学校编码：10384　　分类号_____密级_____
学　　号：9928006　　UDC _____

学　位　论　文

数据挖掘在市场营销中的应用

陈锦秀

指 导 教 师：　叶仰明　教授
申请学位类别：　硕　　士
专 业 名 称：　计 算 机 应 用
论文提交日期：　2002 年 5 月　日
论文答辩日期：　2002 年　月　日
学位授予单位：　厦　门　大　学
学位授予日期：　2002 年　月　日

答辩委员会主席：_____
评　　阅　　人：_____

二〇〇二年五月

摘　要

本文主要研究数据挖掘在市场营销中的应用，目标是为了满足用户需求，自动处理大量的原始交易数据，从中识别重要和有意义的关联规则。将关联规则挖掘应用于市场营销有助于识别顾客购买行为，发现顾客购买模式和趋势，改进服务质量，取得更好的顾客保持力和满意程度，提高货品销量比率，设计更好的货品运输与分销策略，减少商业成本。因而将数据挖掘技术应用于市场营销领域中具有重要的意义。

超量数据是数据挖掘中经常遇到的技术难题。为能从大量数据（特别是大型数据库）中有效地抽取信息，数据挖掘算法必须是高效的。本课题组的研究重点正是针对关联规则的一序列挖掘和更新算法进行探讨，并将其应用到市场营销中。努力实现一个具有实际应用价值的数据挖掘系统。

本文所提出的挖掘及更新算法已经在模拟交易数据库中具体地实现，实验表明这些算法都是有效的。

本书共分为如下五章：

第一章是绪论。概要介绍了数据挖掘的重要性及数据挖掘的过程和方法，从中引出了数据挖掘在市场营销中的应用，最后提出了本文的主要工作要点。

第二章提出基于时间窗口的经常性周期关联规则的增量式更新算法，该算法是针对在给定的最小支持度和最小置信度下，当一个新的事务数据集 db 添加到旧的事务数据库 DB 中时，如何更高效地更新 db∪DB 中的经常性周期关联规则的问题而提出的。并与原有的直接挖掘算法的性能效率进行了比较。

第三章提出了周期性广义序贯模式的交互式更新算法，该算法解决了给定事务数据库 DB，在最小支持度和最小经常性信度发生变化时，如何更高效地更新数据库 DB 中的周期性广义序贯模式的问题。

第四章将关联规则挖掘算法应用于市场营销理论，将市场营销理论中三个基本要素即最近购买时间、购买频率和利润全部考虑进来进行关联规则的挖掘，提出有实际应用价值的挖掘算法及更新算法；

第五章总结了本文的主要工作，并提出进一步工作的设想。

关键词：数据挖掘、经常性周期关联规则、周期性广义序贯模式、基于时间窗口、增量式更新、时间加权、利润加权

Abstract

This thesis presents a description of data mining applied in the marketing. It aims at assisting humans in extracting useful information (knowledge) from the rapidly growing volumes of data. Data mining and knowledge discovery techniques are more important to understand user behavior better, to improve the service provided, and to increase the business opportunities. In response to such a demand, it is valuable to research how to apply data mining techniques to marketing.

With the increasing of sales transactions, the database is dramatically large. In order to mine valuable information in the vast database, the algorithms of data mining should be efficient. So our research group mainly studies how to improve to mine and update association rules and how to apply these algorithms to marketing. We are making efforts to implement a data mining application system.

The mining and updating algorithms in this thesis have been implemented in simulated databases. The experiment result shows that the presented algorithm is of great efficiency.

The whole thesis is made up of the following five chapters:

Chapter One is the preface. It shows the importance of data mining and gives the brief introduction of the process and means of data mining. And then it shows the data mining's application in marketing. At last this chapter gives the topic of the thesis.

Chapter Two puts forward an incremental algorithm based on time_window for updating frequent cyclic association rules. The algorithm supposes the minimal support and conference degree is changeless. It aims at how to update frequent cyclic association rules through re-using the results acquired in the previous process when new data adds. Then we compare the updating algorithm with the direct mining algorithm on efficiency.

Chapter Three proposes an interactive updating technique in order to deal with the maintenance of discovered cyclic generalized sequential patterns resulted from the change of minimal support and minimum frequent confidence. The main idea is to reutilize the results acquired in process with the old minimum support and the old minimum frequent confidence.

Chapter Four describes the application of data mining in marketing, which concerns not only purchasing frequency, but also recent purchasing time and profit. Then considering these factors, we bring forward valuable algorithms to mine and update association rules.

