2003 年 6 月
第 5 卷第 6 期

中国工程科学
Engineering Science

Jun. 2003
Vol.5 No.6

学术论文

基于知识库的知识发现（KDK）的结构模型与挖掘算法研究

杨炳儒，申江涛，陈泓婕
（北京科技大学信息工程学院，北京 100083）

[摘要] 从知识库中发现新知识 KDK（knowledge discovery in knowledgebase）是一个新课题，它的成功将直接作用于大型知识库的构建，并将为解决目前机器学习的瓶颈问题——知识获取起到重要影响。笔者的主要工作是：基于知识库中事实的 KDK 归纳结构、算法及其验证；基于知识库中规则的 KDK 归纳结构、算法及其验证。

[关键词] 基于知识库的知识发现；卡尔纳普归纳逻辑；科恩归纳逻辑；假设评估

[中图分类号] TP182 [文献标识码] A [文章编号] 1009－1742（2003）06－0049－06

1 引言

如何从现有的知识库中进一步的发现出更多的深层次的知识，即 KDK，在国内外基本上无人涉足。从知识库中发现规则对于知识工程和机器学习都是一个重要的问题，因为它的成功将直接作用于知识获取和大型知识库的构建，并且发现规则对于发现大型和通常意义下的知识库可以产生怎样的机器学习程序是很有用的。

知识库中的知识发现（KDK）与数据库中的知识发现（KDD）有所不同，主要表现在：**a.** 发现的基础不同，KDK 针对的对象是知识库，一个真实的知识库一般包含事实库和规则库，它们的结构与数据库有着明显的区别；**b.** 采用的手段不同，知识库中包含着显性的关系，如何针对关系得出更高层次的知识，将采用与数据挖掘完全不同的方法。

笔者在对 KDK 初步描述的基础上，给出基于知识库的事实与规则的两类 KDK 结构模型与挖掘算法[1,2]。

2 关于 KDK 的描述

目前在国际和国内都没有明确提出过 KDK 这个概念，更没有人对它进行详尽的定义。在这里，将通过分析最终采用描述性的方法界定所研究的 KDK 的范围。

总括而论，演绎逻辑系统（包括经典的和非标准的）已在计算机科学技术中占有重要的地位。机器学习、归纳学习与不确定性推理至今仍主要是以演绎逻辑系统作为工具。由于计算机本身是演绎化的产物，而发现的核心是归纳，因此，唯一要做的是归纳的演绎化或演绎的归纳化。可以看出，发现本身既离不开归纳也离不开演绎。

首先，KDK 的目的是为了在真实的大型知识库中发现新的知识。这种发现过程借用 KDD 的语言来说是非平凡的。意即这种发现过程的核心将是归纳，而演绎将作为辅助手段；该过程不同于传统的演绎，它有可能是不保真的。

其次，KDK 能够发现深层次的知识。具体而言就是在已有属性与关系的基础上进一步发现其上的关系，从逻辑的角度上说就是发现谓词间的关系或函词间的关系。

第三，由于知识本身所可能具有的一些属性，如不确定性、非单调性、不完全性等，就决定了 KDK 过程的进行也将是一个复杂的、多方法、多途径的过程。它与知识库的组织、用户对最终寻求

[收稿日期] 2002－12－27；修回日期 2003－02－25
[基金项目] 国家自然科学基金重点资助项目（69835001）；北京市自然科学基金资助项目（4022008）
[作者简介] 杨炳儒（1943－）男，天津市人，北京科技大学教授，博士生导师

的知识类型都紧密相关，采用的推理手段可能涉及很多不同的逻辑领域。

最后，KDK 发现的知识应该是新颖的、有效的、潜在有用的、用户可理解的，这与 KDD 的要求相同。

从以上界定性的描述可以看出：KDK 究其本质应是一种机器学习过程[3]，其目的是获取知识，学习源是知识库，学习手段是用归纳结合演绎的方法，其最终结果将既能够发现事实上的知识，也能发现规则上的知识。因此，在具体的实现中，应该采用两条挖掘线路，其一是利用归纳方法发掘事实之上的规则；另一条线路是通过高阶推理等方法，从规则库中发现规则，即属性与关系之上的关系[4]。

