

数据挖掘发展研究

The Development Research on the Data Mining

张 伟^{1,2} 刘勇国¹ 彭 军^{1,3} 廖晓峰¹ 吴中福¹

(重庆大学计算机科学与工程学院 重庆400044)¹

(重庆教育学院计算机与现代教育技术系 重庆400067)²

(重庆工业高等专科学校计算机系 重庆400050)³

Abstract Mining knowledge from database has been thought as a key research issue in database system. Great interest has been paid in data mining by researchers in different fields. In this paper, data mining techniques are introduced broadly including its definition, purpose, characteristic, principal processes and classifications. As an example, the studies on the mining association rules are illustrated. At last, some data mining prototypes are provided and several research trends on the data mining are discussed.

Keywords Data mining, Knowledge discovery in database, Association rule

1 引言

世纪之交,人类面临着新的问题:不缺数据缺知识。随着数据库技术的成熟和数据应用的普及,人类积累的数据量正以指数速度增长。例如,Wal Mart 公司每天要处理二千万个事务;美国航天局1999年发射的地球观测系统每小时要产生50Gb的图像数据等。毫无疑问,这些庞大的数据库及其中的海量数据是极其丰富的信息源,但是仅仅依靠传统的数据检索机制和统计分析方法已经远远不能满足需要了^[1]。无怪乎 John Naisbert 惊呼:“We are drowning in information, but starving for knowledge”。面对浩瀚无际的数据,人们呼唤从数据汪洋中一个去粗取精、去伪存真的技术,因此,从数据库中发现知识(Knowledge Discovery in Database, KDD)及其核心技术——数据挖掘(DM)便应运而生。

2 概述

2.1 数据挖掘的定义

在文[2]中,知识发现是这样定义的:知识发现是识别出存在于数据库中有用的、新颖的、具有潜在价值的乃至最终可理解的模式的非平凡过程。数据挖掘则是指从数据库的大量数据中揭示出隐含的、先前未知

的并有潜在价值的信息的非平凡过程^[3,4]。可见这两个术语的内涵大致相同。对这两个术语更严格的区分是在“知识发现96国际会议”上,Fayyad, Piatetsky-Shapiro 和 Smyth 指出^[2],知识发现是从数据库中发现知识的全部过程,而数据挖掘则是此全部过程的一个特定的关键步骤。这种定义把数据挖掘的对象定义为数据库。

数据挖掘更广义的说法是^[5],数据挖掘意味着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。实际上,数据挖掘的对象不仅是数据库,也可以是文件系统、或其它任何组织在一起的数据集合,例如WWW信息资源,数据挖掘最新的对象是数据仓库。

2.2 数据挖掘的目的

数据挖掘的任务是从大量数据中发现知识。数据挖掘是知识发现的核心技术,数据挖掘研究的主要目标是发展有关的方法论、理论和工具,以支持从大量数据库中提取有用的和让人感兴趣的知识和模式。

在传统的决策支持系统中,知识库中的知识和规则是由专家或程序人员建立的,是由外部输入的。而数据挖掘的任务是发现大量数据中尚未被发现的知识,是从系统内部自动获取知识的过程。对于那些决策者明确了的信息,可以用查询、联机分析处理(OLAP)或其它工具直接获取;而另外一些隐蔽在大量数据中的

张 伟 博士研究生,主要研究方向为远程教育、人工智能。刘勇国 博士研究生,主要研究方向为计算机网络、人工智能。彭军 博士研究生,主要研究方向为网络安全、人工智能。廖晓峰 教授,主要研究方向为网络安全、人工智能。吴中福 教授、博士生导师,主要研究方向为远程教育、网络通信、人工智能。

关系、趋势,即使是管理这些数据的专家也是没有能力发现的,这些信息对于决策可能又是至关重要的,这类问题就可以用数据挖掘来对付。

数据挖掘发现的知识可以直接提供给决策者,用以辅助决策过程;或者提供给领域专家,修正专家已有的知识体系;也可以作为新的知识转存到应用系统的知识存储机构中,比如专家系统规则库等。

