一种基于加权相似性的粗糙集数据补齐方法

赵洪波[1] 江 峰[1] 曾惠芬[2] 高 宏[3]
(青岛科技大学信息科学与技术学院 青岛 266061)[1] (九江职业技术学院 九江 332007)[2]
(91286部队气象台 青岛 266003)[3]

摘 要 近年来,对不完备数据的处理引起了人们的广泛关注。目前,在粗糙集理论中已经提出了多种不完备数据补齐方法,这些方法通常需要计算决策表中具有缺失值的对象与其他没有缺失值的对象之间的相似性,并以最相似对象的取值来代替缺失值。然而,这些方法普遍存在一个问题,即在计算决策表中对象之间的相似性时假设决策属性对所有条件属性的依赖性都是相等的,而且所有条件属性都是同等重要的,并没有考虑不同条件属性之间的差异性。针对这一问题,引入一个加权相似性的概念,以决策属性对条件属性的依赖性和条件属性的重要性作为权值来计算相似性。基于加权相似性,提出一种新的粗糙集数据补齐算法WSDCA。最后,在UCI数据集上,将WSDCA算法与现有的数据补齐算法进行了比较分析。实验结果表明,所提出的数据补齐方法是有效的。

关键词 粗糙集,不完备数据,数据补齐,相似性,加权相似性
中图法分类号 TP39 **文献标识码** A

Rough Set Approach to Data Completion Based on Weighted Similarity

ZHAO Hong-bo[1] JIANG Feng[1] ZENG Hui-fen[2] GAO Hong[3]
(College of Information Science and Technology, Qingdao University of Science and Technology, Qingdao 266061, China)[1]
(Jiujiang Vocational and Technical College, Jiujiang 332007, China)[2]
(91286 Army Weather Station, Qingdao 266003, China)[3]

Abstract In recent years, much attention has been given to the treatment of incomplete data. By now, many completion methods to incomplete data have been proposed in rough set theory. These methods usually compute the similarities between the object that contains missing values and other objects that do not contain missing values, and use the values of the most similar object to replace the missing values. However, there is a common problem for these methods. That is, these methods assume that the dependencies of decision attribute on all condition attributes are the same, and the significances of all condition attributes are also the same, they ignore the differences between different condition attributes in a decision table. To solve this problem, in this paper we introduced a new notion of weighted similarity, which employs the dependencies of decision attribute on condition attributes and the significances of condition attributes as weights to compute the similarity. Based on the weighted similarity, we proposed a novel rough set data completion algorithm WSDCA. We compared WSDCA with the current data completion algorithms on UCI data sets. And experimental results demonstrate the effectiveness of our method to data completion.

Keywords Rough sets, Incomplete data, Data completion, Similarity, Weighted similarity

1 前言

自20世纪80年代Pawlak提出粗糙集理论以来,粗糙集已在数据挖掘等领域发挥着重要的作用[1-5]。目前,粗糙集中所提出的属性约简、规则生成、离散化等算法都是基于完备的信息表。然而现实世界的数据通常是不完备的,即存在缺失的值。例如,在统计某个公司职员的家庭成员情况时,对于未婚者来说,配偶姓名、子女姓名等信息肯定是缺失的[10]。

针对现实世界中存在的大量不完备数据,我们需要采取相应的办法来加以处理。目前,最常用的处理办法就是对这些不完备数据按照一定的策略进行填补,即将缺失的值补齐。近年来,对于不完备数据的补齐问题逐渐受到重视,已经提出了多种数据补齐方法。大体上看,现有的数据补齐方法可分为两类[10]:(1) 基于统计的方法,例如平均值补齐、条件平均值补齐、组合化补齐、条件组合化补齐等[3];(2) 基于粗糙集的方法[6-9,11-13,15-16],例如ROUSTIDA算法[1]。基于统计学方法既可以对连续型数据补齐又可以对离散型数据补齐,很好地保证了数据的分布规律,但忽略了数据属性之间以及对

到稿日期:2010-12-14 返修日期:2011-03-21 本文受国家自然科学基金(60802042),山东省自然科学基金(ZR2009GQ013,ZR2010FQ027)资助。

