关联规则的基本研究

牛猛
（皖南医学院 教务处，安徽 芜湖 241002）

[摘 要]关联规则是数据挖掘的一个重要的研究领域。简要介绍了关联规则的产生与概述；详细介绍了关联规则的相关定义、性质、步骤与分类情况；阐述了关联规则的多种挖掘方法。

[关键词]关联规则，定义，性质，步骤，分类，方法
doi:10.3969/j.issn.1673-9477.2016.02.039
[中图分类号]G43 [文献标识码]A [文章编号]1673-9477(2016)02-114-04

一、关联规则的产生

关联规则挖掘最早应用于对零售业中的购物篮数据进行分析。零售机构记录了大量的销售记录，这些销售记录被称为购物篮数据(basket data)，而记录顾客购物信息的购物篮数据称为事务(transaction)。

在此情况下，1993 年 Agrawal 等人首先探索了销售记录中项集间的关联问题，并提出了基于频繁项集的 Apriori 算法。

二、关联规则的概述

关联规则挖掘就是从大量数据中挖掘出有价值的描述数据项之间相互联系的相关知识的数据挖掘过程[1]。

关联规则善于找出数据库中满足相关要求的数据属性域之间的相互关系。

三、关联规则的相关定义

（一）项/项目与项集

项/项目是数据库中最小的、不可分割的信息单位，用 i 表示。

项集是项/项目的集合。

设定集合 $I=\{i_1, i_2, \cdots, i_k\}$，则 $i_1, i_2, \cdots, i_k$ 为集合中的项。集合中的项的数目为 k，则集合称为 k-项集。

（二）事务与事务数据库

设集合 I 是由数据库中所有项目构成的集合，即 $I=\{i_1, i_2, \cdots, i_k\}$。一次处理所含项目的集合用 T 表示，T 包含于 I，即 $T \subseteq I$，并且每一个 T 都有唯一的标识 TID，则二元组<TID,T>为数据库中的事务，在不混淆的情况下可简单表示为 T。

事务数据库是所有事务的集合，用 D 表示，即 $D=\{t_1, t_2, \cdots, t_m\}$。

（三）项集的频率

设 U 为项集，$U=\{u_1, u_2, \cdots, u_n\}$，且 $U \subseteq I$，$U \neq \Phi$；D 为包含所有事务的事务数据库，$D=\{t_1, t_2, \cdots, t_m\}$；设 Q 为事务的集合，$Q=\{T_i | T_i \in D, U \subseteq T_i\}$。则项集 U 在事务数据库 D 中的频率（简称为支持计数或计数）为：$P(U)=P(u_1 \hat{} u_2 \hat{} \cdots \hat{} u_n)=|Q|/|D|*100\%$。

（四）关联规则

关联规则是形如 $X \Rightarrow Y$ 的蕴含式，其中 $X \subset I$，$Y \subset I$，并且 $X \cap Y=\Phi$。X 称为关联规则的前提，Y 称为关联规则的结果。关联规则揭示的是前提(X)中的项目出现时，结果(Y)中的项目也跟着出现的规律。

（五）支持度(Support)

支持度描述事务数据库中同时包含前提（X）和结果（Y）的事务数占所有事务数的比值，即包含前提（X）和结果（Y）的事务数在事务集中出现的频率，记为 $Support(X \Rightarrow Y)$，即 $Support(X \Rightarrow Y)=Support(X \cup Y)=P(X \cup Y)$ [2]。

（六）置信度(Confidence)

置信度描述事务数据库中包含前提（X）和结果（Y）的事务数占包含前提（X）的事务数的比值，即在包含前提（X）的事务中出现结果（Y）的概率，记为 $Confidence(X \Rightarrow Y)$，即 $Confidence(X \Rightarrow Y)=Support(X \cup Y)/Support(X)=P(Y/X)$ [3]。

支持度描述了挖掘出的关联规则在整个事务数据库中的有用性（即统计重要性）；置信度描述了挖掘出的关联规则在整个事务数据库中的确定性（即可靠程度）。通常，只有支持度和置信度都比较高的关联规则才是有价值的关联规则。

（七）最小支持度（min_sup）与最小置信度(min_conf)

考虑到实际情况的要求，一般均需要为关联规

[投稿日期]2016-03-20
[基金项目]安徽省高校人文社会科学研究项目（编号：SK2014A416）
[作者简介]牛猛（1982-），男，安徽淮北人，助教，硕士，研究方向：数据挖掘。

