

# KDD 活动的开展及其价值分析

吴颖红

(杭州师范学院图书馆 杭州 310036)

**【摘要】** 介绍了数据库知识发现(KDD)活动的展开要求,着重从它的技术处理流程来分析它的特性及其存在价值与意义。

**【关键词】** 数据库知识发现 技术处理 价值分析 **【分类号】** G250

## The KDD Activity Development and Value Analysis

Wu Yinghong

(Library of Hangzhou Teachers College, Hangzhou 310036, China)

**【Abstract】** This article introduces the KDD activity the request that launch, and puts great emphasis on its characteristic and value by technique processing.

**【Keywords】** Knowledge discovery in database Technique processing Value analysis

知识主要来源于人类智慧的积累。进入网络时代,知识发现则主要依赖数据库。知识发现与信息检索有很多相似之处,但又有些区别。信息检索是从大量信息中查找到特定的信息,而知识发现是在大量似乎无关的数据中发现其中的规律和知识。信息检索主要用途是使用户发现可用的资源,从中找寻满足其检索要求的信息内容,而知识发现则是为了揭示数据库中文档信息的隐含知识而进行的活动。所以,知识发现,一般又被称为数据库知识发现(Knowledge Discovery in Database,简称 KDD),它的存在,使得大量可信、新颖、有效的数据从数据库中提取并成为人们理解的模式的处理过程的构想成了现实。利用崭新的信息处理技术和数据分析工具,提供高于信息检索的数据分析功能,自动地、智能地将大量数据转变为有用的、系统化的知识,是开展数据库知识发现活动的基本动因,也可以认为,它是为了适应新要求而出现的一种新型数据分析技术。

### 1 开展 KDD 活动的基本要求

#### 1.1 能以数据库为知识源对大量数据信息进行有效整理

数据库一般都有规范的结构,因为数据库的创建基本目的是为了机器可读,但网络环境下的数据库由于有各种方式可与因特网链接,使得数据库在不断扩

容情况下,规模日益庞大复杂程度也日益提高,对“海量”数据进行有效整理成为开展知识发现的首要任务。要完成数据库原始数据向有价值知识的转化,就要具备把数据库作为知识源,从大量数据进行提取、过滤、转换、集成的能力,从中发现新的知识,不仅做到知识发现速度的及时,更保证发现过程的高效。

#### 1.2 能够对不同数据类型的信息分类处理

在数据库中建立新的信息资源组织方式,通过对新录用信息、新发现知识的适应结构的系统排查,以确定不同类型数据的分类处理,才能极大地提高检索速度和检索效率。反之,面对瞬息万变的搜索资源,仅靠用户本身的能力来解决搜索专题术语的情况就不能得到改善。机器使用的局限性与分类模式的优势有机结合,能直接影响数据库从输入、检索到输出结果的整个过程,结构严谨的数据库,便于各种信息新技术的应用,不仅能有效分离不同类型数据,也能更有利于知识发现的进展。

#### 1.3 能让用户在使用过程中共同参与

知识发现本身的复杂性在于用户自身也能参与数据挖掘,发现新信息并转化为新知识,也能对有关领域的知识在相应系统的支持下进行评估和选择。这就对系统的性能提出较高的要求,交互性强成为重要的环节。一方面,交互界面接受用户提出的检索、查询要求和数据挖掘策略,另一方面,交互界面把生成的

结果返回给用户。这其中用户参与和系统确认领域知识有效的发现就是一个需要多次交互和多次反复的渐进过程。

#### 1.4 整个发现过程有严格的组织和有效的提取方法

知识发现的一项重要任务,就是按照一定的数据提取,从数据库中发现隐含的、有意义的知识。现在大多数数据发掘都应用于关系数据库和面向对象数据库,它们一般都有完整的结构,可按照预先设定的模式进行组织、存储和存取。如果提取记录与数据库原有记录匹配,正确的术语录用到检索方法当中,这无疑既是简单又有效的一种方法。但如果所选的术语未能匹配,就很少有资源能相对集中,因此,要使KDD活动顺利展开,需要大范围收集数据资源,对之进行严格组织和有效管理。目前主要应用的知识提取检索方法有:查找搜索(解决如何有效率地从数据集合中找所需数据项的方法);线性检索(在集合内逐个排查,效率较低);散列检索(用散列的方法寻找数据元素,可直接访问存储器来查找目标元素,可缩短查询时间);对分检索(从文件中点开始逐层压缩查询范围,也是一种快速查询方法);另外还有全文检索及智能代理技术,一个信息源搜集广泛全面,另一个能进行信息过滤,识别所需的潜在信息的时效性,是一种效率很高的检索方法。