赵洪波 硕士生,主要研究方向为数据挖掘、粗糙集理论等;江 峰 博士,副教授,主要研究方向为人工智能、粗糙集理论等,E-mail:jiangkong@163.net(通信作者)。

象之间的相互关系；基于粗糙集的方法则充分考虑了数据属性之间和对象之间的相互关系，从而使得补齐的数据更接近现实数据，更具有真实性[10]。

随着对粗糙集理论研究与应用的不断深入，越来越多地提出基于粗糙集的数据补齐算法。目前，最常用的基于粗糙集的数据补齐算法有 ROUSTIDA 算法[1] 以及对 ROUSTIDA 算法的一些改进[11-13,15,16]。此类算法的主要思想是：通过计算决策表中具有缺失值的对象与其他对象之间的相似性，并以最相似对象的取值来补齐缺失值。假设决策表中对象 o 在某个属性 a 上的取值空缺，那么这些算法将在决策表中寻找与 o 最相似的对象，并以该对象在属性 a 上的取值作为对象 o 在 a 上的取值。然而，这些算法普遍存在一个问题，即在计算决策表中对象之间的相似性时假设决策属性对于每个条件属性的依赖度都是一样的，而且每个条件属性都是同等重要的，并没有考虑不同属性之间所存在的差异性。在一个给定的决策表中，不同的条件属性对于决策的贡献即属性重要性在很多情况下是不同的。另外，决策属性对于不同的条件属性的依赖度也是不同的。在计算对象之间的相似性时，有必要考虑到这两个因素，将不同的条件属性区别开来。

针对上述问题，本文将重新定义对象之间相似性的概念，提出一个加权相似性的概念。与传统的基于粗糙集的数据补齐方法中所采用的相似性定义不同，加权相似性将充分考虑决策表中不同属性之间的差异性。在计算对象之间的相似性时，不同的条件属性将按照决策属性对其的依赖度以及该属性自身的重要性区别开来。具体来说，对于任意一个条件属性 a，以决策属性对于 a 的依赖性和 a 本身的重要性作为 a 的权重，在计算对象之间的相似性时，不同的条件属性将按照其权重的大小发挥不同程度的作用，从而将不同的条件属性区别开来。通过计算对象之间的加权相似性，本文将给出一种新的粗糙集不完备数据补齐方法。

2 粗糙集理论的基本知识

定义 1(信息表) 信息表是一个四元组 $IS=(U,A,V,f)$。其中[2]，

(1) U 是一个非空有限的对象集合；

(2) A 是一个非空有限的属性集合；

(3) V 是所有属性论域的并，即 $V=\bigcup_{a\in A}V_a$，其中 V_a 为属性 a 的值域；

(4) $f:U\times A\to V$ 是一个信息函数，使得对任意 $a\in A$ 以及 $x\in U$，$f(x,a)\in V_a$。

进一步，属性集 A 又可以划分为两个不相交的子集——条件属性集 C 和决策属性集 D，即 $A=C\cup D$。这种特殊的信息表被称为决策表，简记 $DT=(U,C,D,V,f)$。

定义 2(不可分辨关系) 给定一个决策表 $DT=(U,C,D,V,f)$，对于任意 $B\subseteq C\cup D$，定义由 B 所决定的一个不可分辨关系 $IND(B)$ 为[1,2,4]

$$IND(B)=\{(x,y)\in U\times U:\forall a\in B(f(x,a)=f(y,a))\}$$

可以证明，对任意 $B\subseteq C\cup D$，不可分辨关系 $IND(B)$ 是 U 上的一个等价关系，并且 $IND(B)=\bigcap_{a\in B}IND(\{a\})$。

