第6辑

BOSHIWENCONG

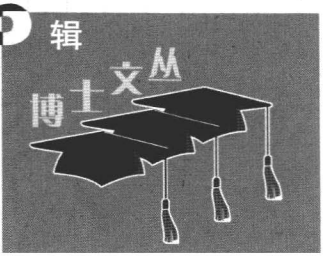
张 诤 著

大规模复杂数据关联规则挖掘方法研究及其应用

Study and Application of Large-scale Complex Data Association Rules Mining Methods

兰州大学出版社

第6辑



张 诤 著

大规模复杂数据关联规则挖掘方法研究及其应用

Study and Application of Large-scale Complex Data Association Rules Mining Methods

图书在版编目(CIP)数据

大规模复杂数据关联规则挖掘方法研究及其应用/张诤著.—兰州:兰州大学出版社,2009.5

(博士文丛.第6辑)

ISBN 978-7-311-03061-2

Ⅰ.大… Ⅱ.张… Ⅲ.数据采集—研究 Ⅳ.TP274

中国版本图书馆CIP数据核字(2009)第068362号

博士文丛(第六辑)

大规模复杂数据关联规则挖掘方法研究及其应用

张 诤 著

兰州大学出版社出版发行

兰州市天水南路222号 电话:8912613 邮编:730000

E-mail:press@onbook.com.cn

http://www.onbook.com.cn

兰州大学出版社激光照排中心排版

兰州德辉印刷有限责任公司印刷

开本:880×1230 1/32 印张:5.625

2009年5月第1版 2009年5月第1次印刷

字数:140千

ISBN 978-7-311-03061-2 定价:220.00元(共十册)

摘　要

随着技术的进步和社会的发展，各领域内的数据库规模不断扩大，而且日益呈现出大样本、多指标、时序性、含噪声等特点。具有这些特点的大规模复杂数据集对现有关联规则挖掘方法的挖掘效率提出了新的挑战。因此，如何从含大量噪声的大规模复杂数据集中提取有效信息，如何针对大规模复杂数据集设计高效的关联规则挖掘算法，如何提高关联规则挖掘的效率和质量，便成为目前关联规则挖掘研究中的一个核心问题。

针对目前关联规则挖掘方法在挖掘大规模复杂数据集时所面临的效率问题，本文从理论模型的构建、理论方法的设计、理论模型与方法的应用与实证三个方面对大规模复杂数据集关联规则挖掘问题展开了深入研究，提出了一套高效的大规模复杂数据关联规则挖掘方法，具有重要的理论意义、现实意义和应用价值。

具体来说，本文的主要研究内容和创新成果如下：

1. 提出了基于数据简约和数据压缩思想的大规模复杂数据关联

规则挖掘模型。

Agrawal关联规则挖掘模型在描述大规模复杂数据关联规则挖掘过程时存在很大的局限性。本文在Agrawal关联规则挖掘模型基础上,结合数据简约和数据压缩思想及其方法,对Agrawal关联规则挖掘模型进行了拓展,提出了基于数据简约和数据压缩思想的大规模复杂数据关联规则挖掘模型,为今后大规模复杂数据集关联规则挖掘研究提供了模型框架。

2. 提出了三种高效的大规模复杂数据关联规则挖掘算法。

(1)提出了基于二元关系矩阵及其运算的频繁项集挖掘算法(Binary Relation Matrix Frequent Itemset Mining, BRMM)。

针对Apriori算法在挖掘大规模复杂数据集时存在的数据结构、连接和剪枝策略、搜索空间等方面的不足,首先将简约、压缩后的事务数据库进一步压缩映射为二元关系矩阵数据结构,在此基础上,提出了一种新的剪枝和连接策略以及搜索空间的动态压缩策略,应用二元关系矩阵数据结构和相应的优化策略,设计了基于二元关系矩阵及其运算的频繁项集挖掘算法,并通过示例和实验对算法的性能进行了分析比较。实验结果表明,无论是在大规模稀疏数据集上还是在稠密数据集上,BRMM算法的挖掘效率都要明显优于Apriori算法。最后,为了适应更大规模关联规则挖掘的需要,将BRMM算法进行了拓展,设计了基于二元关系矩阵及其运算的并行频繁项集挖掘算法(Parallel Binary Relation Matrix Frequent Itemset Mining, PBRMM),并对算法性能进行了分析。分析表明,PBRMM算法同样是一种高效的并行频繁项集挖掘算法。

(2)提出了基于频繁项目关系矩阵的频繁项集挖掘算法(Frequent Item Matrix Mining, FIMM)和相应的事务间关联规则挖掘算法。

针对Apriori算法和FP-Growth算法在挖掘大规模稠密数据集和稀疏数据集时各自存在的不足,提出了空间压缩效率更高、挖掘效率更好的频繁项目关系矩阵数据结构。在此基础上,将Apriori算法和FP-

