# 双库协同机制对知识发现主流发展的驱动*)

周 颖 杨炳儒
(北京科技大学信息工程学院 北京100083)

## The Driving Force of Double Bases Cooperating Mechanism to Knowledge Discovery Main Stream

ZHOU Ying YANG Bing-Ru
(Information Engineering Institute, the University of Science and Technology, Beijing 100083)

**Abstract** The paper, by a research report, summarizes emergence and definition of double bases cooperating mechanism, and introduces its driving force and influence to many sides of main stream of knowledge discovery from structural model to algorithm , from structuring data mining to complex type data mining. The influence also expands to philosophy field. It has been above five years from proposing it to now. Summarizing it makes us learn a thing clearly: its functions are not simply improvement to algorithm, are to bring forward many new structural models and technology methods . It answers those urgent questions in the one paragraph of the paper to a greater extent. So we may say: double bases cooperating mechanism has important driving force to main stream of knowledge discovery.

**Keywords** Double bases cooperating mechanism, KDD*, DFSSM, Maradbcm algorithm, Auto-cognition

## 1 引言

当前KDD发展的主流是寻求在各类数据库和应用问题背景下的高性能、高扩展性的发掘算法。但知识发现系统中一直存在着一些典型的问题急需解决，如：1)新旧知识如何有机地融合在一起并由此产生的知识库实时维护的问题；2)通过用户聚焦确定发掘方向具有局限性，不能体现认知自主性，系统自身能否产生创见意向；3)数据库中数据量增长到一定程度，算法复杂性增大，引起算法失效等问题[1~3]。

为解决以上问题，杨炳儒教授撇开KDD中单一地围绕数据库进行挖掘的孤立且封闭的过程，从问题的总体结构出发，将数据库与知识库有机地联系起来思考解决方案。于1997年从知识发现、认知科学与智能系统等学科交叉结合的角度，以认知自主性为核心，将知识发现视为一个开放和不断进化的系统，研究其系统结构、方法、进程与运行机制，独立地提出了双库协同机制，并构建了将KDD与双库协同机制相结合的KDD*结构模型，在结构和功能上形成了相对于KDD而言的一个开放的、优化的扩体[4~6]。

