计算机科学 1996 Vol. 23 №. 6

KDD 研究现状及发展*)

陈 栋 刘 兵 徐洁磐

(南京大学计算机软件新技术国家重点实验室 南京 210093)

摘 要 A comprehensive summarization of the recent researches on Knowledge Discovery in Databases is presented in this paper. Fundamental ideas and methods are deliberated. Analyses on the current studies demonstrate fruits bore as well as drawbacks existed. Further directions on KDD are also discussed in this paper.

关键词 Knowledge, Discovery, Database, Data Analysis.

计算机技术和信息网络技术的发展把人们推入了信息社会。信息的增长呈现超指数上升，据有关估计，全世界的信息量不到每 20 个月就增加一倍。人们寄予希望的数据库技术在一定程度上帮助了人们有效地利用信息，但目前它只是象文具盒一样作为一种容器，你不可能得到像橡皮要和铅笔配合使用的那样别的东西。尽管利用统计学原理发展的数据分析技术已出现多年，但能够分析发现多种模式的智能技术还远不成熟，结果数据理解与数据产生之间出现了越来越大的距离。

于是，一个新的研究领域——知识发现应运而生。由于蕴藏知识的信息大多存储于数据库中，数据库中的知识发现(Knowledge Discovery in Databases, KDD)，又称数据挖掘(Data Mining)成为研究的焦点。这一研究领域兴起于八十年代初，机器学习和数据分析的理论及实践成为知识发现研究的铺垫。研究人员从各个角度出发作了大量的探索。极大的商业应用前景是推动 KDD 研究的又一主要因素，财务预算、公共决策、医疗分析等众多领域都急需智能的分析工具。GM 通用汽车公司的汽车事故诊断专家系统已初步体现了 KDD 的思想。

本文首先介绍 KDD 的有关概念，指出 KDD 的工作目标，分析、比较现有的方法和思想，并介绍比较成功且有代表性的几个实现系统。

一、数据库中的知识发现

数据库中的知识发现，又称数据挖掘，简单地说，就是对数据库中蕴涵的、未知的、非平凡的、有潜在应用价值的模式的提取。有关 KDD 的一些基本特征如下：

模式：用高级语言表示的表达一定逻辑含义的信息，这里通常指数据库中数据之间的逻辑关系。

知识：一个满足用户兴趣度和置信度的模式。

置信度：一则知识在某一数据域上为真的量度。置信度涉及到许多因素，如数据的完整性、样本数据的大小、领域知识的支持程度等。没有足够的确定性，模式不能成为知识。

兴趣度：一则在一定数据域上为真的知识被用户关注的程度。

有效性：知识的发现过程必须能够有效地在计算机上实现。

非平凡性(nontrivial)：能够以确定的计算过程提取的模式称为平凡知识。平凡的知识(如根据数据库中的薪水字段求得职员的平均薪水)不是 KDD 的目标。在 KDD 中，知识的发现过程都应具有某种不确定性和一定的自由度，也就是要发现不平凡的知识。

一个知识发现系统的基本输入是数据库中的原始数据。发现系统需要特别关注数据本身固有的一些性质。数据库中存在与知识发现相关的一些问题，其中每一个问题在机器学习中都有一定程度的涉及，但少有系统全部透彻地阐述过。总的说来，处理这些问题是对 KDD 的一大挑战。

* 动态数据：大多数数据库的内容将经常变化。在一个在线系统中，必须采用预警机制来保证这些变化不导致错误的发现。

* 噪声和不确定性：错误的数据对于现实世界数据库是在所难免的，主要在于数据采集的各个环

*)国家自然科学基金资助项目，编号为 69573014。

节。另一种不确定性存在于发现的模式只可能在一部分数据上有效。

* 不完整数据：由于不完整的数据域和数据域上值的缺少所造成的不完整数据，当然会影响发现的结果。数据库的设计并不考虑发现，模式的发现、评价、解释很可能需要数据库中不存在的信息。

* 冗余信息：指同一数据在数据库中的多处出现。冗余信息有时会误导知识的发现过程，如此所发现的知识缺乏足够的兴趣度。

* 稀疏数据：数据库中的信息在实例空间中可能是稀疏的，会严重影响发现的效率。

二、发现目标

在这里，主要讨论关系数据库中数据之间的逻辑关系，分析发现的具体目标。概括地，关系数据库中数据之间的逻辑关系可分为域间关系和记录间关系。域间关系指表中各域之间的取值关系。记录间关系指表中记录间的取值关系。从数据间逻辑关系的内涵来讲，又可分为以下几种：

