一种基于关系数据库的频繁项集挖掘算法

王治和

（西北师范大学数学与信息科学学院　兰州 730070）

摘　要　频繁项集的挖掘是数据挖掘中的一个十分重要的组成部分，目前对于事务数据库频繁项集的挖掘算法研究较多。本文根据事务数据库中布尔型频繁项集挖掘的理论和方法，再结合关系数据库的特殊性，利用标准 SQL 语言提出了一种新的在关系数据库中挖掘频繁项集的简易算法。实验证明该算法具有较高的效率。

关键词　数据挖掘，关系数据库，频繁项集，SQL 语言

An Algorithm for Discovering Frequent Itemsets Based on Relational DataBase

WANG Zhi-He

(College of Mathematics and Information Science, Northwest Normal University, Lanzhou 730070)

Abstract　The discovering of frequent itemsets is a most important part of Data Mining. Currently, more research work is done on the algorithm of mining transactional database than on the algorithm of relational database. According to the relational theory and method of Boolean frequent itemsets mining in transactional database, and combining the particularity of mining in relational database, this paper discusses and proposes a new simply algorithm of discovering all the frequent itemsets in relational database by the standard SQL. Experiments prove the algorithm is high effective.

Keywords　Data mining, Relational database, Frequent itemsets, SQL language

1　引言

数据挖掘技术从诞生到现在，其经济价值已经得到了大家的公认，许多大公司已经开始在实际中使用数据挖掘技术[1]。最初的挖掘知识只是针对事务数据库中布尔型关联规则的挖掘，然而时下大量的数据存储在关系数据库中，如何利用关系数据库，并从关系数据库中挖掘知识、得到规则，则成为一个有意义的现实与理论问题。

数据挖掘中最为流行的一项技术就是关联规则(association rule mining ARM)的挖掘，通常我们认为关联规则的挖掘是一个两步的过程：1)找出所有频繁项集；2)由频繁项集产生强关联规则[2]。在这两步中，第二步最容易，而挖掘关联规则的总体性能则由第一步决定，所以对于挖掘频繁项集的算法的研究是有着重大意义的，也因此许多改进挖掘性能的算法主要集中在如何快速有效地挖掘频繁项集上。

目前大部分针对关系数据库中关联规则的挖掘算法都是以 Apriori 算法为核心，与 SQL 语言结合的该类算法大多也只是对 Apriori 算法的关键步骤运用 SQL 语言进行实现和有限扩展，如文[3～5]利用产生候选集其实只需考虑某一部分项目组合的特点改进了 Apriori 算法的关键步骤，这样大大减少了生成 k 项候选集时所要考虑的项目子集数和子集内的项数，但却仍需要产生大量的候选项集。有些算法(如文[6])，虽不产生候选项集，却需要先求得属性组合集并用数组来存储，又增加了存储属性组合集的空间。本文在继承经典 Apriori 算法的优点的基础上，利用 SQL 的集函数及分组语句提出了一种新的关系数据库中频繁项集的挖掘算法，解决了上述问题。

2　关联规则问题描述[2]

设 $I=\{i_1, i_2, \cdots, i_m\}$ 是所有项的集合。D 是所有相关数据事务的集合，其中每个事务 T 是项的集合，使得 $T\subseteq I$。每个事务都有一个确定的标识符 TID。设 A 是一个项集，事务 T 包含 A 当且仅当 $A\subseteq T$，则关联规则是一个形如 $A\Rightarrow B$ 的蕴涵式，其中 $A\subset I, B\subset I$，且 $A\cap B=\varnothing$。规则 $A\Rightarrow B$ 事务集中 D 中成立，具有支持度 s，其中 s 是 D 中事务包含 $A\cup B$ (即 A 和 B 二者)的百分比。规则 $A\Rightarrow B$ 在事务集 D 中具有置信度 c，如果 D 中包含 A 的事务同时也包含 B 的百分比是 c。概率表示即为

$$support(A\Rightarrow B)=P(A\cup B) \quad (1)$$

$$confidence(A\Rightarrow B)=P(B|A) \quad (2)$$

项的集合称为项集(itemset)。包含 k 个项的项集称为 k-项集，项集的出现频率是包含项集的事务数，简称支持数。项集满足最小支持度 min_sup，即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。如果项集满足最小支持度，则称它为频繁项集。对于满足最小支持度阈值和最小可信度阈值的规则称为强关联规则，反之，称为弱关联规则。