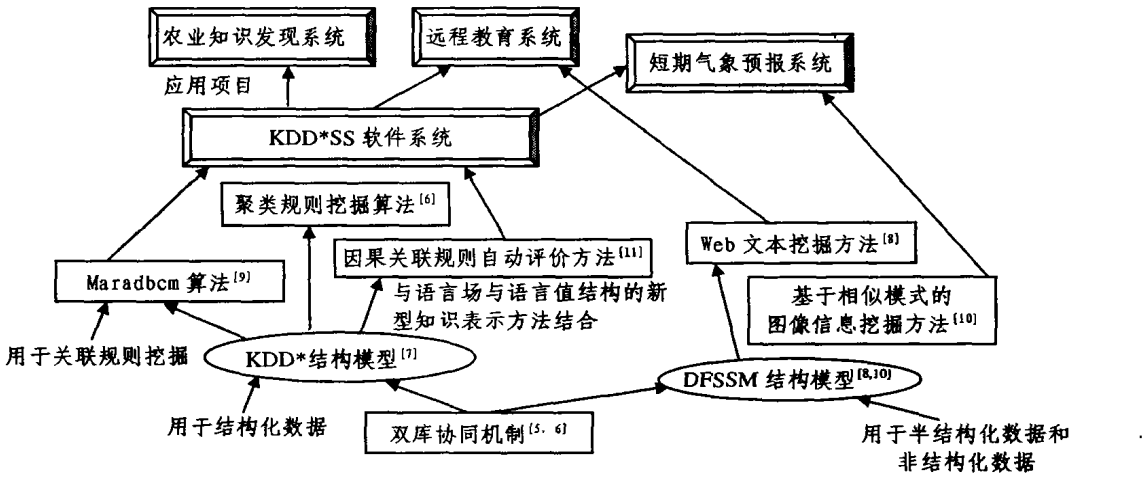


图1 基于双同协同机制提出的结构模型、技术方法、软件系统和实际应用等

经过多年的科学研究，已经由双库协同机制诱导出了适用于结构化数据挖掘的知识发现新结构模型--KDD*（KDD+双库协同机制）和适用于复杂类型数据挖掘的结构模型--DFSSM(Discovery Feature Sub-Space Model，基于复杂类型数据的发现特征子空间模型)；并由此派生出了许多新的技术方法和软件系统等，如图1所示。本文的意图是对双库协同机制产生的影响进行全面的阐述，但限于篇幅本文仅通过对KDD*、DFSSM结构模型和Maradbcm算法的简要介绍阐明原创性的技术方法、原创性的软件系统需要有原创性的理论为基础，从而证明了双库协同机制已经和正在对知识发现主流发展产生重要和深刻的影响，它相对于跟踪性和单一性的算法研究更具宏观与本质的意义。

*)国家自然科学基金重点项目(69835001)，教育部科技重点项目([2000]175)和北京市自然科学基金(4022008)。周 颖 博士生，主要研究方向为知识发现。杨炳儒 教授(首席一级)，博士生导师，主要研究方向为知识发现与智能系统，柔性建模与集成技术。

## 2 内在机理之一——双库协同机制的内涵及意义

### 2.1 双库协同机制的内涵[5,6]

在给定真实数据库和基础知识库的前提下，在数据发掘过程中，具备以下特征的KDD中的运行机制称为双库协同机制：

(1)在真实数据库上，按数据子类结构形式所构成的发掘数据库的可达范畴与基于属性间关系的发掘知识库的推理范畴之间构建范畴间的等价关系；两个范畴的等价关系为定向发掘和定向搜索奠定理论基础。

(2)在KDD聚焦过程中，除依据用户需求确定聚焦外，通过启发协调算法可以形成依发掘知识库中知识短缺而生成的机器自身提供的聚焦方向，进而在数据库中形成定向发掘(算法和进程)。

(3)从获得假设规则到知识评价的过程中产生中断进程，即先不对假设规则进行评价，而是通过中断协调算法到发掘知识库中进行定向搜索(算法和进程)，以期发现产生的假设规则与知识库中原有的知识是否重复、冗余和矛盾，并作相应处理，即对知识库进行实时维护。

### 2.2 意义

双库协同机制基本上解决了数据发掘过程中对领域固有的基础知识库的实时维护，同时在一定程度上解决了认知自主性的问题。即利用启发型协调器，实现了计算机自动发现“知识短缺”，系统自身根据知识短缺产生创见意向，形成定向发掘；对挖掘出来的知识通过中断型协调器，对知识库进行实时管理与维护。