则指定必须要满足的支持度阈限和置信度阈限。当支持度和置信度均大于或等于各自的阈限值时，认为关联规则是有价值的，这两个阈限值分别称为最小支持度阈值(min_sup)和最小置信度阈值(min_conf)。其中，min_sup 明确了关联规则的最低有用性（即最低统计重要性），min_conf 明确了关联规则的最低确定性（即最低可靠程度）。

（八）频繁项集

设 U 为项集，$U=\{u_1, u_2, \cdots, u_n\}$，且 $U \subseteq I$，$U \neq \Phi$ ；设最小支持度为 min_sup，若项集 U 的支持度 Support(U)≥min_sup，则称项集 U 为频繁项集；否则称项集 U 为非频繁项集[4]。若某一项满足最小支持度要求，则称该项为频繁项。

（九）强关联规则

满足 $Support(X \Rightarrow Y) \geqslant min\_sup$，并且同时满足 $Confidence(X \Rightarrow Y) \geqslant min\_conf$ 的关联规则称为强关联规则；其他不满足的关联规则称为弱关联规则[5]。

强关联规则必须同时满足 min_sup 和 min_conf 这两个条件的要求。关联规则挖掘实质上就是挖掘出强关联规则的过程。

四、关联规则的性质

（一）非结合性

若关联规则 $X \Rightarrow Z$ 及 $Y \Rightarrow Z$ 在事务数据库 D 中成立，但关联规则 $X \cup Y \Rightarrow Z$ 在 D 中不一定成立。

（二）不可分解性

若关联规则 $X \cup Y \Rightarrow Z$ 在事务数据库 D 中成立，但关联规则 $X \Rightarrow Z$ 及 $Y \Rightarrow Z$ 在 D 中不一定成立。

（三）不可传递性

若关联规则 $X \Rightarrow Y$ 及 $Y \Rightarrow Z$ 在事务数据库 D 中成立，但不能推出关联规则 $X \Rightarrow Z$。

（四）可扩展性

设有项目集 L，A，B，且 $B \subseteq A \subset L$，若关联规则 $A \Rightarrow (L-A)$ 不满足最小可信度条件，则 $B \Rightarrow (L-B)$ 也不满足最小可信度条件。

即：对项目集 L，C，D，且 $C \subseteq D \subset L$，C=0，若关联规则 $(L-D) \Rightarrow D$ 成立，则关联规则 $(L-C) \Rightarrow C$ 也成立。

五、关联规则的步骤

（一）找出所有频繁项集

从资料集合中找出所有频率大于或等于最小支持度的频繁项集。找出所有频繁项集是挖掘的前提，决定了挖掘的整体效率，也是现在研究的重点。

（二）根据频繁项集和最小置信度产生强关联规则

在挖掘出的所有频繁项集的基础上，列出所有可能的关联规则，根据支持度和置信度同时大于或等于 min_sup 和 min_conf 的原则生成强关联规则。

（三）关联规则挖掘需注意的问题

1.明确目标

首先要明确关联规则挖掘的目标，即用户需要解决什么问题；然后确定挖掘的数据类型和采用的算法；之后制定挖掘的步骤，确定每一步的操作和实现的目标，确保挖掘的有效进行。

2.数据准备

数据准备非常重要，准备的好坏将直接影响到数据挖掘的准确性和效率。关联规则挖掘通常适用于变量取离散值的情况。若变量取连续值，则应先进行数据离散化（即将区间内不同范围的连续值分别对应于不同的具体的离散值）。数据离散化是数据准备的重要工作，是挖掘的前提，离散化的合理性将直接影响挖掘的准确性和效率。

3.确定最小支持度和最小置信度

要选择合适的最小支持度和最小置信度阈值，不能太小，也不能太大。太小会获得过多的规则，而其中大部分可能都是无用的，这样会明显降低挖掘的效率；太大则会获得极少的规则甚至是找不到规则。

4.理解关联规则

关联规则挖掘仅仅是一种能够找出满足要求的规则的分析方法，挖掘出的结果无法判断是否具有实际意义。需根据实际情况，确定哪些规则有意义，哪些规则没有意义；有意义的被采用，没有意义的不能被采用。

