第34卷 第22期 计 算 机 工 程 2008年11月
Vol.34 No.22 Computer Engineering November 2008
·软件技术与数据库· 文章编号：1000—3428(2008)22—0057—03 文献标识码：A 中图分类号：TP311.52

基于三维坐标的关联规则可视化新技术

易先卉，彭 黎
(湖南大学软件学院，长沙 410082)

摘 要：关联规则可视化技术中普遍存在界面紊乱、产生歧义等问题。该文提出一种新的关联规则可视化方法 ARVMiner，利用三维坐标可视化技术改进现有可视化技术的不足。采用 Java 3D 可视化技术实现了基于 ARVMiner 的关联规则可视化系统原型。实验表明，该系统能够有效、有序地显示大量多种关系的关联规则，用户可以根据给定的约束条件进行有选择的挖掘。

关键词：关联规则；可视化；数据挖掘

New Visualization Technique for Association Rules Based on Three-dimensional Coordinate

YI Xian-hui, PENG Li
(Software College, Hunan University, Changsha 410082)

【Abstract】 There are many common problems such as confusion and ambiguity in the visualization results. A new novel method ARVMiner for visualizing association rules is introduced, which uses three-dimensional coordinate to improve those existed visualization techniques. A system for visualizing association rules based on ARVMiner is implemented with Java 3D. The system is not only efficiently and orderly visualizes all kinds of association rules, but also filters the mined results by constraint condition.

【Key words】 association rule; visualization; data mining

1 概述

关联规则挖掘是数据挖掘领域中的一个重要研究方向，也是数据挖掘的主要方法之一，其目的是找出大量数据中项集之间隐含的关系网，从中发现一些对用户有用的信息，因此，有效地表示关联规则挖掘结果很重要。可视化技术充分利用了图形和图像的表达能力以及人对色彩和形状的敏锐感知能力，使用户可以更加方便地对结果进行观察和分析。将可视化技术应用到关联规则挖掘结果的表示中，是关联规则挖掘研究的一个新进展。近年来，文献[1-6]提出了多种可视化技术来支持用户对关联规则的观察和分析。但这些可视化技术普遍存在很多不足之处，如不能高效地描绘多对多关系的关联规则(前项和后项中都包含多个项的规则)，只能描绘一对一或多对一关系的关联规则(前项中包含多个项，后项中只包含一个项的规则)，关联规则的支持度和置信度不能清楚地标注，且关联规则数量较多时，会出现可视化界面紊乱，容易产生歧义。

2 关联规则的基本定义

1993年，文献[7]中提出了挖掘顾客交易数据库中项集间的关联规则问题，阐述了解决该问题的核心算法——Apriori算法。

设$I=\{i_1,i_2,\cdots,i_m\}$为数据项集合，D为与任务相关的交易数据库，其中的每一个交易T是一个数据项子集，即$T\subseteq I$，每一条交易记录存在一个识别编号TID。A为数据项集合，当且仅当$A\subseteq T$时，称交易T包含A。

定义 1 关联规则是具有“$A\Rightarrow B$”形式的蕴涵式，其中，A为前项集(antecedent)且$A\subset I$；B为后项集(consequent)且$B\subset I$，并且$A\cap B=\Phi$。

定义 2 规则“$A\Rightarrow B$”的支持度s为D中同时包含A和B的事务数与总的事务数的比值，描述为：$s(A\Rightarrow B)=P(A\cup B)$，最小支持度记为$min_sup$。

定义 3 规则“$A\Rightarrow B$”的置信度c为D中同时包含A和B的事务数与只包含A的事务数的比值，即D中包含A的事务中也包含B的百分率，描述为：$c(A\Rightarrow B)=P(B|A)$，最小支持度记为min_conf。同时满足最小支持度和最小置信度的关联规则称为强关联规则。

定义 4 数据项的集合称为项集(itemset)，包含k个数据项的项集称为k-项集。如果一个项集满足最小支持度，称该项集为频繁项集(frequent itemset)。

定义 5 核心算法 Apriori 的中心思想是由频繁(k-1)-项集构建候选k-项集。首先找到所有的频繁1-项集；然后扩展频繁(k-1)-项集得到候选k-项集；最后剪除候选k-项集中不满足最小支持度的项集，生成频繁k-项集，依此类推，直至没有新的频繁项集被发现，每次寻找频繁项集都会对数据库作一次完全的扫描。

3 现有的关联规则可视化技术

一幅关联规则挖掘结果可视化图中至少要体现5个参数：规则的前项，规则的后项，前项与后项间的关系，规则的支持度，规则的置信度。目前用于关联规则的可视化技术主要有如下5种：