笔者将以归纳逻辑的研究为基础，探讨 KDK 结构模型的构建，在其构建中，逻辑的参与将主要作用在知识库的语义构造和归纳假设的证实过程。

3 基于事实的 KDK 建模与挖掘算法

3.1 基于属性的知识库建库

在 KDK 系统中，选用了基于属性的知识表示方法，其最终形式为产生式规则，这主要是基于以下几点考虑：

1) 下一步的工作中，将把 KDK 的结果与数据库和 KDD 过程相连接，这就要求必须为知识库和数据库的协调留下接口；而基于属性的建库方式使得知识库与数据库在结构上相似，便于协调。

2) 基于属性的建库方式可以有效的借用数据库的现有功能，便于存储大容量的知识，解决目前知识库容量小的问题。

3) 基于属性的知识库建库曾有人研究，并有实例证明基于这种方式的推理是可行和有效的[5, 6]。

4) 产生式规则的知识表示方法具有模块化、清晰、便于理解等优点，尤其重要的是它提供的是一种粗框架的表示，可以在产生式内部结合基于属性的具体表示方法。

定义 1 在相关于论域 X 的知识库中，称“一个属性词＋一个属性程度词”这样形式表示的知识结点为知识素结点。

定义 2 相应于论域 X 的知识结点，是指如下形式的合式公式：

$$\theta_0 a_1 \theta_1 a_2 \cdots \theta_{m-1} a_m \theta_m,$$

其中 a_i 为某一知识素结点，$\theta_i \in J$，$i = 0, 1, \cdots, m$。这里 J 是由符号“∧”，“∨”，“not”，“(”和“)”等 5 个符号及其任意组合以及 NOP（空）而形成的集合；但 θ_i 在其中取值要使该公式有意义；只有 θ_0 和 θ_m 可以取空。显然，知识素结点是知识结点的一种特殊形式。

3.2 基于卡尔纳普归纳逻辑的 KDK 建模

根据对概率的不同解释及其度量的不同处理，现代归纳逻辑分为贝叶斯派和非贝叶斯派。在贝叶斯派中又可分为逻辑贝叶斯派和主观贝叶斯派。其中较有代表性的有凯恩斯的概率逻辑、卡尔纳普的归纳逻辑、柯恩的归纳支持逻辑、勃克斯的因果陈述句逻辑等。在我们的研究中，选择了卡尔纳普归纳逻辑作为理论基础。该理论基础将主要作用于知识库的语义构造和 KDK 的假设评价中。

3.2.1 知识库的语义构造

定义 3 用以完全描述给定知识库的某给定个体域的一切可能状态的语句集称为一个状态描述。

定理 1 知识库中的任一知识结点都能写成若干个状态描述的析取式。(证明略)

3.2.2 KDK 的模型构建

定义 4 m 是关于某状态描述的正则测度函数，当且仅当 m 满足以下条件：

1) 对于知识库中的每一个状态描述，m 值为正实数；

2) 知识库中某给定个体域的所有状态描述的 m 值之和为 1。

定理 2 设 m 是知识库中某一状态描述的正则测度函数，则每一个状态描述的 $m \in [0, 1]$。(证明略)

定义 5 m 是知识库中关于某状态描述的正则测度函数，扩展 m 为知识库中关于知识结点的正则测度函数 M：

1) 对于任何在知识库中不成立的知识结点 j，$M(j) = 0$；

2) 对于任何在知识库中非不成立的知识结点 j，$M(j) =$ 知识结点中所有状态描述的 m 值之和。

定义 6 若 m 是知识库中的关于知识结点的正则测度函数，c 为知识库中知识结点的二元函数，且 $c(h, e) = M(h, e)/M(e)$，则称 c 为知识库中的正则确证函数。

