# 面向业务流设计的TPRAR算法研究*)

冀常鹏 包 剑 刘建辉
(辽宁工程技术大学电子与信息工程系 阜新123000)

**摘 要** 通过对网络业务进行分析来达到对网络性能进行评价和优化变得日益重要。提出了一种新的基于业务流分析方法,采用关联规则挖掘的算法对全网业务进行关联关系分析。该方法在Apriori算法的基础上,引入时态约束和路径约束,使之更加体现网络业务的特性。使用该方法,可以充分利用已有数据,分析已有业务的行为,且可以预测未来业务的行为。并指出该分析方法是以业务为分析单位的,运用该方法可以从业务角度对网络状况和网络行为进行分析。

**关键词** 业务流设计,关联规则,时态路径约束关联规则

## Study on Path Restraint Associated Algorithm for Network Traffic

JI Chang-Jeng BAO Jian LIU Jian-Hui
(Department of Electronic and Information Engineering, Liaoning Technical University, Fuxin 123000, China)

**Abstract** It becomes more and more important to evaluate and optimize the network performance by analyzing network traffic. Traffic flow of network is analyzed using a method based on time and path restrained algorithm. The TPRAR could be used to analyze the traffic of the global network using data mining algorithm. This algorithm is based on Aprior algorithm, and it also introduces time restraint and path restraint. It can represent the traffic character more realistically. Based on the existing data traffic it can forecast the behavior of future traffic. The paper points out that this analytical method is based on traffic and can be used to analyze network state and network behavior in terms of traffic.

**Keywords** Traffic engineering, Association rules, Time and path restrained association rules

## 1 引言

网络业务流是网络中具有一定业务属性的数据包组成的数据流。业务流设计是网络设计的一部分,主要是进行网络性能评价和网络性能优化,而网络业务流分析则是进行网络性能评价和性能优化的基础。如今对于网络业务流分析,人们提出了各种各样的分析方法,主要有排队论理论和马尔可夫链方法等,这些方法均是基于对数据包进行的分析,即基于包的分析方法。考虑到业务间具有一定的时间相关性和业务的传输与路径其有很强的依赖性,为此我们对以业务流为单位的具有时态路径约束的关联规则分析方法进行研究。这种方法可对全网业务进行相关性分析,即采用关联规则挖掘的算法对全网业务进行分析。

## 2 关联规则及相关算法

关联规则表示数据库中一组对象之间某种关联关系的规则。关联规则挖掘的对象一般是事务型数据库,典型的如超级市场利用前端收款机收集存储了大量的售货数据,这些数据是一条条的购买事务记录,每条记录存储了事务处理时间、顾客购买的物品、物品的数量及金额等。这些数据中常常隐含形式如交易项之间是否存在某种关联关系:在购买铁锤的顾客当中,有70%的人同时购买了铁钉。这些关联规则很有价值,管理人员可以根据这些规则更好地规划商场,如把铁锤和铁钉这样的商品摆放在一起,能够促进销售。

关联规则定义为:$I=\{i_1,i_2,\cdots,i_m\}$是一项集(其中$i_1,i_2,\cdots,i_m$为项);$D=\{T_1,T_2,\cdots,T_n\}$是一事务集(其中$T_1,T_2,\cdots,T_n$为事务),且$T_i\subseteq I$。一个关联规则是一个蕴含式$X\Rightarrow Y$,其中$X\subset I$,$Y\subset I$,且$X\cap Y=\phi$,设$X$为$I$的子集,如果$X\subseteq T$,则称事务$T$包含$X$。如果事务集$D$中有$s$%的事务包含$X$,则$X$在$D$中的支持度为$s$,记为support($X$)。关联规则$X\Rightarrow Y$在$D$中的支持度为support($X\cup Y$),置信度记为$confidence(X\Rightarrow Y)=\frac{support(X\cup Y)}{support(X)}\times$

*)该课题得到辽宁省教育厅高等学校科学研究项目(编号:202183385)资助。冀常鹏 硕士,副教授,主要从事研究方向:计算机通信与网络技术,Web信息系统等方面的研究。

100%。

关联规则算法可以分为两部分：一是找到所有事务支持度在最小支持度之上的项的集合——大项集(频繁项集)；二是用大项集去产生需要的规则。一般情况下，关联规则算法研究主要集中在寻找大项集。

## 3 面向业务流设计的新关联规则算法

对于网络业务的分析和控制，一个很重要的前提是测得业务的传输时间区间(包括起始时间和终止时间)，有了业务的传输时间区间，就可在传输时间区间内对业务进行调节和控制。另一个很重要的前提是测得业务在传输过程中经过哪些路径，有了这些路径，就可以在路径上选择点对业务进行调节和控制。为此在关联规则算法 Apriori 的基础上增加了时态和路径约束，用以提高业务的相关性，即具有时态和路径约束的关联规则(Time and Path Rest rained Association Rules，TPRAR)算法。