## 3 基于双库协同机制的KDD*结构模型

### 3.1 KDD*结构模型(图2)

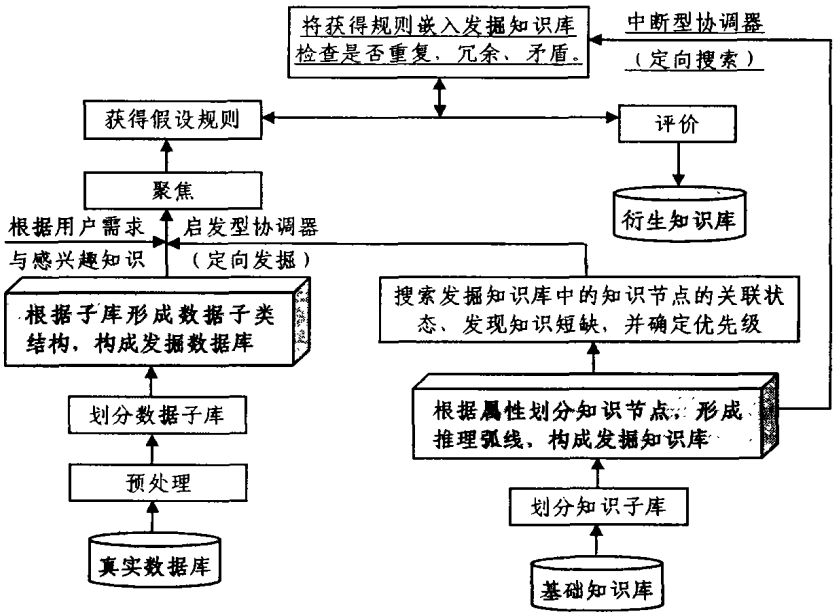


图2 KDD*的总体结构模型图

图2中正常部分是原流行的KDD结构模型，阴影部分是为实现启发式协调器的准备工作和启发式协调器本身的工作，带下划线的字是中断式协调器部分。从图中可以看出：KDD*与原KDD模型有很大区别--KDD根据用户需求聚焦，没有考虑已有知识库，并将挖掘到的规则一并送入评价系统进行评价。很明显，这不符合人的认知规律，更不能体现认知自主性。

### 3.2 KDD*的特征[7]

KDD*相对于KDD而言，是KDD与双库协同机制相融合的一种知识发现的新结构，它具有以下特征：

(1)KDD*有机地沟通与融合了KDD*新发现的知识与基础知识库中固有的知识，使它们成为一个有机整体；即实现了“用户的先验知识与先前发现的知识可以耦合到发现过程中”。

(2)在知识发现过程中，KDD*对于冗余性的、重复性的、不相容的信息作出了实时处理，有效地减少了由于过程积累而造成的问题的复杂性，同时为新旧知识的融合与合成提供了先决条件；实现了“知识与数据库同步进化”。

(3)在数据库数据的积累过程中，虽然知识库的结构具有一定的稳定性，但它也是随着数据的积累而不断进化的，并且这种进化的能力是双库协同机制本身所具有、无须领域专家的干预。

(4)KDD*改变与优化了知识发现的过程与运行机制，实现了“多源头”聚焦，减少了评价量。从认知科学的角度看，KDD*强化并提供了知识发现的智能化程度，提高了认知自主性(这将是今后相当长一个阶段内保持的研究基调)，较有效地克服了领域专家的自身局限性，实现了“采用领域知识辅助初始发现的聚焦”。

(5)作为KDD*的核心技术——双库协同机制的研究，揭示了在一定的建库原则下，知识子库与数据子类结构之间的对应关系，为实现“限制性的搜索”而减小搜索空间、提高发掘效率提供了有效的技术方法。

## 4 发现特征子空间模型DFSSM

为了便于复杂类型数据的知识发现的研究，文[10]从知识发现的角度给出模式的定义。

**定义1** 模式(Pattern)是知识发现过程中的一种知识表征方式，是具体或抽象的客观对象的量化描述，是知识发现过

程中的基本运算单元。模式参与知识的发现过程并表征所获得的知识。

在知识发现过程中，模式不仅是知识的表征方式，同时也是知识发现过程中最基本的操作单元。由于客观对象都存在于一定的物理空间，而 Hilbert 空间可以很好地描述和刻画客观对象在状态空间中的性质和结构，因而把模式定义为 Hilbert 空间中的矢量。用模式的矢量定义可以定量地表征复杂类型数据的复杂多变及具有的不确定的状态和行为，即具有非线性动力学性质和特征，从而可以用模式这一概念来描述复杂类型数据的知识表示和知识发现过程，及其挖掘结果的可视化展现，同时用模式的变化来刻画其整体知识发现过程的发展和演变规律[10]。

由双库协同机制派生的发现特征子空间模型 DFSSM——基于复杂类型数据的知识发现系统的总体结构模型如图3所示。DFSSM 主要分为如下几个部分[8]：

1）*复杂类型数据的知识表示及数据预处理过程* DFSSM 方法主要通过在高维的 Hilbert 空间进行特征抽取，形成原始数据集，然后在此基础上进行特征变换（对于文本数据类型、多媒体等数据类型可以采用空间层次分解方法，如小波分析处理），构造维数适中的特征子空间，在该特征子空间可以利用矩阵的奇异值分解变化和近似计算方法来构造模式。

