**西 安 邮 电 大 学**

**毕 业 设 计（外文翻译）**

题 目： Research and implementation of frequent

itemset mining algorithm

学 院： 计算机学院

专 业： 软件工程

班 级： 软件1401

学生姓名： 代栋

导师姓名： 孟彩霞 职称： 教授

An Improved Apriori Algorithm for Association Rules

## Introduction

Data mining is a subject of the application in the field of endeavors which to provide policy support for the fields based on information analysis. Data mining extract the unknown, implied and useful information and knowledge from a large number of structural and the structure of data. As one of the most popular method of data mining which signify local patterns, association rules provide a very simple and valuable description. Apriori algorithm is the most classic one to found frequent item set [1]. Nowadays, there are a lot of impoved Apriori algorithm.For example,according to the improved Apriori algorithm, which scans the database only once by using arrays to store data and the new algorithm sorts the frequent itemsets from small to large [2]. For instance, Coenen and Leng proposed Apriori-TFP (Total-from-Partial) based on Apriori improvment, which used to P (partial sup-port tree) and T (total support tree) to decrease computer time and storage space [3], and as literature 4 proposed to improve the the way of candidate connection [4], and as others put forward the the improvement of the algorithm [5, 6]. After that improved Apriori algorithm by appied widly in the aspects. For example, Literature 11 the improved algorithm can quickly find the corelation of the stock plate, it has a certain guiding role in the stock market analysis and investment decisions [7]. As literature 12 popose a model of intrusion detection system based on class-association rule to improve the detection rate of intrusion detection system, experiments show that the proposed method could detect intrusions efficiently in the network [8], and there is a aviation safety reports analysis application and web recommender system in another literature [9, 10].

Of course, improved Apriori althorigthm is also put into use in teaching, as literature propose the measure quality of teaching [11]. But there is less in the curriculum. In the teaching process, there should have the small tests to control the situation of the students. Therefore, there is large number of students in the usual results of the feedback at every semester or the end of the year. However, some important potential value has not been utilized. For example, there are close touch with comparison in the final stages of review which can be connected to a review class in a more easy to learn knowledge framework to review. This can also guide the next instruction, it is likely to cause other links between the teaching processes, and we should pay attention to the strengthening of knowledge leading. Comprehensive, we hope to find a course in connection with the close to where it is more and more conducive of the structures to construct the course of the intellectual system. Therefore, we can dig up by the rules of the largest number of project assembly to complete. But the traditional rules which are iterated to produce the maximum number of project assembly are not entirely suitable. So in accordance with computer programming charateristic this paper achieve knowledge system establish based on analysis of the classical Apriori algorithm in mining association rule. The improved algorithm adopts matrix to express database, just scans database once, cancels a great number of linking operations in finding frequent item sets dimension by dimension but finds out the highest dimension frequent item sets directly from high dimension itemsets, the algorithm is improved efficiently.

## 2. Research Theory

In 1993, Agrawal and several people are the first carried up with to excavate the database transaction set of rules and associated issues to find between different goods of the rules, then they put forward Apriori algorithm [12]. These basic concepts are below.

Definition 1 is item: a field in a transaction database, it usually uses the lowercase im as the mark. For example, basic knowledge is the Item in the C++ Programming.

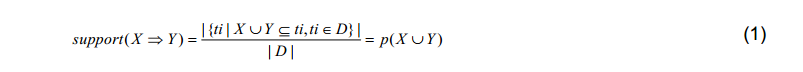
Definition 2 is transaction: Correspond to a record in the course results database, it usually uses the lowercase ti as the mark, ti= {il，， ， i2 … ip}. Every transaction has a unique identifier that calls TID. There are a transaction such as {basic knowledge of C++, basic knowledge of object-oriented, destructor and constructor, and so on }.

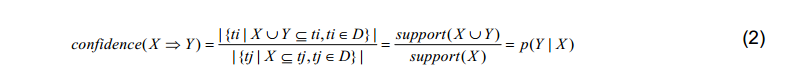
Definition 3 is Itemset: The database for all the items collection mark the upper I, the any subset of x is called the itemset of D, it means that equivalently the set of i1，， ， i2 … im.

Definition 4 is dimensionality of itemset: the number of items which Itemset contains, it will be called K-Itemset If the number is K.

Definition 5 is association Rules: if both X and Y are itemsets, and X∩Y=φ, then there is association rules X=>Y, and its meaning that X has also led to y, X is premise and Y is conclusion.

Definition 6 is support: there is X=>Y among X belong to I and Y also belong to I, at the same time X∩Y=φ, the support of Rule is support (X=>Y). It also means on the deal set at once the percentage of X and Y, such as formula (1).