六、关联规则的分类

分类方法有多种，分类标准不同，所属类型也不同，主要有以下几种。

（一）根据涉及到的值的类型，可分为布尔型关联规则(Boolean Association Rule)和量化型关联规则(Quantitative Association Rule)[6]

布尔型关联规则涉及到的值均是离散的、种类化的，它能揭示这些值之间的关系。

量化型关联规则涉及到的是量化的值或属性之间的关系，又被称为数值型关联规则。在量化型关联规则中，将项或属性的量化值划分为不同的区间。

（二）根据涉及到的数据维数，可分为单维关联规则(Single-Dimensional Association Rule)和多维关联规则(Multi-dimensional Association Rule)[7]

若要处理的数据只涉及一个维，称为单维关联规则。单维关联规则忽略了现实中数据的多个不同属性，仅需考虑单个属性的相互关系。

若要处理的数据涉及超过一个维，称为多维关联规则。多维关联规则必须考虑多个属性及属性之间的相互关系。

（三）根据规则中涉及到的抽象层次，可分为单层关联规则(Single-level Association Rule)[8]和多层关联规则(Multi-level Association Rule)[9]

在单层关联规则中，涉及到的数据只有一个抽象层次。不用考虑不同的抽象层次，只需要处理同一抽象层次的项或属性间的关联。

在多层关联规则中，涉及到的数据有多个不同的抽象层次。必须要考虑来自不同抽象层次的数据或属性间的关联。

七、关联规则的方法

（一）基于多循环方式的挖掘方法

是最基本挖掘方法，有 AIS 算法[10]、Apriori 算法、AprioriTid 多维关联算法、AprioriHybrid 混合算法[11]、Partition 分割算法及 Sampling 抽样算法等。其中，Apriori 算法是使用逐层搜索的迭代方法，找到频繁项集，挖掘关联规则的单维、单层的宽度优先算法；分割算法 Partition 通过对数据库进行分割，从而减少挖掘过程中 I/O 操作次数；抽样算法 Sampling 先对数据库进行抽样，然后对抽样数据进行挖掘，从而提高挖掘的效率。目前国内的主要研究方向是对 Apriori 算法进行改进。

（二）并行挖掘方法

包括计数分布 CD、候选分布 CaD、数据分布 DD[12]、PDM 以及 DMA 等方法[13]。虽然它们通常被用于分布式数据库的挖掘，但也可被认为是并行挖掘方法。

（三）增量式更新方法

主要有两种情况：一是在给定 min_sup 和 min_conf 的基础上，当数据库记录发生变化时（如添加或者删除记录），如何生成关联规则；二是在给定数据库的基础上，当 min_sup 和 min_conf 发生变化时，如何生成关联规则。此类算法包括 NIUA、FUP、IUA 以及 PIUA 等算法[14]。

（四）基于约束的挖掘方法

为了便于用户参与、指导以及控制挖掘的过程，增加挖掘的交互性，设置相应的约束条件。在约束条件下，进行关联规则挖掘的方法称为基于约束的挖掘方法。约束类型多样。此类算法包括 CFG 算法、Direct 算法等算法。

（五）基于多值属性的挖掘方法

一般通过将多值属性的每一个类别转化为一个属性，从而将多值属性挖掘转化为布尔型挖掘。通常可分为基于数量的挖掘方法和基于类别的挖掘方法。此类算法包括 SA 算法和 Equi-Depth Partitioning 算法等。

参考文献：

[1]朱祥玉，侯德文，陈希.对关联规则挖掘 Apriori 算法的进一步改进[J].信息技术与信息化，2005(6):81-83.

[2]郑涛.基于超市交易信息数据挖掘的城市居民消费行为研究[J].科技情报开发与经济，2010(19):136-138.

[3]文拯.关联规则算法的研究[M].中南大学，2009.

[4]刘寒冰.数据挖掘中的关联规则算法研究[M].河北工程大学，2007.

[5]高明.关联规则挖掘算法的研究及其应用[M].山东师范大学，2006.

[6]R.Ng.L.V.S. Lakshmanan , J.Han and A.Pang.Exploratory Mining and Pruning OPtimizations of Constrained Associations Rules [C]. Proc. of 1998 ACM-SIGMOD Conf. on Management of Data，Seattle，Washington，June 1998:13-24.

[7]秦锋，杨学兵.一种基于 APRIORI 性质的多维关联规则挖掘算法的研究[J].安徽工业大学学报，2003，20(2):141-144.

