学校编码：10384　　分类号_____密级_____
学号：X200343023　　UDC _____

厦门大学

硕 士 学 位 论 文

数据挖掘中关联规则及其更新算法的研究

The Research of Association Rule and its Updating Algorithm

蔡进

指导教师姓名：薛 永 生 教授
专 业 名 称：计 算 机 应 用

论文提交日期：2007 年 5 月
论文答辩时间：2007 年 6 月
学位授予日期：2007 年 月

答辩委员会主席：_____
评 阅 人：_____

2007 年 6 月

厦门大学学位论文原创性声明

兹呈交的学位论文，是本人在导师指导下独立完成的研究成果。本人在论文写作中参考的其他个人或集体的研究成果，均在文中以明确方式标明。本人依法享有和承担由此论文而产生的权利和责任。

声明人（签名）：

年　　月　　日

厦门大学学位论文著作权使用声明

本人完全了解厦门大学有关保留、使用学位论文的规定。厦门大学有权保留并向国家主管部门或其指定机构送交论文的纸质版和电子版，有权将学位论文用于非赢利目的的少量复制并允许论文进入学校图书馆被查阅，有权将学位论文的内容编入有关数据库进行检索，有权将学位论文的标题和摘要汇编出版。保密的学位论文在解密后适用本规定。

本学位论文属于

1、保密（　），在　　年解密后适用本授权书。

2、不保密（　）

（请在以上相应括号内打“√”）

作者签名：　　日期：　　年　月　日

导师签名：　　日期：　　年　月　日

摘要

数据挖掘就是从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中，提取隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识的过程。数据采掘工具还可以对现有的大型数据库直接挖掘，以发现数据库中潜在的知识。它在决策支持系统、专家系统和智能信息系统等各个方面有着重要的作用。数据挖掘可以采用关联规则，决策树，神经网络，遗传算法，规则推理,模糊逻辑等方法，其中以关联规则最为常用。

本文主要介绍了数据挖掘的定义、特点、分类，讨论了数据挖掘的产生和发展，并完成了关联规则典型算法的研究以及关联规则的更新算法的开发的基础研究工作。本文论述的内容主要分为以下几部分：数据挖掘技术，关联规则，经典关联规则算法的研究，关联规则的更新算法研究，关联规则的存储。

本文着重研究了如下几个方面的内容：

1．关联规则研究：关联规则自被提出以来，众多研究者对其进行了深入广泛的研究，主要涉及两个方面：一方面是算法方面，包括新算法的提出和对以往算法的改进；另一方面是对关联规则概念的拓宽和延伸。本文详细地介绍了关联规则的有关概念和定理。同时为了能够改进现有的更新算法，本文比较详细地研究了典型的关联规则挖掘算法 Apriori 算法，FP 算法；以及关联规则增量式更新算法 FUP，FUP2 算法。其中主要对现有的典型算法的算法思想、实现细节、性能等方面做了比较详细的研究。

2．关联规则更新算法研究：在论述关联规则的更新问题过程中,对于大型数据库中关联规则的发现常常需要对数据库进行大量重复的扫描工作，因此发现强壮规则的代价是很大的。然而在实际应用中，数据库不是静止的，它会随着数据记录的增加而不断地改变，而根据用户需要，常常要求发现的关联规则能反应数据库的当前状态，从而为决策支持，企业管理提供理论依据。那么在旧的数据库中的关联规则怎样有效地被更新的问题就成为每一个完善的数据挖掘系统应该考虑的问题。所谓有效地更新，是指既能利用原有的知识，又能及时发现新知识，而不是重新在更新后的数据库上运行原来的挖掘算法。本文是在考虑最小支持度不变的情况下，在交易数据库增加记录时关联规则的高效更新问题。提出一种新