正则确证度表示了知识结点 h 和 e 共同成立的

可能世界的个数与 e 成立的可能世界的个数的比值。它代表了知识结点间的逻辑关系，根据卡尔纳普的逻辑体系，尽管 h 和 e 本身是涉及到事实经验的，但只要有了适当的确证度理论，就可以抛弃事实根据，只凭语义和语形分析得出确证度。

定义 7　KDK 过程模型为一个四元组 $M = \langle W, R, m, c \rangle$，其中：

1）W 为知识结点集，可理解为可能世界集；

2）$R \subseteq W \times W$，可理解为认知通达关系；

3）M^* 为知识库中的正则测度函数，$M^* = M_s/M_t$，M_s 为特定知识结点状态描述的正则测度函数，M_t 为全部知识结点状态描述的正则测度函数；

4）c 为知识库中的正则确证函数。

说明：以上模型是针对 KDK 的归纳过程而建，KDK 的目的是针对已知的两个知识结点，通过知识库的归纳得出两个或更多知识结点之间的关系，即上述模型中的 R（认知通达关系）。在这个模型中，定义了知识库中的正则测度函数和知识库中的正则确证函数，对关系 R 进行了量化的表示。据此量化值可知对于归纳结果的可信任程度。

3.2.3　对于利用卡尔纳普归纳逻辑构建 KDK 结构模型的几点说明　卡尔纳普归纳逻辑[7]虽然是一种很庞大、理论较完备的归纳逻辑，但它在逻辑界中引发的争议也较多。普遍认为，卡尔纳普归纳逻辑有其先天缺陷和不可克服的问题。而 KDK 的系统中采用了卡尔纳普归纳逻辑作为建模的基础，这主要是基于以下几点考虑：

1）KDK 系统的知识库采用了基于属性的知识表示方法，而基于卡尔纳普的知识评价方法也是基于属性的。因此，在形式上类似。

2）现代逻辑界普遍认为卡尔纳普归纳逻辑因为暗示了归纳问题的一种先验主义的解决方式，因此是一种证明的逻辑，而不是一种发现的逻辑。在 KDK 过程中，要借鉴卡尔纳普逻辑的主要原因是用于假设规则的评价，因此是一种较适当的方法。

3）卡尔纳普逻辑作为一种严密的数理逻辑有它的缺陷，主要在于它在处理无限世界中的问题时，它不能给出所有的状态描述。而应用于计算机实现时，因为计算机处理的只能是有穷世界，卡尔纳普的障碍不会影响 KDK 系统的实现。

4）卡尔纳普在处理正则测度函数时使用了平权的方法，这也是逻辑界争论的焦点，因为他不能给与平权制一个合理的解释。在 KDK 实现时，因为处理的是有穷世界，可以给不同的状态描述以不同的权重。这个权重可以用主客观相结合的方法给定，即首先由多专家参与共同给出加权均值，再通过扫描数据库用统计方法计算出权重，最后将两者结合。结合的方法可采用 $\alpha A_1 + (1-\alpha)A_2$ 的公式，此公式中 A_1 表示主观性因子，A_2 表示客观性因子，α 为主观性权值。这样就可以极大地减少个人的主观因素的影响。

3.3　利用 KDK 模型从事实库中发现规则

图 1 给出了基于卡尔纳普归纳逻辑的 KDK 算法流程图。

4　基于规则的 KDK 建模与挖掘算法

4.1　归纳假设的产生

4.1.1　广义概念格的定义与基本性质

定义 8　若有形式背景 $K = (U, D, R, S)$，其中 U 是规则集合，D 是规则特征属性的集合，R 是 U 与 D 之间的一个二元关系，即 $R \subseteq U \times D$，S 是规则支持度、可信度、综合指标的集合。则在此形式背景下，存在偏序集合与之相对应，并且这个偏序集合产生唯一的格结构，称为 K 一格结构。