定义 3(相似关系) 给定一个决策表 $DT=(U,C,D,V,f)$，对于任意 $B\subseteq C$，定义由 B 所决定的一个相似关系 $N(B)$ 为[1,2,4]

$$N(B)=\{(x,y)\in U\times U:\exists a\in B(f(x,a)=f(y,a))\}$$

定义 4(相对正区域) 给定一个决策表 $DT=(U,C,D,V,f)$，对任意 $B\subseteq C$，定义决策属性集 D 的 B 正区域 $Pos_B(D)$ 为[1,2]

$$Pos_B(D)=\bigcup\{Y\mid Y\subseteq X,X\in U/IND(D),Y\in U/IND(B)\}$$

定义 5(属性依赖度) 给定一个决策表 $DT=(U,C,D,V,f)$，对于任意 $B\subseteq C$，决策属性集 D 对条件属性集 B 的依赖度定义为[1,17,18]

$$\gamma_B(D)=Card(Pos_B(D))/Card(U)$$

式中，$Card(M)$ 表示集合 M 的基数。

定义 6(属性重要性) 给定一个决策表 $DT=(U,C,D,V,f)$，对于任意条件属性 $a\in C$，属性 a 在集合 C 中相对于决策属性集 D 的重要性定义为[1]

$$SGF(a,C,D)=\gamma_C(D)-\gamma_{C-\{a\}}(D)$$

3 基于加权相似性的不完备数据补齐算法

现有的基于粗糙集的数据补齐算法普遍存在一个问题，即在计算决策表中对象之间的相似性时假设决策属性对于每个条件属性的依赖度都是一样的，而且每个条件属性都是同等重要的，并没有考虑不同属性之间的差异性。而在计算对象之间的相似性时，非常有必要将不同的条件属性区别开来。因为任意两个对象之间的相似性主要是通过比较这两个对象在每个条件属性上的取值而得到的。如果我们在计算相似性时，采用同一种方式来处理决策表中的所有条件属性，明显是不合理的，而且最终将导致数据补齐结果存在偏差。

为了解决此问题，本文将重新定义对象之间相似性，提出一个加权相似性的概念。与传统的基于粗糙集的数据补齐算法中所采用的相似性定义不同，加权相似性将充分考虑决策表中不同条件属性之间的差异性，在计算对象之间的相似性时不同的条件属性将按照其权重的大小发挥不同的作用。但是，如何在决策表 $DT=(U,C,D,V,f)$ 中有效地度量 C 中各个属性之间的差别呢？即如何为 C 中的每个属性合理地分配一个权值，这是我们在实际应用中需要考虑的问题。

由上一节所给出的定义 5 和定义 6 可知，在一个给定的决策表 $DT=(U,C,D,V,f)$ 中，对于任意条件属性 $a\in C$ 来说，可以采用以下两种方法来度量 a 在整个决策表决策分类中所发挥的作用：决策属性集 D 对 a 的依赖度以及 a 在条件属性集 C 中相对于 D 的重要性。这两种度量方法分别从不同的角度刻画了 a 对于整个决策表决策分类的影响，前者着重考虑属性 a 作为单个属性对于决策分类的影响，而后者则着重考虑属性 a 作为条件属性集 C 中的一员，它在决策分类过程中相对于 C 中的其他成员而言所起的作用[1,17,18]。

可以看出，上述两种度量方法是互补的。为了更加全面有效地度量属性 a 在整个决策表决策分类中所发挥的作用，有必要同时引入这两种度量方法来计算 a 的权值。

下面将以属性依赖性和属性重要性作为属性权值计算的依据，从而给出一种决策表中对象之间相似性的新定义。

定义 7(加权相似性) 给定一个决策表 $DT=(U,C,D,V,f)$，对于任意属性 $a\in C$，令 $\gamma_{\{a\}}(D)$ 和 $SGF(a,C,D)$ 分别表示决策属性集 D 对于 a 的依赖度以及 a 在 C 中相对于决

策属性集 D 的重要性。对于任意 $x,y\in U$，x 与 y 之间的加权相似性可定义为

$$WS(x,y)=\sum_{a\in C}(1+\gamma_{\{a\}}(D)+SGF(a,C,D))\times h_a(x,y)$$

式中，$h_a:U\times U\to\{0,1\}$ 是一个从 $U\times U$ 到 $\{0,1\}$ 的函数，使得对任意 $(x,y)\in U\times U$，如果 $f(x,a)=f(y,a)$，则 $h_a(x,y)=1$；否则 $h_a(x,y)=0$。

