文章编号:1004-5694(2001)增-0070-04

决策支持分析新技术——数据挖掘

印 勇
(重庆大学 通信工程学院,重庆 400044)

摘 要:数据挖掘是目前国际上智能信息处理和决策支持分析领域的最前沿的研究方向之一。因此对数据挖掘的基本概念、关键技术以及主要研究内容作一个综合性的介绍,并指出了数据挖掘技术的研究方向。

关键词:数据挖掘;数据库知识发现;决策支持
中图分类号:TP311.133 文献标识码:C

New Technology of Decision Support Analysis — Data Mining

YIN Yong
(*College of Communication Engineering, Chongqing University, Chongqing 400044, China*) **Abstract**: Data mining is one of the most advanced research directions of intelligent information processing and decision support analysis in the modern world. This paper gives a comprehensive introduction about the basic ideas of data mining, main techniques involved, and the content of the main researches.

Key words: data mining; knowledge discovery in databases; decision support

0 引 言

近十几年来,人们利用信息技术生产和搜集数据的能力大幅度提高,千万个数据库被用于商业管理、政府办公、科学研究和工程开发等,信息增长呈现指数上升。这一势头仍将持续发展下去。于是,一个新的挑战被提了出来:在这被称之为信息爆炸的时代,信息过量几乎成为人人需要面对的问题。如何才能不被信息的汪洋大海所淹没,从中及时发现有用的知识,提高信息利用率呢?要想使数据真正成为一个企业的资源,只有充分利用它为企业自身的业务决策和战略发展服务才行,否则大量的数据可能成为包袱,甚至成为垃圾。因此,面对"人们被数据淹没,人们却饥饿于知识"的挑战,新的数据处理技术——数据挖掘(Data Mining)技术便应运而生了。

面对海量的存储数据,如何从中发现有价值的信息或知识,成为一项非常艰巨的任务,数据挖掘就是为迎合这种要求而产生并迅速发展起来的。数据挖掘是指从大型数据库或数据仓库中提取出隐含的、先前未知的、对决策有潜在价值的知识和规则。数据挖掘是数据库发展与人工智能技术相结合的产物,是目前国际上数据库和信息决策领域的最前沿的研究方向之一,引起了学术界和工业界的广泛关注。世界上许多公司(如 IBM、Informix、Oracle 等)都投入巨资对其进行研究。第一本关于数据挖掘和知识发现的国际学术杂志《Data Mining and Knowledge Discovery》已于 1997 年 3 月创刊,不同领域的研究学者都对数据挖掘有极大的兴趣。目前研究的主要目标是发展有关的理论、方法和工具,以支持从大量数据中提取有价值的知识和模式。

本文将对数据挖掘的基本概念、关键技术以及

* 作者简介:印勇(1963-),男,重庆市人,重庆大学副教授,博士,研究方向为智能信息处理,数据仓库和数据挖掘技术。

主要研究内容作一个综合性的介绍。

1 数据挖掘的概念[1,2]

数据挖掘(Data Mining)就是从大量的、不完全的、有噪声的、模糊的、随机的数据中提取人们感兴趣的知识和规则的过程，这些知识和规则是隐含的、先前未知的、对决策有潜在价值的有用信息。通过数据挖掘，有价值的知识、规则或高层次的信息就能从数据库的相关数据集合中抽取出来，为决策提供依据，从而使数据库作为一个丰富可靠的资源为知识归纳服务。

数据挖掘和知识发现(KDD)有密切的联系。知识发现是指从数据库中发现有用知识的整个过程，数据挖掘是这一过程中的一个特定步骤，如图1所示。知识发现包括数据选择、预处理、数据转换、数据挖掘、模式解释和知识评价等多个步骤，是应用特定数据挖掘算法和评价解释模式的一个循环反复过程，并要对发现的知识不断求精深化，使其易于理解；数据挖掘是知识发现过程中的一个关键步骤，它利用特定的数据挖掘算法从数据中抽取模式，不包括数据的预处理、领域知识结合及发现结果的评价等步骤。

