国际知识发现与数据发掘工具评述*)

International Survey of Knowledge Discovery and Data Mining Tools

欧阳为民　郑　诚　张　燕

（安徽大学计算中心　合肥230039）

摘　要　数据库中的知识发现是一个正在迅速发展的新兴领域，受到了学术研究界和企事业单位的广泛重视。在过去几年间，知识发现工具主要用于研究环境；而现在，复杂的工具产品正不断出现。在本文中，我们概述常见知识发现任务及其解决方法，并按照知识发现与数据发掘工具的一般特征、数据库连接性和数据发掘特征这三大项指标分析研究13种具有代表性的软件产品，这些产品有的是研究原型，有的是商品化的。最后，我们指出为了有效地满足用户需求，以及为了解决那些尚未解决或尚未充分解决的问题，知识发现软件所应该具有的重要特性。

关键词　数据库中的知识发现，数据发掘

1. 引言

在当今这样一个信息时代里，电子数据管理方法正迅速地不断涌现。用于收集，管理数据的功能强大的数据库系统几乎在所有的大中型企业中得到了应用。几乎每一项事务都会生成一条计算机记录。这些大量数据中包含了富有价值的信息，如趋势和模式，可以用于改进和优化事务决策。然而，今天的数据库包含的数据如此之多，几乎不可能为获得有价值的决策支持信息而进行人工分析。在许多情况下，为了准确地为系统行为建模，有必要同时考虑数百个彼此独立的属性。所以，数据分析人员在其分析能力方面需要帮助。

现在人们已广泛认识到这种对于从大量数据中自动提取有用知识的需求，并且，这种需求最终导致了迅速发展的自动分析与发现工具市场。知识发现和数据发掘是从大型数据库中发现隐藏的策略信息的技术。自动发现工具能够分析原始数据，并为分析专家或决策者表达所提取出来的高层次信息，而不是要求分析专家自己去进行这种发现与表达。

在过去的几年间，知识发现与数据发掘工具主要用于实验和研究环境，而今天已经处在面向主流商业用户的复杂工具迅速涌现的阶段。几乎每个月都有新的工具面市。Meta 公司1996年估计数据发掘市场的规模将从1996年的5亿美元发展到2000年的80亿美元。

本文讨论分析大量数据和将其转化为有价值的决策信息的过程，描述了多种分析任务以及完成这些任务的相关技术。也描述用于评价知识发现工具的三大特性，并以此评述现有的软件工具。最后总结中提出真正实用的发现工具所应该具有的某些特性，从而为进一步研究指出了方向。

2. 知识发现和数据发掘

2.1　知识发现过程

关于术语 KDD 和数据发掘还有一些模糊之处。这两个术语经常交换使用，我们用 KDD 这一术语是指把低级别的数据转化为高级别数据的整个过程。KDD 可定义为：在数据中发现有效、新颖、潜在有用并最终可理解模式的非平凡的过程。而数据发掘则通常定义为从数据中提取模式。尽管数据发掘是知识发现的核心步骤，但这一步通常只占整个过程的很小的一部分（大约15%-25%），知识发现的其它步骤包括：

- 理解应用领域和数据发掘过程的目标
- 获取或选择目标数据集
- 数据净化，预处理和变换
- 模型研制与假设建立
- 选择合适的数据发掘算法
- 结果解释与可视化
- 结果测试和证实
- 已发现知识的使用与维护

2.2　数据库问题

任何实际的知识发现过程都不是线性的，而具有较高的重复性和交互性，任何一步都可能对前面的几

*)本课题得到国家自然科学基金（№69975001）和安徽省教委科研基金资助。欧阳为民　博士，教授，主要研究方向：KDD，机器学习，人工智能及其应用。

步产生影响，这样就产生了多种多样的循环往复。这些工具需要和数据库系统或数据仓库紧密结合，以便于数据选择、预处理、后处理以及数据变换。当前可用的许多工具都是来自人工智能或统计学界的通用工具。这些工具通常独立于数据源，需要大量的时间用于数据的输入输出、预处理、后处理以及数据变换。然而，人们显然希望利用现有的DBMS来实现知识发现工具和分析数据库之间的紧密联系。为了评述知识发现工具，我们将检查下列特性：