由定义 7 可知，在计算两个对象之间的加权相似性时，不同的条件属性将按照决策属性对其的依赖度以及该属性自身的重要性区别开来。具体来说，对于任意条件属性 a，以决策属性对于 a 的依赖性和 a 本身的重要性作为 a 的权重。在计算对象之间的相似性时，不同的条件属性将按照其权重的大小发挥不同程度的作用，从而被区别开来。

通过计算对象之间的加权相似性，本文将提出一种新的粗糙集不完备数据补齐方法。可以看出，本文所提出的基于加权相似性的数据补齐方法是对传统的基于粗糙集的数据补齐方法的一种有效改进。

定理 1 给定一个决策表 $DT=(U,C,D,V,f)$，对于任意 $x,y\in U$，如果 x 与 y 在属性集 C 上具有相似关系，即 $(x,y)\in N(C)$，则 $WS(x,y)>0$；否则 $WS(x,y)=0$。其中 $WS(x,y)$ 表示 x 与 y 之间的加权相似性。

证明：如果 x 与 y 在属性集 C 上具有相似关系，由定义 3 可知，至少存在一个属性 $a\in C$，使得 $f(x,a)=f(y,a)$。再由定义 7 可知，至少存在一个属性 $a\in C$，使得 $h_a(x,y)=1$。另一方面，由定义 5 和定义 6 可知，$\gamma_{\{a\}}(D)\geqslant 0$ 且 $SGF(a,C,D)\geqslant 0$，因此 $1+\gamma_{\{a\}}(D)+SGF(a,C,D)>0$。这样，至少存在一个属性 $a\in C$，使得 $(1+\gamma_{\{a\}}(D)+SGF(a,C,D))\times h_a(x,y)>0$。因此，我们可以得出 $WS(x,y)>0$。

反过来，如果 x 与 y 在属性集 C 上不具有相似关系，则由定义 3 可知，对于任意 $a\in C$，都有 $f(x,a)\neq f(y,a)$。再由定义 7 可知，对于任意 $a\in C$，都有 $h_a(x,y)=0$。这样，对于任意 $a\in C$，都有 $(1+\gamma_{\{a\}}(D)+SGF(a,C,D))\times h_a(x,y)=0$。因此，我们可以得出 $WS(x,y)=0$。