 Definition 7 is confidence: there is X=>Y among X belong to I and Y also belong to I, at the same time X∩Y=φ, the confidence (X=>Y) of Rule is the proportion of transactions number of include of X,Y and X. such as formula (2).

 Definition 8 is the minimum support and the minimum confidence of association rules, The former, namely sup\_min, is measure the lowest importance of rules requirement and the latter, namely confmin, is he minimum reliability of rules requirement.

Definition 9 is strong association rules, if rules is X=>Y, and meet support (X=>Y)>=sup\_min and confidence (X=>Y)>= confmin , then this rules is strong association about X=>Y,otherwise is weak association.

According to the classical Apriori algorithm, There are two most important quality for associated rules: The first one is that the set of frequent Itemsets is frequent Itemsets, the other one is that the superset of unfrequent Itemsets is unfrequent Itemsets. In 1994, Agrawal is the first carried up with the famous Apriori algorithms.

## 3. Improvment Results and Analysis

The classical Apriori algorithm used k of frequently items to generate K+1 candidate in connections through cutting to have frequent and a set of items, until not frequent items. Apriori algorithm is generated every length of the assembly that will scan a database, and to be a candidate, so the number of scans the database are decided by the most frequent set of items. So the paper improves a new algorithm against the problem.

## 3.1 Improvment Method

There is a great of research about Apriori algorithm in improvement. To sum up, if the number of K frequent itemset, it is LK <=K,this set is the largest frequent itemset which less or equal to K, it is support the maximum items number all the transaction, so in the improvment algorithm the largest frequent itemset k, according to above what the conclusions to determine, is used, through only pay attention to the affairs of item number which greater or equal to k to find the largest frequent Lk. If not find frequent itemset K then k-1, k is the hypothetical item number of the largest frequent itemset, and then repeated according to the above method. There is an analog transactional database as Table 1, each of the transaction T is a student record, from I1 to I7 represent different knowledge, sup\_min is 4.

When Scanning the database record C[n] there is C[1] is 1, C[2] is 5, C[3] is 8, C[4] is 4, C[5] is 2, then set each of the transaction Flag1 and change data storage in order to item as keyword, whether or not transaction include of this item by 0 or Flagi. If every transaction the maximum length is n, then use n binary to represent its transaction set, this n system storage by item as keyword shown in the following Table 2.

If set C[4]+C[5]> sup\_min, then presuppose the maximum frequent itemset number is 4. Beacause sup\_min is 4, the Flag 2 of l1 is 3, it meaning there is three transactions in the I1, so I can’t be frequent itemset, of course, it isn’t the largest one. Now it should only to judge I2, I4, I5,I6, I7 after delete I1 and its data record. Presuppose that the maximum frequent itemset number k is 4, then if the transaction item is equal or greater k, then change transaction value to 1, or 0.

There are olny the transaction of I2，I4，I5，I6 which sup\_min is satisfactory, then solve the mixed set of I2，I4，I5，I6. Now that the data as binary to store, so it will be finished with∩operation of binary. If number of elements is equal or greater to minimum support, then deduce the set of I2，I4，I5， I6, it is the largest frequent itemset, and then In the set of 0 0 0 0 1 0 0 0 0 1 0 0 0 0 1 0 1 there are involving four 1, it’s conform to the minimum support. Usually, with the assumption that the maximum number of items frequent itemsets k equals the number of items. Typically, in the solution process, such as seeking I2, I4 intersection set. Their transaction∩2,4, if the number "1" is less than 4, that does not meet the minimum support required, then the number of items for consideration when less than k, there is no maximum frequent itemsets, then do not need to continue to seek common ground. If you do not find 4 frequent itemsets, then it will assume the item number of frequent item sets the maximum number of items k minus 1, in accordance with the above method to obtain the same. According to the nature of Apriori algorithm, a subset of frequent itemset are frequent itemset, so the maximum frequent itemset: I2，I4，I5，I6 to produce some two frequent itemset: I2, I4 and I2, I5 and I2, I6 and I4, I5 and I4, I6 and I5, I6, there are there frequent itemset: I2, I4, I5 and so on.

## 3.2 Improved Algorithm Efficient Analysis

According to the method of the above improvement Arpriori algorithm, the experimental simulation data from C++ examination in the computer programming. Improve the efficiency of the algorithm compared before and after. As shown in the Figure 1, this algorithm is the number of items in the mining maximum frequent set, no access to all databases, but only concerned about the affairs of the number of items that is greater than the number of items which is equal to the assumed maximum frequent item sets the number of items. Meanwhile, not the transaction as the keyword, but the project as a keyword on data storage, and simplified step by step, finally stored in binary form, which can not only save storage space, but also through the use of binary "and" obtain the maximum frequent item sets operation simplifies the algorithm, thus enhancing the efficiency of the algorithm.