存取各种数据源的能力：在很多情况下，待分析数据是分散存储的，在进行有意义的分析之前，必须对这些数据进行收集、检查和整合，直接存取不同数据源的能力可以极大地降低数据变换的代价。

在线/离线数据存取：在线数据存取是指可对数据库直接进行查询，并且与其它事务并发运行。离线数据存取是对一个数据源的快照进行查询，在许多情况下，包括了一个从数据源到按照知识发现工具指定的数据格式进行输入/输出的处理。当人们不得不面对不断变化的知识和数据时，在线或离线数据存取问题就变得特别重要。例如，在金融市场中，快速改变的市场条件可能会使以前已发现的规则和模式无效。

潜在的数据模式：如今，许多工具都是以单一表格的形式作为其输入，其中每个样例(记录)都具有固定数目的属性。也有一些工具是建立在关系模型上，并且允许对数据库进行查询。面向对象和非标准的数据模型，如多媒体、空间或时态数据模型，大大超出了当前KDD技术的能力范围。

表/行/属性的最大数：是发现工具处理能力的理论限制。工具可适当处理的数据库规模：待分析数据的预计量是选择发现工具的一个重要因素。表/行/属性的最大个数在理论上有限制，而计算时间、存储需求、表达和可视化能力等方面也存在实际限制。例如，即使在理论上行的最大个数没有限制，将所有数据保存在内存中对于大型数据源来说也是不适合的。

工具可处理的属性类型：某些发现工具对输入数据的属性类型作了限制。例如，基于神经网络的工具通常要求所有属性都是数值型的。而其它的方法可能无法处理连续(实数)值。因此，当我们选择一种分析工具时要考虑出现在数据源中的属性类型。

查询语言：查询语言是用户与知识和数据库之间的界面。它允许用户处理数据和知识，引导发现过程。一些工具没有查询语言，人的交互作用被限制为对一些过程参数的指定上。其它有查询语言的工具允许通过以查询语言描述的公式化查询语句存取数据/知识。这可能是SQL一类的标准语言，或者是一种面向应用的特殊语言。数据和知识查询也可能通过图形用户界面(GUI)进行。

本文所评论的工具在上述各个特性方面均有很大的区别，因此工具的选择取决于特别的应用需求和考虑，如数据的格式与规模、发现过程的目标、最终用户的需求和训练等等。

2.3 数据发掘任务

KDD过程的核心是从数据中提取模式的数据发掘方法，这些方法可以有不同的目标，这取决于KDD过程的期望结果。应该注意的是，具有不同的目标的不同方法可能会被相继应用。例如，为了确定哪些顾客喜欢购买新产品，商务分析员可能会首先利用聚类方法来划分顾客数据库，然后应用回归方法对每个划分簇预测购买行为。

大部分的数据发掘目标可如下分类：

数据处理：根据KDD过程的目标和需求，分析员可以对数据进行选择、过滤、聚类、抽样、净化和变换。自动处理一些典型数据处理任务并将其和谐地集成到整体处理过程可以削除，或者至少降低例程编制和数据输出输入的需求，从而提高分析员的工作效率。

预测：已知数据项和预测模型，预测该数据项特定属性的值。例如，已知一个信用卡事务的预测模型来预测某特定事务是否欺诈。预测也可用于证实已提出的假设。

回归：已知一数据项目集，回归是分析一些属性值对同一个项目中其它属性值的依赖性，自动生成一个预测模型来对新记录的这些属性值进行预测。例如，已知一信用卡事务数据集，为新事务建立一个模型来预测新事务的欺诈可能性。

分类：已知一个预定义的分类集，确定一个特定的数据集属于这些分类中的哪一个。例如，给定与医疗反应相对应的病人分类，确定一个新病人最可能的医疗反应。聚类：已知一个数据项目集，将该集合划分成一个类集，使得类内相似性最大，而类间相似性最小。聚类对于寻找相似项目群很有用。例如，已知一个顾客数据集，确认一个具有相拟购买行为的顾客子群。