这里，规则集合 U 可仅以规则序列号的集合来表示。规则特征属性集合 D 就是知识素结点的集合。R 是此规则所具有的特征属性，或说是该规则的条件和决策属性的包含。

定义 9　当且仅当三元组（X，Y，S）满足性质：

$X = g(Y), g(Y) = \{x \in U \mid \forall_y \in D, xRy\}$，

$Y = f(X), f(X) = \{y \in D \mid \forall x \in U, xRy\}$

时，称三元组 (X, Y, S) 关于 R 是完备的，且有 $f(\Phi) = D, g(\Phi) = U$。

定义 10　由定义 8 和定义 9 所诱导的格 L 称为广义概念格。

定理 3　所有广义概念格中的结点都是最大扩展序偶。（证明略）

定理 4　这种最大扩展是偏序集中的一种闭包。对于偏序集（U，=）中的闭包，有 $h: U \to U$，性质如下：

1）$\forall x \ \forall y (x = y \Rightarrow h(x) = h(y))$；

2）$\forall x (h(x) = x)$；

3）$\forall x (h(h(x)) = h(x))$，

$h(x)$ 称为 x 的 h 闭包。若 $x = h(x)$，则称 x 是

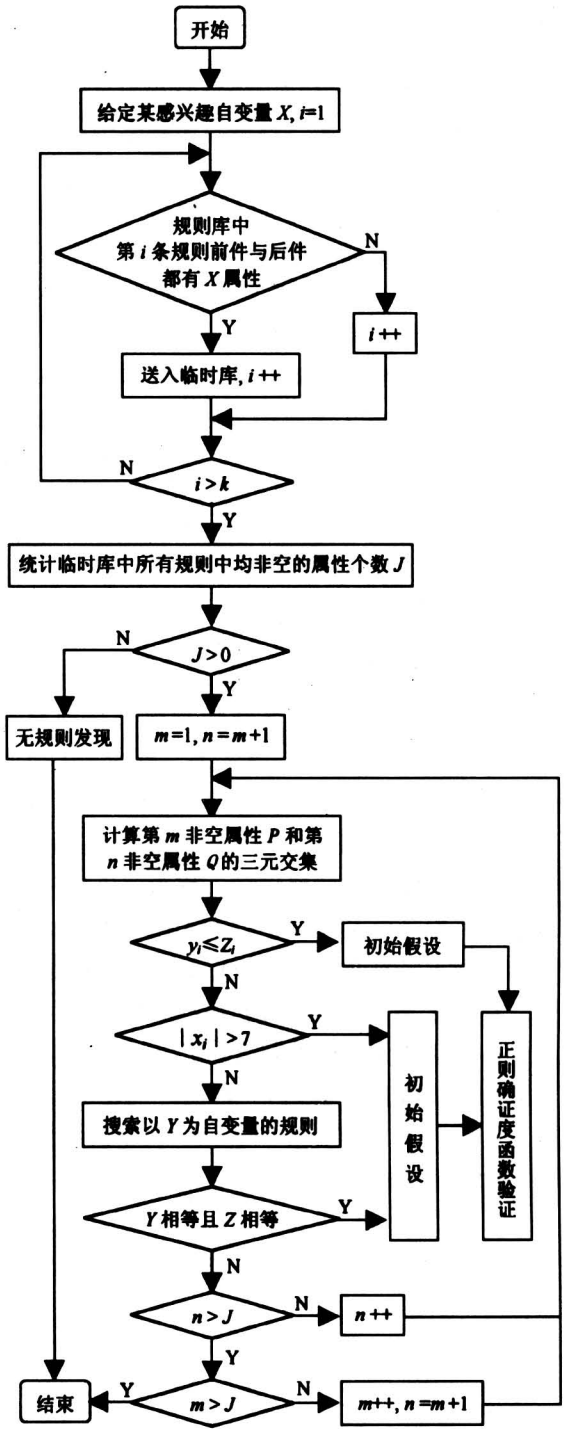

图1 基于卡尔纳普逻辑的 KDK 算法流程图

Fig.1 Flow chart of KDK algorithm based on induction logic of Carnap

h 近似。(证明略)