(1)基于表的可视化技术

该方法的基本思想就是用文字化的表结构描述关联规

作者简介：易先卉(1978－)，女，硕士研究生，主研方向：数据挖掘与可视化，计算机应用技术；彭 黎，副教授、博士

收稿日期：2007-12-14 **E-mail：**rj_yxh@hnu.cn

则。表中的每一行描述一条关联规则，每一列分别描述关联规则中的参数，包括规则的前项、后项、支持度和置信度。

此方法的优点是能够利用表的基本操作对规则的参数值进行排序或者过滤出前项和后项中包含特定项目的规则。它的缺点是，规则采用文字描述，不能利用人脑对色彩和图像敏锐的感知能力，不利于用户方便深入地对结果进行观察和分析。

(2)基于二维矩阵的可视化技术

该方法的基本思想是用一个二维矩阵的行和列分别表示规则的前项和后项，并在对应的矩阵单元画图，可以是柱状图或条形图等。不同的图形元素(如颜色或高度)可以用来描述关联规则的不同参数，如规则的支持度和置信度[1]。二维矩阵法的优点是易于可视化一对一关系的关联规则，但在可视化多对一以及多对多关系的规则时，二维矩阵法的局限性就显现出来了。

很难判断图 1 表示的是一条关联规则 $A+B \Rightarrow C$，还是 2 条关联规则 $A \Rightarrow C$ 和 $B \Rightarrow C$。将规则的前项(后项)作为一个单独的实体标志在二维矩阵中来解决多对一、多对多的歧义问题，但是这种方式会使二维矩阵变得很大，特别是生成的规则较多时，不利于视图的生成。

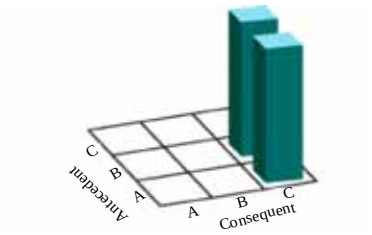

图 1　有歧义的二维矩阵

此外，二维矩阵法还存在一个严重的问题，当存在大量关联规则需要可视化时，后面的图形会被前面的图形遮蔽。

(3)基于有向图的可视化技术

有向图法[2]是另一种流行的关联规则可视化技术。其基本思想是有向图中的节点代表项，连接 2 个节点的边代表项间的关联。图 2 共显示了 3 条规则，左图表示 2 条关联规则 $A \Rightarrow C$ 和 $B \Rightarrow C$，右图表示一条关联规则 $A+B \Rightarrow C$。当只需显示少量项(节点)和少量关联规则(边)，此方法非常有效。但当项数和关联规则数量增多时，有向图很快变得十分紊乱。此外，有向图法不能清楚地标注支持度和置信度等关联规则参数值。

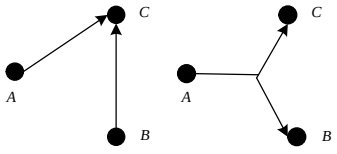

图 2　关联规则的有向图表示

(4)基于平行坐标的可视化技术

平行坐标技术[3]是最早提出的以二维形式表示 n 维空间的数据可视化技术之一。它的基本思想是将 n 维数据空间用 n 条等距离的垂直平行轴映射到二维平面上，每条轴线都对应于一个属性维。坐标轴的取值范围从对应数据维属性的最小值到最大值均匀分布(名词性属性依次在数据维上标出即可)，这样数据库中的每一条数据记录都可以用一条折线表示在 n 条垂直平行轴上，如图 3 所示。它的优点是表达多维数据关系非常直观，易于理解。缺点是表达维数决定于屏幕的水平宽度，当维数增加，引起垂直轴靠近，辨认数据的结构和关系变得困难。坐标间的依赖关系很强，垂直平行轴的维数次序会影响数据之间关系的查找，而且多维结构也是复杂的。另外，关联规则参数值标注也不太方便。目前已有一些对平行坐标的改进技术[4-5]。

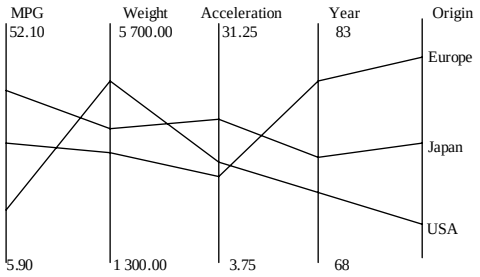

图 3　基于平行坐标的关联规则可视化图

(5)基于规则多边形的可视化技术

规则多边形表示法[6]非常适用于多维关联规则挖掘结果的显示，例如有以下多维关联规则：

age(“30~40”)∧income(“40k~50k”) ∧
occupation (“businessman”) ⇒ buy(“laptop”)