连接(关联)分析：已知一个数据项目集，确定属性和项目之间的相互关系，例如一个模式的出现意味着另一个模式的出现。这种关系可以是同一个数据项目的属性间的关联(买牛奶的顾客中有64%也买面包)，也可以是不同数据项目之间的关联(任何时候某样东西下降5%，那么4个星期后另一样东西就上升12%)。

模型可视化：可视化在人们理解、解释已发现知识方面起着重要的作用。可视化技术包括从简单的散点图和直方图到3维动画。

探索性数据分析(Exploratory Data Analysis)：EDA是一种设法从数据集中识别令人感兴趣模式的

交互式分析技术，无需预先设定假设和模型。目前已有很多专门研制的支持数据探索分析的软件包，人们希望将这些方法集成到 KDD 系统中。

2.4 数据发掘方法

由上可知，数据发掘不是一种单一的技术，任何一种能从数据中得到更多信息的方法都是有用的。不同的方法服务于不同的目标，每种方法都有自己的优缺点。然而，大多数的数据发掘方法可以分成如下几类。

统计方法：从历史上看，统计工作主要集中在测试预先的假说以及使模型适合于数据等研究上。统计方法通常依赖于一个明确的潜在概率模型。此外，人们假定这些方法是由统计学家来使用的；为了生成候选假说和模型，人的干预是必要的。

基于事例的推理：基于事例的推理(Case-Based Reasoning) 是一种试图通过直接使用过去的经验和方法来解决一个给定问题的技术。事例是以前遇到过或解决过的一个特定问题。给定某个特定的新问题，CBR 检测一系列已存储的实例，并寻找与之相似的事例。如果存在相似的事例，其解决方案就应用于这个新问题，并且将该问题加入到事例库中，以备将来参考引用。

神经网络：神经网络(Neural Networks)是一种模仿人脑的系统模型。人脑由数以百万计的通过突触相互联系的神经元组成，NN 也是由大量的模拟神经元组成的，它们通过类似于人脑神经元的方法相互联系。象人脑一样，神经元相互关联的强度可能会改变(或被学习算法改变)，以响应某个已出现的刺激或某个以获得的输出，这就使得网络能够"学习"。

决策树：一棵决策树是这样一棵树，该树的每个非终端点均表示被考察数据项目的一个测试或决策。根据测试结果，选择某个分支。为了分类一个特定数据项目，我们从根结点开始，一直向下判定，直到到达一个终端结点(或叶子)时为止。当到达一个终端结点时，一个决策便形成了。决策树也可解释成一种特殊形式的规则集，其特征是规则的层次组织关系。

规则推理：规则描述了一个数据项目中某些属性之间或数据集中数据项目之间的统计相关性。关联规则一般形式为 $X_1,\cdots,X_n \Rightarrow Y[C,S]$，意思是属性 $X_1,\cdots,X_n$ 能预测 Y，其信任度和支持度分别为 C 和 S。

贝叶斯信念网络：贝叶斯信念网络(Bayesian Belief Networks)是概率分布的图形化表示。BBN 是有向无环图，其结点表示属性变量，边表示属性变量间的概率依赖性。与各结点相关的是描述相应结点与其父结点之间关系的条件概率分布。

遗传算法/演化程序设计：遗传算法和演化程序设计均是模仿自然演化规律的算法优化策略。在一个彼此相互竞争的潜在问题解法集中，最佳解法被选出来并与其它解法相互结合。为此，我们期望方法集中总体优势变得越来越好，类似于一个有机体种群的进化过程。遗传算法和演化程序设计可用于数据发掘，以构造变量之间的依赖性假设，其形成可以是关联规则或其它内部机制。

模糊集：一种表达和处理不确定性的重要方法。不确定性以多种形式发生在今天的数据库模型中，如不精确、不完全、不典型、不一致、含糊，等等。模糊集利用不确定性使系统的复杂性变得可处理。当精确输入不可能或太昂贵时，模糊系统就是一种强有力的模型方法。