分类(Classification) 是最基本的一种认知形式，在这里属于一种记录间关系，早在KDD研究之前，ID_3、Knn、Bayesian等诸多算法运用了概率统计、决策树、多元代数空间等理论，使得分类算法的研究就取得了长足进展。

函数依赖 是数据库设计中不可完全避免的，同时也是数据内在联系的体现。在这里，函数依赖属于域间关系。

数据相关 指数值型取值而言，分析域间相关系数，根据领域知识给出描述函数。

体现以上关系的模式可以有多种表示形式，例如自然语言描述、形式逻辑描述以及可视化描述。自然语言描述易于理解，形式逻辑描述有利于推导，可视化描述直观明了。目前，以自然语言描述和形式逻辑描述结合运用为多见。

规则(Rule)是常见的知识表示形式，介于自然语言描述和形式逻辑描述之间，从而成为知识发现的首选目标，分类、约束、关联等都可以以规则表示。

实际的应用有时需要反映数据库中数据的变化情况，如销售趋势等。这时可以采用扩充原数据模型的方法，建立与之相关的、用于知识发现的新数据模型和数据集。

三、知识发现的方法

3.1 发现语言

知识发现的操作需要确定数据范围、发现目标和用户兴趣等方向性指示，于是需要一种类SQL的语言来描述，文〔1〕给出了以下形式的发现语言：

```
discover association rules
from sales_transactions T,sales_item I
where T.bar_code=I.bar_code and I.category=
    "food" and I.storage_period<21
with interested attributes category,content,brand
```

以上的操作基于一个超级市场的销售情况数据库，意在发现"milk"和"bread"之间的关联。标准的SQL语言嵌入其中形成类SQL的知识发现语言。在处理过程中，类SQL语言首先被转换成标准的SQL查询，确定知识发现的工作范围和发现目标。

3.2 发现系统的知识库

知识发现系统不可能完全自主，引导发现的领域知识总是受欢迎的，因为效率的提高是肯定的。同样地，知识发现不同阶段的结果也具有相互启发的作用，因此，需要设计一个知识库存储领域知识和阶段知识，协调并促进发现算法的实施。

领域知识、阶段知识、结果知识对于发现系统来说，是在不同时间得到的。领域知识在发现过程执行前就已明确，阶段知识是在发现算法实施过程中产生并反馈的，有可能在反馈的同时也作为一种结果知识呈现给终端用户，结果知识是系统的工作目标，是逻辑完整、满足兴趣度的模式。三者之间的相互关系决定了它们的表示形式要一致。

3.3 发现方法

KDD的发展经历了短短的十年时间，虽然还未达到实用化的程度，但探索者们已从各自的背景出发做了大量的工作。发现算法是从数据中抽取知识的过程，鉴于数据本身的性质，符号推理、信息论、遗传算法、统计原理等知识被运用到知识发现的算法设计中。

* 基于符号推理的知识发现算法。符号推理是KDD中应用的最为广泛的技术，文〔2〕、〔1〕、Ashok S.〔1995〕都运用了层次概念上的符号推理，以概念包容关系的分类树为基础，逐层分析归纳出相关规则。

* 基于信息论思想的知识发现算法。Kn-MED中David K.Y.C.〔1991〕运用了信息论中最大熵的相关理论。最大熵的方法在科学、工程的许多应用中已被成功地用作标准。Kn-MED迭代地划分结果空间，产生一个层次模式以提供最小信息丢失的分析和合成。在传统的模式识别中，两个主要问题引起了相当的关注。一个是欧拉空间中数据表示的尺度和

距离，另一个是混合多种类型数据的复合类型数据的分析。Kn-MED 运用信息论的方法着重就以上两点展开工作。

* 基于进化计算的知识发现算法。Knn(K nearest neighbours)是一种代数空间分类方法。测试集的每一个特征作为分类空间的一维。当前数据向量的归属依赖于它的最近的 K 个向量的归属。传统的 Knn 算法对各维一视同仁，不区分其重要性，而且计算量很大。W. F. Punch 等人把遗传算法结合于 Knn 提出了 GA-Knn 的算法，通过弯曲特征空间以使类间相似最小，类内相似最大。这样改善了特征提取和分类优化的性能。