Chapter Five is the concluding remark. It summarizes the work of this thesis and gives some ideas of further research.

Key Words data mining, frequent cyclic association rules, cyclic generalized sequential patterns, time_window, incremental updating, time-weighted, profit-weighted

目　录

第一章 绪论

1．1 引言

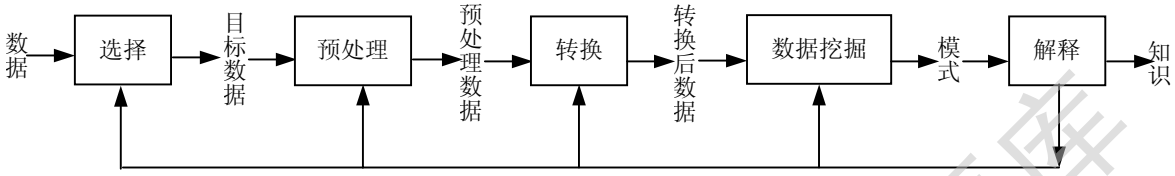
随着个人计算机的普及和 Internet 的迅速发展，社会的信息化程度越来越高。人们在日常生活中每天都要面对大量的信息数据，所以经常会遇到这样的情况:超市的经营者希望能从过去几年的销售记录中分析出顾客的消费习惯和行为，以便及时变换营销策略；保险公司想知道购买保险的客户一般具有哪些特征；医学研究人员希望从已有的成千上万病历中找出患某种疾病的病人的共同特征，从而为治愈这种疾病提供一些帮助等等。对于这些问题，现有信息管理系统中的数据分析工具无法给出解决办法。虽然数据库管理系统（DBMS）可以高效实现数据录入、检索和维护等管理功能，但不能发现数据中的关联和规则，也不能根据现有的数据预测未来的发展趋势。所以，迫切需要一种能够智能地自动地把数据转换成有用信息和知识的技术和工具。需求是发展之母，数据库管理系统和人工智能中机器学习两种技术的发展和结合，促成了在数据库中发现知识（KDD）这一新技术的诞生。1989 年 8 月，在美国底特律召开的第 11 届人工智能联合会议的专题讨论会上，首次提出了 KDD。它是一门交叉性学科，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、专家系统等领域，内涵极为广泛，理论和技术难度很大，从而使针对大型数据库的 KDD 技术一时还难以满足应用需要。于是，1995 年的（美）计算机学会（ACM）会议提出了数据挖掘（data mining）概念，它形象地把大型数据看成是存放有价值信息的矿藏，通过有效的知识发现技术，从中挖掘或开采出有用的信息。

我们知道，原有数据库技术只是将数据有效地组织和存储在数据库中，并对这些数据做一些简单的分析，大量的隐藏在数据内部的有用信息我们无法得到。而机器学习、模式识别、统计学等领域却有大量的提取知识的方法，但没有和实际应用中的海量数据结合起来,很大程度上只是对实验数据或学术研究发挥作用。数据挖掘（Data Mining）从一个新的角度将数据库技术、机器学习、统计学等领域集合起来，从更深层次中发掘存在于数据内部的有效的、新颖的、具有潜在效用的乃至最终可理解的模式。

1．2 什么是数据挖掘

数据挖掘（Data Mining），也叫数据开采，数据采掘等，是按照既定的业务目标从海量数据中提取出潜在、有效并能被人理解的模式的高级处理过程。

也有一些文献把数据挖掘称为知识抽取（knowledge extraction）、数据考古学（data archaeology）、数据捕捞（data dredging），等等。多数人认为数据挖掘是 KDD 过程中的关键步骤（见下图）。



KDD 过程

数据挖掘与传统数据分析工具的主要区别在于它们探索数据关系时所使用的方法。传统数学分析工具使用基于验证的方法，即用户首先对特定的数据关系作出假设，然后使用分析工具去确认或否定这些假设。这种方法的有效性受到许多因素的限制，如提出的问题和预先假设是否合适等。与分析工具相反，数据挖掘使用基于发现的方法，运用模式匹配和其他算法决定数据之间的重要联系。数据挖掘处理的数据规模十分庞大，由于数据变化迅速，因此要求数据挖掘能快速地做出相应反应以随时提供决策支持。此外，数据挖掘所发现的规则是动态的，它只反映了当前状态的数据库具有的规则，随着不断地向数据库中加入新数据，需要随时对其进行更新。为了更好地提高挖掘效率及价值，从什么角度以及采用什么方法来进行数据挖掘也显得很重要。具体来说，有以下几种主要的数据挖掘方法：