粗糙集：用一个集合的上下界来定义。下界中的每个成员都是这个集合的成员，而上界的每个非成员也一定是这个集合的非成员。粗糙集中的上界由下界和边界区域的并构成。边界区域的成员可能(但是不能肯定)是这个集合中的成员。因此，粗糙集可以被看成是一个有三级成员函数(是，否，可能)的模糊集。象模糊集那样，粗糙集是处理数据不确定性的一种数学概念。与模糊集类似，粗糙集很少单独使用，而通常是与规则推导、分类、聚类等其它方法一起组合使用。

3. 通过特性分类机制研究知识发现和数据发掘工具

如第2节所述，知识发现和数据发掘工具在数据选择、预处理、集成和变换等等方面需要与数据库系统或数据仓库紧密结合。在数据模型、数据规模、查询支持等方面，并不是所有工具都具有相同的数据库特征。不同的工具可以完成不同的数据发掘任务，并可利用不同的方法达成其目标。一些工具需要或支持与用户较多的互动，而另一些工具则较少互动；一些运行于一个独立的体系结构，而另一些则工作于客户/服务器体系结构。为刻画所有这些不同点，我们将知识发现和数据发掘工具的特性分为一般特征、数据库连接性和数据发掘特征三大类。

3.1 一般特征

知识发现和数据发掘工具的一般特征如表3.1所示。

(1)产品：软件产品名称和开发商。

(2)产品状况：产品开发状况。P==商业版，A==Alpha 版，B==Beta 版，R=研究原型。

(3)法律状况：PD=公用领域，F=免费软件，S==共享软件，C==商业软件。

(4)学术许可：针对商业产品，如果有特别免费或用于学术研究的降价许可，给予说明。

(5)演示：如果有演示版本，给予说明，D=可从网

上下载的演示版本，R＝需申请可获得的演示版本，U＝未知。

(6)体系结构：计算机体系结构，S＝独立，C/S＝客户/服务器，P＝并行处理。

(7)操作系统：支持软件运行的操作系统。

3.2 数据库连接性

知识发现和数据发掘工具的数据库连接性如表3.2所示。

(1)数据源：说明待分析数据的可能格式。F＝ASCII码文本文件，D＝Dbase文件，P＝Paradox文件，F＝Foxpro文件，Ix＝Informix，O＝Oracle，Sy＝Sybase，Ig＝Ingres，A＝MS Access，OC＝Open database connection(ODBC)，SS＝MS SQL Sever，Ex＝MS Excel，L＝Lotus 1-2-3。

(2)数据库连接：数据库连接的类型。Onl.＝在线：能直接对数据库进行查询，并且能与其它事务并发运行，Offl.＝离线：通过对数据源的快照进行分析。

(3)规模：软件可以适当处理的记录的最大个数。S＝小(10,000条记录以下)，M＝中(10,000到1,000,000条记录)，L＝大(超过1,000,000条记录)。

(4)模型：待分析数据的数据模型。R＝关系型，O＝面向对象型，1＝单一库表。

(5)属性：软件可以处理的属性类型。Co＝连续型，Ca＝类别型(离散数值)，S＝符号型。

(6)查询：说明用户如何构造对知识库进行查询和引导发现过程。S＝结构化查询语言，Sp.＝面向特殊应用的界面语言，G＝图形用户界面，N＝不可用，U＝未知。

3.3 数据发掘特征

知识发现和数据发掘工具的数据发掘特征如表3.3所示。

(1)发现任务：产品所设计的知识发现任务。Pre.＝数据预处理(抽样过滤)，P＝预测，Regr＝回归，Cla＝分类，Clu＝聚类，A＝关联分析，Vis＝模型可视化，EDA＝探索性数据分析。

(2)发现方法：用于发现知识的方法类型。NN＝神经元网络，GA＝遗传算法，FS＝模糊集，RS＝粗糙集，St.＝统计方法，DT＝决策树，RI＝规则推理，BN＝贝叶斯网络，CBR＝基于事例的推理。

