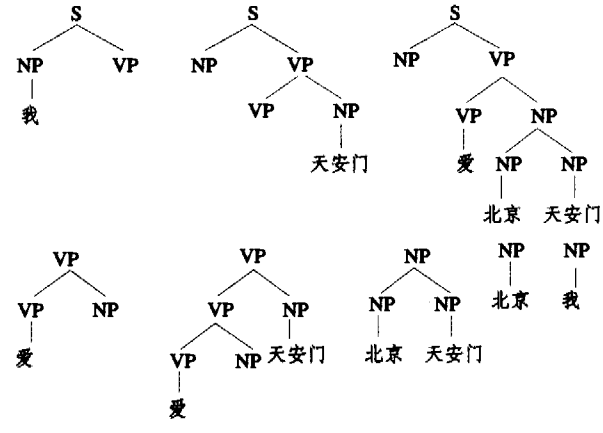

图1 一个简单的STSG G的基本树集合

对于输入句子“我爱北京天安门。”，通过不同的派生过程为其生成不同的句法分析树，从而为其获取句法分析森林的过程如图2所示。图中每一条弧(i, j)表示跨越输入句子中

维普资讯 http://www.cqvip.com

第 i 个位置至第 j 个位置之间所有单词的一条边，包含短语、子句甚至整个输入句子的派生过程中所利用的基本树。在该森林中，不同的派生过程可生成相同的句法分析树，基于四种不同的派生生成两棵不同的分析树。而且，通过不同的派生过程(可能具有不同的概率)，两棵分析树分别被生成两次。

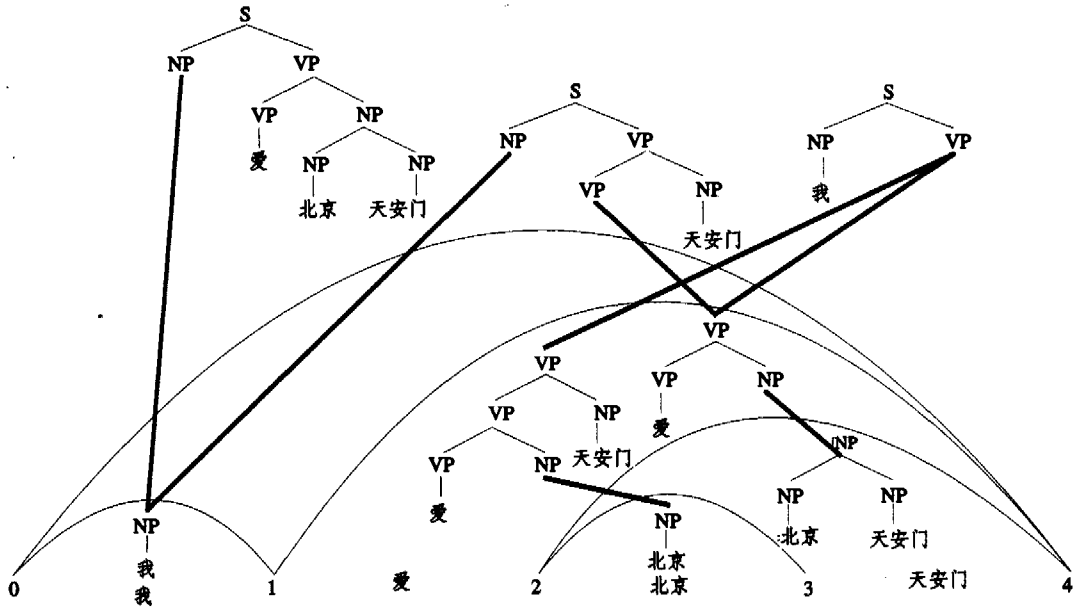

图 2　句法分析森林的获取过程

4　基于 Monte Carlo 的句法分析消歧

根据 DOP 概率模型，一棵句法分析树的生成需要考虑其所有派生，其生成概率应该等于所有派生的概率之和。实际上，许多句子都具有歧义性，即存在不同的句法分析树，则其生成概率应该为其所有分析的概率之和。然而，输入句子可能具有按指数级别增长的大量句法分析结果，而每一种分析也可能对应着按指数级别增长的大量派生。为找到输入句子的最可能句法分析，通过比较分析概率并不是一种有效的途径。即使可以决定一种分析的概率，求取该分析所有派生的概率之和也不是一种有效的方法。因此，多项式时间的句法分析并不能保证多项式时间的分析消歧，需要使用启发式或者近似求解的方法加以解决。

