学术论文

基于复杂类型数据的发现特征子空间模型（DFSSM）的研究

杨炳儒，唐 菁
（北京科技大学信息工程学院，北京 100083）

[摘要] 探讨围绕知识发现领域中较为宏观、较为重大的问题。首先，根据复杂类型数据（包括 Web 数据、多媒体数据、空间数据、时间序列数据等）所具有的非线性动力学性质和特征，采用模式（定义为 Hilbert 空间中的矢量）来定量地表征复杂类型数据的多变性及具有的不确定状态和行为，并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；其次，以知识发现系统内在机理的研究为基础，构造了复杂类型数据知识发现系统的总体结构模型——发现特征子空间模型 DFSSM；最后，用基于 Web 的文本挖掘系统和基于图像信息（气象云图）的知识发现系统作为实例进行了验证，结果表明 DFSSM 方法对于非结构化的文本数据及图像数据类型的知识发现过程具有指导性作用。因此，该结构模型具有较好的实用性与普适性，有望拓展到其他复杂类型数据的知识发现过程中。

[关键词] 复杂类型数据；数据挖掘；文本挖掘
[中图分类号] TP182 [文献标识码] A [文章编号] 1009－1742（2003）01－0056－06

1 引言

Internet 作为信息时代的基础构架，涵盖了更为广泛的信息类型和内容，有着更为广泛的应用范围。各类应用的 Web 数据库、电子邮件和网页中包含着丰富的信息资源如文本、视频、音频、图形图像等，同时也存在着各种不同类型的应用处理和信息服务的需求。

不同学科领域的研究和应用都需要对这些信息进行分析和处理，从中抽取出隐藏着的可满足不同需求的规律和人类孜孜以求的知识。随着信息量的增大和信息类型的复杂化，面对这些纷杂多变的信息，传统的数据分析方法和分析工具难以奏效。因而，驱动了知识发现理论和技术方法的产生、发展和深入研究以及广泛应用[1,2]。

目前，KDD 不仅是信息科学中的一个重要的理论研究课题，而且被许多工商界人士看作是一个能带来巨大回报的重要领域。所以如此，是因为知识发现克服了传统分析方法的不足，适应于处理大规模信息量、复杂的以及具有非线性特征的数据类型，顺应人类的认知过程，采用广泛的分析方法和处理手段，实现各类知识和信息的获取[3]。

知识发现的应用对象从结构化数据发展到半结构化及非结构化的复杂类型数据，包括关系数据库、面向对象数据库、空间数据库、推理数据库、多媒体数据库、时态数据库、文本数据、Web 访问日志、图形图像数据及音频和视频数据等。

目前国内外知识发现的研究主要是以知识发现的任务描述、知识评价与知识呈现为主线；以基于各种理论的有效知识发现算法研究为中心；以及更加广泛的应用研究为主要特点[3~5]。具体体现在结构化数据和非结构化的复杂类型数据的挖掘技术

[收稿日期] 2002－06－17；修回日期 2002－08－29
[基金项目] 国家自然科学基金重点资助项目（69835001）；国家教育部科技重点资助项目（教技司［2000］175）
[作者简介] 杨炳儒（1943－），男，天津市人，北京科技大学信息工程学院教授，博士生导师

和应用的研究上以及软件产品和应用系统的开发中。知识发现机理和一般性框架的理论研究日益受到重视，复杂类型数据的挖掘形成了新的研究热点。

在结构化数据的挖掘理论与方法上已经进行了较为深入地研究，提出了双库协同机制[1,2]、双基融合机制与信息扩张机制及其诱导的新结构模型，作为对 KDD 系列性研究中所提出的新研究方向（经查询检索证实），即内在机理的研究；形成了基础—机理—模型—算法—软件—应用的研究体系。基于以上的研究成果，笔者着重从知识发现内在机理的角度探讨了基于复杂类型数据的知识发现系统的一般化的结构模型——发现特征子空间模型（DFSSM），从而可望对于非结构化类型数据的挖掘技术和应用起到一定的指导性作用；同时采用基于文本数据类型和图像数据类型的知识发现结构模型作为该总体结构模型的两个应用实例，以验证其有效性和普适性；进一步扩展到其他复杂类型的数据——例如图形、音频和视频等多媒体数据、空间数据、时间序列数据的挖掘过程中。