#### 1.5 通过知识更新保持发现的动态性

不断发展的网络技术为KDD的发展创造了无限契机。尽管信息技术日新月异,但在大规模的数据库内进行有效检索仍有一定的困难,系统要能适应发展变化的情况,通过知识及时更新保持发现的动态性,才能提供有效的决策支持。鉴于知识发现本身就是个动态发展过程,把它的技巧运用到数据库检索中,就能设计出兼容性更强更为复杂的模式处理方法,使两者能够相互补充、相互促进、共同发展。

## 2 KDD 活动的技术处理过程

知识发现是一个能从大量数据中提取出隐藏在其中的有用信息的高级处理过程,它从数据集中识别出以模式来表示的知识。高级的处理过程是指一个多步骤的处理过程,多步骤之间相互影响、反复调整,形成一种螺旋式上升过程。其中有可能多次反复,如图1所示。

#### 2.1 准备阶段

了解KDD相关领域的有关情况,熟悉有关的背景知识,并弄清楚用户的要求。

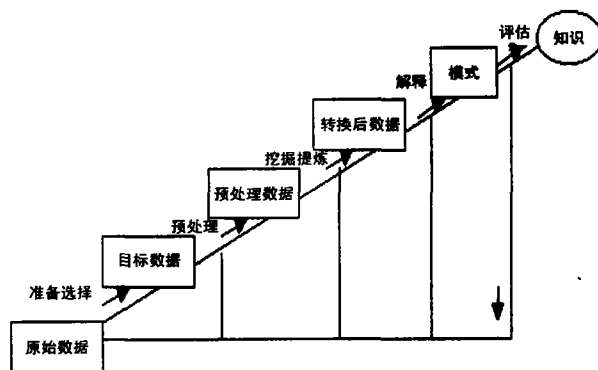


图1 KDD技术处理流程

#### 2.2 数据的选择与提取

根据用户的要求从数据库中选取与KDD有关的数据,从中进行知识提取,利用对数据库的操作完成一些数据处理。目前,随着数据库技术的不断发展,数据处理的方法也在不断完善并趋于成熟。在数据库的知识发现中,利用现有的一些数据库技术和专门针对数据库的一些启发式方法,可以用来提取数据库的一些特征知识。

#### 2.3 数据预处理

主要是对前一阶段产生的数据进行再加工,检查数据的完整性及数据的一致性,对其中的无效数据进行过滤,对丢失的数据进行填补。其中包括:

(1) 数据缩减。对经过数据预处理的数据,根据知识发现的任务对数据进行再处理,主要通过投影或数据库中的其它操作减少数据量。

(2) 确定知识发现的目标。根据用户的要求,确定KDD发现的是何种类型的知识,因为知识发现的要求不同,KDD在具体的知识发现过程中就会采用不同的知识发现算法。

(3) 确定知识发现的算法。根据既定目标,选择最为合适的知识发现算法。包括选取合适的模式和参数,并且使知识发现算法与整个知识发现的评判标准相一致。

#### 2.4 数据挖掘

运用选定的知识发现算法,从数据中提取出以后所需要的知识,这些知识可以用一种特定的方式表示或使用一些常用的方式表示,如产生式规则等。一般而言,知识发现是数据挖掘的结果,它通常可表现为概念(Concept)、规则(Rules)、规律(Regularities)、模式(Patterns)、约束(Constraints)、可视化(Visualizations)等多种形式。这些知识被发现后既可以直接提供给决策者,用以辅助决策过程,也可以提供给该领域专家,修正专家已有的知识体系,还可作为新的知识转存到应用系统中作为决策的依据。

(1) 数据挖掘的任务。各学科领域数据挖掘的发现任务

都不相同,但从知识发现的角度出发也有共通之处。都是对大型数据库中的海量业务数据进行抽取、转换、分析和模型化处理,从中提取辅助决策的关键性数据和隐藏的预测性信息。它通过发现数据间的潜在模式,找出人们可能忽视的信息,便于以理解和观察的形式反映给用户,并给予基于知识的决策意见和结论。