## 3.3 Improved Algorithm Application Found

In the C++ programming examination in the computer by the VC. There are 10 knowledges, respectively: c + + preliminary knowledge, object-oriented basic knowledge of the constructor and destructor, the class static members, class friend, class templates, operator overloading, inheritance and derived more state and virtual functions, input and output streams. Then each student has a total score, in accordance with the Arabic numerals 1-11 sequential number. Each knowledge point first determine the passing score, if the student has mastered a certain knowledge point, then the knowledge of the corresponding point in the student affairs records may correspond to the database as follows.

# 一种改进的关联规则Apriori算法

## 1、引言

数据挖掘是面向信息分析领域的政策支持领域的一个应用课题。数据挖掘从大量的结构和数据结构中提取未知的、隐含的和有用的信息和知识。关联规则作为表示局部模式的最常用的数据挖掘方法之一，提供了一种非常简单和有价值的描述。Apriori算法是发现频繁项集的最经典算法[1]。

目前已有很多Apriori算法，如改进的Apriori算法，只使用数组来扫描数据库来存储数据，新算法将频繁项集从小到大排序[2 ]。例如，Coenen和Leng提出了基于先验改进的Apriori TFP（总计从部分），用于P（部分SUP-端口树）和T（总支持树）以减少计算机时间和存储空间[3 ]，并且如文献4所提出的改进候选连接的方式[4 ]，AND等人提出了算法的改进〔5, 6〕。改进后的Apriori算法在各个方面都得到了广泛的应用。例如，文献11改进后的算法能够快速找到股票板块的相关性，对股票市场分析和投资决策具有一定的指导作用[ 7 ]。由于文献12 POPSE是一种基于类关联规则的入侵检测系统模型，提高了入侵检测系统的检测率，实验表明，该方法能够有效地检测网络中的入侵，并有一个航空安全报告分析。S应用和Web推荐系统在另一篇文献〔9, 10〕中的应用。

当然，改进的Apriori算法也在教学中投入使用，因为文献提出了教学的测量质量〔11〕。但是课程中的内容较少。在教学过程中，应该有一些小的测试来控制学生的情况。因此，在每学期或年末的反馈中，通常会有大量的学生出现反馈。然而，一些重要的潜在价值尚未被利用。例如，在复习的最后阶段有比较密切的联系，可以在一个更容易学习的知识框架中与复习课联系起来复习。这也可以指导下一步的教学，很可能会导致教学过程之间的其他环节，而我们应该注意加强知识的引导。综上所述，我们希望能找到一个与之密切相关的课程，使之越来越有利于结构体系的构建。因此，我们可以按规则挖掘出最大数量的工程装配来完成。但是迭代生成最大装配数量的传统规则并不完全适用。因此，根据计算机编程特点，本文在分析关联规则挖掘的经典Apriori算法的基础上，实现了知识系统的建立。改进的算法采用矩阵表示数据库，只扫描一次数据库，取消了大量的频繁项集维数的链接操作，但直接从高维项集中找出最大维度频繁项集，算法是改进的。ED有效。

## 2、理论研究

在1993，阿格拉沃尔和几个人最先携带了挖掘数据库事务集的规则和相关问题来查找不同商品之间的规则，然后提出了Apriori算法[12 ]。这些基本概念如下。

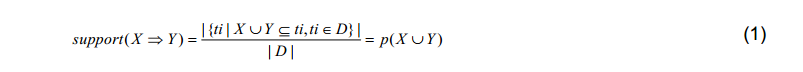
定义1是项：事务数据库中的字段，它通常使用小写的IM作为标记。例如，基本的知识是在C++编程项目。

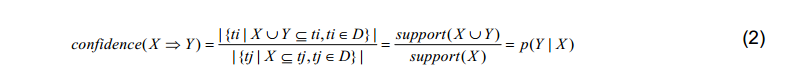
定义2是事务：对应于课程结果数据库中的一个记录，它通常使用小写的TI作为标记，Ti＝{IL，I2，IP}。每个事务都有一个调用TID的唯一标识符。有一个交易如{ C++基础知识、面向对象的基本知识，析构函数与构造函数，等等}。

定义3是项集：所有项目集合的数据库标记上i，x的任何子集被称为D的项集，这意味着I1、I2、I2…IM的集合。

定义4是项集的维数：项目集包含的项的数量，如果数字是k，则称为k项集，

定义5是关联规则：如果x和y都是项集，x y＝y，则有关联规则x=＞y，并且它的意思是x也导致y，x是前提和Y是结论。

定义6是支持：X中属于x=＞y，属于I，Y也属于I，同时x y＝ω，规则的支持是支持（x=＞y）。这也意味着立即处理X和Y的百分比，如公式（1）。定义7是置信度：x属于y，x属于y，y也是i，同时x（y＝y），规则的置信度（x=＞y）是x、y、x的包含数的比例，如公式（2）。