2　基于结构化数据类型的知识发现系统的研究

2.1　新的知识发现系统——KDD* 总体结构模型

根据认识与逻辑发展的必然，笔者 1 于 1997 年从知识发现、认知科学与智能系统交叉结合的角度，提出了基于双库协同机制的 KDD* 新知识发现系统，并建立了知识发现系统的一般性框架，其总体结构模型如图 1 所示。

2.2　KDD* 系统的特征

KDD* 系统区别于一般 KDD 系统的特征如下[1,2,6]：

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的整体；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）KDD* 改变与优化了知识发现的过程与运行机制，实现了多源头聚焦与减少评价量。

4）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（这将是今后相当长的一阶段内保持的研究基调），较有效地克服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。

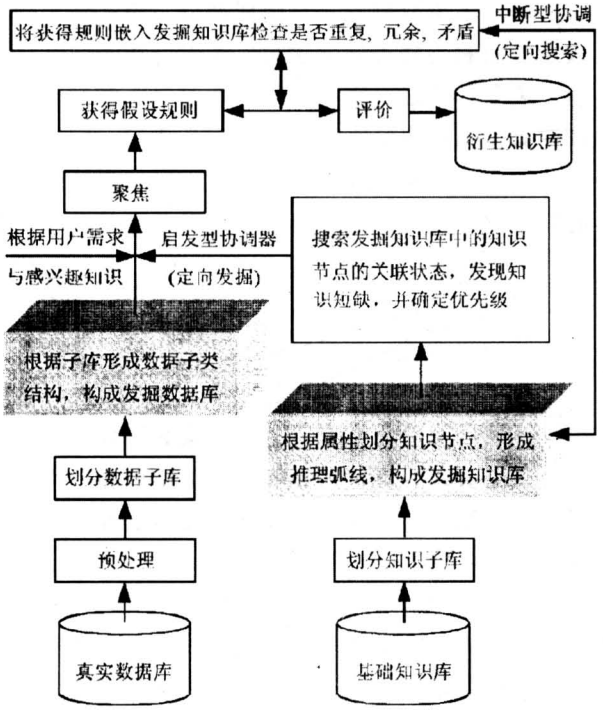

图 1　KDD* 的总体结构模型图

Fig.1　The overall structural model of KDD*

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。

6）对 KDD 主流发展——构造高效可扩展的挖掘算法、知识表示与评价方法等，产生重要影响（如优于 Apriori 算法的 Maradbcm 算法的产生）。

7）进一步拓展了知识发现系统的新结构模型（KD（D&K），ESKD，…），并对复杂类型数据的挖掘（文本挖掘，多媒体挖掘，乃至 Web 挖掘）的算法、结构、机理、体系与应用产生重要影响。

3　基于非结构化与半结构化的复杂类型数据的知识发现系统研究

3.1　基于复杂类型数据的知识表示方法

对于复杂类型数据来说，如果采用原有的知识表示方法，那么随着问题复杂程度的增加和数据量

的扩大，其表征能力和系统的可管理性变得越来越差。因而有必要针对复杂类型数据的知识发现问题寻找一种有效的知识表征方式。

在人的认知过程中模式起着重要的作用。模式存在于人类认知和思维的全过程，是人类认知和思维过程中不可分割的一个组成部分。模式作为知识表示方法和心理表征方法被广泛地应用于知觉、注意、表象、记忆、思维解释、学习等心智过程，发挥着其他知识表征方式无法取代的作用。同时模式参与知觉、注意、表象、记忆、思维解释、学习等心智过程，这些过程是模式的映射和变换过程。在此过程中模式的维数不断降低，表征能力逐渐加强。因此正是利用模式，人类可以表示客观事物的复杂结构和性态，利用模式使得认知过程变得有效而且易于理解。但是目前数据挖掘和知识发现理论中没有给出明确的模式定义和相应的解释，也没能充分利用模式的这种优势和人的认知机理。为了便于复杂类型数据的知识发现的研究，笔者从知识发现的角度可给出模式的定义[7]。

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。

1）*该定义从知识发现的角度将模式定义为知识表示的一种方式* 模式本身可以表示复杂的知识结构，通过模式的映射或组合可以表示不同抽象级的知识。同时模式的可变性适于表达知识的多样性，如同在人工智能中同一种知识可以分别用产生式规则、框架和语义网络等多种知识表示方式进行表征；再从知识的表征和认知过程来看，知识的模式表示同人类思维中的表象和意象表征方式存在一致性，符合认知过程的心理学特征。因而，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。

2）*把模式定义为 Hilbert 空间中的矢量* 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[7]。