定理5 设广义概念格节点 $C_1(X_1, Y_1, S_1)$和 $C_2(X_2, Y_2, S_2)$，若 $Y_1 \leqslant Y_2 \Leftrightarrow X_2 \subseteq X_1$，则有 $C_1 \leqslant C_2 \Leftrightarrow X_2 \subseteq X_1$。(证明略)

4.1.2 产生式规则的数据库表示 若有产生式规则知识(包括领域知识或者 KDD 挖掘结果知识) $X_1 \wedge X_2 \wedge \cdots \wedge X_n \Rightarrow Y_1 \wedge Y_2 \wedge \cdots \wedge Y_m$ (sup, conf)，则可视为一个五元组(序号，条件，结果，支持度，可信度)，其中 X_n，Y_m 为语言变量，表示若有条件 X_1，X_2，…，X_n 发生，则有决策属性 $Y_1 \wedge Y_2 \wedge \cdots \wedge Y_m$ 成立，规则支持度为 sup，可信度为 conf。其表示形式如表1所示。

表1 产生式规则知识

Table 1 Producing pattern rule knowledge

ID	X_1	X_2	X_3	X_4	Y_1	…	sup	conf
1	2	3	4	3	1	…	1	1
2	4	4	2	1	5	…	2	2
⋮	⋮	⋮	⋮	⋮	⋮		⋮	⋮

4.1.3 产生式规则的广义概念格表示 由广义概念格定义，可构造匹配的形式背景(U，D，R，S)，如表2所示：$U = \{1, 2, \cdots\}$ 为规则序号，$D = \{a_1, a_2, a_3, a_4, a_5, b_1, b_2, \cdots\}$ 为知识素结点的集合，R 为每条规则的条件知识素结点和决策知识素节点，S 为此条规则的支持度。可以此形式背景得到相应的广义概念格，图2为其所对应的广义概念格的哈斯图。

表2 产生式规则知识的形式背景

Table 2 The pattern background of producing pattern rule

R	a	b	c	d	e	…	sup
1	2	3	4	3	1	…	s_1
2	4	4	2	1	5	…	s_2
3	1	0	0	1	0	…	s_3
4	0	1	1	0	0	…	s_4
5	0	1	0	0	1	…	s_5
⋮	⋮	⋮	⋮	⋮	⋮		⋮

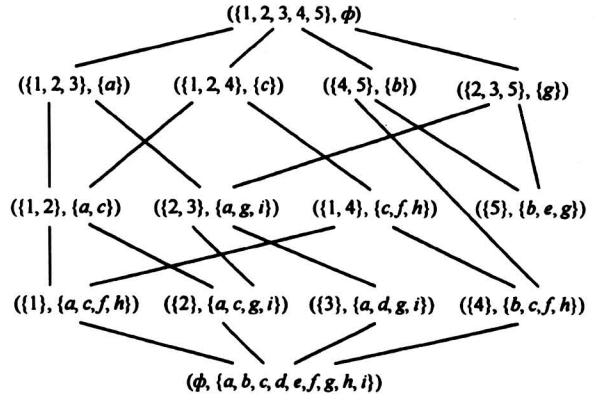

图2 表2所对应的广义概念格哈斯图

Fig.2 Table 2 broad sense concept metre

4.1.4 广义概念格结构产生式规则的批量式生成

在基于广义概念格的生成算法中（算法从略），建立边（结点间联系）过程和通常的算法区别不大，但增加了支持度信息，并实时得到可信度信息；只有满足规则支持度和可信度要求的结点才会被加入格中。为加快后续规则知识的发现过程，增加队列 First A，记录知识素结点的首次出现。

此外，还提出广义概念格的快速增量式更新算法（算法从略），其思路是对索引树结点进行分类，从而设计出一种基于树的快速增量式广义概念格（产生式规则）的生成算法。

4.2 基于科恩归纳逻辑的归纳假设评价体系

由 4.1 节给出的方法和归纳学习的方法都可以得到归纳假设。而这两种规则的获取方法从本质上来说都是一种归纳，而归纳是一个不保真的过程，因此对此归纳假设的评价是一个非常必要和十分困难的任务[8]。