2.3 数据挖掘的特点

1)数据挖掘与机器学习比较 数据挖掘与机器学习都是从数据中提取知识,其区别在于:机器学习主要针对特定模式的数据进行学习;数据挖掘则是从实际的海量数据源中抽取知识,如数据仓库。由于数据挖掘

处理的数据量非常巨大,数据的完整性、一致性及正确性都难以保证,因此,数据挖掘算法的效率、有效性和可扩充性都很重要。

2)数据挖掘与传统的数据库查询系统比较 数据挖掘技术与传统的数据库查询系统相比较,也有显著的不同:首先,传统的数据库查询是被动的查询,用户必须要清楚查询的内容,而数据挖掘则可以辅助用户发现隐含的信息;其次,传统的数据库查询一般生成严格的结果集,但数据挖掘可能并不生成严格的结果集;第三,通常情况下,传统的数据库查询只对数据库的原始字段进行,而数据挖掘则可能在数据库的不同层次上挖掘知识规则。

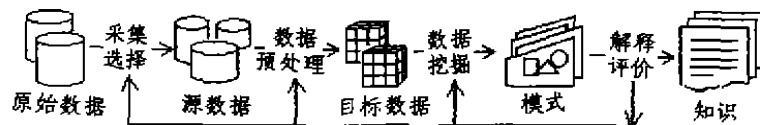


图1 数据挖掘流程图

2.4 数据挖掘的过程

由数据挖掘过程图^[6,7]可知(见图1),知识发现有以下四个主要的阶段:采集选择、数据预处理、数据挖掘和解释评价。数据挖掘是知识发现的一个特定的关键阶段,知识发现则是其中的一个或多个阶段的反复过程。

采集选择的目的是辨别出需要分析的数据集合,缩小处理范围,然而实际系统中收集到的原始数据通常是“脏”的,即数据存在杂乱性、重复性及不完整性;数据预处理可以处理数据中的遗漏及清洗脏数据,从而提高数据挖掘的质量,数据预处理应该包括数据集成(Data Integration)、数据清理(Data Cleaning)、数据变换(Data Transformation)和数据简化(Data Reduction)等几方面的功能;数据挖掘阶段进行实际的挖掘操作,它要先决定是进行发现型(Discovery-Driven)的数据挖掘,还是验证型(Verification-Driven)的数据挖掘,然后选择合适的工具,进行发现知识的操作及证实发现的知识;解释评价这一步骤的任务不仅是把结果表达出来(例如采用信息可视化方法),还要对信息进行过滤处理,如果不能令决策者满意,需要重复以上数据挖掘的过程。

2.5 数据挖掘的分类

从不同的视角看,数据挖掘有以下几种分类方法^[8]:

1)按挖掘的数据库分类 数据挖掘基于的数据库类型有:关系型、事务型、面向对象型、主动型、空间型、时间型、文本型、多媒体、异质数据库和遗留系统等。

2)按挖掘的知识分类 这种分类方法有:总结(Summarization)规则挖掘、特征规则挖掘、关联规则挖掘、分类规则挖掘、聚类规则挖掘、趋势分析、偏差分析、模式分析等。如果以挖掘知识的抽象层次划分,又有原始层次的数据挖掘、高层次的数据挖掘和多层次的数据挖掘等。

3)按采用的技术分类 最常用的数据挖掘技术是^[9]:

A)模糊和粗集方法:应用模糊和粗集理论进行数据查询排序和分类是数据挖掘的重要方法;

B)人工神经网络:从结构上模仿生物神经网络,是一种通过训练来学习的非线性预测模型,可以完成分类、聚类、特征挖掘等多种数据挖掘任务;

C)遗传算法:是一种新的优化技术,基于生物进化的概念设计了一系列的过程来达到优化的目的。这些过程有基因组合、交叉、变异和自然选择。为了应用遗传算法,需要把数据挖掘任务表达为一种搜索问题而发挥遗传算法的优化搜索能力;