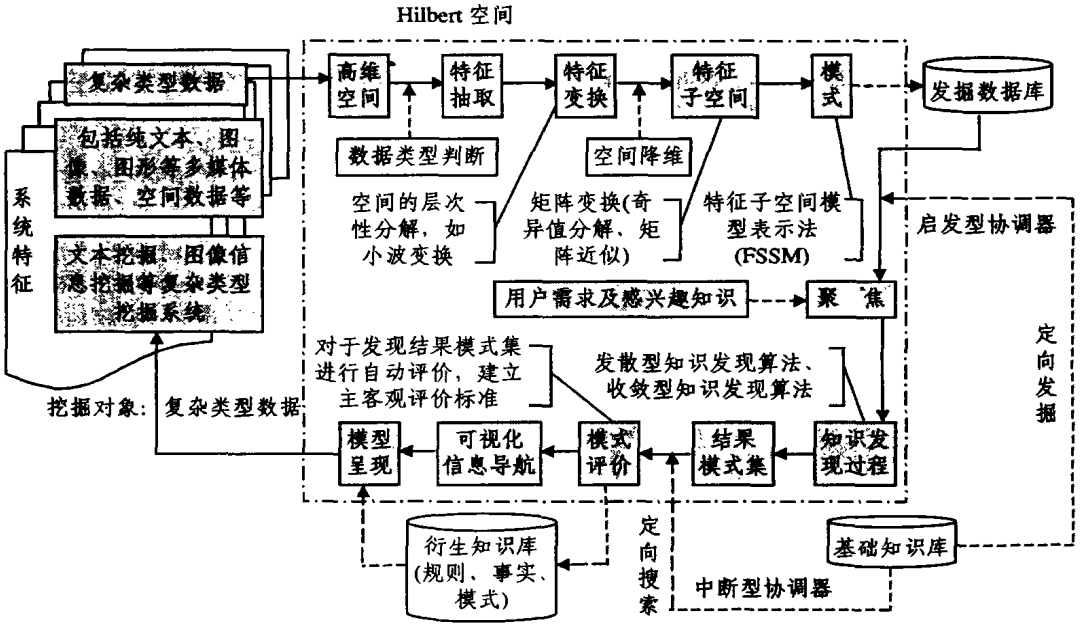


图3 基于复杂类型数据的知识发现系统的总体结构模型

2）*复杂类型数据的知识发现过程* 基于模式的知识发现同形象思维十分相似，它包含着比较、研究、推测、预测并遵从抽象化和具体化的法则。在关系数据库中，属性与属性之间则是相互独立的，结构化的知识发现就是建立在此基础之上。知识发现过程是以属性为基本的信息单元参与知识发现的全过程，并以属性与属性之间的关系来表征知识。但是对于文本、多媒体数据、空间数据、时间序列数据等复杂类型的数据来说，难以用独立的属性来对其进行表征，而是用属性的集合以及集合之间的关系来进行描述。模式可以很好地表征这种数据的集合及其元素之间的关系。由于模式表示的是一个相对来说独立的概念，模式可以同客观对象的组织结构建立联系，也可以表示十分抽象的概念且更具可理解性。在复杂类型数据的知识发现过程中模式（或子模式）作为一个整体，参与知识发现的过程。

基于模式的知识发现过程是一个发现新模式或对模式进行某种确证的过程。由于模式是定义在 Hilbert 空间中，因而基于模式的知识发现同空间变换紧密地联系在一起。可同分类、聚类、相似模式等收敛性的知识发现算法及预测、时序等发散性的知识发现算法相结合来完成各种类型的知识发现。同样在结构化数据的知识发现中运用模式可以发现不同抽象层次的知识。