4.2.1 归纳假设概率的确定

规则 1　若假设与事实相符合，则它是正确的。

规则 2　若假设没有反例，则它是正确的。

规则 3　假设的正确性程度由其所有的相关变量的变化范围内无反例区域的大小决定。

下面直接确定归纳假设概率 P_i 值。

定义 11　相关变量 v 是假设 H 的相关变量，如果 v 的变化可能改变 H 的值。

定义 12　一个相关变量是以某一属性（有限个）不同值为定义域的变量。这些值就称为它的变素。其中有且仅有一个是默认变素。

定义 13　检验特征函数 Vari ($H(x)$)，为标记归纳假设 $H(x)$ 经受检验状况的函数。

Vari ($H(x)$) = 1 表示 $H(x)$ 通过某一检验。Vari ($H(x)$) = −1 表示 $H(x)$ 未能通过某一检验。Vari ($H(x)$) = 0 表示不能判断 $H(x)$ 能否通过某一检验。

规则 4　$\square^{\bar{i}} H(x) \wedge \text{Vari}(H(x)) = 1 \Rightarrow \square^{\overline{i+1}} H(x)$。

若 $H(x)$ 通过由增加一个相关变量构成的检验 t_{i+1}，则它的 P_I 值加 1。这条规则是确定 P_I 的主要规则，称为主规则，其余规则称为从规则。

规则 5　$\square^{\bar{i}} H(x) \wedge \text{Vari}(H(x)) = 0 \Rightarrow \square^{\bar{i}} H(x)$。

若 $H(x)$ 通过 t_i，但无法判断它是否通过 t_{i+1}，则有 $\square^{\bar{i}} H(x)$。

规则 6　$\square^{\bar{i}} H(x) \wedge \text{Vari}(H(x)) = -1 \Rightarrow \square^{\bar{i}} H(x)$。

若 $H(x)$ 通过 t_i，但未能通过检验 t_{i+1}，则有 $\square^{i} H(x)$。

$H(x)$是否能经受住某一次的检验，无法作出直接的回答，此时系统将从已知条件和 $H(x)$ 出发，进行演绎推理，以确定 Vari ($H(x)$) 的值。

在归纳评价时，按相关变量及其检验序列，采用演绎和归纳交替的办法。首先给出归纳假设作为从目标，从目标出发构造演绎的与/或树。若演绎成功，则返回成功的可信度；若演绎不成功而又可选取主规则，则开始进行归纳，根据归纳结果返回相应的 P_I 值，并将此值赋给相应的与/或树。然后，利用返回的 P_I 值，通过计算将结果返回给根目标，从而得出目标成立的 P_I 值。最后若无规则可用，则采用直接提问的方式使评价进行下去。

定义 14　相关域函项 r 是一个从假设集到相关变量的有穷序列集的映射。$r[H] = (v_1, v_2, \cdots, v_n)$ 指已知 H 的全部相关变量为 $v_1, v_2, \cdots, v_n$。

定义 15　归纳度量函项 m 是一个从相关变量的有穷序列集到$U\{-1, 0, 1\}^n$ 的映射。

若设归纳假设 H 有全称条件形式$r[H] = (v_1, v_2, \cdots, v_n)$，并有 domain ($v_i$) = ($v_i^k$, v_i^1, $\cdots$, v_i^n)，$i = 1, \cdots, n$。其中 v_i^k是 v_i 的默认变素。对 H 的检验由一系列实验 t_i 组成，每个测试实际上是对 H 的一个与相关域有关的蕴涵进行的，因此必须借助于某个域有关的判别准则来决定测试结果对 H 是支持、否定还是中立，并以此形成证据。记 m 的第 i 个分量为$m_i[H]$，则 r 函项与 m 函项在 H 上的值刻画了假设 H 经过检验所获得的归纳评价。对给定的问题域，通过比较两个假设的 r 函项与 m 函项的值就可以比较它们可靠性的优劣。特别是，对任一假设 H 若 $\forall i$ ($m_i[H] = 1$)，则此归纳假设是完全可靠的。