* 基于统计方法的知识发现算法。启发式搜索与统计方法相结合是 KOSI(Zhong，1995)中算法的特色，KOSI 中的搜索被分成三种。启发搜索-1 用来发现精确的函数关系。当启发搜索-1 失败，启动启发搜索-2 来发现强函数关系。另一种是基于回归分析的搜索，这是一种统计推理，主要用来依据属性计算对假设函数关系进行评价，寻找可能的弱函数关系。

除了上述的几种知识发现算法外，有些知识发现系统依据系统目标，还设计了特定的知识发现算法。Reduction 算法是 KEPLER 系统(Wu，1991)中的发现算法，用于搜索数据蕴涵的函数关系。多元变量之间的复杂规律是 Reduction 致力解决的难题，它把多元规律分散成二元规律来处理。二元规律的发现相对来说要简单得多，因为二元规律的展现形式是很有限的。Reduction 中启用了原函数的概念。原函数指一个公式中不可分的部分。把公式划分成不同的部分，每一个变量仅出现在一个部分，这样通过发现部分就可以发现整个公式，Reduction 就是通过发现原函数来发现复杂的多元公式的。

四、实现系统的分析

4.1 系统模型

图 1 是 Christopher M. [1993]描述的 KDD 模型。

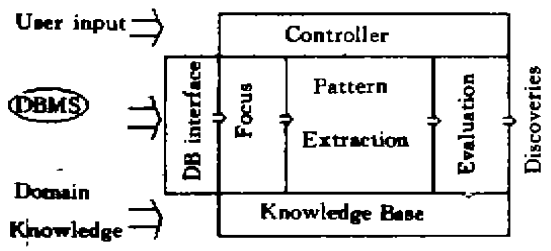

图 1 KDD 系统模型示意图

通用的全自动化的 KDD 系统的实现还远不可能。一般来说，用户必须提供一定的属于系统的控制机制的交互式指导：* 选择要展开工作的数据范围；* 鉴别相关的域；* 细化目标。

4.2 系统介绍

尽管 KDD 的研究进行得如火如荼，但研究者们的出发点大都局限于一个相对狭小的领域，其发现目标也不全尽，且不说其系统的思想深度和效率了。众多的实验性系统都基本上体现了以上模型的结构，但具体不同的工作却各有千秋。下面着重谈谈现有系统的模型及组成的特点。

4.2.1 Explorer。是由德国国家计算机科学研究中心开发的一个知识发现系统。系统主要用于概念化的数据分析和寻找有趣的数据间关系。

Explorer 提供刻画数据间关系的框架，该框架称为模式模板(Pattern Template)或断言原型(Statement Type)，具有较强的模式表示功能。模式模板有三种：①rule searcher 表达具有某种共同特性的一类数据对象。②change detector 表达对象特性的变化。③trend detector 表达对象特性的发展趋势。

对模式模板的实例化就得到了“事实”。“事实”是否有趣依赖于这样一些因素：①可信度：基于统计学和信息论的方法用于计算“事实”的可信度。②可用性：“事实”是否与用户的目标相关。③新颖性：新“事实”与已有知识间距离。④通用性：“事实”涉及到的数据对象所占百分比。

在判定有趣“事实”的过程中涉及到的策略和计算方法主要依赖于系统的领域知识。作为一个交互式系统，Explorer 允许用户根据需要对上述过程中采用的策略和方法作适当调整。

4.2.2 KDW。是一个对数据库作交互式分析的界面友好的工具集，涉及数据库访问、创建新域、定义聚焦(focus)、规划数据和结果、实施发现算法、处理领域知识等。其最新版本嵌入了扩展的命令解释器，以使用户交互地控制发现过程。

KDW 通过基于 SQL 的查询界面直接访问 DBMS。它的知识库包括与数据库相关的域分组、记录分组、函数依赖等信息，用来聚焦(focus)并指导如何从数据库中访问信息。KDW 中的控制唯一地由用户掌握。KDW 中、结果的评价留给了用户。