D)决策树:用树形结构来表示决策集合,这些决策集合通过对数据集的分类产生规则。典型的决策树方法有分类回归树(CART),典型的应用是分类规则的挖掘;

E)最近邻技术:通过K个最与之相近的历史记录的组合来辨别新的记录,有时也称这种技术为K-最近邻方法。这种技术可用作聚类、偏差分析等数据挖掘任务;

F)规则归纳:通过统计方法归纳、提取有价值的

If-Then 规则,规则归纳的技术在数据挖掘中被广泛使用,例如关联规则的挖掘:

G)可视化技术:采用直观的图形方式将信息模式、数据的关联或趋势呈现给决策者,决策者可以通过可视化技术交互地分析数据关系。

3 数据挖掘举例

数据挖掘例子很多,以关联规则的数据挖掘为例。关联规则的数据挖掘被分解为下面两步^[2-4]:1)首先找一个支持度大于给定值的大数据项集;2)用这个大数据项集产生规则。即是:对给定的 L,如果 A 是 L 的非空子集,有

$$\frac{\text{Support}(L)}{\text{Support}(A)} \geq \text{Conf}_{\min}$$

则产生形式为 $A \Rightarrow L-A$ 的规则。在这两步中,第一步是算法的关键,一旦找到了数据项集,关联规则的导出是很自然的。

实例 设数据库中有4次交易 $T_1 = \{A, B, C\}$, $T_2 = \{A, B, D\}$, $T_3 = \{A, D, E\}$, $T_4 = \{A, B, D\}$ 。设最小 Support 为 0.5,最小 Confidence 为 0.8,求其有效规则。

求解 第一步,产生大数据项集。计算 $\text{Support}(\{A\})$, $\text{Support}(\{B\})$, $\text{Support}(\{C\})$, $\text{Support}(\{D\})$, $\text{Support}(\{E\})$, $\text{Support}(\{A, B\})$, $\text{Support}(\{A, C\})$, $\text{Support}(\{A, B, C\})$ 等,要求满足最小 Support 为 0.5,因此,可得大数据项集为 $\{A\}$, $\{B\}$, $\{D\}$, $\{A, B\}$, $\{A, D\}$, $\{A, B, D\}$ 。

第二步,对每个大数据项集,产生所有比最小 Confidence 大的规则。由此可得有效的规则为: $B \Rightarrow A$ 和 $D \Rightarrow A$ 。

4 数据挖掘系统或原型介绍

下面仅列出部分数据挖掘系统或系统原型:

• Quset 由 IBM Almaden 研究中心的 R. Agrawal 等人研究开发,可以从数据库中发现多种知识,包括关联规则、分类规则、时序模式、模式匹配与分析等;

• DBMiner 由加拿大 Simon Fraser 大学的 J. Han 等人研究开发,是一个交互式、多层次挖掘系统,主要挖掘特征规则、分类规则、关联规则和预测等;

• IMACS 由 AT&T 的 Brachman 等人研究开发,用于知识库的建构;

• SKICAT 由日本的 Kayyad 等人研究开发,主要用于天体数据的分类;

• Explora 由 GMD 的 Klossgen 研究开发,这是一个多模式、多策略发现的辅助发现系统。

5 数据挖掘研究的发展趋势

数据挖掘的研究还很不成熟,其应用还有很大的局限性。数据挖掘研究和发展方向主要体现在以下几个方面:

1) 算法效率和可伸缩性 数据挖掘是直接面向海量数据库系统的,这类数据库通常有上百个属性和数百万个记录,并且数据表之间包含复杂的关系,这就必然导致数据挖掘过程中搜索维数和搜索空间的激增,同时这也增加了出现不确定性、错误和病态模式的可能性,因此,提高算法的效率以及其有规模伸缩性是它们在实际应用中必须面对的巨大挑战。