(3)人的交互作用：说明在多大程度上需要或支持人与发现过程的交互作用，A＝自治的，G＝由人引导发现过程，H＝高度交互。

在表3.1、3.2和3.3中，我们利用三大类特征研究刻画了43种现有的知识发现和数据发掘工具，尽管现在人们已普遍认识到存取多样化数据源能力的重要性，但表3.2表明当前大多数工具仍然只支持少数几种数据格式。出人意料的是少数工具能同时分析若干张数据库表。几乎所有产品都能够分析连续型、离散型和符号型属性，大多数是通过属性类型的转换来实现，即将符号变量转变为数值型变量或将连续属性范围分解为若干区间。

从表3.3我们知道大多数工具使用"标准的"数据发掘技术，象规则推理技术，决策树，统计学方法，来自其它富有希望领域中的技术，如模糊集、粗糙集和遗传算法这些方法在知识发现软件的应用还仅仅只是个开始。

4. 结论和进一步研究

知识发现可以定义为从大型数据集中自动发现新颖有用的信息。数据发掘是知识发现核心过程中的一步，用于从大量数据中提取模式与关系。

今天，许多企业正在积极地收集、存储大规模数据。它们中的大多数已经认识到这些数据作为商业决策信息源的潜在价值。对更好的决策支持需求的急剧增长导致知识发现和数据发掘的不断涌现。我们在本文中对常见知识发现任务、解决这些任务的方法和采用这些方法的软件工具进行了评述。无论KDD如何快速地增长，它也仍是一个新生领域。成功的数据发掘应用仍是一个漫长而艰苦的过程。下面列举了一些尚未解决或至少尚未很好解决的问题：

不同技术的集成。现有工具或者仅用某种单一的技术或者仅用有限的几种技术来进行数据分析。我们一开始就指出在数据分析方面没有所谓的最好技术，因此问题不是某一种技术比另一种更好，而是哪一种技术更适合要解决的问题。一个真正有用的工具必须为解决不同的问题提供不同的解决方法。

扩展性本质上是源自不同技术处理不同问题的各自优势。随着已提出的技术及其应用的不断增加，我们越来越清楚地看到任何算法都不是万能的。因此，重要的是提出一种体系结构，使得新方法易于合成，已有方法便于援用。

与数据库的无缝联接。当前的数据分析产品大体上分为两类。第一类是数据分析和报表系统，由RDBMS开发商提供，有时也可能有在线数据分析处理OLAP能力。这些系统与数据库紧密连接，通常利用了DBMS的处理能力和扩放能力。它们局限于测试用户所提供的假设，不能自动提取模型和模式。第二类是模式发现工具，它能自治地探测数据中的模式。这些工具倾向于离线存取数据库。这些工具一般没有充分的数据处理能力，而将数据预处理完全交给用户。这就会导致大量费时的重复性的I/O处理。