KDW 中的提取算法包括：①线性相关类的集簇；②决策树分类算法；③偏离检测；④依赖分析。

4.2.3 INLEN[3]。是综合了数据库、知识库、机

器学习技术的工具集成系统，用于数据分析、搜索关系和规则。INLEN 系统由一个既存储有事实的关系数据库和又存储有规则、约束、决策树、方程的知识库组成。它使用了三组操作：数据管理操作（DMOs），知识管理操作（KMOs），知识产生操作（KGOs）。这里着重谈一下知识产生操作。

KGOs 基于知识段（knoweledge segment）实施复杂的推理以产生新的知识。KGOs 使用了以下操作：

Cluster 运用元组上的概念集簇来创建分组，给出标志分组的规则集。

Rulestruct 在规则集上实施概念集簇，返回一个复合的 KS，由附带分组信息的原始规则集和解释分组的新规则集组成。

Diff（differentiate）归纳标志两类对象差距的规则，输出由所产生的规则集和以关系表表示的对象类组成。

Char（characterize）发现描述一个分类中所有对象的特征规则（characteristic rule），不关心该类与其它之间的差别。结果输出包括初始类加上产生的描述信息。

Atest 测试一组决策规则以检查一致性和完整性。

Varsel 发现在区分各类对象时最重要的属性

Esel 发现对于给定分类最具有代表性的一些对象，结果返回一描述输入类的规则、被选中的对象和被拒绝的对象。

Varcon 事实数学操作把变量组合成有用的成份。结果 KS 由新的合成变量、一个确定初始表的规则、数学算子和所创建的变量组成。

Treecon 把一组规则组织成一个决策树，这对于知识的存储和使用是一个有效的方法。

Discor 发现属性值之间的相关，以标准的统计相关来实现，返回一个表。

Dismon 发现属性间的单调相关。Dismon 要被 Diseq 调用。

Diseq 发现描述数值型数据的一组方程，生成适用这些方程的条件。

Stanal 实施统计分析以确定各种统计特性。

4.2.4 KOSI 系统。从数据库中发现函数关系，其关键所在是通过扩展启发式搜索并与一些统计方法相结合来增强处理不确定性的能力。

KOSI 运用两种类型的启发、模型库机制和多种领域知识来控制多重搜索，用以函数关系的产生和评价。KOSI 支持数值类型和符号类型的发现，运用了新的统计方法，如用 AIC 标准来挑选优化模型；用基于 Householder 转换的最小二乘法来做回归分析；用 SCT 算法对时序数据进行集簇。而且，这些统计方法与启发搜索、概念知识相结合以控制发现过程。KOSI 给用户相当大的自由度来选择要进行的分析过程。

KOSI 中的搜索被分成三种。启发搜索-1 和启发搜索-2 用于属性计算以产生新的属性，另一种是基于回归分析的搜索，主要用来依据属性计算对函数关系进行评价。

KOSI 系统是一个层次结构，Level-1 是原始控制层，包括一个模型库、用户界面、全局管理。Level-2 是元控制层，分为两部分：控制属性计算和预处理的领域知识；控制评价和预处理的领域知识。Level-3 是对象层，分为三部分：预处理包括数据收集、聚焦和生成集簇；属性计算包括两种启发搜索；回归分析评价包括回归分析和 AIC 标准。

4.2.5 KEPLER。是一个交互式的、领域无关的发现系统，领域知识主要包含在函数原形中。函数原形以及优先级的选择是领域相关的。KEPLER 的数据源于实验或观测结果，其目的是导出一个数量规律。它可以自动地发现原形库定义范围内的复杂规律，甚至涉及多元方程。

KEPLER 系统可以发现多元方程、检查数据正确性、发现部分关系并提供友好的交互环境，其缺陷是：需大量的规整形式的数据、缺乏利用领域知识、原形库需增强以涵盖更广泛的数学公式（如微分方程等）。

五、KDD 应用前景及进一步的工作方向

KDD 的起源也同样是它的应用领域。最初的应用有发现光谱测定规则（Buchanan and Mitchell，1978），发现诊断大豆疾病的规则（Michalski，1980），发现药物对肺炎病人的副作用规律（Blum，1982）。KDD 的应用远不只这些方面，随着 KDD 技术的进展，对 KDD 的应用也扩展到更多的领域，诸如医药：生物医药、药物副作用、遗传基因分析。金融：信用评定、破产预测、股票市场预测、非法使用信用卡预警。农业：大豆和西红柿疾病分类。社会：人口统计数据分析、选举趋势。市场营销：产品分析、销售预测。工程：自动诊断专家系统、CAD。军事：作战数据分析。科学：气象数据分析、超导研究。