3）*采用超图模型* 当超图模型的节点集表示为矢量时就建立了超图模型同模式的联系。超图模型不仅可以用形象化的方式来表示知识结构，简化复杂的知识结构，使得领域专家通过可视化途径进行模式的操作；同时模式的超图模型同面向对象技术有着很好的对应关系，易于采用面向对象技术编程实现模式的可视化。

3.2 基于复杂类型数据的知识发现系统的总体结构模型

基于以上提出的复杂类型数据的知识表示方法——模式表示方法，同时借鉴结构化数据类型的知识发现系统的总体结构模型，我们给出如下的基于复杂类型数据的知识发现系统的总体结构模型DFSSM。（见图2所示）

DFSSM 主要分为如下几个部分：

3.2.1 *复杂类型数据的知识表示及数据预处理过程* 为了全面地表征待挖掘对象，在高维空间中构建其表示方式。由于 Hilbert 空间可以很好地描述和刻画挖掘对象在状态空间中的性质和结构，所以在此空间进行特征抽取、特征变化及特征子空间的选取等一系列的操作。最终用模式来表征复杂类型数据，使得后续的各种处理过程可以参考结构化知识发现过程。

在数据预处理阶段，首先将判断复杂数据的类型（如文本数据、多媒体数据、空间数据及时间序列数据等），然后选择合适的特征抽取工具，进行复杂数据对象的特征抽取操作，形成原始的特征表征方式。该数据表征方式是构建在高维数据空间（Hilbert 空间）中，由 Hilbert 空间定义可知 Hilbert 空间是一个完备的线性赋范空间，所以它必然是一个线性空间。在线性空间中存在线性变换，通过线性变换可以构建子空间，并可以利用子空间来对原始空间进行描述。其中空间变换成为从不同的角度分析和观察原始空间的有利工具。同时从原始空间到子空间，其维数将减少，更加适合于知识发现过程。在此提出了发现特征子空间模型 DFSSM 方法。相对于传统的向量空间模型 VSM 方法而言，它将特征表征中的特征抽取、变换及映射过程融合成一个整体；其适用的挖掘对象范围更加广泛；同时简约了特征子集的选取过程，提高了发掘效率。

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。

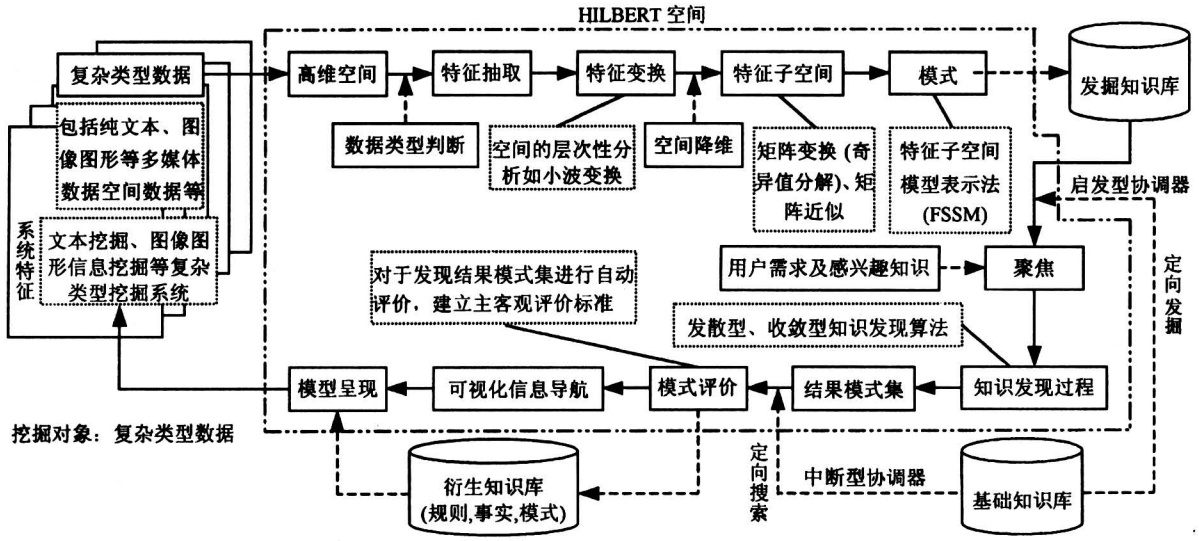

图 2　基于复杂类型数据的知识发现系统的总体结构模型 DFSSM

Fig.2　The overall structural model of knowledge discovery system based on complex type data