既支持分析专家又支持初学者。大多数工具是针

表3.1 一般特性

Product	Prod Status	Legal Status	Acad. Lic	Demo	Arch	Operating System							
						MF	Unix	Mac	Dos	W3x	W9x	WNT	OS/2
Alice (Isoft)	P	C	N	D	S						x	x	
AutoClass (Nasa)	P	PD	-	D	S,P		x		x				
Bayesian Knowl Disc (Open U.)	B	F	-	D	S		x	x					
BrainMaker (Cal.Sc Software)	P	C	N	N	S			x	x	x	x	x	
Brute (Univ. of Washington)	P	C	U	N	S		x						
BusinessMiner (Business Objects)	P	C	N	N	S					x	x	x	
Claudien (K.U.Leuven)	P	C	Y	D	S		x						
Clementine (Integral Solutions Ltd.)	P	C	N	N	S		x					x	
CN2 (Univ. of Texas)	R	PD	-	D	S		x						
CrossGraphs (Belmont Research)	P	C	N	N	S		x	x		x	x	x	
Cubist (RuleQuest)	P	C	D	D	S		x				x	x	
C5.0 (RuleQuest)	P	C	D	D	S		x			x	x	x	
Darwin (Thinking Machines)	P	C	N	N	C/S,P,S		x						
Data Surveyor (Data Destillenes)	P	C	S	R	C/S,P		x	x		x	x	x	x
DBMiner (SFU)	R	C	S	N	C/S		x				x	x	
Delta Miner (Bissantz)	P	C	S	D	C/S,P						x	x	
Decision Series (NeoVista)	P	C	N	N	C/S,P		x						x
DI Diver (Dimensional Insight)	P	C	N	D	C/S,S		x	x		x	x	x	
Ecobweb (Tel Aviv Univ.)	P	PD	-	D	S		x						
IDIS (Information Discovery)	P	C	N	N	C/S,S	x	x	x		x	x	x	x
IDN(Nasa)	P	C	D	S	S		x						
Intelligent Miner (IBM)	P	C	S	R	C/S/P,S	x	x			x	x	x	x
KATE-Tools (AcknoSoft)	P	C	N	N	C/S,S		x	x		x	x	x	x
Kepler (GMD)	R	C	S	R	S		x				x		
KnowledgeSEEKER	P	C	N	D	C/S,S		x			x	x	x	
MineSet (Silicon Graphics)	P	C	S	N	C/S,S		x						
MLC++	P	F	-	D	S		x					x	
Mobal (GMD)	P	F	-	D	S		x						
ModelQuest (AbTech)	P	C	N	N	S		x				x	x	
MSBN (Microsoft)	P	F	-	D	S						x	x	
MVSP (KCS)	P	S	D	D	S				x	x	x	x	
PolyAnalyst (Megaputer)	P	C	F	D	C/S,S						x	x	x
PVE (IBM)	P	C	N	N	S		x						
Q-Yield (Quadrillion)	P	C	F	D	S						x	x	
Ripper (AT&T)	P	C	F	D	S		x						
Rosetta (NTNU)	R	C	S	D	S						x	x	
Rough Enough (Troll Data)	R	F	-	D	S				x	x	x	x	
Scenario (Cognos)	P	C	N	R	S						x	x	
Sipina-W (Univ. of Lyon)	P	S	-	D	S					x			
SuperQuery (AZMY)	P	C	N	D	S						x	x	
ToolDiag (UFT-S)	P	F	-	D	S		x		x				x
Weka (Univ. of Waikato)	P	C	F	D	S		x						
WizRule (WizSoft)	P	C	N	D	S						x	x	