的增量更新算法 EUFIA，该算法主要是想办法尽量使用系统的先验知识，减少新的候选集的数量，同时提高算法的执行效率。经反复的实验表明该算法具有优点：不需要扫描原数据库便可快速获得一些有用的局部规则，同时根据需要，最多只扫描 1 次原数据库也能获得更新后的全局频繁项集。因此 EUFIA 是一个有实际应用价值的高效的算法。

3．关联规则的存储：对于挖掘出来的大量关联规则来讲，显然这些规则不是很容易处理和存储的，当然也不容易被使用者消化理解。因此,"规则爆炸"本身也成为数据挖掘中的一个很严重的问题。本文论述了对国际流行的关联规则存储方法进行的一些基础的研究工作。此外，为避免关联规则存储时存入冗余规则，提出了：若一条规则能从一个简单规则集合中获取,则我们认为该规则是冗余规则,可以不必存储。但是，与其相关的算法，作者仍在进一步研究之中。

本文的研究工作为关联规则的进一步研究以及对现有的关联规则如何有效更新提供了一定的理论依据和方法。

关键词：数据挖掘；关联规则；增量式更新；强频繁项集；次频繁项集；弱频繁项集；

Abstract

Data mining means a process of nontrivial extraction of implicit previously unknown and potentially useful information from data in database. Data mining tools can directly mining in large database for discovering potential knowledge in database . It plays an important role in decision support system, expert system and management information system. It may exploit all kinds of tools, such as, association rules, decision trees, artificial ne ural network, genetic algorithm, rule induction, and fuzzy logic .In above tools, association rules are mostly used.

It is mainly introduced in the thesis that the definition, characteristic, classification, production and development of data mining. At the same time, it has finished the research of the typical algorithms of association rules and the development of association rules updating.The structure is: the technology of data mining, association rules, the research of typical algorithms of association rules,the research of the algorithms of association rules updating, the storage of association rules.

It is studied in the dissertation that the contents of the following several respects emphatically.

1. The research of association rules: Since the association rules are proposed, numerous researchers have carried on extensive research of deepening to it, involve two respects mainly: On one hand, namely algorithm respect, including the proposition of new algorithm and improvement of the past algorithm. On the other hand, widening and extending to the concept and theorem of association rules. This thesis has introduced the concepts and theorems of association rules in detail. At the same time in order to improve the existing updating algorithm, it studied the typical association rules algorithm Apriori, algorithm FP in detail and association rules incremental updating algorithm FUP, FUP2. Method, realizing and performance of the existing typical algorithm mainly among them have been studied more in detail.

2. In the course of research the association rules updating, the mining of association rules often needs a large amount of repeated scanning work in a large-scale database, so the cost of finding the strong association rules is very large.

But in practical application, the database is not static, it will be changing constantly with increase that the data are written down, and according to the needs of

user , they often require the association rules to reflect the present state of the database , thus to the decision support, business administration they can offer the theoretical foundation. How the association rules should been effectively updated in the old database become newer problem in every perfect data mining system. What is called, it can find new knowledge in time not only can utilize existing knowledge but also to mean, but not operate the original algorithm of association rules in the database after being newer again. In this dissertation under the condition that consider that mini supports degree not to be changed, the association rules are updated high-efficiently when the trade records are added. It puts forward a kind of new incremental updating algorithm EUFIA, this algorithm mainly tries every possible means to use priori knowledge, reduce the quantity of the new cand idate sets, improve the efficiency of the algorithm at the same time. It is indicated through the repeated experiment that this algorithm has the advantages: it need not rescan the original database to discover newly generated local frequent itemsets and can discover newly generated frequent itemsets more efficiently and need rescan the original database only once to get overall frequent itemsets in the final database if necessary. So EUFIA is a high-efficient algorithm with actual using value.

3. Obviously a large number of mined association rules are not very easy todeal with and store, certainly it is difficult that a person uses, digests andunderstands. So, "the rule explodes ", a problem is becoming serious in the data mining system. This dissertation introduces some basic research work of the storing method of association rules following the international prevailing. In addition, in order to prevent the association rules from storing in the redundant rule while storing, it has been proposed: If a rule can be obtained from a simple rules set, then this rule is a redundant rule and needn't store. However, instead of relevant algorithms, the author is still in studying further.