3　关系数据库中的实现方法

关系数据库是当前使用最广泛的一种数据库，它只有表这一种数据结构[7]。表有多个属性，每个属性有多个值，这就使得在其上进行数据挖掘有了一定的难度。目前有一种思路就是通过某种方法将源数据库转化为布尔型数据的挖掘数据库[5]，然后就可以利用布尔型数据单值型的特点来进行下一步的挖掘工作。可是这样就必须在挖掘前进行大量的数据转换处理，还要另外存储挖掘数据库，增加了时间和空间开销。但实际上如果考虑采用 SQL 这一比较成熟的关系数据库结构查询语言，我们发现直接对关系数据库进行数据挖掘不但可行而且比较简单和实用。以超市数据库 Transaction 为例，关联规则的支持度可以简单地利用 SQL 语句来实现。如执

王治和　副教授，主要研究方向：数据库技术及应用。

行下列语句：

```
select beer,water,count(*)from transaction
group by beer,water
having count(*)>=min_sup
```

就可得到啤酒(beer)、矿泉水(water)两个属性满足最小支持度(min_sup)的频繁2-项集，同理可以得到三个、四个乃至所有可行的频繁k-项集。

而且由于采用了count(*)集函数，可以在得到不同k项集的同时就得到对应的支持数并一次就舍弃了非频繁集，不用再像Apriori算法先产生候选的k项集，然后再计算每个候选的支持数进行取舍。这样既节约了分次操作的时间，又节省了存储多余无用候选项的空间，这正是本文算法的优越性所在。

4 算法描述及实现[5,6,8]

针对关系数据库的特点和上面的分析思路，给出了一个挖掘所有满足最小支持度的频繁项集的算法，并描述了相应的关系数据库上的实现。

算法描述：

```
INPUT：数据源(Datasource，简称D)和最小支持度阈值(min_sup)
k=1
Lk=generate_frequence_sets(k)
//生成频繁1-项集
While Lk≠∅//当(k-1)项集不为空时生成频繁k-项集
{
k = k+1;
Lk= generate_frequence_sets(k);
}
```

以上算法在实现时，L_k 为临时表，每个项对应着表中的一条记录，这样既便于用SQL操作来实现算法，又可以减少算法对计算机内存的依赖。

产生 L_k 的过程是，利用标准的SQL查询语言的集函数count()及分组语句group by，对源数据库和 L_{k-1} 直接进行检索产生 k 属性组合的表 L_k，插入 L_k 中的项集的条件是它们的支持度不小于最小支持度。该步骤是整个算法的关键环节，因为源数据库相对于 L_k 来说要大得多。但由于使用标准的SQL语句实现查询，可以利用服务器资源进行查询工作，只要把结果返回即可，而且因为直接由 L_{k-1} 来得到 L_k 比用候选集来说要考虑的项目大大减少，扫描数据库的次数也少了许多，所以它的执行速度要快很多。具体实现描述如下：

```
generate_frequence_sets(k):
if (k = 1)
{ create table L1(item1 char,frequent int) //L1表有两个属性，item1的记录为D中所有属性(即项目)，frequent的记录为对应属性的支持数
for (int i=1; i<=attibutecount; i++)
//attibutecount为总属性(即总项目数)
  { insert into L1 select from D where item = i
// count(*)用于得到项集对应支持数
   group by item
   having count(*)>=min_sup//若某1项集的支持数小于最小阈值则可直接丢弃该记录，以减少频繁2项集的项数
}
}
else
{create table Lk(item1 char,item2 char,…,
itemk char,frequent int) //建立频繁k-项集的存储结构
insert into Lk
select p. item1,…, p. itemk-1, q. itemk-1 as itemk,count(*)//产生k属性组合的表Lk
from Lk-1 p, Lk-1 q ,D D1,…,D Dk,
//以下where条件用于产生D中所有属性的某个k组合
where p. item1 = q. item1 and … and p. itemk-2= q. itemk-2 and p. itemk-1< q. itemk-1
and D1. item= p. item1 and…and Dk-1. item = p. itemk-1 and Dk. item = q. itemk-1
and D1. TID= D2. TID and …and Dk-1. TID = Dk. TID // TID为D中记录主码
group by p. item1, p. item2,…, p. itemk-1, q. itemk-1
having count(*)>=min_sup //通过聚集操作得到当前频繁k-项集并存入Lk}
```