(2) 数据挖掘的方式。数据挖掘是在一些事实或观察数据的集合中寻找模式的决策支持过程,它从理论上和技术上承继了信息处理和数据分析、结论提取等领域的成果,同时又涵盖了其它许多领域如人工智能(Artificial Intelligent)、模式识别(Pattern Recognition)、统计学(Statistics)等的发现结果。数据挖掘有两种类型:第一,根据发现知识的种类分类。主要包括关联规则、分类规则、特征规则、聚类规则、归总规则、趋势分析、偏差分析等等。第二,根据采用的挖掘技术分类。数据挖掘技术是人工智能领域的一个新的重要分支,它可以综合利用各种人工智能代理技术,比较常用的有:粗集方法(Rough Set)、神经网络(Neural Network)、决策树归纳法(Decision Tree Induction)、最近邻技术(Nearest Neighbor)、规则归纳(Rule Induction)、可视化(Visualization)、聚类法(Clustering)、数据仓库(Data Warehouses)等等。

### 2.5 模式解释

对发现的模式进行解释,在此过程中,为了取得有效的知识,可能返回到前面的某些步骤以反复提取,从而提取出更为有效的知识,发现更有价值的知识。根据数据挖掘的结果可产生的模式有很多,如分类模式、聚类模式、关联模式、序列模式等等。在解决实际问题时,经常同时使用多种模式来降低问题的复杂性,提供给用户较大的灵活性和比较强的分析能力。数据库知识发现提供多种途径产生同种模式,在实际应用中效果将更显著。

### 2.6 知识评价

将发现的知识以用户能了解的方式呈现给用户,这期间也包含对知识的一致性的检查,以确定本次发现的知识与以前发现的知识不相抵触。知识评价有利于提高和改善知识发现的质量,也有助于选择知识发现的应用系统。知识发现的评价需要很多方面的共同支持:

(1) 系统支持。允许多种系统运行,便于合理评价。

(2) 强大的数据存取、处理能力的支持。好的挖掘工具可以使用SQL语句直接从数据库管理系统中读取数据,简化了数据准备工作,充分利用数据库的优点读取数据。

(3) 可视化的程度。可视化工具提供直接、简洁的方式表达信息,它的种类、质量和灵活性直接影响到KDD的展开和自解释性。

(4) 易操作性。操作性能直接影响到用户的使用,优化的

界面能为用户节省时间,提高效率。有些工具还提供数据挖掘的嵌入技术,通过嵌入到应用程序,不仅缩短开发时间,将模式运用到已存在或新增加的数据上,也可把模式导出到程序或数据库中。

(5) 数据挖掘的可扩展性程度。鉴于知识发现的更新速度,对数据挖掘的扩展性提出了更高要求,知识发现活动的展开不仅要能与传统查询工具、可视化工具、联机分析工具兼容,更能以自身的一些优点(如并行计算等)与传统工具相集成,与数据库或数据仓库以组件形式集成于一个信息处理环境中,极大地提高它的效率。

## 3 KDD活动的特性及其价值意义

从数据库知识发现的技术处理过程可以看出,数据库知识发现主要有以下几个特性:第一,它是从现实世界中存在的一些具体数据中提取知识,这些数据在此之前早已存在,对现实世界很有意义;第二,它使用的数据来源于数据库,处理的数据量很大,因此,在知识发现的过程中学习算法效率和可扩充性就显得尤为重要;第三,由于它所处理的数据来自于现实世界,数据的完整性、一致性、正确性都很难保证,所以如何将这些数据加工成为可以接受的数据需要进行进一步深入的研究;第四,利用目前的数据库技术所取得的研究成果来加快学习过程,提高效率必定将成现实;第五,数据库知识发现处理的数据来自实际的数据库,而与这些数据库数据有关的还有其它一些背景知识,如边缘学科、交叉学科等都与之有很强联系,合理运用效果会很显著。随着DBS在各行各业的迅速普及,即使在科研领域,目前的很多研究也是在大量的数据基础上进行的,而以数据为处理对象,知识发现系统无疑将帮助人们更好的了解数据的含义,更好的利用数据。

数据库知识发现的理论意义在于不仅提供了各种学科领域情报分析研究的科学性,而且促进了文献信息研究的进一步深入,丰富和完善了网络信息的研究内涵。同时,数据库知识发现的活动本身也为鉴定和评估数据库提供了模式和趋向。KDD工作与数据库的研究工作是相辅相成、互相促进的。它的实际价值主要体现在两个方面:首先,知识发现通过对数据库的技术挖掘,借鉴了信息检索的查询技术,进一步提高知识发现的效果。其次,可以利用网络挖掘的成果来提高网络信息检索的精准度和效率,改善检索结果的组织结构。比如,它发现统计数据过程的相关性,能够暗示出新的类目或子类目的相互关联,使得数据