3.2.2　*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。利用不同层次的模式可获得反映事物的共性或本质的规律，通过模式操作来形成分类、聚类、相似等形式以反映事物内在的本质或规律。

在关系数据库 R 中，对于规范化以后的数据库模式来说，任何一个非主属性的完全函数依赖于每个候选关键字，并且不存在任何非主属性传递函数依赖于 R 的某个候选关键字；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。知识发现过程中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更具有可理解性。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。同结构化数据的知识发现相类似，基于模式可以进行关联、分类和聚类以及预测等类型的知识发现。

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。

3.2.3　*模式的评价*　由于经过数据挖掘处理后，将形成大量假设模式集，它们需要进行相应的评价，才能够存放到知识库中，并为模式的使用奠定基础。

构造模式评价函数，同时结合评价的主客观标准，采用定量的方式来评估结果模式集中有效的、新颖的、潜在可用的及最终可理解的模式，并把它存放到知识库中。

对于文本挖掘来说，其评价函数可以采用查全率、查准率及信息估值等客观指标，也可以定义用户感兴趣度等主观指标。

3.2.4　*模式的解释与呈现*　由于模式本身的可视性不强，不能够让用户快速、准确地从模式集合中获取其所需要的知识。因为对于知识库中的模式进

行解释和呈现就成为用户获取知识的一种有效方式。

在此，结合超图模型来表示相应的模式，用图形的方式直观地反映模式集。超图模型不仅可以用形象化的方式来表示知识结构，简化复杂的知识结构，使得领域专家通过可视化途径进行模式的操作；同时模式的超图模型同面向对象技术有着很好的对应关系，易于采用面向对象技术编程实现模式的可视化。

对于文本挖掘来说，采用可视化信息导航机制给用户提供简明、多视角的知识获取方法；使得用户能够更快的接受信息并根据自己的兴趣度对所反馈的挖掘结果进行有目的的查询和浏览。

3.2.5 双库协同机制——两个协调器的构建 目前国内外同行对于知识发现系统内在机理的研究甚少。我们在该方向进行了深入、广泛地研究和探讨，并取得了一定的研究成果。因此，提出了基于复杂类型数据的知识发现系统总体结构模型的双库协同机制——启发型协调器和中断型协调器的构建[1,2]。

启发型协调器的主要目的是为系统的聚焦提供另一个途径。在经典知识发现进程中，系统的聚焦通常是由用户提供感兴趣方向，知识发现系统沿此方向进行挖掘。但如果仅沿此方向行进，大量数据中的潜在的、也许会对用户有用的信息往往会被忽略掉。为尽可能多地搜索到对用户有用的信息，以弥补用户或领域专家自身的局限性，提高机器的认知自主性，而构造了启发型协调器。启发型协调器是通过启发协调算法来实现的。算法的实质是通过寻求知识短缺产生创见意向，使系统产生自动聚焦。

中断型协调器的主要目的是实时地、尽早地将重复、矛盾、冗余的知识淘汰掉，从而做到只对那些有可能成为新知识的假设进行评价，最大限度地减少了评价工作量。传统的知识发现系统，对KDD过程产生的假设直接进行评价，被接受的知识归并到知识库时，由知识库管理系统负责对知识库的一致性、冗余性进行检查，对矛盾和冗余的知识进行处理，形成新的知识库。此方式的缺点是：形成许多无意义的假设评价和由于问题的大量积累而加重一致性、冗余性检查的负担。在实际的专家系统中，最终成为新知识的假设占原假设的比例是很小的（发现新知识是困难的），大量假设会是重复和冗余的，因此中断型协调器的引入将提高知识发现系统的效率，利于知识库的实时维护。

通过两个协调器的构造，实现了基于双库协同机制的复杂类型数据的知识发现过程。它能够更好地解决知识库中的规则、事实及模式与数据库中的数据协同进化的问题，从而改变了传统KDD固有的运行机制，在结构与功能上形成了相对于KDD而言的一个开放的、优化的扩体。

4 实例验证

DFSSM将在很大程度上对复杂类型数据的知识发现过程起到指导作用。下面以基于Web的文本挖掘系统和基于图像信息的知识发现系统为例，说明该总体结构模型的实用性和有效性。

在我们承担的国家教育部科技重点项目——智能化、个性化的现代远程教育系统的关键技术研究中，基于复杂类型数据的知识发现系统的总体结构模型的基础上，构建了一个适用于现代远程教育的文本挖掘系统。它能够充分利用Web站点（远程教育站点）上积累的丰富文本信息，更好地服务于远程教育。该系统的核心知识发现方法主要是采用了收敛性的知识发现算法——分类算法。