4.1　Monte Carlo 方法概述

Monte Carlo 方法也称为随机模拟方法，有时也称作随机抽样技术或统计试验方法[11,12]。其基本思想是：首先建立一个概率模型或过程的观察或抽样试验来计算所求参数的统计特征，最后给出所求解的近似值，而解的精确度可利用估计值的标准误差来表示[13,14]。在 DOP 框架中，尽管不存在找到最可能句法分析的确定性的多项式算法，但仍然可以利用 Monte Carlo 方法来估计最可能句法分析，同时具有尽可能小的误差。

基于 Monte Carlo 的句法分析消歧算法是一种概率算法，属于受限误差概率多项式时间算法类。该算法不能在多项式时间内决定一个输入句子最可能的句法分析，但是可在多项式时间内估计最可能的句法分析，并且能够确保估计最可能句法分析的误差概率任意小。在基于 Monte Carlo 的消歧过程中，首先利用一种随机优先搜索，基于各种派生的概率，从句法分析森林中随机选择一种派生。然后，通过迭代生成足够多的大量随机派生样本，能够尽可能精确地估计最可能的句法分析，即从这些随机派生中所产生的具有最高频率的句法分析。根据大数定律，生成频率最高的句法分析收敛于最可能的句法分析，即该收敛属性可用于估计句法分析树的 DOP-概率。

4.2　派生集合的随机抽样

有关句法分析森林中派生的抽样，是通过采取一种自底向上的宽度优先搜索(或者自顶向下的深度优先搜索)方式来实现派生的随机选择。基于所构建的描述句法分析森林的带标短语图结构，在其中的每一个共享结点处随机选择一种派生。该选择过程基于每一种派生的概率进行。如果某一种派生的概率大于另一种派生的概率，则前者被选择的几率大于后者。一旦随机选择到达 S 结点，可通过回溯在选择路径上每一共享结点处所做出的随机选择，来生成有关输入句子的随机派生，从而可以获取由 S 结点开始的所有派生，构成对应输入句子的派生集合，即样本空间，从中进行直接的随机抽样。

对于一个输入句子而言，可能存在按指数级别增长的大量派生，对应着随机抽样时间也将按指数级别增长。因此，需要建立一种约束准则，即在具有歧义的每个结点处进行自底向上的抽样过程中，必须将每一共享结点处的不同派生数目约束为一个常量。从而，确保对于为输入句子生成一种随机派生与为其找到最可能派生，两者渐进时间复杂度量级相等，即可在立方级别的时间内进行随机派生的抽样。

有关派生集合的随机抽样算法如算法 1 所示。

算法 1　派生集合的随机抽样算法

FUNCTION Random_Sampling(G: STSG, S: 输入语句, Ω: 基本树库, Ψ: 带标短语结构图)
DEFINE
　　Num(S): 输入句子 S 中的单词个数;
　　Φ: 弧所包含的基本树中可替换根结点集合;
　　Num(Φ): 集合中可替换根结点个数;
　　∑: 对应输入语句的所有派生集合(句法分析森林);
BEGIN
1. 基于 STSG G、基本树库 Ω 与输入语句 S，利用可替换根结点链接相关的基本树，构建描述链接关系的弧——二元组(i, j)，存储于带标短语结构图 Ψ 中;
2. 从基本树库 Ω 中抽取各基本树的根结点信息，构建基本树库 Ω 的可替换根结点集合 Root(Ω);
3. For i = 1 to Num(S)
　{

维普资讯 http://www.cqvip.com

```
4.      For j =0 to(Num(S)-i)
        {
5.          从带标短语结构图 Ψ 中抽取弧——二元组(j,j+i);
6.          构建该弧所包含的基本树中可替换根结点集合 Φ;
7.          For k =1 to Num(Φ)
            {
8.              随机选择根结点 Φ[k]的一种派生,消除其它派生;
9.              IF(Φ[k]= =S) THEN
                {
10.                 回溯选择路径上每一共享结点处所做出的随
机选择;
11.                 生成对应于输入句子的由 S 结点开始的随机
派生,保存在Σ中;
                }
            }
        }
    }
12. RETURN Σ;
END
```