接下来，将给出不完备数据补齐算法 WSDCA。

算法 1 WSDCA

输入：不完备决策表 $DT=(U,C,D,V,f)$，其中 $|U|=n$，$C=\{a_1,a_2,\cdots,a_m\}$，$D=\{d\}$

输出：完备的决策表 $DT'=(U,C,D,V',f')$

(0)将决策表 DT 中的所有空缺值暂时替换为一个特殊值 $*$。

(1)对于条件属性集 C，执行如下操作：

(1.1)根据 U 中对象在属性集 C 上的取值，按照值域 V_C 上的一个给定次序(例如字典序)，对 U 中的所有对象进行基数排序[19]；

(1.2)求出划分 $U/IND(C)$；

(1.3)计算决策属性集 D 相对于 C 的正区域 $Pos_C(D)$。

(2)对于任意属性 $a\in C$，循环执行如下操作：

(2.1)根据 U 中对象在 $C-\{a\}$ 上的取值，按照值域 $V_{C-\{a\}}$ 上的一个给定次序，对 U 中的对象进行基数排序；

(2.2)求出划分 $U/IND(C-\{a\})$；

(2.3)计算 D 相对于 $C-\{a\}$ 的正区域 $Pos_{C-\{a\}}(D)$；

(2.4)根据论域 U 中对象在属性 a 上的取值，按照值域 V_a 上的一个给定次序，对 U 中对象进行基数排序；

(2.5)求出划分 $U/IND(\{a\})$；

(2.6)计算决策属性集 D 相对于 $\{a\}$ 的正区域 $Pos_{\{a\}}(D)$；

(2.7)计算决策属性集 D 对于属性 a 的依赖度 $\gamma_{\{a\}}(D)$；

(2.8)计算属性 a 在 C 中相对于决策属性集 D 的重要性 $SGF(a,C,D)$；

(2.9)计算属性 a 的权值。

(3)对于任意 $x\in U$，循环执行如下操作：

对于任意 $a\in C$，如果 $f(x,a)=*$，则

(I)在 U 中找出所有在属性 a 上的取值不等于 $*$，并且在决策属性 d 上的取值等于 $f(x,d)$ 的对象，令集合 S 来表示所有这些对象；

(II)计算对象 x 与 S 中每个对象之间的加权相似性；

(III)在 S 中找出与 x 的加权相似性最大的对象 max；

(IV)令 $f(x,a)=f(\max,a)$。

(4)算法结束，返回补齐后的决策表 DT。

在算法 1 中，我们采用了一种预先对 U 中对象进行基数排序，然后再求划分 $U/IND(B)$ 的方法[19]，此方法的时间复杂度为 $O(|B|\times n)$。因此，通过使用此方法可以有效降低计算划分 $U/IND(B)$ 的复杂度，进而降低整个算法的复杂度。

在最坏的情况下，算法 1 中步骤(0)的时间复杂度为 $O(m\times n)$，步骤(1)的时间复杂度为 $O(m\times n)$，步骤(2)的时间复杂度为 $O(m^2\times n)$，步骤(3)的时间复杂度为 $O(m\times n\times k)$，其中 m 和 n 分别表示集合 C 与 U 的势，k 表示不完备决策表 DT 中空缺值的个数。因此，当 $k<m$ 时，算法 1 的时间复杂度为 $O(m^2\times n)$；当 $k>m$ 时，算法 1 的时间复杂度为 $O(m\times n\times k)$。另外，算法 1 的空间复杂度为 $O(m+n)$。

据统计，在实际使用的数据库中，空缺值占整个数据库的比例一般不超过 4%到 5%[10]，因此本算法的时间复杂度是可以接受的。

4 实验

下面通过实验来验证算法 WSDCA 对于不完备数据的补齐能力。实验所采用的数据集有两个：Voting 数据集和 Roth 数据集[20]。在 Voting 数据集上，分别比较 Mean-Completer、Conditioned-Mean-Completer 和 WSDCA 这 3 种数据补齐算法的性能，其中 Mean-Completer 和 Conditioned-Mean-Completer 算法均来自于 ROSETTA 软件[21]。在 Roth 数据集上，分别比较 Conditioned-Combinatorial-Completer 算法、Conditioned-Mean-Completer 算法和 WSDCA 这 3 种数据补齐算法的性能，其中 Conditioned-Combinatorial-Completer 算法也来自于 ROSETTA 软件[21]。另外，对于经过不同补齐算法补齐之后的数据集，分别采用 WEKA 软件中所提供的各种分类算法(包括 Id3、J48 和 J48graft 分类器)进行分类测试[22]，以便于比较不同的补齐算法对于数据集分类性能的影响。

实验的基本流程如下：

(1)使用 ROSETTA 中的数据补齐算法对不完备的 Voting和 Roth 数据集进行补齐；

(2)基于 Delphi 开发工具实现 WSDCA 算法，并利用该算法对 Voting 和 Roth 数据集进行补齐；

(3)将前面两步所产生的、由不同的数据补齐算法补齐之后的数据集分别导入到 WEKA 中，并且采用 WEKA 中的多种分类算法进行分类性能测试，从而可以比较各种数据补齐算法的优劣。

4.1 Voting 数据集

首先对 Voting 数据集进行实验。该数据集包括 435 个

对象、16 个条件属性和 1 个决策属性，并且含有 392 个缺失值[20]。实验中，我们直接补齐 Voting 数据集中的缺失数据。具体的实验步骤如下：

首先，利用 ROSETTA 软件中的 Mean-Completer 算法和 Conditioned-Mean-Completer 算法分别对 Voting 数据集进行补齐操作，从而得到 Voting-MC-Completed 和 Voting-CMC-Completed 这两个补齐之后的数据集。

其次，利用 WSDCA 算法对 Voting 数据集进行补齐操作，从而得到 Voting-WSDCA-Completed 这个补齐之后的数据集。

最后，利用 WEKA 中的 Id3、J48 和 J48graft 这 3 种分类器分别对 Voting-MC-Completed、Voting-CMC-Completed 和 Voting-WSDCA-Completed 这 3 个数据集进行分类训练和测试，其中测试模式为 Percentage split(66% for training)，即从每个数据集中随机抽取 66%的数据作为训练集，其余 34%的数据作为测试集。