（1）、关联规则挖掘

顾名思义，挖掘关联规则就是发现隐藏在大数据集中的关联性或相关性。即给定一组 Item 和一个记录集合，通过分析记录集合，推导出 Item 间的相关性，有一个关联规则的典型例子就是“90%的客户在购买面包的同时也会购买牛奶”，其直观意义为顾客在购买某些商品的时候有多大倾向会购买另外一些商品。

（2）、多层次序贯模式分析

序贯模式分析和关联规则分析法相似，其目的也是为了挖掘出数据之间的联系，但序贯模式分析的侧重点在于分析数据间的前后（因果）关系。在医疗保险行业，该方法具有非常好的效果。保险公司利用序贯模式分析法可以预测用户投保后最常采取的医疗措施，从而识别可能的欺诈行为。

（3）、分类分析（Classifiers）

分类分析时首先为每一个记录赋予一个标记（所谓标记是指一组具有不同特征的类别），即按标记分类记录，然后检查这些标定的记录，描述出这些记录的特征。利用它可以分类新记录，实际上它就是一种模式。

举一个简单的例子，信用卡公司的数据库中保持着各持卡人的记录，并根据信誉程度（标记），将持卡人分作三类：良好，普通，较差。这一过程实际就是将持卡人记录标定为三类。分类分析法检查这些记录，然后给出一个对信誉等级的显式描述：

“信誉良好的用户是指那些收入在 25000 以上，年龄在 45 到 55 岁之间，居住在 XYZ 地区附近的人士”。

（4）聚类分析（Clustering）

与分类分析法不同，聚类分析法的输入集是一组未标定的记录，也就是说此时输入的记录还没有被进行任何分类。其目的是根据一定的规则，合理地划分记录集合，并用显式或隐式的方法描述不同的类别。由于聚类分析可以采用不同的算法，所以对于相同的记录集合可能有不同的划分。例如，运用序贯模式的分析方法找出几类重要的用户群，他们具有如下的购物模式：在购买了某些商品后购买微波炉。运用聚类分析法就可以找出具有该购物模式并且尚未购买微波炉的用户，他们，就是市场销售人员所要争取的对象。

（5）、决策树方法

决策树方法是利用信息论中的互信息（信息增益）寻找数据库中具有最大信息量的字段，建立决策树的一个结点，再根据字段的不同取值建立树的分支；在每个分支子集中重复建立树的下层结点和分支的过程，即可建立决策树。国际上最早的、也是最有影响的决策树方法是 Quinlan 提出的 ID3 方法，它对越大的数据库效果越好。

在数据挖掘和知识发现中应用的人工智能技术还有邻近搜索方法（Nearest Neighbor Method）、集合论的粗集方法（Rough Set）、规则推理（Rule Induction）、模糊逻辑（Fuzzy Logic）、遗传算法、公式发现、Bayesian 网络等等。由以上所述可知，数据挖掘的核心技术是人工智能、机器学习、统计等，但它并非多种技术的简单组合，而是一个不可分割的整体，还需要其它技术的支持，才能挖掘出令用户满意的结果。

1．3 数据挖掘的过程

数据库中的数据挖掘是一个多步骤的处理过程，一般分为：

问题定义：了解相关领域的有关情况，熟悉背景知识，弄清用户要求；

数据提取：根据要求从数据库中提取相关的数据；

数据预处理：主要对前一阶段产生的数据进行再加工，检查数据的完整性及数据的一致性，对其中的噪音数据进行处理，对丢失的数据进行填补；

知识提取：运用选定的知识发现算法，从数据库中提取用户所需要的知识，

这些知识可以用一种特定的方式表示或使用一些常用的表示方式；

知识评估：将发现的知识以用户能理解的方式呈现，如某种规则，再根据实际情况对知识发现过程中的具体处理阶段进行优化，直到满足用户要求。

1．4 数据挖掘的应用

数据挖掘是一个年轻而又非常活跃的研究领域，目前面临的问题，除了基础理论和技术方面的外，更重要的是开发和应用。数据挖掘被广泛地应用于市场营销、银行业、生产销售、制造业、经济业、保险业、医药业、电信业等各个应用方向，其中零售业是数据挖掘的主要应用领域，这是因为零售业积累了大量的销售数据，顾客购买历史记录，货物进出，消费与服务记录，等等。其数据量在不断地迅速膨胀，零售数据为数据挖掘提供了丰富的资源。零售数据挖掘可有助于识别顾客购买行为，发现顾客购买模式和趋势，改进服务质量，取得更好的顾客保持率和满意程度，提高货品销量比率，设计更好的货品运输与分销策略，减少商业成本。因而将数据挖掘技术应用于市场营销领域中具有重要的意义。本课题组的研究重点正是针对关联规则的挖掘和在实际商业数据库中的应用进行探讨。努力实现一个具有实际应用价值的数据挖掘系统。