4.3 句法分析概率的估计

在通常情况下,考虑句法分析消歧作为寻求最可能句法分析的计算过程。实际上,该过程主要针对各种可能的句法分析进行概率评估。当通过统计方法估计句法分析的概率分布时,必须建立一种评价机制,即利用概率估计中所存在的显著错误的概率,来评价概率估计的可信度。因此,基于派生集合的随机抽样算法,从句法分析森林中随机选择一种派生,通过该派生所产生的句法分析树,形成有关最可能句法分析的初始候选者。之后,通过不断地随机抽样大量派生,来计算可信度较高的候选者,并从中选择产生频率最高的句法分析。

假设$\{(e_1,p_1),(e_2,p_2),\cdots,(e_n,p_n)\}$为所获取分析 e_1, e_2, …,e_n 的概率分布,其中一种句法分析 e_i 的概率为 p_i。基于该句法分析在包含 N 个独立样本的样本空间中出现的频率,来对其进行概率估计。由于 $0<p_i\leqslant 1$,则估计概率的方差为 $p_i(1-p_i)/N$,且该值一般小于或等于 $1/(4N)$。因此,标准误差 σ 为方差的平方根,经常小于或等于 $1/(2\sqrt{N})$。如果给定 σ 的上界,可通过 $N\geqslant 1/(4\sigma^2)$ 来计算 N 的下界。例如,如果 $N\geqslant 100$,可获得标准误差为 $\sigma\leqslant 0.05$。根据这些参数的设定,可从概率分布中选择具有最高概率的句法分析。

4.4 随机抽样的统计受控

在句法分析消歧过程中,不可避免地会存在选择错误的情形,即通过抽样派生所生成的频率最高的句法分析,实际上并非最可能的句法分析。因此,需要考虑随机抽样的统计受控问题。在确定的样本数目之中,能够确保抽样集合中出现频率最高的句法分析树,就是实际中的最可能句法分析结果。

从计算错误概率的角度出发,即选择错误的句法分析树作为最可能句法分析的概率,该概率被称为最可能句法分析误差或者 MPP 误差,其上界由 $\sum_{i\neq 0}(1-(\sqrt{p_0}-\sqrt{P_i})^2)^N$ 决定。其中,不同的 i 值表示对应于各种不同句法分析的索引,$i=0$ 时表示对应于最可能句法分析的索引;p_i 表示索引为 i 的句法分析的概率;N 为随机抽样的派生数。从该式可以看出,随着 N 的不断增大,错误概率的上界不断变小,发生错误现象的可能性也在不断减低。但是,如果存在一个索引为 i 的句法分析,其 p_i 接近于 p_0,即可能存在几乎相同的不同句法分析,则必须使 N 非常大,才能达到减小错误上界的目的。

特别需要注意的是,可能存在一种情形,如果最可能句法分析不唯一,则随机抽样与概率估计过程将不会收敛于一种结果。在一种较为极端的情形中,如果一些最可能句法分析的概率非常接近却不相等,各种分析几乎同样非常可能,则必须考虑整个最可能句法分析集合作为相关结果。

有关完整的最可能句法分析的估计算法如算法 2 所示。

算法 2 最可能句法分析的估计算法

```
FUNCTION MPP_Estimation(Σ:句法分析森林,Error_Threshold:
误差阀值)
DEFINE
    Num(Σ):句法分析森林中的句法分析数;
    MPP:最可能句法分析;
    σ:句法分析的标准误差;
    ε:MPP 误差;
BEGIN
1. 计算句法分析的标准误差,存放于 σ 中;
2. 计算 MPP 误差,存放于 ε 中;
3. While((σ≤Error_Threshold)&&(ε≤Error_Threshold))
  {
4.      Σ= Random_Sampling(G,S,Ω,Ψ);//通过派生集合的随
        机抽样,来获取句法分析森林
5.      For i =0 to Num(Σ)
        {
6.        MPP=具有最大频率的句法分析;
7.        重新计算 σ 与 ε;
        }
  }
8. RETURN MPP;
END
```