3）*模式的评价*

4）*模式的解释与呈现*

5）*双库协同机制——两个协调器的构建*

## 5 挖掘关联规则的新算法——Maradbcm 算法[9]

### 5.1 概述

Maradbcm 算法赖以产生的理论基础是双库协同机制与 KDD* 结构模型。它依循 KDD* 的思路构造了关联规则的挖掘算法。其中启发型协调器的功能是通过搜索知识库中“知识结点”的不关联态，以发现“知识短缺”，产生“创见意向”，从而启发与激活真实数据库中相应的“数据类”，以产生“定向发掘进程”，即完成了计算机自动聚焦。具体实现上是利用有向超图表示知识库中的知识，通过计算可达矩阵发现“知识短缺”，进而用规则强度阈值进行剪枝并形成聚焦；中断型协调器的功能是当从真实数据库的大量数据中经聚焦而生成规则（知识）后，使 KDD 进程产生“中断”，用 SQL 语言或计算有向超图的可达矩阵来判断知识库中对应位置有无此生成规则的重复、冗余、矛盾、从属、循环等。若有，则取消该生成规则或相应处理后返回 KDD 的“始端”；若无，则继续 KDD 进程，即知识评价。

### 5.2 比较

Maradbcm 算法与 Apriori 算法的主要共同点是两者在本质上都是基于统计方法的；两者的主要区别在于以下5个方面：

（1）基于的学术思想不同：Maradbcm 算法是基于内在机

理研究，具体而论是基于“知识短缺”(利用有向超图)进行“定向挖掘”；而 Apriori 算法是基于组合论的数据库全局搜索。

(2)基本流程(或基于的模型)不同：Maradbcm 算法的流程见 KDD* 结构模型图，它是一条一条短缺知识的挖掘；而 Apriori 算法是整体性一并挖掘。

(3)基础不同：Maradbcm 算法是基于规则强度，它考虑了主观和客观两个方面，即考虑了用户的聚焦(感兴趣度)，并涵盖了 Apriori 算法的支持度阈值。

(4)发现知识的量不同：Maradbcm 算法考虑了知识库，从而能真正发现新颖的、用户感兴趣的知识，这正符合了 KDD 定义；而 Apriori 算法是把满足条件的规则全部挖掘出来；另外，Maradbcm 算法克服了 Apriori 算法的两大缺点：遗漏重要的规则和数据库的全局搜索。

(5)Maradbcm 算法可融入 KDD 中形成新的开放型的结构模型——KDD*，整个算法实现的运算背景是 KDD* 结构；而 Apriori 算法是原有的封闭系统 KDD。

文[9]就 Maradbcm 算法和 Apriori 算法进行了对比实验，试验结果证实：Maradbcm 的速度、挖掘的知识量、发现的意外规则的量、有效性、时空复杂性、可扩展性都是优于后者的。

## 6 对哲学概念与 KDD 核心概念的影响

有关双库协同机制的研究，不仅给知识发现的结构模型、技术方法带来了重要的贡献，同时也引发了我们在哲学方面的思考：

(1)*质量观范畴的深层思考* 根据质量互变规律可知：量变引起质变，质变引起量的扩张，量变过程中发生部分质变等。因而，随着数据库中数据量的剧增，知识发现过程中所获取的知识在反映事物的本质上可能会存在着部分质变。这就要求对于某些目前无法理解和尚未引起重视甚至不相容的知识应当予以充分的注意，要适度保存一些矛盾知识和一些目前无法应用的知识。在对知识库进行实时维护时，要肯定事物发展的相对稳定性，研究知识之间的关系，重视知识的评价。

(2)*认知自主性引申了对人机交互的辩证思考* 通过人机交互，人类智能与机器智能的协同，以及机器智能与人机交互关系的辩证分析和研究，引申了对于知识发现、认知科学与智能系统交叉结合的深入理解。认知自主性是知识发现的核心概念和知识发现智能化的体现。通过启发型协调器可产生创见意向，通过定向发掘可使计算机在无外界干预和指导的条件下主动发掘知识。然而，在目前的研究阶段，在聚焦、发掘方法与评价等环节上又确实离不开领域专家，因而人机交互是必不可少的。随着机器智能、自然语言理解等研究的进展，必然使提高认知自主性进一步成为研究的主导基调。