2) 处理不同类型的数据和数据源 现今的数据库模型大都是关系数据库模型,在这类数据库模型上执行数据挖掘的能力是非常重要的。随着数据库应用范围日益扩大,数据库的规模和功能日益完善,新的数据库模型不断出现,在数据库中包含着大量复杂的数据类型,所以,数据挖掘系统对这些数据库的操纵能力是至关重要的。更进一步,把企业网和国际互联网上的众多数据源连接成一个大型分布、异构的数据库,从这些包含不同语义的格式化和非格式化数据中发现知识无疑是对数据挖掘的重大挑战。

3) 数据挖掘系统的交互性 数据挖掘过程中操作者的适当参与是必不可少的。一方面,交互界面接收用户的检索、查询要求和数据挖掘策略,为用户提供方便的手段来表达其要求和策略是这方面的关键;另一方面,交互界面又把生成的结果传递给用户,由于生成的结果多种多样,因此,准确而直观地描述挖掘的结果和友好而高效的用户界面一直是这方面研究的重要课题。

4) 私有性和安全性 数据挖掘能从不同的角度、不同的抽象层次上看待数据,这将潜在地影响数据的私有性和安全性^[10]。研究数据挖掘可能导致非法的数据入侵,同样是实际应用过程中亟待解决的问题。

5) 网络与分布式环境下的数据挖掘 随着 Internet 的不断发展,网络资源日益丰富。这就要求分散的技术人员各自独立地处理分离的数据库的工作方式应是可协作的。因此,研究适应网络与分布式环境下的数据挖掘工具、技术及系统是必然的。

参考文献

- 1 Shortland R, Scarfe R. Digging for Gold (Data Management). IEE Review, 1995, 41(5): 213~217
- 2 Fayyad U. Knowledge Discovery and Data Mining towards a Unifying Framework. KDD'96Proc. 2nd Intl. Conf. on Knowledge Discovery & Data Mining, AAAI Press, 1996
- 3 Chen M-S, Han J, Yu P S. Data Mining: An Overview from a Database Perspective. IEEE Trans on Knowledge and Data Eng. 1996, 8(6): 866~883
- 4 Piatetsky-Shapiro G, Frawley W J. Knowledge Discovery in Database. AAAI/MIT Press, 1991 (下转第94页)

(2) HasBall=A and IsNear(Gate)

(3) Score=Score+1 or IsBallBelong=Goalkeeper. 其中 IsBallBelong=GoalKeeper 表示球被守门员救起。

3) 对手模型。了解对手是对抗活动的基本前提,建立对手模型也成为了机器人足球赛的一个基本问题。基于 BDI 模型,在 Anand S. Rao 等人工作基础之上,本文给出了对手 Agent 模型和相应算法,以达到在动态开放环境下,实时跟踪对手的思维状态,为决策提供支持的目的。

将 Agent 自身视为 Agent1,将被跟踪的 Agent 视为 Agent2,则 Agent1 所建立的 Agent2 对手模型元组描述为 $O = \langle B, D, I, I_p, I_w, H \rangle$, O 在 Agent1 的 BDI 模型中应是一组 Belief,即 $Bel(\text{Agent1}, O)$,各元组分别为:

B 任意 $\Psi \in B = \{ \text{Bel}(\text{Agent2}, \Psi) \}$,集合 B 包含了 Agent1 知道的 Agent2 所具有的 Belief;

D 任意 $\Psi, Des(\text{Agent2}, \Psi) = \{ \Psi \in D$, Agent2 实际的集合是 D 的一个子集,即 D 包含了 Agent2 所有可能的 Desire;

I_s 任意 $\Psi \in I_s = \{ \text{Intention}(\text{Agent2}, \Psi) \}$, I_s 表示可以确定的 Agent2 的 Intention,包括先验的 Intention 和通过观察、通信等手段获得的 Intention;

I_p 对手的可能 Intention,通过已有 BDI 信息所推断出当前 Agent2 可能具有但无法确定的 Intention;