表3.2 数据发掘特征

Product	Data Sources													DB Conn		Size	Mod	Attributes			Queries
	T	D	P	F	Ix	O	Sy	Ig	A	OC	SS	EX	L	onl	offl			Co	Ca	S	
Alice (Isoft)	x	x	x	x		x			x	x		x	x		x	L	I	x	x	x	S,G
AutoClass (Nasa)	x	x													x	L	I	x	x	x	N
Bayesian Knowl.Disc (Open U.)	x														x	L	I		x		N
BrainMaker (Cal Sc.Software)	x	x										x	x		x	S	I	x	x		N
Brute (Univ. of Washington)	x														x	S	I	x	x	x	N
BusinessMiner (Business Objects)	x											x			x	M	R	x	x	x	S,G
Claudien (K.U Leuven)	x														x	S	R		x	x	U
Clementine (Integral Solutions Ltd.)	x				x	x	x	x		x		x	x	x	x	M	R	x	x	x	G
CN2 (Univ. of Texas)	x														x	S	I		x	x	N
CrossGraphs (Belmont Research)	x			x		x				x				x	x	M	R	x	x	x	Sp
Cubist (RuleQuest)	x														x	M	I	x	x	x	N
C5.0 (RuleQuest)	x														x	M	I	x	x	x	N
Darwin (Thinking Machines)	x	x	x	x		x	x							x		L	R	x	x	x	S
Data Surveyor (Data Destilleries)		x	x	x		x	x			x				x		L	R	x	x	x	S
DBMiner (SFU)										x				x		L	R	x	x	x	S,G
Delta Miner (Bissantz)										x		x			x	L	R	x	x	x	S,G
Decision Series (NeoVista)	x	x			x	x	x			x				x		L	R	x	x	x	S
DI Diver (Dimensional Insight)	x									x				x	x	M	R	x	x	x	S,G
Ecobweb (Tel Aviv Univ.)	x														x	S	I	x	x	x	N
IDIS (Information Discovery)	x	x				x	x				x			x	x	L	R	x	x	x	S,G
IDN(Nasa)	x														x	S	I	x	x	x	N
Intelligent Miner (IBM)	x	x				x	x							x	x	L	R	x	x	x	S,G
KATE-Tools (AcknoSoft)	x					x			x	x	x			x	x	M	R	x	x	x	S,G
Kepler (GMD)	x														x	M	R	x	x	x	S,G
KnowledgeSEEKER	x	x	x						x	x	x	x	x		x	L	I	x	x	x	S,G
MineSet (Silicon Graphics)	x				x	x	x							x	x	L	R	x	x	x	S,G
MLC++	x														x	L	R	x	x	x	N
Mobal (GMD)	x														x	S	R		x		S,G
ModelQuest (AbTech)	x														x	L	R	x	x	x	G
MSBN (Microsoft)	x														x	S	R	x	x	x	G
MVSP (KCS)		x	x									x	x		x	M	I	x	x		N
PolyAnalyst (Megaputer)	x	x	x	x	x	x	x	x		x	x	x			x	M	R	x	x	x	G
PVE (IBM)	x	x													x	M	I	x	x	x	U
Q-Yield (Quadrillion)	x									x				x		S	R	x	x	x	S
Ripper (At&T)	x														x	L	I	x	x	x	N
Rosetta (NTNU)	x									x					x	S	R	x	x	x	S,G
Rough Enough (Troll Data)	x		x											x	x	S	R	x	x	x	S
Scenario (Cognos)	x	x	x	x								x		x	x	M	R	x	x		S,G
Sipina-W (Univ. of Lyon)	x	x	x										x		x	M	I	x	x	x	G
SuperQuery (AZMY)	x	x	x	x					x	x		x	x	x		M	R	x	x	x	D,A
ToolDiag (UFES)	x														x	S	I	x			N
Weka (Univ. of Waikato)	x														x	M	R	x	x	x	N
WizRule (WizSoft)	x	x	x							x				x		M	I	x	x	x	N

表3.3 数据发掘特征

Product	Tasks								Methods									Interaction		
	Pre	P	Regr	Cla	Clu	A	Vis	EDA	NN	GA	FS	RS	SI	DT	RI	BN	CBR	A	G	H
Alice (Isoft)																				
AutoClass (Nasa)																				
Bayesian Knowl Disc (Open U)																				
BrainMaker (Cal Sc Software)																				
Brute (Univ of Washington)																				
BusinessMiner (Business Objects)																				
Claudien (K U Leuven)																				
Clementine (Integral Solutions Ltd)																				
CN2 (Univ of Texas)																				
CrossGraphs (Belmont Research)																				
Cubist (RuleQuest)																				
C5 0 (RuleQuest)																				
Darwin (Thinking Machines)																				
Data Surveyor (Data Destilleries)																				
DBMiner (SFU)																				
Delta Miner (Bissantz)																				
Decision Series (NeoVista)																				
DI Diver (Dimensional Insight)																				
Ecobweb (Tel Aviv Univ)																				
IDIS (Information Discovery)																				
IDN(Nasa)																				
Intelligent Miner (IBM)																				
KATE-Tools (AcknoSoft)																				
Kepler (GMD)																				
KnowledgeSEEKER																				
MineSet (Silicon Graphics)																				
MLC++																				
Mobal (GMD)																				
ModelQuest (AbTech)																				
MSBN (Microsoft)																				
MVSP (KCS)																				
PolyAnalyst (Megaputer)																				
PVE (IBM)																				
Q-Yield (Quadrillion)																				
Ripper (AL&T)																				
Rosetta (NTNU)																				
Rough Enough (Troll Data)																				
Scenario (Cognos)																				
Sipina-W (Univ of Lyon)																				
SuperQuery (AZMY)																				
ToolDiag (UFES)																				
Weka (Univ of Waikato)																				
WizRule (WizSoft)																				