The research work of this thesis offers certain theoretical foundation and methods effectively to the existing association rules for further research and the association rules effectively being updated.

Key words: data mining; association rules; incremental updating; strong frequent itemsets; inferior frequent itemsets; weak frequent itemsets

目　录

Contents

第一章　绪论

随着计算机、网络和通讯等信息技术的高速发展，信息处理在整个社会规模上迅速产业化，而这种产业化在技术上就表现为整个社会规模的大规模数据操作的产业化,这其中又包含了数据产生、采集、传输、检索及其分析综合等等环节。近些年来,商务贸易电子化,企业和政府事务电子化的迅速普及都产生了大规模的数据源，同时日益增长的科学计算和大规模的工业生产过程也提供了海量数据；而日益成熟的数据库系统和数据库管理系统都为这些海量数据的存储和管理提供了技术保证；另一方面，计算机网络技术的长足进步和规模的爆炸性增长，则为数据的传输和远程交互提供了技术手段，特别是国际互联网更是将全球的信息源纳入了一个共同的数据库系统之中。这些都表明人们生成、采集和传输数据的能力都有了巨大增长，为步入信息时代奠定了基础。在这些能力迅速提高的同时,我们看到数据操纵中的一个重要环节：信息提取及其相关处理技术却相对地大大落后了。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。因此，一门新兴的的自动信息提取技术：数据采掘和知识发现,应运而生并得到迅速发展[1][2]。它的出现为自动和智能地把海量的数据转化成有用的信息和知识提供了手段。

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 5OGb 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。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（Data Mining）便应运而生了。

1.1 KDD 技术

1.1.1　KDD的定义

KDD 是从大量数据中提取出可信的、新颖的、有效的并能被人理解的模式的处理过程，这种处理过程是一种高级的处理过程[1][2]。

数据：数据是指一个有关事实 F 的集合（如学生档案数据库中有关学生基本情况的各条记录），它是用来描述事物有关方面的信息，一般来说这些数据都是准确无误的。

模式：对于集合 F 中的数据，我们可以用语言 L 来描述其中数据的特性。表达式 E? L,E 所描述的数据是集合 F 的一个子集 FE。只有当表达式 E 比列举所有 FE 中元素的描述方法更为简单时，我们才可称之为模式。如："如果成绩在 81-90 之间，则成绩优良"可称为一个模式，而"如果成绩为 81、82、83、84、85、86、87、88、89 或 90，则成绩优良"就不能称之为一个模式。

处理过程：KDD 是一个多步骤的处理过程，包括数据预处理、模式提取、知识评估及过程优化。我们说这个过程是非繁琐的，主要是指这个处理过程的大部分阶段是系统自动进行的而无需人工干涉。

可信：通过 KDD 从当前数据所发现的模式必须有一定的正确程度，否则 KDD 就毫无作用。可以通过新增数据来检验模式的正确性，我们用 c 表示模式 E 的可信度，c = C(E,F)

新颖：经过 KDD 提取出的模式必须是新颖的，至少对系统来说应该如此。模式是否新颖可以通过两个途径来衡量：其一是得到的数据，通过对比当前得到的数据和以前的数据或期望得到的数据之间的比较来判断该模式的新颖程度；其二是通过其内部所包含的知识，通过对比发现的模式与已有的模式的关系来判断。通常我们可以用一个函数来表示模式的新颖程度 N(E,F)，该函数的返回值是逻辑值或是对模式 E 的新颖程度的一个判断数值。

潜在作用：提取出的模式应该是有意义的，这可以通过某些函数的值来衡量。用 u 表示模式 E 的有作用程度，u = U(E,F)。

可被人理解：KDD 的一个目标就是将数据库中隐含的模式以容易被人理解的形式表现出来，从而帮助人们更好地了解数据库中所包含的信息。当然一个模式是否容易被人理解，这本身就很难衡量，比较常用的方法是对其简单程度进行衡量。我们假定模式 E 的简单度（可理解度）s 可用函数 S(E,F)来衡量。