1．5 本文的主要工作

本文主要研究数据挖掘在市场营销中的应用，目标是为了满足用户需求，自动处理大量的原始数据，从中识别重要的和有意义的关联规则。在课题的研究过程中，主要针对以下几个方面进行探讨：

1、提出基于时间窗口的经常性周期关联规则的增量式更新算法，并与原有的直接挖掘算法的性能效率比较（只考虑购买频率）；

2、提出周期性广义序贯模式的发现及交互式更新算法，并将两者效率进行比较；

3、将关联规则挖掘算法应用于市场营销理论，将市场营销理论中三个基本要素即最近购买时间、购买频率和利润全部考虑进来进行关联规则的挖掘，提出有实际应用价值的挖掘算法；

4、将上述的 1 和 2 中的更新算法的思想应用于时间利润约束加权关联规则的挖掘中，对其进行增量式更新。

5、通过对关联规则挖掘系列算法的研究，将其应用到实际的市场营销中，具体实现一个实用的挖掘系统，其中不仅包括对一般的关联规则的挖掘和更新维护，也包括对多层及经常性周期关联规则的发现，使其有一定的理论价值和实际应用价值。

另外对广义序贯模式和周期性广义序贯模式的增量式更新算法我们也均已完成，结果也已经发表或被录用（参见参考文献[19]、[20]，本人均为第一作者）。所以虽然这些工作也是在研究生期间所做的，但因篇幅所限本文就不介绍。

第二章 关联规则的挖掘及更新

在事务数据库中采掘关联规则是数据采掘领域中的一个非常重要的研究课题。它是由 Agrawal R.等人首先提出的，目的是要在交易数据库中发现各项目之间的关系。有一个关联规则的例子就是“90%的客户在购买面包的同时也会购买牛奶”，其直观意义为顾客在购买某些商品的时候有多大倾向会购买另外一些商品。关联规则的应用主要包括顾客购物分析、目录设计、商品广告邮寄分析、追加销售、仓储规划、网络故障分析等。

关联规则的采掘问题可形式化描述如下：

设 $I=\{i_1,i_2,\ldots,i_m\}$是由 m 个不同的项目组成的集合。给定一个事务数据库 D，其中的每一个事务 T 是 I 中一组项目的集合，即 $T\subseteq I$。每一个事务有唯一的一个标识符，称作 TID。设 A 是一个项集，事务 T 包含 A 当且仅当 $A\subseteq T$。关联规则是形如 $A\subseteq B$ 的蕴涵式，其中 $A\subset I$，$B\subset I$，并且 $A\cap B=\Phi$。规则 $A\Rightarrow B$ 在事务集 D 中成立的条件是：①它具有支持度 s，即事务数据库 D 中至少有 s%的事务包含 $A\cup B$（即 A 和 B 二者）。②具有置信度 c。即在事务数据库 D 中包含 A 的事务至少有 c%同时也包含 B。可用概率如下表示：

$$support(A\Rightarrow B)=P(A\cup B)$$

$$confidence(A\Rightarrow B)=P(B|A)$$

关联规则的采掘问题就是在事务数据库D中找出同时满足用户指定的最小支持度阈值(min_sup)和最小置信度阈值(min_conf)的关联规则。

项目的集合称为项集。包含 k 个项目的项集称为 k-项集。集合{牛奶，面包}是一个 2-项集。项集的出现频率是包含项集的事务数，简称为项集的频率、支持计数或计数。如果项集的出现频率大于或等于 min_sup 与 D 中事务总数的乘积，则项集满足最小支持度 min_sup。如果项集满足最小支持度，则称它为频繁项目集（frequent itemset），有时也称之为大项目集。频繁 k-项目集的集合通常记作 L_k。

2.1 采掘关联规则的一般步骤

Agrawal R.等人在首先提出了关联规则的采掘问题并给出解决此问题最原始的算法 AIS 之后，该问题得到了国际人工智能和数据库等领域学者的密切关注，提出了多种算法。所有的采掘算法不论它是采用什么数据结构，其复杂程度、效率如何，它们都可以分为如下几个步骤：