(下转第67页)

量政府信息,仍然是必须重点考虑的问题。

#### (4) 行业障碍

IT部门对信息的形式处理与传送关注很多,但很少认识到专业知识的限制。他们在信息内容的呈递、描述与规范上,没有专门的培训或训练。他们也很少考虑概念之间的逻辑关系与知识的映射等。这必然给数据的转换带来问题。

### 5 国家电子数据中心的搭建

同其它信息工程的建设一样,国家电子数据中心的建构必须统一领导、整体规划、合理分工、协调行动。然而,与其它信息工程不一样的是,国家电子数据建设无论是在资金与操作要求上,还是在行动日程、站点选择上,都有自己的特色。本文主要对其设备、地点选择与人员配备及领导问题详细阐述。

#### (1) 设备配置

在设备的选择上,国家电子数据中心最合理的建构是与地方的档案中心合作共享一些设备。当地的档案中心有长期的数据存储经验,以及法定的数据档案存储权力。而国家电子数据中心与地方档案中心都要求特定的环境条件,都使用元数据对信息进行访问与检索。因此,两者可结合使用。电子数据处理容量应该被限制到主要的记录系列,比如电子文献、网页以及带档案性的电子信件等。最好开始就安装容量为几千G字节的多重处理器群。

#### (2) 地点选择

在地点的选择上,电子数据中心的建立应该选择合适的地点,有着良好的自然环境。建筑电子数据中心时,应该设置有计算机实验室与讨论室,也要留有足够的空间为学术团体使用。此外,必须预留公共使用的面积,比如休息厅、电梯、楼梯等。在此过程中,要合理预算整个项目的开支。

#### (3) 人员配备

在人员的配备上,电子数据中心的建构应该包括以下人员:档案中心的工作人员、学术团体的专家以及硬件与软件设施的销售商。其中,学术团体的专家主要参与研究、指导与分析。同时配备有相应的设备经理、规划人员以及系统管理人员,并且合理预算薪水。

#### (4) 领导班子的组建

在领导问题上,应该成立领导委员会与国家电子数据存档规划组。前者包括档案中心职员、政府机关相关人员、地方主管人员、会计审核员、司法人员、信息产业厅人员、大学研究人员、地方各协会商会等在内统筹指挥委员会,主要对数据中心的设计、发展与建构检查与指导。后者主要负责制定电子数据存储规则、处理程序及相关标准,其目的是帮助实现国家电子数据中心应该实现的功能。国家案卷保管人员将领导规划组,其包括电子存储领域的专家、数据存档专家以及信息处理专家。

#### 参考文献:

- 1 D Stamoulis, D Gouscos, P Georgiadis, D Martakos: Revisiting public information management for effective e-government services, Information Management & Computer Security, 2001(9): 146-153
- 2 Philip Coombs: The crisis in electronic government record keeping: A strategy for long-term storage. Library Computing, 1999(18): 196-202
- 3 刘家真. 数字信息保存的策略. 情报学报, 2000(4)
- 4 Pamela A Houghtaling: Building the government's electronic records archive. Signal, Falls Church, 2000(53): VG6-VG13
- 5 Ernest Perez: MetaDatabases, GILS, and Megaproblems. Econtent. 1999(22): 75

(作者E-mail: pan81706@sohu.com)

(上接第31页)

库的有效价值能得到进一步的澄清和扩充。KDD的开展其实是对网络资源极大限度的开发和利用,而且根据用户定义的知识发现策略,发现的知识必须是可理解的。惟有如此才能把发现的知识明确表达,加以掌握和利用,才真正体现出被发现知识的价值。它的开展不仅是知识管理的需要,更是提高网络服务水平、构建学科决策支持系统的需要,它也顺理成章地成为进行科学研究的有力工具。

#### 参考文献:

- 1 严良文等. 数据库中知识发现的实现技术研究. 化工装备技术, 2002(1)
- 2 熊新阶等. 数据库中的知识发现. 黄冈师范学院学报, 2002(9)
- 3 黄晓斌. 基于网络的文献知识发现系统研究. 情报科学, 2003(2)
- 4 肖牧安等. 数据挖掘与知识发现的理论方法及技术分析. 交通与计算机, 2002(1)
- 5 赵丹群. Web资源发现工具的技术分析. 情报学报, 1999年(增刊)
- 6 Han Jiawei, Micheline Kamber. 数据挖掘概念与技术. 北京:机械工业出版社, 2001

(作者E-mail: jftsg@sohu.com)