根据复杂类型数据知识发现系统的总体结构模型DFSSM方法，并结合具体的挖掘算法，得到文本挖掘系统的总体结构模型、工作流程及其挖掘方法[8]，具体说明如下：

1）特征抽取及特征变换 对Web上采集到的挖掘目标样本进行特征预处理（分词和词频统计处理），然后采用特征子空间模型法（DFSSM），将特征抽取、变换及特征子集的选择融为一体，用模式来表示文本中间表示形式。

2）文本挖掘过程 对于Web文本的中间表示形式采用向量空间的距离测度分类算法（收敛性的知识发现算法）进行分类挖掘处理，也可以结合聚类和关联挖掘算法。最终得到潜在的模式集。

3）模型质量评价 对挖掘得到模式集进行评价，将符合一定标准的知识或模式呈现给用户。其中使用的客观评价指标主要是查全率（recall）和查准率（precision）。

4）信息呈现及信息导航 将反馈的结果用可视化的方式（树型结构和图形结构）进行显示，同时对用户提供信息导航功能，从而在极大的程度上方便用户有效地浏览和获取信息。

5）双库协同机制　当用模式表征了挖掘对象后，通过启发型协调器来搜索知识库中知识结点的不关联态，以发现知识短缺，产生创见意象，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，提高其认知自主性及智能化程度。对于分类算法生成的假设规则（知识），利用中断型协调器使 KDD 进程产生中断，从而搜索知识库中对应位置有无此生成规则的重复、冗余与矛盾（定向搜索进程）。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，即评价与结果入库。

经过该 Web 文本挖掘系统处理后，其分类挖掘结果导航界面如图 3 所示。

图 3　Web 文本挖掘系统分类挖掘结果导航界面

Fig.3　The navigation interface of results from web text mining system

除此之外，我们还承担了国家气象局“大城市环境气象信息系统研究：气象数据的知识发现系统”课题的研究和实现。课题的主要任务是针对复杂多变的天气系统和海量的气象数据，利用知识发现的理论与方法，从大量云图资料中寻找隐含的，先前未知的，可用于地面的温度、压力、湿度等气象要素预测和辅助气象预报决策的知识与规律。因此，在复杂类型数据的知识发现系统总体结构模型的基础上，我们构建了一个基于相似模式的气象云图数据的知识发现与短期气象预测系统原型，即图像数据类型的知识发现系统原型。实验结果表明，该系统原型提高知识发现效率、减少数据噪音的干扰；具有广泛的适应性，能够很好地求解复杂数据类型的知识发现问题。

5　结论

笔者提出并具体构造了基于复杂类型数据的知识发现总体结构模型 DFSSM，该结构模型不仅给出了复杂类型数据知识发现的进程、发现线路，而且给出了一般化的发现方法——DFSSM 方法；只要将 DFSSM 方法与具体类型的挖掘算法相结合，即可得到知识发现的结果。由于复杂类型数据具有非线性动力学性质和特征，很难用普通的知识表示方式进行描述。因此，采用了模式（定义为 Hilbert 空间中的矢量）来定量地表征复杂类型数据的复杂多变及具有的不确定行为和状态。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。

在深入探讨基于复杂类型数据的知识发现总体结构模型之后，用基于 Web 的文本挖掘系统和基于图像数据类型的知识发现系统作为实例进行了验证，结果表明该总体结构模型对于非结构化的复杂类型数据的知识发现过程具有一定的指导作用，具有较好的实用性与有效性。今后，将更加深入地研究其内在机理，并且不断地扩充和完善该结构模型及其方法。

参考文献

[1] 杨炳儒，王建新，KDD 中双库协同机制的研究(I)[J]. 中国工程科学，2002，4(4)：26～32

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

[3] Piatetsky-Shapiro G, Frawley W J. Knowledge discovery in databases [M]. AAAI/MIT Press, 1991, 166～175

[4] Fayyad U M, Piatetsky-Shapiro G, Smyth P, et al. Advances in knowledge discovery and data mining [M]. AAAI/MIT Press, 1996. 20～32

[5] Piatetsky-Shapiro G, Fayyad U, Smith P. From data mining to knowledge discovery: an overview [A]. In Fayyad U M, et al. Advances in Knowledge Discovery and Data Mining [C]. AAAI/MIT Press, 1996. 1～35

（下转第 68 页）