具体的实验结果如表 1 所列。

表 1　Voting 数据集上的结果

采用不同补齐算法补齐之后的数据集	分类精度(%)		
	Id3	J48	J48graft
Voting-CMC-Completed	95.2703	95.2703	95.2703
Voting-MC-Completed	94.596	96.6216	95.9459
Voting-WSDCA-Completed	97.2973	97.973	97.973

在表 1 中，Voting-CMC-Completed、Voting-MC-Completed 和 Voting-WSDCA-Completed 分别表示采用 Conditioned-Mean-Completer 算法、Mean-Completer 算法和 WSDCA 算法补齐之后的 Voting 数据集。Id3、J48 和 J48graft 分别表示 WEKA 中所提供的 3 种分类算法。

从表 1 可以看出，采用 WSDCA 算法补齐之后的数据集 Voting-WSDCA-Completed 在 Id3、J48 和 J48graft 这 3 种分类器下的分类精度明显都要高于另外两个补齐算法所补齐的数据集。因此，对于 Voting 数据集，采用 WSDCA 算法进行数据补齐，可以获得更好的分类性能。

4.2　Roth 数据集

Roth 数据集中包括 160 个对象、4 个条件属性和 1 个决策属性，但不含有缺失值[20]。实验中，为了验证数据补齐算法的性能，从 Roth 中随机删除了一些条件属性值(占整个数据集的 10%)，从而人为地得到一个不完备数据集“Roth 空缺数据集”。我们对该数据集中的缺失数据进行补齐。具体实验步骤如下：

首先，利用 ROSETTA 中的 Conditioned-Mean-Completer 和 Conditioned-Combinatorial-Completer 算法对 Roth 空缺数据集进行补齐，从而得到 Roth-CMC-Completed 和 Roth-CCC-Completed 这两个补齐之后的数据集。

其次，利用 WSDCA 算法对 Roth 空缺数据集进行补齐操作，从而得到 Roth-WSDCA-Completed 这个补齐之后的数据集。

最后，利用 WEKA 软件中的 Id3、J48 和 J48graft 这 3 种分类器分别对 Roth-CMC-Completed、Roth-CCC-Completed 和 Roth-WSDCA-Completed 这 3 个数据集进行分类训练和测试，其中测试模式依然为：Percentage split(66% for training)。

具体的实验结果如表 2 所列。

表 2　Roth 数据集上的结果

采用不同补齐算法补齐之后的数据集	分类精度(%)		
	Id3	J48	J48graft
Roth-CMC-Completed	61.111	64.8148	64.8148
Roth-CCC-Completed	60	68	68.8
Roth-WSDCA-Completed	77.778	75.9259	75.9259

从表 2 可以看出，采用 WSDCA 算法补齐之后的数据集 Roth-WSDCA-Completed 在 Id3、J48 和 J48graft 这 3 种分类器下的分类精度明显都要高于另外两个补齐算法所补齐的数据集。因此，对于 Roth 数据集，采用 WSDCA 算法进行数据补齐，同样可以获得更好的分类性能。

结束语　本文针对现有的基于粗糙集的数据补齐方法在对象相似性定义上所存在的缺陷，提出了一种加权相似性的概念，并给出了一种基于加权相似性的不完备数据补齐算法 WSDCA。与传统的数据补齐方法中所采用的相似性定义不同，加权相似性充分考虑到决策表中不同属性之间的差异性。在计算对象之间的相似性时，不同的条件属性将按照决策属性对其的依赖度以及该属性自身的重要性被严格区别开来。我们通过实例演示了 WSDCA 算法的执行过程，并且通过在真实数据集上的实验验证了 WSDCA 算法的有效性。与现有算法相比，WSDCA 算法能够显著提高补齐之后数据的分类性能。