I_w 表示在 Agent2 所有可能的 Intention 中 Agent1 所关心的部分,可将 I_w 称为 Intention 的观察窗口,对于该集合之外 Intention 的,模型不考虑;

$H = \{ a_i \}$, a_i 表示 Agent2 在特定时间所执行的特定动作。 H 记录一定时间内的对手事件。

约定 O 中的 B, D, I 满足典型的 BDI 约束关系;对于任意 $\Psi \in I_p, B, D, I, \{ \Psi \}$ 也满足典型的 BDI 约束关系,但模型中, B, D, I_p 无须满足 DBI 约束,即 I_p 集中的 Intention 可以相互矛盾。

模型的实现算法:

Step0 初始化操作,将模型各元素置初始状态;

随后,系统按 Step1 至 Step4 顺序循环执行。

Step1 获得外界信息,修正对手模型中的 B, D, I_s 元素。

Step2 根据 B, D, I 的约束关系,修正 I_p ,消除思维状态的内部矛盾。

Step3 将可能的 Intention 加入 I_p 。

Step4 在 I_p 中存在多个对手 Intention,它们均是不确定的,为了有效跟踪对手,应对其进行排序,以便为自身决策提供依据。对可能意图的排序的方法应依不同具体情况而定,通常,我们可考虑意图实现的代价和对手的历史信息等因素。

参考文献

- Asada M. Robocup: Today and tomorrow—What we have learned. Osaka University, 1998
- Veloso M, Stone P, Han K. The CMUnited-97 robotic soccer team: Perception and multi-agent control. Carnegie Mellon University, 1998
- 洪炳熔, 薄喜柱. 机器人足球比赛策略仿真系统的开发. 哈尔滨: 哈尔滨工业大学计算机系智能机器人研究室, 1999
- Marsella S. On being a teammate: Experiences acquired in the design of Robocup teams, University of Southern California, 1998
- Kitano H, Asada M. Robocup: The Robot World Cup Initiative
- Tambe M. Building agent teams using an explicit teamwork model and learning. Artificial Intelligence, 1999 (110): 215~239
- Stone P, Veloso M. Task decomposition, dynamic role assignment, and low-bandwidth communication for real-time strategic teamwork. Artificial Intelligence, 1999(110): 241~273
- Wenger B B. Cooperation without deliberation: A minimal behavior-based approach to multi-robot teams. Artificial Intelligence, 1999(110): 293~320
- Haddadi A, Sundermeyer K. Belief-Desire-Intention Agent Architectures, in Foundations of Distributed Artificial Intelligence, O'Hare G. M. P and Jennings N. R. (eds), John Wiley & Sons Inc, 1996: 169~185
- Rao A S. Multi-Agent Mental-State Recognition and its Application to Air-Combat Modelling, in Workshop on Distributed Artificial Intelligence, 1994: 283~304
- Han J. Conference tutorial notes: Data Mining Techniques, In: Proc. of ACM 'SIGMOD Int'l. Conf.' 96 on Management of Data (SIGMOD'96) Montreal, Canada, June, 1996
- Data Mining: Discovering Hidden Value in Your Data Warehouses. Pitwt Software. WWW files, 1996. Available at <http://www.pitwt.com/dmpaper/dmindex.html>
- C Zhenxin. Knowledge Discovery and System-user Adversary. Kybernetes: An International Journal of Cybernetics and General Systems, 1997, 26(4): 425~441
- Weldon Jay-Louise. Data Mining and Visualization. Database Programming and Design, 1996, 9(5): 21~24
- Fayyad U M, Piatetsky-Shapiro G, Smyth P. From Data Mining to Knowledge Discovery in Databases. AI Mag., 1996, 17: 37~54
- Fayyad U M, Piatetsky-Shapiro G, Smyth P. The KDD Proces for Extracting Useful Knowledge from Volumes of Data. Commun. ACM, 1996, 39: 27~41

(上接第81页)