(3)*算法复杂性与逻辑问题所引申的思考* KDD 的发展对算法的复杂性、算法的有效性与数据的约简和降维等问题提出了新的挑战。由于 KDD 所处理的数据量为海量数据，往往使得在局部数据环境中成立的算法在处理海量数据时失效。这就要求发掘算法具有高性能与高可扩展性，因而应对算法的时空复杂性、建模复杂性等进行深入研究。复杂性(特别是 NP 难问题)算法的深入研究将为提高知识发现的效率提供有力的理论基础。

认知逻辑与认证逻辑等在知识发现的研究中已经成为评价算法、发掘算法研究的理论支撑点。这一方面拓展了上述逻辑研究的应用范围，另一方面又由 KDD 的发展为其研究提供若干新的生长点。将哲学、逻辑科学与信息科学紧密结合，可能会诱导出思维科学与方法论中的若干新命题。

**结论** 本文以研究报告的形式综述了双库协同机制的提出及其定义，并重点介绍了它对 KDD 从结构模型到技术方法、从结构化数据挖掘到复杂类型数据挖掘等方面的推动和影响，这种影响甚至广及哲学领域。事实上，由于这一原创性的理论，我们才能提出优于 Apriori 算法的 Maradbcm 算法，文中论述了二者除同是基于统计方法外，在其余方面是完全不同的；文中介绍的其他模型、方法等与原有方法相比也都有了质的飞跃。从双库协同机制提出至今已逾5年，对其进行一个总结，可由此清楚地看出：双库协同机制对知识发现的主流发展起到了重要的驱动作用。

知识发现内在机理的研究还包括双基融合机制和信息扩张机制，这两个机制的研究也对知识发现的主流发展产生一定的推动作用。如构建了分别基于其上的 KDK* 和 KD(D&K)结构模型；解决了 KDK 中的难点问题，即归纳评价的问题，方法是将知识发现结果嵌入数据库的发掘过程中加以验证 。信息扩张机制拟解决的问题是：在数据库不断动态扩张的过程中，寻求其与知识库的动态关联关系；寻求支持度、可信度和充分性因子的动态变化规律；发现过程的“不动点”与“突变性”等问题。

## 参 考 文 献


1 Anand S S, Bell D A, Hughs J G. EDM: A General Framework for Data Mining Based on Evidence Theory. Data & Knowledge Eng., 1996,18:189～223

2 Piatetsky-shapiro G, Matheus C J. Knowledge Discovery Workbench for Exploring Business Databases. International Journal of Intelligent Systems,1992,7:675～686

3 Yoon J P, Kerschberg L. A Frame Work for Knowledge Discovery and Evolution in Databases. IEEE Trans. on Knowledge and Data Eng., 1993,5:973～979

4 知识工程研究所主页. http:\\www.ustb.edu.cn\kdd\default.htm,实名：知识工程研究所

5 杨炳儒，王建新. KDD 中双库协同机制的研究(I). 中国工程科学，2002，4(4):41～51

6 杨炳儒，王建新，孙海洪. KDD 中双库协同机制的研究(II). 中国工程科学，2002，4(5):34～43

7 Yang Bingru. A KDD related open system -- KDD*. 计算机科学，2000,27(2):83～87

8 杨炳儒，唐菁. 基于复杂类型数据的发现特征子空间模型 DFSSM 的研究. 中国工程科学，2002,4(9)

9 杨炳儒，孙海洪. 基于双库协同机制的挖掘关联规则算法 Maradbcm. 计算机研究与发展，2002,39(10)

10 张德政. 基于相似模式知识发现方法的研究与应用：[博士学位论文]. 北京科技大学，2002

11 杨炳儒，綦艳霞. KDD 中因果关联规则的评价方法. 软件学报，2002,12(6)