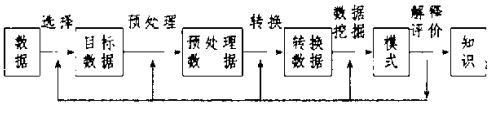

图1 KDD处理过程

2 数据挖掘的特点及分类

数据挖掘的特点：① 处理的数据规模十分巨大，一般为GB级，甚至TB级；② 由于用户不能形成精确的查询要求，因此需要靠数据挖掘技术来寻找其可能感兴趣的东西；③ 在某些应用中，要求数据挖掘对数据的迅速变化作出快速响应，以提供决策支持信息；数据挖掘既要发现潜在规则，还要管理和维护规则。而规则是动态的，当前的规则只能反映当前状态的数据库特征，随着新数据的不断加入，规则需要随之更新；④ 数据挖掘中规则的发现基于统计规律，发现的规则不必适合于所有数据，而且当达到某一阈值时，便认为有此规则。因此，利用数据挖掘技术可能会发现大量规则。

数据挖掘根据所挖掘的数据库类型、挖掘的知识类型有不同的分类方法。

· 根据挖掘的数据库类型分类：如果基于关系数据库的数据挖掘，称为关系数据挖掘；如果基于面向对象数据库的数据挖掘，称为面向对象数据挖掘；此外，还有事务数据库、主动数据库、演绎数据库、时间数据库、文本数据库、空间数据库、多媒体数据库、历史数据库、互联网信息库等数据挖掘。

· 根据挖掘的知识类型分类：按挖掘的知识类型可分为总结规则、关联规则、特征规则、分类规则、偏差规则、聚类规则及时序规则等数据挖掘；如果按知识的抽象层次可分为原始层次知识、高层次知识和多层次知识的数据挖掘等。

3 数据挖掘常用技术

① 决策树方法：利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。国际上最有影响和最早的决策树方法是由Quinlan研制的ID3方法，后人又发展了其它的决策树方法[3~5]。

② 规则归纳方法：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，其中以关联规则挖掘的研究开展得较为积极和深入[6~9]。

③ 神经网络方法：从结构上模拟生物神经网络，以模型和学习规则为基础，建立三大类多种神经网络模型：前馈式网络、反馈式网络、自组织网络。这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[10]。

④ 遗传算法：模拟生物进化过程的算法，由繁殖(选择)、交叉(重组)、变异(突变)三个基本算子组成。为了应用遗传算法，需要将数据挖掘任务表达为一种搜索问题，从而发挥遗传算法的优化搜索能力。

⑤ 粗糙集(Rough Set)方法：Rough集理论是由波兰数学家Z. Pawlak在八十年代初提出的，是一种处理含糊和不精确性问题的新型数学工具[11]。它特别适合于数据简化、数据相关性的发现、发现数

据意义、发现数据的相似或差别、发现数据模式、数据的近似分类等，近年来已被成功地应用在数据挖掘和知识发现研究领域中[12]。

⑥ K-最邻近技术：这种技术通过 K 个最与之相近的历史记录的组合来辨别新的记录。这种技术可以作为聚类[13]、偏差分析[14]等挖掘任务。

⑦ 可视化技术：将信息模式、数据的关联或趋势等以直观的图形方式表示，决策者可以通过可视化技术交互地分析数据关系。可视化数据分析技术拓宽了传统的图表功能，使用户对数据的剖析更清楚。

4　数据挖掘的发现任务

数据挖掘所涉及的学科领域和方法很多，以下 4 种是非常重要的发现任务。

· 数据总结：其目的是对数据进行浓缩，给出它的紧凑描述。数据挖掘主要关心从数据泛化的角度来讨论数据总结。数据泛化是一种把数据库中的有关数据从低层次抽象到高层次上的过程。

· 分类分析：其目的是学会一个分类函数或分类模型(也称作分类器)，该模型能把数据库的数据项映射到给定类别中的某一个。

· 聚类分析：聚类是把一组个体按照相似性归成若干类别，即“物以类聚”。它的目的是使属于同一类别的个体之间的距离尽可能地小，而不同类别的个体间的距离尽可能地大。

· 关联分析：发现数据库或数据仓库中数据项之间潜在的关联关系。

5　数据挖掘技术的研究方向

当前，数据挖掘技术的研究正方兴未艾，其研究正在蓬勃发展。预计在 21 世纪还会形成更大的高潮，研究焦点可能会集中到以下几个方面：研究专门用于知识发现的数据挖掘语言，也许会像 SQL 语言一样走向形式化和标准化；寻求数据挖掘过程中的可视化方法，使得知识发现的过程能够被用户理解，也便于在知识发现过程中的人机交互；研究在网络环境下的数据挖掘技术，特别是在 Internet 上建立 DMKD 服务器，与数据库服务器配合，实现数据挖掘；加强对各种非结构化数据的挖掘，如文本数据、图形图像数据、多媒体数据。