用规则多边形表示法可视化，如图 4 所示。该方法可以清楚地表示关联规则的各个维，用户能够从图中简单直观地获知从特定条件最终会得到哪种后续结果。然而，规则多边形表示法在需要同时显示多条多维关联规则时，图形会出现重叠，效果会明显降低。此外支持度、置信度的标注也不太方便。

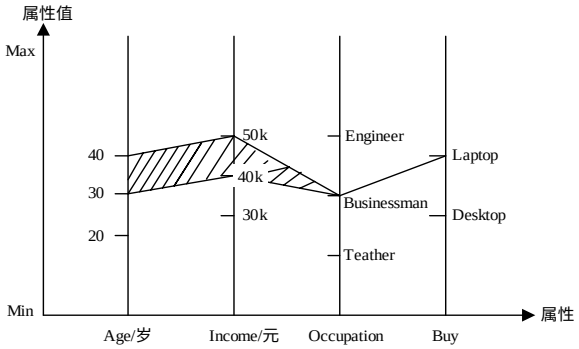

图 4　基于规则多边形的多维关联规则

4　新的关联规则可视化技术 ARVMiner

现有关联规则可视化技术大部分利用二维图形来描绘关联规则，如果将关联规则可视化的结果用三维图形表示，可以改善现有可视化技术的不足，本文提出一种新的可视化技术 ARVMiner，它利用 Java 3D 可视化技术的特点，将关联规则的 5 个参数用三维坐标的颜色、图形及三维属性表示出来。这种可视化技术能有效地表示一对一、一对多、多对一及多对多关系的关联规则，而且还可以可视化多维关联规则，关联规则的支持度和置信度的标注也容易实现，当关联规则较多时，也不会出现界面紊乱和产生歧义等问题。

图 5 中给出了对蘑菇数据库(mushroom)进行挖掘后的多维关联规则表，其最小支持度、置信度的取值为 0.9，采用 ARVMiner 技术后的可视化结果如图 6 所示，其中，X 轴表示出现在每条规则中的不同项；Z 轴表示每条规则；Y 轴表示每条规则对应的支持度和置信度。方格内淡色立方体表示前项、深色立方体表示后项，方格末端深色长形立方体表示支持度、淡色长形立方体表示置信度。黑色数字对应每条水

平线 Y 值，用来标注支持度和置信度的大小。

Mining Result

Antecedent	Consequent	Support	Confidence
veil-color=white	gill-attachment=free	0.9731659	0.9977284
gill-attachment=free	veil-color=white	0.9731659	0.99898916
veil-type=partial	gill-attachment=free	0.97415066	0.97415066
gill-attachment=free	veil-type=partial	0.97415066	1.0
veil-type=partial, veil-color=white	gill-attachment=free	0.9731659	0.9977284
gill-attachment=free, veil-color=white	veil-type=partial	0.9731659	1.0
gill-attachment=free, veil-type=partial	veil-color=white	0.9731659	0.99898916
veil-color=white	gill-attachment=free, veil-type=partial	0.9731659	0.9977284
veil-type=partial	gill-attachment=free, veil-color=white	0.9731659	0.9731659
gill-attachment=free	veil-type=partial, veil-color=white	0.9731659	0.99898916
ring-number=one	veil-type=partial	0.9217135	1.0
veil-type=partial	ring-number=one	0.9217135	0.9217135
veil-color=white	veil-type=partial	0.9753816	1.0
veil-type=partial	veil-color=white	0.9753816	0.9753816

图 5　蘑菇数据库的多维关联规则

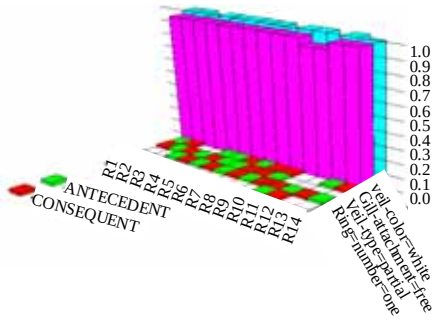

图 6　基于 ARVMiner 技术的多维关联规则三维图

5　基于 ARVMiner 的关联规则可视化原型系统

关联规则挖掘算法常产生大量的关联规则，将所有的规则一次性显示出来，容易造成界面紊乱。在大多数情况下，用户只对关联规则的一部分感兴趣，例如用户只想知道包含某些特定项的规则。为了使用户更加有效地提取感兴趣的关联规则，本文用 Java 3D 可视化技术开发了基于 ARVMiner 的关联规则可视化系统，此系统在可视化的过程中提供了友好的人机交互界面。用户可以给定约束条件，选择感兴趣的项进行挖掘，其挖掘结果也可以按照支持度(置信度、前项或后项)排序显示。例如，用户想知道前项中包含“ veil-type= partial ”且支持度和置信度都大于 0.9 的关联规则，挖掘出来的关联规则按置信度递减的顺序排列的可视化结果如图 7 所示。