定义8是关联规则的最小支持度和最小置信度，前者是衡量规则需求的最低重要性，后者是规则要求的最小可信度。

定义9是强关联规则，如果规则是x=＞y，并且满足支持（x=＞y）＞SUP\_min和置信度（x=＞y）＞COMMIN，则该规则是关于x=＞y的强关联，否则是弱关联。

根据经典的Apriori算法，关联规则有两个最重要的质量：第一个是频繁项集的集合是频繁项集，另一个是非频繁项集的超集是非频繁项集。1994，阿格拉沃尔是第一个用著名的Apriori算法进行的算法。

## 3、改进结果与分析

经典的Apriori算法使用频繁项的K来生成K+ 1候选在连接中通过切割来具有频繁和一组项，直到不频繁项。Apriori算法生成的每一个组件的长度将扫描数据库，并作为候选，因此扫描数据库的数量由最频繁的项集决定。因此，本文针对这一问题，提出了一种新的算法。

## 3.1、改进方法

Apriori算法在改进方面有很大的研究。综上所述，如果k个频繁项集的数目是LK <k，该集合是最大或最小的频繁项集，它支持所有事务的最大项目数，因此在改进算法中最大频繁项集k，根据以上结论得到威慑。矿井，是通过，只注意事项数量的事务，或大于或等于K，以找到最大的频繁LK。如果没有找到频繁项集k，则k-1，k是最大频繁项集的假设项数，然后根据上述方法重复。有一个模拟事务数据库作为表1，每个事务T是一个学生记录，从I1到I7代表不同的知识，SUP\_min是4。

当扫描数据库记录C[N]时，C[1]为1，C[2]为5，C[3]为8，C[4]为4，C[5]为2，然后设置事务FLAG1中的每一个，并将数据存储改变为项目为关键字，无论该项的事务是否包含0或FRAGI。如果每个事务的最大长度为N，则使用N个二进制表示其事务集，该N个系统按项存储为下表2所示的关键字。

如果集合C[4]＋C[5]>SUP\_min，则预设最大频繁项集数为4。因为SUP\_min是4，L1的标志2是3，这意味着在I1中有三个事务，所以我不能频繁的项目集，当然，它不是最大的项目集。现在只能判断I2，I4，I5，I6、I7在删除I1后及其数据记录。假定最大频繁项集数k为4，则如果事务项等于或大于k，则将事务值更改为1或0。

I2、I4、I5、I6的事务处理是满意的，然后求解I2、I4、I5、I6的混合集。现在把数据作为二进制来存储，这样就完成了相交运算。如果元素的数目等于或大于或小于最小，则推断出I2、I4、I5、I6的集合，它是最大频繁项集，然后在集合中的0个0、0、0、1、0、0、0、0、1、0、0、0、0、γ等，涉及到γ，它符合最小支持度。通常，假设项目频繁项集的最大数目k等于项目的数量。通常，在求解过程中，如求I2、I4相交集。它们的事务2，4，如果数字“1”小于4，则不满足所需的最小支持，那么在考虑小于k的项的数量时，没有最大频繁项集，则不需要继续寻求共同的基础。如果你找不到4个频繁项集，那么它将假定项项的频繁项集最大项数k减1，按照上述方法得到相同。根据Apriori算法的性质，频繁项集的子集是频繁项集，因此最大频繁项集：I2、I4、I5、I6产生两个频繁项集：I2、I4和I2、I5和I2、I6和I4、I5和I4、I6和I5、I6，存在频繁项集：I2、I4、I5等。

## 3.2、改进算法的有效性分析

根据以上改进arpriori算法的方法，从计算机编程C++考试的模拟实验数据。提高了算法前后的效率。如图1所示，该算法是挖掘最大频繁集中的项的数目，不访问所有数据库，而只关心大于等于假定最大频繁项的项的数目的事务的数目设置的数目。项目。同时，不以事务为关键字，而是将项目作为关键字进行数据存储，并逐步简化，最后以二进制形式存储，这不仅可以节省存储空间，而且可以通过使用二进制“和”获取最大频繁项集操作简化ALG。从而提高了算法的效率。

## 3.3、改进算法的应用发现

在C++编程考试计算机VC。有10的知识，分别为：C++的初步知识，面向对象的基本知识，构造函数和析构函数，类的静态成员，类的友元，类模板、运算符重载、继承和派生更多的状态和虚函数、输入输出流。然后每个学生都有一个总分，按照阿拉伯数字1-11顺序编号。每个知识点首先确定通过分数，如果学生掌握了某个知识点，那么学生事务记录中对应点的知识可以对应于数据库。