6　结束语

大量数据的产生和收集导致了信息爆炸，随着信息处理量的剧增，数据挖掘技术的应用为实时和深层次地分析这些数据提供了可能，数据挖掘使用户能够从大量繁杂的数据中挖掘出真正有价值的信息和知识，将给用户带来更大的利益。

参　考　文　献

[1] HAN J. Data mining techniques[C]. Proceedings of ACM SIGMOD International conference'96 on Management of Data (SIGMOD'96). Monteral, Canada, 1996.

[2] AGRAWAL R, IMIELINSKI T, SWAMI A. Database mining: a performance perspective[J]. IEEE Transactions on Knowledge and Data Engineering, 1993, 5(6): 914-925.

[3] QUINLAN J R. Induction of decision trees [J]. Machine Learning, 1986,(1):81-106.

[4] PAGALLO G, HAUAALER D. Boolean feature discovering in empirical learning[J]. Machine Learning, 1990,(5):71-99.

[5] BRODLEY C E, UTGOFF P E. Multivariate decision trees[J]. Machine Learning, 1995,(19): 45-77.

[6] SRIKANT R, AGRAWAL R. Mining generalized association rules[C]. Proceedings of the 21th International Conference on Very Large Databases. Zurich, Switzerland. 1995: 407-419.

[7] AGRAWAL R et al. Mining associstion rules between sets of items in large databases[C]. Proceedings of ACM SIGMOD Conference on Management of Data. Washington, DC. 1993: 207-216.

[8] ASHOK SAVASERE, OMIECINSKI E, NAVATHE S. An efficient algorithms for mining association rules in large databases [C]. Proceedings of the 21th International Conference on Very Large Databases. Zurich, Switzerland. 1995: 432-444.

[9] AGRAWAL R, SHAFER J C. Parallel mining of association rules[J]. IEEE Transacation on Knowledge and Data Engineering, 1996, 8(6): 962-969.

[10] LU Hongjun, RUDY Setiono, LIU Huan. Effective data mining using neural networks [J]. IEEE Transacation on Knowledge and Data Engineering, 1996, 8(6): 957-961.

[11] PAWLAK Z et al. Rough sets[J]. Communications of the ACM. 1995, 38(11): 89-95.

[12] ZIARKO W. Ed. Rough sets, fuzzy sets and knowledge iscovering[C]. Proceedings of RSKD'94 Workshop(Bauff). Springer—Verlag, Berlin. 1994: 208-216.

[13] FISHER D. Optimization and simplication of hierarchical clustering[C]. Proceedings of the 1st International Conference on Knowledge Discovering and Data Mining (KDD'95). Montreal, Canada. 1995: 118-123.

[14] ARNING A, AGRAWAL R, RAGHAVAN P. A linear method for deviation detection in large databases[C]. Proceedings of the 2nd International Conference on Knowledge Discovering and Data Mining (KDD'96). Portlan, Oregon. 1996: 182-187.

(上接69页)

供电路调度、阻塞控制、网络规划等决策依据。该系统已经用于实际工作中。

4 结束语

本文在分析了本地网网管系统现状的基础上，结合网管工作的实际需求，提出以本地网网管系统的管理信息库为数据源，建立基于数据仓库的网络管理决策支持系统的思路和实现方案，并初步实现了在本地网网管与监控系统基础上的应用，取得初步成效。然而，研究工作仅仅是初步的，许多问题特别是相应的数据仓库模型和数据挖掘方法需要进一步探讨与研究。

参 考 文 献

[1] [美]TOM HAMMERGREN. 数据仓库技术[M]. 北京：中国水利水电出版社，1998.

[2] 王珊等. 数据仓库技术与联机分析处理[M]. 北京：科学出版社，1999.

[3] 王广清，杨学良. 数据仓库技术及其在电信计费领域应用的探讨[J]. 计算机工程与应用，1999，(9).

[4] 张伟民. 数据仓库技术在电信管理网中的应用[J]. 电子工程师，2000.