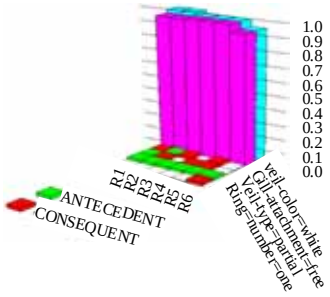

图 7　基于 Java 3D 技术的约束性关联规则三维图

另外，当显示的规则较多时，为了更加清楚地识别每条规则的参数，用户可利用鼠标的左、中、右键分别来控制三维图形的旋转、缩放和平移。此原型系统集成了 3 种关联规则挖掘算法：Apriori[8], FPgrowth[9], Closure[10]。可选择挖掘的测试数据库有mushroom.db, T20_AT5_I5_P5_AP5.db(手动生成，其中，T20 表示事务数为 20；AT5 表示事务的平均大小为 5；I5 表示项数为 5；P5 表示模式数为 5；AP5 表示模式平均大小为 5)。用户在挖掘某一数据库之前可以先用这 3 种算法进行测试，根据扫描数据库时间与扫描数据库次数的比值确定哪种算法更适合选中挖掘的数据库。

6　结束语

本文在分析现有关联规则可视化技术的优缺点的基础上实现了基于 ARVMiner 技术的关联规则可视化系统原型。实验表明，该系统不仅能够有效地解决目前关联规则可视化技术的许多问题，而且还提供了友好的人机交互界面。用户可以选择不同的关联规则挖掘算法对不同数据库进行挖掘。还可以根据给定的约束条件对挖掘结果进行过滤，从而方便用户对结果进行观察和分析。今后将对此原型系统数据库进行拓展，使其成为现实生活中应用广泛的数据库。另外，不断开发新的可视化技术和关联规则挖掘算法，集成到此系统原型中，使其成为一个集多种可视化技术、挖掘算法和数据库为一体的关联规则可视化应用系统。

参考文献

[1] Wong Pak Chung, Whitney P, Thomas J. Visualizing Association Rules for Text Mining[C]//Proceedings of the 1999 IEEE Symposium on Information Visualization. Richland, USA: [s. n.], 1999: 120-123.

[2] Hetzler B, Harris W M, Havre S, et al. Visualizing the Full Spectrum of Document Relationships[C]//Proceedings of the 5th Int’l Conf. of Society for Knowledge Organization. Wurzburg, France: Verlag, 1998: 168-175.

[3] Li Yang. Visualizing Frequent Itemsets, Association Rules, and Sequential Patterns in Parallel Coordinates[C]//Proceedings of Int’l Conf. on Computational Science and Its Applications. Montreal, Canada: [s. n.], 2003: 21-30.

[4] Li Yang. Pruning and Visualizing Generalized Association Rules in Parallel Coordinates[J]. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(1): 60-70.

[5] 吉根林，韦素云，曲维光. 基于平行坐标的关联规则可视化新技术[J]. 计算机工程, 2006, 32(24): 87-89.

[6] Han Jiaochan, Cercone N. RuleViz: A Model for Visualizing Knowledge Discovery Process[C]//Proceedings of KDD’00. Boston, USA: [s. n.], 2000: 244-253.

[7] Agrawal R, Imielinski T, Swami A. Mining Association Rules Between Sets of Items in Large Databases[C]//Proceedings of the ACM SIGMOD Int’l Conf. on Management of Data. Washington, USA: [s. n.], 1993: 207-216.

[8] Agrawal R, Srikant R. Fast Algorithms for Mining Association Rules[C]//Proceedings of 1994 Int’l Conf. on Very Large Data Bases. Santiago, Chile: [s. n.], 1994: 487-499.

[9] Han Jiawei, Pei Jian, Yin Yiwen. Mining Frequent Patterns Without Candidate Generation[C]//Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data. Dallas, USA: [s. n.], 2000: 1-12.

[10] Cristofor D, Cristofor L, Simovici D. Galois Connection and Data Mining[J]. Journal of Universal Computer Science, 2000, 6(1): 60-73.