1.1.2 KDD的过程

KDD 处理过程：

知识发现包含如下几个步骤[3]：

- 数据清理：去处噪音和不相关的数据。
- 数据合并：不同的数据源合并成一个。
- 数据筛选：提取出与挖掘任务相关的数据。
- 数据转换：把数据转化合并成正确的格式，用于数据挖掘。
- 数据挖掘：选用选定的知识发现算法，从数据中提取出用户所需要的知识，这些知识可以用一种特定的方式来表示或使用一些常用的表示方式，来产生规则。
- 模式解释：对发现的模式进行解释，在此过程中为了取得更为有效的知识，可能会返回前面的某个处理步骤反复提取，从而提取更有效的知识。
- 知识输出：将发现的知识以用户能理解的方式呈现给用户。这期间也包含对知识一致性的检查，以确定本次发现的知识与以前所发现的知识是否相抵触。

1.1.3　KDD的研究历史与现状

一、 研究历史

从数据库中发现知识（KDD）一词首次出现在 1989 年举行的第十一届国际联合人工智能学术会议上。到目前为止，由美国人工智能协会主办的 KDD 国际研讨会已经召开了 8 次，规模由原来的专题讨论会发展到国际学术大会（见表 1），研究重点也逐渐从发现方法转向系统应用，注重多种发现策略和技术的集成，以及多种学科之间的相互渗透。1999 年，亚太地区在北京召开的第三届 PAKDD 会议收到 158 篇论文，空前热烈。IEEE 的 Knowledge and Data Engineering 会刊率先在 1993 年出版了 KDD 技术专刊。并行计算、计算机网络和信息工程等其他领域的国际学会、学刊也把数据挖掘和知识发现列为专题和专刊讨论，甚至到了脍炙人口的程度。

二、 出版物及工具

此外，在 Internet 上还有不少 KDD 电子出版物，其中以半月刊 Knowledge Discovery Nuggets 最为权威（http://www.kdnuggets.com/subscribe.html）。在网上还有许多自由论坛，如 DM Email Club 等。至于 DMKD 书籍，可以在任意一家计算机书店找到十多本。目前，世界上比较有影响的典型数据挖掘系统有：SAS 公司的 Enterprise Miner、IBM 公司的 Intelligent Miner、SGI 公司的 SetMiner、SPSS 公司的 Clementine、Sybase 公司的 Warehouse Studio、RuleQuest Research 公司的 See5、还有 CoverStory、EXPLORA、Knowledge Discovery Workbench、DBMiner、Quest 等。读者可以访问 http://www.datamininglab.com.网站，该网站提供了许多数据挖掘系统和工具的性能测试报告。

三、 国内现状

与国外相比，国内对 DMKD 的研究稍晚，没有形成整体力量。1993 年国家自然科学基金首次支持我们对该领域的研究项目。目前，国内的许多科研单位和高等院校竞相开展知识发现的基础理论及其应用研究，这些单位包括清华大学、中科院计算技术研究所、空军第三研究所、海军装备论证中心等。其中，北京系统工程研究所对模糊方法在知识发现中的应用进行了较深入的研究，北京大学也在开展对数据立方体代数的研究，华中理工大学、复旦大学、浙江大学、中国科技大学、中科院数学研究所、吉林大学等单位开展了对关联规则开采算法的优化和改造；南京大学、四川联合大学和上海交通大学等单位探讨、研究了非结构化数据的知识发现以及 Web 数据挖掘。

1.2 数据挖掘的定义

一、 技术上的定义及含义

数据挖掘（Data Mining）就是从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中，提取隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识的过程[2][3][4]。

与数据挖掘相近的同义词有数据融合、数据分析和决策支持等。这个定义包