结束语 本文所介绍的 DOP 框架下的句法分析消歧算法,以 Monte Carlo 数值方法作为主要构建基础。其中,通过基于句法分析空间的近似估计,以删除相对不可能的句法分析,即所谓的伪句法分析,从而允许快速且可靠地近似估计句法分析森林中各棵树的 DOP-概率,以此来计算最可能的句法分析。

实际上,人类对于语句的理解,在经历相当长的一段时间之后,感知一些不合理的分析,通常具有很大难度来,甚至根本不能感知某些不合理的分析[15]。而基于 Monte Carlo 的分析消歧过程与人类的感知过程非常匹配,其中使用可能分析的整个样本空间,并倾向于仅抽样最可能的一些样本。在经过相当长的处理时间之后,也许仅抽样出非常不可能的分析,并且不能保证在有限时间内找到所有分析。

另一方面,在人类的感知过程中,经常可能存在一种极端情形,即针对一个句子感知到多种不同分析,并且这些分析的合理性等同(主要归因于消歧信息的缺乏)[16]。而在 Monte Carlo 消歧过程中,对应着如果样本分布中存在着几乎同样可能的多种分析,则存在多种最可能分析,造成抽样数目变大,使得样本估计不能仅收敛于唯一一种分析。但是,由于 Monte Carlo 消歧不要求针对不同的样本序列给出相同结果,则在该极端情形中,保持多种可能性选择,从而有效地进行最可能分析的估计。

综上所述,Monte Carlo 消歧方法可根据描述问题的概率密度函数(Probability Density Function, PDF),通过进行大量的随机抽样,获取样本估计结果,并且当样本数量增长至无穷大时,使得估计器的计算最终收敛于所期望的结果。同时,这种方法具有一个绝对错误估计,该估计值以 $N^{-1/2}$ 递减,而在其它方法的最好情形中通常是以 $N^{-1/M}$ 递减,因此该属性也成为利用 Monte Carlo 方法进行 DOP 分析消歧的主要优势所在。

参 考 文 献

1 Scha R, Bod R. Computationele Esthetica. Informatie en Informatiebeleid, 1993, 11(1)

2 Bod R. Data-oriented Parsing(DOP). In: Proc. of COLING'92, Nantes, France, 1992

3 Bod R. A Computational Model of Language Performance: Data Oriented Parsing. Computational Linguistics in the Netherlands 1991, Amsterdam, 1992

4 Bod R. Data Oriented Parsing as a General Framework for Sto-

维普资讯 http://www.cqvip.com

chastic Language Processing. Sikkel K, Nijholt A, eds. Parsing Natural Language, TWLT6, Twente University, 1993
5 Bod R. Using an Annotated Corpus as a Stochastic Grammar. Proceedings of EACL'93, Utrecht, The Netherlands, 1993
6 Bod R. Mathematical Properties of the Data Oriented Parsing Model. Preprints Third Meeting on Mathematics of Language, Austin, Texas, 1992
7 Sima'an K. Computational Complexity of Probabilistic Disambiguation by Means of Tree-Grammars. In: Proc. COLING-96, Copenhagen, 1996
8 Bod R. Monte Carlo Parsing. Recent Advances in Parsing Technology. Kluwer Academic Publishers
9 Sima'an K. An Optimized algorithm for Data Oriented Parsing. In: Proc. International Conference on Recent Advances in Natural Language Processing, Tzigov Chark, Bulgaria, 1994
10 Goodman J. Parsing Algorithms and Metrics. In: Proc. of th 34th Annual Meeting of the ACL, June, 1996
11 Green P J. Overview: Monte Carlo methods. In: Complex Stochastic Systems and Engineering. Titterington D M, ed. IMA Conference Series, No. 54, Oxford Univ, Press, 1995. 183~190
12 Hammersley J, Handscomb D. Monte Carlo Methods. Chapman and Hall, 1964
13 Robert C, Casella G. Monte Carlo Statistical Methods. New York: Springer-Verlag, 1999
14 Mayer H. Symposium on Monte Carlo Methods. New York Wiley: 1956
15 Nicol J, Pickering M. Processing Syntactically Ambiguous Sentences: Evidence from Semantic Priming. Journal of Psycholinguistic Research, 1993
16 MacDonald M, Pearlmutter N, Seidenberg M. Lexical Nature of Syntactic Ambiguity Resolution. Psychological Review, 1994, 101: 676~703