① 预处理与采掘任务有关的数据。根据具体问题的要求对数据库进行相应的操作，从而构成规格化的数据库 D。

② 针对 D，求出所有满足最小支持度的项集，即大项集。由于一般情况下我们所面临的数据库都比较大，所以此步是算法的核心。

③ 生成满足最小置信度的规则，形成规则集 R。

④ 解释并输出 R。

2.1.1 Apriori 算法：使用候选项集找频繁项目集

我们简要描述 Agrawal R.等人提出的寻找所有频繁属性序列集的 Apriori 算法，它是一种最有影响的挖掘布尔关联规则频繁项集的挖掘算法，它利用性质：频繁项集的所有非空子集都必须是频繁的。在第 k 次迭代后（k>1），它根据频繁 k-项集，形成频繁(k+1)-项集候选，并扫描数据库一次，找出完整的频繁(k+1)-项集。

令 Fre[k]为频繁 K-项集的集合，而 C[k]为候选 k-项集（即可能的频繁项集）的集合。Apirori 算法需对数据库作多次遍历，每次遍历均由两个阶段构成。第一、利用上次（第 k-1 次）遍历所得到的频繁(k-1)-项集 Fre[k-1]生成候选 k-项集 C[k]。候选生成算法 Apriori-gen 保证 C[k]是所有频繁 k-项集的超集。第二、对数据库作一次遍历，对其中的每个元组确定它支持 C[k]中的哪些候选，并在相应候选的 count 域中累计支持数。遍历结束后，检查候选集 C[k]，确定哪些候选是频繁的，从而构成频繁 k-项集 Fre[k]。该算法反复进行，直到 Fre[k]为空时为止。

已知频繁(k-1)-项集 Fre[k-1]，候选生成算法 Apriori-gen 返回所有频繁 k-项集的超集。该候选生成的算法思想是基于这样一个性质：频繁项集的子集均是频繁的。该候选生成算法 Apriori-gen 也分为如下两步：

（1）链接（join），即 Fre[k-1]与自己链接生成 C[k]。

insert into C[k]

select $p.item_1$, $p.item_2$, ……, $p.item_{k-1}$, $q.item_{k-1}$

From $p \in Fre[k-1]$, $q \in Fre[k-1]$

Where $p.item_1=q.item_1, p.item_2=q.item_2, \ldots, p.item_{k-2}=q.item_{k-2}$,

$p.item_{k-1}<q.item_{k-1}$

（2）修剪（prune），删除 C[k]中的任何一个候选 c，如果 c 中存在一个长度 k-1 属性子集不属于 Fre[k-1]，即不是频繁的。

Apriori 具体算法如下：

Algorithm Apriori

Input：事务数据库 D；最小支持度阈值 min_sup；

Output：D 中的频繁项目集 Fre。

Begin

Fre_1={frequent 1-itemsets};

For ($k=2; Fre_{k-1} \neq \Phi; k++$)

{ C_k=apriori_gen(Fre_{k-1},min_sup);
For each transaction t∈D
{ C_t=subset(C_k,t);
For each candidate c∈C_t
c.count++; }
Fre_k={c∈C_k |c.count≥min_sup}
}
Fre=∪$_k Fre_k$;
End;

例 2.1.1：已知如下表所示的事务数据库 D，数据库中 4 个事务，即|D|=4。

TID	Items
T100	1, 3, 4
T200	2, 3, 5
T300	1, 2, 3, 5
T400	2, 5

设最小支持度计数为 2，则可以根据 Apriori 算法寻找出 D 中的频繁项目集：

C_1

项目集	支持度计数
{1}	2
{2}	3
{3}	3
{4}	1
{5}	3

扫描 D，对每个候选计数

比较候选支持度计数与最小支持度计数

L_1

项目集	支持度计数
{1}	2
{2}	3
{3}	3
{5}	3

由 L_1 产生候选 C_2

并扫描 D，对每个候选计数

C_2

项目集	支持度计数
{1 2}	1
{1 3}	2
{1 5}	1
{2 3}	2
{2 5}	3
{3 5}	2

比较候选支持度计数与最小支持度计数

L_2

项目集	支持度计数
{1 3}	2
{2 3}	2
{2 5}	3
{3 5}	2

由 L_2 产生候选 C_3

并扫描 D，对每个候选计数

C_3

项目集	支持度计数
{2 3 5}	2

比较候选支持度计数与最小支持度计数

L_3

项目集	支持度计数
{2 3 5}	2