推动 KDD 的研究不仅源于学术上的动机，更是

实际工作的需要。数据复杂性使得需要更多的领域知识，巨大的数据库对算法的效率提出更高的要求，不断变化的环境和信息种类(如多媒体信息)需要新的发现方法，复杂的问题可能需要多种发现策略协作。对社会领域、商业领域的数据库的发现要当心可能是违法的。

自主性(Autonomy)和普适性(Versatility)是KDD系统所追求的两个标准，自主性要求更多的领域知识，而普适性则要求相对地领域无关。这种看上去的不和谐可以得到一定程度的改善。首先，一些领域知识，比如数据依赖，是可以从数据库中自动提取的。其次，我们不妨设计更好的算法从用户那里得到知识，这自然需要功能强大的交互式工具。再次，多重发现的各种结果及阶段知识都可用来指导进一步的发现。

数据库中的数据，尤其是非数值型的数据，很难给出一种贴切的表示使之在数据空间中有合适的位置。而且，由于数据的随机性和数据间的相对弱关联，数据空间可能呈现出稀疏状态。这些都影响发现的准确性和效率。

参考文献

[1]Jiawei Han, Yongjian Fu, Discovery of Mutiple-Level Association Rules from Large Databases, In Proc. 21th Intl. Conf. on VLDB Zurich, Swizerland, 1995

[2]Ramakrishnan S. et al., Mining Generalized Association Rules, Same to[1]

[3]Hongjun Lu et al., NeuroRule: A Connectionist Approach to Data Mining, Same to[1]

[4]David K. Y. Chiu et al., Information Discovery through Hierarchical Maximum Entropy Discretization and Synthesis, in G. Piatetsky-Shapiro and W. J. Frawley (eds.), Knowledge Discovery in Database, AAAI/MIT Press, 1991

[5]Jiawei Han et al., Knowledge Discovery in Database: An Attribute-Oriented Approach, In Proc. 18th Intl. Conf. on VLDB Vancouver, Canada, 1992

[6]W. J. Frawley et al., Knowledge Discovery in Databases: An Overview, AI Magazine, Fall, 1992

[7]C. J. Matheus et al., System for Knowledge Discovery in Databases, IEEE Trans. Knowl. Data Eng., 5(6), 1993

[8]Rakesh Agrawal et al., Database Mining: A Performance Perspective, Same to[7]

[9]Kenneth A. Faufman et al., Mining for Knowledge in Databases: Goals and General Description of the INLEN System, Same to[4]

(上接第29页)

总之，这类帮助用户查询和检索的工具也是当前研究开发的热点，很多软件也正在不断地开发和完善之中。

这里，举一个简单的例子。当用户在Internet上查找信息时，基本上有两种情况：一是已知要找信息的URL，这时比较简单，只需运行相应的浏览器，键入URL即可，如要浏览WWW国际会议的相关信息，只要输入http://www.elsevier.nl/cgi-bin/ID/www94，然后通过超链接进行浏览并可将需要的论文和相关信息拷贝到自己的用户中。

另一种情况是不知道URL地址，这时就需要借助一些检索工具来找出相应的URL或其它线索。上文提到的工具都可做到这一点，用户可以利用这些检索工具，输入相应的关键词，就可得到有关的URL或其他帮助信息。Netscape公司将一些常用的检索工具做成了一个文档，用户可以通过访问这一文档来激发相应的工具，其URL是http://www.netscape.com/home/internet-search.html。

五、今后工作

Internet的用户接口除了应在用户的视觉效果和可用性方面要进行改进之外，关键的是要智能化。因为Internet的巨大的信息量使得一个人穷尽其一生也是无法理解的，这就要求我们开发出具有智能化的接口，为用户用好Internet提供帮助。

自然语言的接口是很有必要的，不仅是对用户查询的理解，而且要对网上的文章进行理解，这样才能更好地同用户的意愿相匹配，这种自然语言理解能力并不需要十分强大，但却是十分必要而有意义的。此外，对网络上新信息的学习和发现能力，也很重要。

总之，我们如何把AI、DAI和相关学科中的一些有用的方法和知识应用到Internet这一飞速发展的领域中，在引进技术、引进人才之后如何更好地利用高科技引进信息为我所用，是我们面临的一个课题和挑战。