5 实验结果

为了验证算法的效率和性能，我们选用了两个实验数据库(experiment1，experiment2)和一个真实数据库(本校2001级学生秋季选课库choose)，主要参数如表1所示。采用C++ Builder 6实现了该算法以及文[4,5]所用算法，C/S结构，数据库系统是微软的SQL-Server 2000，服务器端的CPU为P4 2GHz，内存为256M，客户端的CPU为赛扬1GHz，内存为128M。

表1 参数表

数据库名称	项数	事务数	记录数
experiment1	5	9	23
experiment2	15	100	658
choose	259	2233	15728

每个数据库分别在最小支持度为0.2，0.4，0.6这三个取值下测试了本文算法的执行时间(对数据库choose考虑到实际情况及背景知识，取了与前两个库不同的支持度)，并与文[4,5]算法进行了比较。从图1、2可以看出，本文算法的执行效率明显优于同样使用SQL语言挖掘的文[4,5]算法。当最小支持度逐渐增大时，两算法执行时间都有下降趋势，这是因为过滤掉的项目增多，生成的频繁项集也大大减少，两算法的执行时间也趋于相近。但图3对于真实数据库的实验测试结果表明：在最小支持度较小，生成最大频繁项集较大时(如最小支持度为2%时，最大频繁项集为频繁8-项集)，本文算法的执行效率是文[4,5]算法的4倍多。

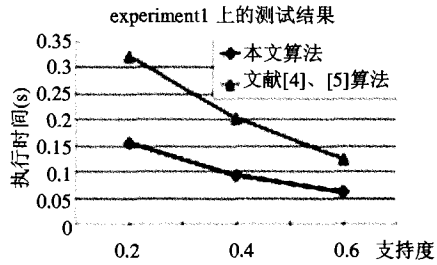

图1 experiment1执行时间

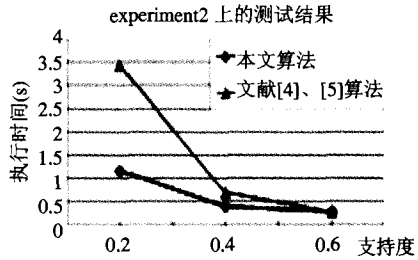

图2 experiment1执行时间

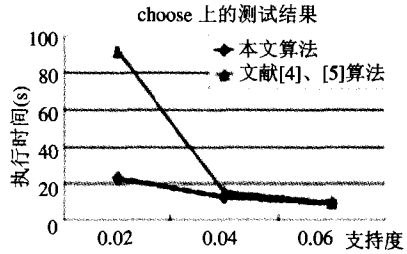

图3 choose执行时间

(下转第198页)

所生成的图像大小为 57×50，入射面光源离散化为 100 个独立点光源。

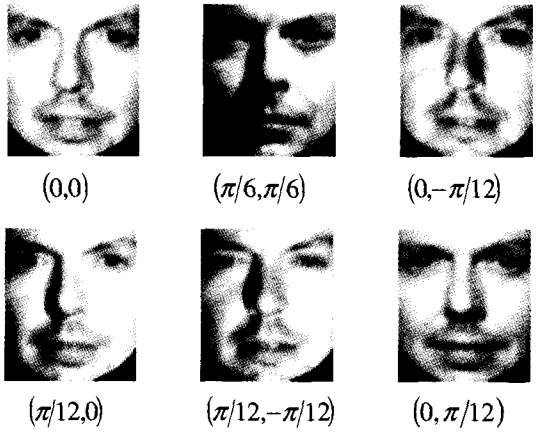

图 4　所使用的部分基图像及其光照属性

表 1　奇异值分解所得矩阵∑的对角元素

1	2	3	4	5
25151	2749	1324	637	1.17e−12

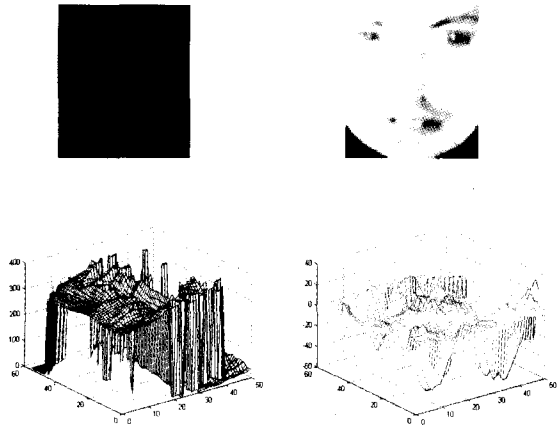

图 5　抽取出的环境光照图像、综合反射率图以及用偏移角和仰俯角表示的目标表面法向

图 6　拍摄图像及算法合成图像：光线方向为(π/12，π/12)