Growth 算法的优点进行了结合，提出了宽度优先和深度优先相结合的基于频繁项目关系矩阵的频繁项集挖掘算法和相应的关联规则挖掘算法，并通过示例和实验对算法的性能进行了分析比较。实验结果表明，无论是在大规模稀疏数据集上还是在稠密数据集上，FIMM 算法的挖掘效率均要明显优于 Apriori 算法和 FP-Growth 算法，说明 FIMM 算法对于具有不同特点的数据集具有更好的适应性。

(3)提出了基于时间约束频繁项目关系矩阵的频繁项集挖掘算法(Temporal Frequent Item Matrix Mining, TFIMM)和相应的关联规则挖掘算法。

针对具有时间约束的事务间关联规则挖掘问题和现有事务间关联规则挖掘算法 E-Apriori、EH-Apriori、FITI 在挖掘大规模复杂数据集时存在的不足，对频繁项目关系矩阵数据结构进行了拓展，提出了时间约束频繁项目关系矩阵数据结构。在此基础上，将宽度优先算法和深度优先算法的优点进行了结合，提出了基于时间约束频繁项目关系矩阵的频繁项集挖掘算法和相应的关联规则挖掘算法，并通过示例和实验对算法性能进行了分析比较。实验结果表明，无论是在大规模稀疏数据集上还是在稠密数据集上，TFIMM 算法的挖掘效率均要明显优于 FITI 算法，说明了 TFIMM 算法对于不同特点的数据集具有更好的适应性，是一种高效的事务间关联规则挖掘算法。

3. 对中国 A 股市场 2001—2006 年各行业板块间的板块关联效应进行了挖掘分析。

选取了 Wind 金融数据库中 2001—2006 年中国 A 股市场所有股票 1444 个交易日的开盘价、最高价、最低价、收盘价、成交量、成交额、换手率、成交均价行情数据为初始样本点。首先根据本文提出的基于数据简约和数据压缩思想的关联规则挖掘模型，利用中信证券二级行业风格指数和涨跌幅指标对初始数据进行了简约和压缩处理，在此基础上，利用本文提出的基于频繁项目关系矩阵的关联规则挖掘算法和基于时间约束频繁项目关系矩阵的关联规则挖掘算法，对中国 A 股市

场各行业板块间的板块联动效应和轮动效应进行了挖掘分析，一方面验证了本文所提模型和算法的有效性；另一方面也实证了2001—2006年，中国A股市场中各行业板块间的确存在板块联动效应，但不存在明显的板块轮动效应。

关键词：关联规则　频繁项集　数据简约　数据压缩　股票市场

Abstract

With the development of technology and society, database in various fields expands more and more, and with the characteristics of large sample, multiple variable, temporal and noisy. This large-scale complex database take new challenges to efficiency of association rules mining. Therefore, how to extract effectively information from large-scale complex database including noise, how to design an efficient algorithm of association rules for mining large-scale complex database, how to improve efficiency and quality of Association Rules Mining, will become a core of the problem of Association Rules Mining.

In view of the efficiency problem of the current association rules mining method for mining large-scale complex datasets, this paper studies large-scale complex datasets association rules mining from the construction of theoretical model, the design of algorithm and the application of theoretical model and algorithm, presents a set of efficient methods of as-

sociation rules mining for mining large-scale complex data, so, this research has important theoretical, practical significance and application value.

In this paper, the main research content and innovative achievements are as follows:

1. A new theoretical model of association rules mining based on the data reduction and data compression for mining large-scale complex database is presented.

Current Agrawal association rules mining model is limited to describe large-scale complex database association rules mining process. This paper, based on Agrawal association rules mining model, combining data reduction and data compression methods and theory, expands Agrawal association rules mining model, presents a new theoretical model of Association Rules Mining based on the data reduction and data compression for mining large-scale complex database, provides a model framework for future study about large-scale complex data association rules mining.

2. Three efficient algorithms for large-scale complex data association rules mining are presented.

(1) The frequent itemset mining algprithm based on binary relation matrix and its computation (Binary Relation Matrix Frequent Itemset Mining, BRMM) is presented.

Aiming at the flaw of Apriori algorithm when mining association rules in large-scale complex data, first, the transaction database is mapped to binary data matrix structure. Base on it, applying binary matrix data structure and the corresponding optimization strategy, the paper presents a new frequent itemset mining algprithm based on binary relation matrix and its computation. And then analysis and comparison of the algorithm performance through example and experiment are done.

Experimental results show: Whether in large dense dataset or sparse dataset, BRMM algorithm is obviously superior to Apriori algorithm. Finally, in order to adapt to the needing of more large-scale association rules mining, this paper designs a parallel algorithm based on the BRMM algorithm (Parallel Binary Relation Matrix Frequent Itemset Mining, PBRMM) and analyses the algorithm performance through example. The result shows PBRMM is also an efficient parallel algorithm.

(2) The association rules mining algorithm based on frequent item matrix (Frequent Item Matrix Mining, FIMM) is presented.