参 考 文 献

[1] 王国胤. Rough 集理论与知识获取[M]. 西安：西安交通大学出版社，2001

[2] 刘清. Rough 集及 Rough 推理[M]. 北京：科学出版社，2005

[3] Han Jia-wei，Kamber M. 数据挖掘概念与技术[M]. 范明，孟小峰，等译. 北京：机械工业出版社，2001

[4] 张文修，姚一豫，梁怡. 粗糙集理论与概念格[M]. 西安：西安交通大学出版社，2006

[5] 梁循. 数据挖掘算法与应用[M]. 北京：北京大学出版社，2006

[6] 潘巍，王阳生，杨宏戟. 粗糙集理论中新的针对不完备信息系统的处理方法研究[J]. 计算机科学，2007，134(16)：158-161

[7] 焦娜，苗夺谦，张红云. 多决策表缺失属性补齐算法的研究[J]. 计算机科学，2009，36(1)：142-145

[8] 张德喜，李晓宇. 绝对信息量不完备信息系统的数据补齐算法[J]. 计算机工程与应用，2006，42(22)：155-157

[9] 王希雷. 一种不完备决策表的数据补齐方法[J]. 天津科技大学学报，2007，22(3)：62-65

[10] 张星，郝伟. 不完备或缺失数据及其填补方法研究[J]. 福建电脑，2007，4：32-33

[11] 孟军，刘永超，莫海波. 基于粗糙集理论的不完备数据填补方法[J]. 计算机工程与应用，2008，44(6)：175-177

[12] 王国胤. Rough 集理论在不完备信息系统中的扩充[J]. 计算机研究与发展，2002，39(10)：1238-1243

[13] Kryszkiewicz M. Rough set approach to incomplete information system [J]. Information Sciences，1998，112(14)：39-49

[14] Kryszkiewicz M. Rules in incomplete information system [J]. Information Science，1999，113(3/4)：271-292

[15] 李萍，吴祈宗. 基于概率相似度的不完备信息系统数据补齐算法[J]. 计算机应用研究，2009，26(3)：881-883

(下转第 190 页)

是，LA-think 规则的功能是通过在词库中提取给定命题因子的接续因子来驱动导航(navigation)。因为词库里的每一个因子通常都有一个以上的可能接续词，所以 LA-think 必须能够在评估外部和内部刺激、已遍历频率、已知程序以及主/述位结构等等的基础上对备选可接续词进行智能选择。对于有语言功能的主体来说，导航是说者概念化的过程，即说者选择说什么和怎么说的过程。

当词库中的动词命题因子，如"军"(仍以"晋师军于庐柳"为例)由外界或者内部刺激激活时，LA-think 的起始状态 STs 也被激活。相关规则包里的规则启动，导航开始。根据"军"命题因子的接续属性[arg]和[mdr]的值，导航激活"师"和"于"命题因子，接下来再分别根据这两个命题因子的接续属性[mdr]和[arg]的值激活"晋"和"庐柳"。如下所示：

[verb:军]	[noun:师]	[verb:于]	[noun:晋]	[noun:庐柳]
[sem:vi]	[sem:cn]	[sem:prep]	[sem:nms]	[sem:nmc]
[arg:师]	[mdr:晋]	[mdd:军]	[mdd:师]	[fnc:于]
[mdr:于]	[fnc:军]	[arg 庐柳]	[prn:1]	[prn:1]
[prn:1]	[prn:1]	[prn:1]		

LA-think 的导航结果成为 LA-speak 的输入。语言生成过程中 LA-think 导航遍历到的命题因子的核心属性值被实现为外部符号，即相应的汉字和标点。除了依存于语言的词汇化过程之外，用于中文的 LA-speak 语法系统还必须根据被激活命题因子的相关属性值和彼此之间的句法语义关系来处理语序问题(如果用于其他语言，如英语，还涉及到功能词析出和一致关系等问题)，最后输出正确完整的句子。

结束语 数据库语义学的方法论原则是自然语言的字面组合性，其经验论原则是自然语言交流的时间线性，其本体论和操作性原则是语言和语境信息在认知主体内部的相互匹配。

数据库语义学的基础之一——左结合语法，其本身就是一种区别于范畴语法、依存语法和短语结构语法的特殊句法分析方法。数据库语义学的另一个基础——词库数据库也以其独特的数据结构区别于一般网络数据库，为左结合语法提供理想的运行条件。