4.2.2 归纳假设的评价算法

1) 给出归纳假设 H；

2) 确定可能证伪的所有可能的因子即相关变量，记为相关变量集合 V；

3) 确定相关域函项 r；

4) 确定归纳度量函项 m；

5) 检验控制所有 N 个变量，使其所有可能的

组合逐个出现；

6）计算归纳概率 P_I。

将此算法应用于基于广义概念格所发现的假设归纳规则，在给定最小归纳概率阈值后，经验证得到的规则数目减少一半，而正确率大大提高了。从而验证了基于规则的 KDK 模型构建与基于科恩归纳逻辑(做了改进)的归纳假设评价算法的有效性。

5 结语

1）在 cyc 数据库中对文献［5，6］提到的基于事实的 KDK 算法（包括归纳评价算法）的有效性进行了验证，在 cyc 知识库的 3 个大的集合中运行了 TANEL：Person，Organization 及 Geoplitical Entitys。从 Person 集合中，证实了 146 个假设。它们中间有 18 个 Imp，17 个 Transfers Through，19 个继承，24 个决策，68 个 $P \wedge R \rightarrow Q$。所有证实的结论中，正确率为 85 %。

为验证所提出的新的 KDK 算法的有效性，同样选取了 cyc 知识库。从 Person 集合中，同样得到了 146 个假设。经过正则确证度函数的评估后，证实了 137 个假设，正确率提高到 93 %。

2）对于基于规则的 KDK 算法（包括归纳评价算法），也得到了类似的有效性的实验验证结论。

3）对于基于模式 KDK 算法（包括归纳评价算法），将做为下一阶段重点研究对象。在下一步的工作中，还将考虑数据库及 KDD 参与 KDK 挖掘过程，即探索知识发现系统内在机理中的第二个机制——双基融合机制。

参考文献

［1］ 杨炳儒. 知识工程与知识发现［M］. 北京：冶金工业出版社，2000

［2］ 杨炳儒，申江涛. 关于 KDD 的一类开放系统 KDD＊的研究［J］. 计算机科学，2000，27（2）：83～87

［3］ 洪家荣. 归纳学习——算法 理论 应用［M］. 北京：科学出版社，1997

［4］ DeJong G F，Mooney R. Explanation-based learning：an alternative view［J］. Machine Learning Journal，1986，1（2）：145～176

［5］ Shen W M. Machine learning with the cyc knowledge base［R］. Technical Report ACT-CYC-224-90，MCC，1990

［6］ 石纯一，郝继刚，王建伟. 基于解释的机器学习方法［M］. 北京：清华大学出版社，1997

［7］ 王雨田. 归纳逻辑与人工智能［M］. 上海：中国纺织大学出版社，1995

［8］ 杨炳儒，蔡艳霞. KDD 的因果关联规则的评价方法［J］. 软件学报，2002，13（6）：1142～1147

Research on the Structure Model and Mining Algorithm for Knowledge Discovery Based on Knowledge Base (KDK)

Yang Bingru, Shen Jiangtao, Chen Hongjie

(*School of Information Engineering, University of Science and Technology Beijing, Beijing* 100083, *China*)

[**Abstract**] Knowledge discovery in knowledge base (KDK) is a brand-new task. Its success will directly act on the construction of large knowledge base, and, at present, it is important to the solving of the bottleneck of machine study—discovering knowledge. The main work of this paper is: The inductive structure of KDK based on the facts in knowledge base, and its algorithm and experimental verification; The inductive structure algorithm of KDK for the rules in knowledge base and its experimental verification.

[**Key words**] knowledge discovery based on knowledge base; induction logic of Carnap; induction logic of L. J. Cohen; evaluation of hypothesis