图 7　使用算法生成的复杂光照条件下的图像

致谢　感谢以色列 Amnon Shashua 教授所提供的帮助。

参 考 文 献

1　沈沉，沈向洋，马颂德. 基于图像的光照模型研究综述. 计算机学报，2000，23(12)：1261～1269

2　McMillan L，Bishop G. Plenoptic modeling：An image-based rendering system. In：Proceedings of the SIGGRAPH 95，1995. 39～46

3　Mukaigawa Y. Photometric Image-Based Rendering for Virtual Lighting Image Synthesis. In：2nd IEEE and ACM International Workshop on Augmented Reality. San Francisco，1999. 115～124

4　Mukaigawa Y，Miyaki H. Photometric Image-Based Rendering for Image Generation in Arbitrary Illumination. In：Proceedings of ICCV01，2001，2：652～659

5　董再励，王建刚，徐心平. 一种基于立体视觉的多视点建模方法. 中国图象图形学报，1997，2(7)：461～463

6　于洪川，吴福朝，阮宗才，韦穗. VR 环境图像生成中几项关键技术研究. 计算机研究与发展，1999，36(11)：1349～1357

7　徐丹，王平安. 变化光照的对象图像合成. 软件学报，2002，13(4)：501～509

8　Shashua A，Riklin-Raviv T. The Quotient Image：Class-Based Re-Rendering and Recognition with Varying Illuminations. IEEE Transactions on Pattern Analysis and Machine Intelligence，2001，25(2)：129～139

9　Basri R，Jacobs D W. Lambertian Reflectance and Linear Subspaces. IEEE Transactions on Pattern Analysis and Machine Intelligence，2003，25 (2)：218～233

10　Epstein R，Yuille A L，Belhumeur P N. Learning Object Representations from Lighting Variations. In：Proceedings of the Internelational Workshop on Object Representation for Computer Vision. Lecture Notes in Computer Science. Springer-Verlag，1996. 179～199

11　Ramamoorthi R. Analytical PCA Construction for Theoretical Analysis of Lighting Variability in Image of a Lambertian Object. IEEE Transaction on Pattern Analysis and Machine Intelligence，2002，24(10)

12　Wong Tien-Tsin，Heng Pheng-Ann，Or Siu-Hang，Ng Wai-Yin. Image-based Rendering with Controllable Illumination. In：Proceedings of the Eighth Eurographics Workshop on Rendering，1997. 13～22

13　Zhang Zhi-yong，Pan Zhi-geng，Zhang Ming-min. Spherical Harmonic Descriptor For Gray-level Image Similarity Matching. In：Proceedings of the third international conference on Image and Graphics，2004. 164～167

14　http://cvc. yale. edu/people/faculty/belhumeur http://www.cog. rown. edu/～tarr/stimuli. html

(上接第 160 页)

结论　通过分析和实验证实：本文提出的这种新的在关系数据库中挖掘频繁项集的算法，利用标准 SQL 语言实现频繁项集的挖掘是快速有效的，该算法实现简单，适用范围较广，有一定的实用性。

参 考 文 献

1　Schuster A，Wolff R，Trock D. A high-performance distributed algorithm for mining association rules [J]. Knowledge and Information Systems，2004

2　范明，孟小峰，等译. 数据挖掘—概念与技术(M). 北京：机械工业出版社，2001

3　Agrawal R，Shim K. Developing tightly-coupled data mining Applications on a Relational Database System [A]. In：Proc. of the 2nd Int'l Conference on Knowledge Discovery In Databases and Data Mining [C]. Portland，Oregon，1996. 287～290

4　周剑雄，王明哲. 基于关联规则的数据挖掘技术的快速算法(J). 计算机工程，2003，7(12)

5　杨炳儒，孙海洪，等. 利用标准的 SQL 查询挖掘多值型关联规则及其评价 (J). 计算机研究与发展，2002，3(3)

6　王芳，王万森. 关系数据库中关联规则挖掘的一种高效算法 (J). 微机发展，2004，9(9)

7　萨师煊，王珊. 数据库系统概论. 第三版. 北京：高等教育出版社，2000

8　董淳，王敏慧，等. 关系表中联系规则挖掘的设计和实现(J). 计算机工程，1999，1(1)