遵循自然语言语表的时间线性组合本身大大降低了计算的复杂程度。LA 语法操作过程中的基于模式匹配的规则方法也保证了较低的计算复杂度。处理修饰语句法歧义的语义重叠(semantic doubling)和共指推理等方法也将先天论(naturlism)和转换语法(TG)的指数复杂度降到了线性复杂度[16]。目前该理论方法在德语、英语和汉语的分析实验中都取得了可喜的成果(参见 http://www.linguistik.uni-erlangen.de/clue/de/forschung.html)。角色转换模型以及左结合的句法语义分析方法更准确形象地反映了自然语言理解与生成的实际过程，以命题因子命名的特征结构也更方便地在词库中存储和提取数据。但是由于实践性操作量上的不足和相关技术的不完善性，数据库语义学理论本身及其应用都还有待进一步研究和提高。

参 考 文 献

[1] Austin J L. How to Do Things With Words[M]. Oxford, England: Oxford University Press, 1962

[2] Bar-Hillel J. Language and Information—Selected Essays on Their Theory and Application[M]. Mass: Addison Wesley and Jerusalem Academic Press, 1964

[3] Chomsky N. Syntactic Structure[M]. The Hague: Mouton & Co, 1957

[4] Fraser N. Prolegmena to a Formal Theory of Dependency Grammar[S]. UCL WPL, 1990, 2: 298-319

[5] Halliday M A K, Hasan R. Cohesion in English[M]. London: Longman, 1976

[6] Hausser R. Left-associative Grammar and the Parser NEWCAT [R]. IN-CSLI-85-5. Center for the Study of Language and Information, Stanford/CA: Stanford University, 1985

[7] Hausser R. Foundations of Computational Linguistics, Human-Computer Communication in Natural Language[M]. Berlin, New York: Springer-Verlag, 1999/2001

[8] Hausser R. Turn Taking in Database Semantics[M]// Kangassalo H, et al., eds. Information Modeling and Knowledge Bases XVI. Amsterdam: IOS Press Ohmsha, 2005

[9] Hausser R. A Computational Model of Natural Language Communication; Interpretation, Inference, and Production in Database Semantics[M]. Berlin, Heidelberg, New York: Springer, 2006

[10] Hausser R. Comparing the Use of Feature Structures in Nativism and in Database Semantics[M]// Jaakkola H, Kiyoki Y, Tokuda T, eds. Information Modelling and Knowledge Bases XIX. Amsterdam: IOS Press Ohmsha, 2007

[11] Mann W C, Thompson S A. Rhetorical Structure Theory: A theory of text organization[M]// Polanyi L, ed. The Structure of Discourse. Ablex, 1988

[12] Montague R. Formal Philosophy[M]. New Haven, CT: Yale University Press, 1974

[13] Quillian M. Semantic Memory[M]// Minsky M, ed. Semantic Information Processing. MIT Press, 1968: 227-270

(上接第 170 页)

[16] 刘伟. 基于粗集理论不完备数据的改进算法[J]. 吉林师范大学学报：自然科学版，2007，28(3)：113-114

[17] 王小菊，蒋芸，李永华. 基于依赖度之差的属性重要性评分[J]. 计算机技术与发展，2009，19(1)：68-70

[18] 韩忠华，刘春光，王长涛，等. 基于属性依赖度分析的粗糙集数据挖掘方法应用[J]. 沈阳建筑大学学报：自然科学版，2009，25(5)：1010-1013

[19] 徐章艳，刘作鹏，杨炳儒等，一个复杂度为 max($O(|C||U|)$, $O(|C|^2|U/C|)$) 的快速属性约简算法[J]. 计算机学报，2006，29(3)：391-399

[20] Bay S D. The UCI KDD repository[EB/OL]. http://kdd.ics.uci.edu, 1999

[21] Øhrn A. Rosetta Technical Reference Manual[EB/OL]. http://www.idi.ntnu.no/aleks/rosetta, 1999

[22] Hall M, Frank E, Holmes G, et al. The WEKA data mining software: an update[J]// SIGKDD Explor. Newsl., 2009, 11(1): 10-18