Aiming at the flaw of Apriori algorithm and FP-Growth algorithm when mining association rules in large dense dataset or sparse dataset, this paper proposes frequent item matrix structure that is more efficient in store space and better in mining efficiency. Based on it, this paper gathers the advantage of Apriori algorithm and FP-Growth algorithm, proposes the association rules mining algorithm based on frequent item matrix. And then analysis and comparison of the algorithm performance through example and experiment are done. Experimental results show: Whether in large dense dataset or sparse dataset, FIMM algorithm is obviously superior to Apriori algorithm and FP-Growth algorithm. The result shows FIMM algorithm has better adaptability for the different characteristics of dataset.

(3) The association rules mining algorithm based on temporal frequent item matrix (Temporal Frequent Item Matrix Mining, TFIMM) is presented.

Aiming at the problem of inter-transaction association rules mining and the flaw of current algorithm like E-Apriori, EH-Apriori, FITI, this paper expands the frequent item matrix data structure, proposes a time-constrain frequent item matrix data structure. Based on it, this papers

propose association rules mining algorithm based on temporal frequent item matrix combining advantage of breadth first search and depth first search. And then analysis and comparison of the algorithm performance through example and experiment are done. Experimental results show: Whether in large dense dataset or sparse dataset, FIMM algorithm is obviously superior to Apriori algorithm and FP-Growth algorithm. The result shows TFIMM algorithm has better adaptability for the different characteristics of dataset and it is a excellent inter-transaction association rules mining algorithm.

3. Analysis Action association relations between various industry share index in Chinese A-share market during 2001 to 2006 are done.

The paper selects various industry share index data in 1,444 trading days of Chinese A-share market from Wind financial database in 2001 to 2006. First, according to theoretical model based on data reduction and data compression, reduction and compression initial data according CITIC Securities industry style index and extent variable of rise and fall are done, and then analyses Action association relations between various industry share index in Chinese A-share market during 2001 to 2006, using association rules mining algorithm based on frequent item matrix and temporal frequent item matrix. One side, the validity of proposed model and algorithm is veried, and the other side, the Action association relations does exist in same trading day between various industry share index in Chinese A-share market during 2001 to 2006 is confirmed, but there is no obvious Action association relations of various industry share index in Chinaese A-share market between different trading day in 2001 to 2006.

Keywords: Association Rule; Frequent Itemset; Data reduction; Data compression; Share Market

目　录

第一章　绪　论

1.1　研究的背景及意义

早在1982年，趋势大师约翰·奈斯比(John Naisbitt)在他的首部著作《大趋势》[1]中就提到："人类正被信息淹没，却饥渴于知识。"计算机硬件技术的稳定进步为人类提供了大量的数据收集设备和存储介质；数据库技术的成熟和普及已使人类积累的数据量正在以指数方式增长；Internet技术的出现和发展已将整个世界连接成一个地球村，人们可以穿越时空般地在网上交换信息和协同工作。在这个信息爆炸的时代，面对着浩瀚无垠的数据海洋，人们呼唤着一个去粗取精、去伪存真的能将浩如烟海的数据转换成知识的技术。数据挖掘(Data Mining，DM)就是在这个背景下产生的。

数据挖掘，又称为数据库中的知识发现(Knowledge Discovery in

Database, KDD),是一门交叉性学科,涉及机器学习、模式识别、统计学、智能数据库、知识获取、数据可视化、高性能计算、专家系统等多个领域,在最近十几年获得了广泛的关注。

关联规则挖掘(Association Rules Mining)问题是 R. Agrawal 等人于 1993 年在文献[2]中首先提出来的。关联规则最初是在超市数据环境下提出的,其动机是想发现顾客购买的各种商品之间是怎样彼此联系的。这些联系通过关联规则算法进行挖掘,可以找出顾客购买行为模式,如在超市里,67%的顾客在购买啤酒的同时也会购买尿布。发现这样的规则,可以应用于商品货架设计、存货安排以及根据购买模式对用户进行分类。

正因为关联规则挖掘具有很大的商业价值,关联规则挖掘问题提出之后,关联规则以其知识表述的简单性、易理解性和实用性成为数据挖掘领域最重要的研究热点之一,是当前数据挖掘研究领域最活跃、研究最深入的一个重要分支,关联规则也是数据挖掘研究的主要知识模式之一。

由于关联规则挖掘的目标是大规模的交易数据库,因此,从关联规则挖掘诞生之日起,关联规则挖掘的效率问题就是人们关注的焦点。而关联规则挖掘中的频繁项集挖掘算法是整个关联规则挖掘问题的核心,如何提高频繁项集挖掘的效率,也是近十多年来关联规则挖掘研究的主线。1994 年,Agrawal 和 Verkamo 提出了关联规则挖掘的经典算法 Apriori[3],该算法对于包含短频繁模式的稀疏数据集挖掘是行之有效的,但也存在需多次扫描数据库、候选项集生成量大、候选项集支持度低、计算时间长等缺点。特别是在挖掘包含长频繁模式的稠密数据集时,算法效率明显下降。2000 年,Pei J. 和 Han J. 等人[4]提出了一种基于频繁模式树(FP-Tree)数据结构的频繁模式增长算法 FP-Growth,它是一种挖掘包含长频繁模式稠密数据集的有效算法,但该算法在挖