(上接第 166 页)

2) $MinSup(K_1, K_2)$ 表示结点 K_1、K_2 的最近公共父结点；

3) 若 K_i 是 K_j 的子结点，则记 $K_i < K_j$。

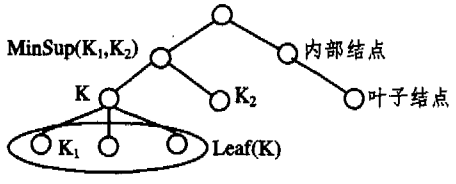

图 1 范例类的层次图

为计算类间相似性，需为类层次树中的每个内部结点 K_i 分配一个数值 $s_i \in [0 \cdots 1]$，用来表示以该结点为根的子树中所有子结点相似度的最小值；并且要求满足如下的条件：若 $K_i < K_j$，则要求 $s_i \geqslant s_j$。

进一步规定，若 x, y 为内部结点 K_i 的叶子结点，则 x, y 结点的类间相似性需满足如下的条件：$\forall x, y \in Leaf(K_i) \Rightarrow Sim_{Inter}(x, y) \geqslant s_i$。

3.2.1 类间相似性计算模型

假设有两个结点 K_i、K_j，类间相似性的计算主要有三种情形：

1) 两个结点 K_i、K_j 都为叶子结点；

2) 结点 K_i 为叶子结点，节点 K_j 为内部结点；

3) 结点 K_i、K_j 都为内部结点。

为简便起见，可将这三种情形统一起来，利用其公共父结点的相似度值加以表示。可表示为：

如果 $K_i = K_j$，则 $Sim_{Inter}(K_i, K_j) = 1$；

如果，$K_i \neq K_j$，则 $Sim_{Inter}(K_i, K_j) = s_{MinSup(K_i, K_j)}$。

3.2.2 范例间相似性计算

在计算类内相似性和类间相似性之后，就可计算两范例的相似性，通过采用类内相似性和类间相似性的乘积加以计算。即：

$$Sim(q, c) = Sim_{Inter}(q, c) \times Sim_{Inter}(q, c)$$

结束语 本文结合 CBR 的具体特点，先给出范例推理中面向对象的范例表示一般方法。这种表示方法，可通过类的层次信息，表示范例间复杂的结构信息。它一方面克服了属性-值对表示法无法表示复杂范例的缺陷，另一方面通过约束规则将领域知识与范例的表示结合起来。同时在范例的相似性计算上，区分两类不同的相似性即类内相似性和类间相似性，使得范例相似性计算尽可能地准确和科学。

参 考 文 献

1 Schank R. Dynamic Memory: A Theory of Reminding and Learning in Computers and People. Cambridge, UK: Cambridge University Press, 1982
2 Juell P, Paulson P. Case-based systems. Intelligent Systems, IEEE, 2003, 18(4): 60 ~67
3 Watson I, Marir F. Case-Based Reasoning: A Review. The Knowledge Engineering Review, 1994, 9(4): 327~54
4 Zheng Quan, Wang Junpu, Wu Xiuqing, at el. Intelligent Control and Automation. In: Proceedings of the 3rd World Congress on, 2000, 1: 242 ~245
5 Fu Yonggang, Shen Ruimin. GA based CBR approach in Q&A system. Expert Systems with Applications, 2004, 26(2): 167~170
6 Smyth B, Keane M T. Remembering To Forget: A Competence-Preserving Case Deletion Policy for Case-Based Reasoning Systems. In: International Joint Conference on Artificial Intelligence (IJCAI), 1995. 377~383
7 Watson I, Perera S. A hierarchical case representation using context guided retrieval. Knowledge-Based Systems, 1998, 11: 285~292
8 Graham I. 面向对象的方法：原理与实践. 北京：机械工业出版社，2003
9 曹元大，徐漫江. 面向对象知识表示在专家系统开发工具中的应用. 北京理工大学学报，2000，20(6)：688~692
10 钱梅，汪斌，吴耿锋，等. 面向对象模型对复杂结构知识的知识表示方法. 计算机工程与科学，2002，24(3)：62~65
11 Pree W, Gamma E. Design Patterns for Object-Oriented Software Development. Addison-Wesley Pub Co, 1995