[8]王文清，乔雪峰.带有时态约束的多层次关联规则的挖掘[J].北京理工大学学报，2003(1):57-90.

[9]程继华，施鹏飞.多层次关联规则的有效挖掘算法[J].计算机学报，1998(11):1037-1041.

[10]R.Agrawal , T.Imielinski , and A.Swami. Mining association rules between sets of items in large database[C]. Proceedings of the ACM SIGMOD International Conference on Management of Data(SIGMOD'93), Washington，DC，1993.ACM Press Publisher，1993:207-216.

[11]R.Agrawal，R.Srikant. Fast Algorithms for Mining Association Rules[C]. Proceedings of the 20th International Conference on Very Large Databases(VLDB'94)，Santiago，Chile,1994:487-499.

[12]R.Agrawal , J.C.Shafer.Parallel Mining of Association Rules[J].IEEE Transactions on knowledge and date engineering，1996，8(6):962-969.

[13]D.W.Cheung.Efficient Mining of Association Rules in Distributed Databases[J].IEEE Transactions on

Knowledge and Data Engineering，1996(6):910-921.
[14]冯玉才，冯剑琳.关联规则的增量式更新算法[J].软件学报，1998(4):301-306.

[责任编辑　王云江]

The basic research of association rules

NIU Meng
(Dean's Office, Wannan Medical College, Wuhu 241002, China)

Abstract: The association rule is an important area of research for data mining. This paper briefly introduces its generation and overview, and also detailedly introduces its related definitions, property, procedure and classification. Additionally, this paper expounds its different kinds of mining methods.

Key words: association rule; definition; property; procedure; classification; method

（上接第 106 页）

（四）完善教师评价机制

改变职称评价标准是各院校首先考虑的重大问题。目前的应用技术院校职称评定多是以学术标准为主，兼顾理论教学水平，缺少对操作的考察。若要提高教师的实践能力，在职称评定标准方面，院校就必须弱化学术方面的要求，比如论文发表数量、综合外语能力等，而要强化实践性的要求，比如某项技术的操作水平、理论在生活中的应用能力等。这是以外在的激励制度来促进教师的意识及行为的转化。

其次，在给教师进行晋升或者是奖励的评定时，应更多侧重于成果的实质性评价。比如在技术教学上的成绩，在技术更新上的研究等。注重质的管理，才能使教师取得根本上的改变与进步。

各院校在薪酬的发放方面，要实现薪酬的合理化和优化，吸引更多的优秀教学人才，激励教师对自身的不断完善。

师资队伍水平的高低决定了院校办学水平的优劣，也对人才培养的质量起着举足轻重的影响。因此转型后的应用技术学院必须建立一支优秀的双师型师资队伍，才能满足为社会培养高技术型人才以及推动企业技术进步这一重大的社会责任。

参考文献：

[1]董建梅.挑战与应对：新建本科院校师资队伍转型思考[J].衡水学院学报，2014(12):84－85.
[2]彭磊.产学研合作背景下的应用型大学师资队伍建设研究[J]．中国建设教育，2013(2):24－26.
[3]詹学文，詹秋文．地方应用型本科高校师资队伍建设研究[J]．黄山学院学报，2013(4):102－104.
[4]张海龙，王占礼．应用型本科人才培养的师资队伍研究与对策建议[J]．中国大学教学，2011(5):35－38.
[5]张淑敏．我校师资队伍建设的几点建议[J]．东北财经大学学报，2003(1):58－62.

[责任编辑　王云江]

Thoughts on the construction of teaching staff in the transformation of normal universities to technology-applied types

LIU Zhuo
（Department of Personnel, Fujian Jiangxia University, Fuzhou 350108, China）

Abstract: Under the demand of the society and the government's call, some ordinary undergraduate colleges and universities began to apply the technology-applied transition. However, in this process, the traditional teaching staff construction inertia leads to biased recruitment consciousness, teachers'lack of practical experience, part-time teachers being rare, and unsuitable evaluation, impeding the transition phenomenon.Therefore, colleges and universities need to take some new measures to adapt to the construction of dual-mode teachers, in order to ensure the smooth progress of the transformation. On the basis of analyzing the current situation and existing problems of the current university faculty construction, this paper puts forward the basic ideas and concrete measures for the construction of the teaching staff in the transformation and development of the technology-applied type.

Key words: teaching staff; dual-mode teacher; practice; mechanism.