对分析专家的，典型的最终用户，如商场主，工程师或企业主，他们缺少复杂的数据分析技能，但对他们自己所从事的领域有深刻的理解，此外，他们对能够清晰快速地回答他们的日常业务问题感兴趣。

最终用户需要使用简单的能有效解决其业务问题的工具，现有的软件包在指导分析过程和以便于用户理解的方式表达分析结果方面缺乏充分的支持。

管理变化的数据。在许多应用中，数据都不是静态的，而是不断变化发展的，不断变化的数据使得以前的发现模式不再有效，而必须以新的模式取而代之，当前解决这一问题的唯一方法是在周期性时间段内重复相同的分析过程。显而易见，人们需要对不断变化的模式进行增量式更新的方法和识别与管理发生于知识库中的时态变化模式的策略。

非标准数据类型。今天的数据库不仅包括像数值和字符串这样的标准数据，也包括大量的非标准的和多媒体数据，如自由格式文本，音频，视频和图像数据，时态、空间以及其它数据类。这些包含特殊模型的数据类型用标准分析方法不能很好地处理。因此，这些应用需要特殊的，常常是面向特定领域的方法和算法。

一方面仍有一些根本问题有待解决，另一方面，已有很多重要的成功范例被报道，并且其结果和优点令人印象深刻。虽然当前的方法仍然依赖于相当简单且能力有限的几种技术，但却能获得有用的结果，KDD技术的巨大潜力已在范围广阔的应用领域中令人信服地得到了展示，前景广阔的实际应用需求使我们期待该领域有一个健康美好的未来。

（上接第118页）

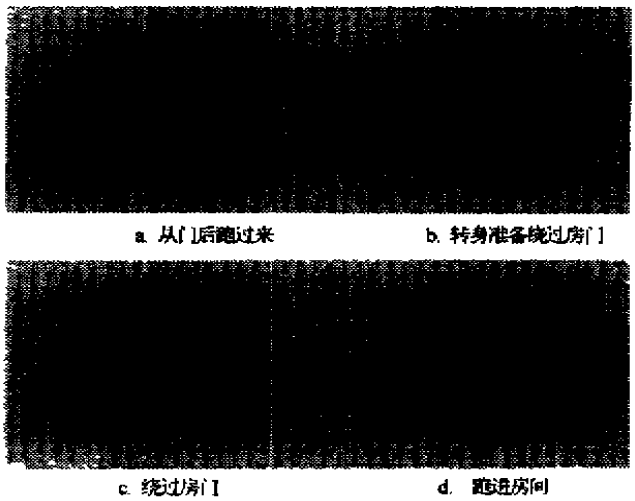

图5(a)-(d) 为 intelligent agent 跑进房间的仿真结果

结论 在基于 Intelligent Agent 的动画系统中，为使 Intelligent Agent 在 VR 环境中智能性地模拟人类的行为，我们在其高级行为控制部分采用进行 Intelligent Agent 的路径规划，在多边形膨胀法建模的基础上，利用多边形的无碰公切线够造切线图，通过 A* 算法搜索从初始位置到目标位置的无碰路径。动画结果表明 C-空间法路径规划在 Intelligent Agent 的 3D 人体动画中是切实可行的，它在 Intelligent Agent 的智能行为控制中发挥了重要作用。

参 考 文 献

1 Gilbert E, Johnson D W. A Fast Procedure for Computing the Distance Between Complex Objects In Three-Dimensional Space. IEEE Journal of Robotics and Automation. 1988, 4(2): 193～195

2 Fukumoto M, Mase K, Suenaga Y. Real-time Detection of Pointing Actions for A Glove Free Interface. In IAPR Workshop On Machine Vision Applications, 1992. 473～476

3 Tomas Lozano-Perez. Spacial Planning: A Configuration Space Approach. IEEE Tr. Computers, 1983, 32(2): 108～120