### 3.1 时态约束

在 TPRAR 算法中，对于业务的时态进行约束分为两步：时间区间的延展和时间区间的合并。

第一步：时间区间的延展。时间区间的延展是指将其区间两个端点向外扩张，以期使两个时间区间能够相遇或交叠，然后再合并为同一个时间区间。在 TPRAR 算法中，时间区间的延展采用的是两个端点同时按比例全向外扩展，扩展比例由时间区间扩展因子 f 决定。

第二步：时间区间的合并。在 TPRAR 算法中，根据两个时间区间$[a_1,a_2](a_1<a_2)$、$[b_1,b_2](b_1<b_2)$的4个端点的不同，区间的合并可分为以下几种情况。

(1)当 $a_1> a_2$或 $b_1>a_2$时，两个区间不可合并。

(2)当$(a_1\geqslant b_1,b_2\geqslant a_2)$或$(b_1\geqslant a_1,a_2\geqslant b_2)$时，两个区间可合并，合并后区间为$[a_1,a_2]$或$[b_1,b_2]$。

(3)当$(b_1\leqslant a_1\leqslant b_2,a_2\geqslant b_2)$或$(a_1\leqslant b_1\leqslant a_2,b_2\geqslant a_2)$时，两个区间可合并，合并后区间为$[a_1,b_2]$或$[b_1,a_2]$。

### 3.2 路径约束

在 TPRAR 算法中，对于路径约束进行了定义：路径是指由源地址、目的地址和测得该业务的各测量点地址组成的集合。

设 $P_1=\{S_1,M_1,D_1\}$，$P_2=\{S_2,M_2,D_2\}$，其中 $S_1,S_2$为源地址；$D_1$、$D_2$为目的地址；$M_1$、$M_2$为业务流经的测量点的地址组成的集合：$M_1=\{m_{11},\cdots,m_{1i}\}$、$M_2=\{m_{21},\cdots,m_{2t}\}$。

完全约束：当$(S_1=S_2,D_1=D_2)$或$(M_1\subseteq M_2$或$M_2\subseteq M_1)$时，则 $P_1$与 $P_2$完全约束，路径的合并有两种方式：一种是取 $P_1$和 $P_2$的交集；另一种是取 $P_1$和 $P_2$的并集。

部分约束：当$(S_1\neq S_2,D_1=D_2)$或$(S_1\neq S_2,D_1\neq D_2,M_1\cap M_2\neq\phi)$时，则 $P_1$与 $P_2$部分约束，合并后的路径是 $P_1$和 $P_2$的交集。

无约束：当 $P_1\cap P_2=\phi$时，则 $P_1$与 $P_2$无约束。

### 3.3 具有时态路径约束的关联规则算法 TPRAR

TPRAR 算法是在 Apriori 算法的基础上提出的，它比 Apriori 算法的数据库中的记录多了两个字段：一个是属性值为时间区间的时态字段，用以体现时态约束；另一个是属性值为地址的路径字段，用以体现路径约束。TPRAR 算法的具体思想如下：

第一步：生成候选大项集

(1)第一遍扫描：先将样本表 $S$ 中所有的记录的时间区间字段按时间区间扩散因子 $f$ 进行时间区间扩展，形成新的时间区间字段，然后取样本表 $S$ 中所有的项作为候选集表 $C_1$(项个数为1)。

(2)不是第一遍扫描，用候选集生成函数生成表 $C_n$ 先链接：从大项集表 $L_{n-1}$(项个数为 $n-1$)中找出记录 $l_i$ 和 $l_j$，$l_i$ 与 $l_j$ 的前 $n-2$项相同且 $l_i$ 的第 $n-1$项小于 $l_j$ 的第 $n-1$项，将 $l_i$ 的 $n-1$项和 $l_j$ 的第 $n-1$项合并生成 $C_n$，形成表 $C_n$ 然后修剪：对于 $C_n$ 中的所有记录 $c$，若 $c$ 中存在一个长度为 $n-1$的子集不属于 $L_{n-1}$，则删除 $c$。

第二步：添加时间区间字段和路径字段

(1)添加时间区间字段：对于候选集 $C_n$ 中的所有记录 $c$，如表 $S$ 中事务 $t$ 的项包含 $c.itemset$，则给 $c$ 加上 $t$ 的时间区间字段，形成 $c'$的集合。

(2)添加路径字段：对于所有记录 $c'$，如表 $S$ 中事务 $t$ 的项包含 $c'.itemset$ 和 $c'.timezone$，则给 $c'$加上 $t$ 的路径字段，形成新的 $c'$的集合，作为计算支持度的候选集 $C'_n$。

第三步：计算支持度，形成大项集表。对于 $C'_n$ 中的每个记录 $c'$计算支持度 $s$，若 $s$ 大于 $minsup$，则将 $c'$插入 $L$ 中，形成大项集表 $Ln$ 支持度的计算分为两部分：①找到 $S$ 表中记录的项集中包含 $c'.itemset$ 的记录集；②比较记录集中的每条记录的时间区间和路径字段，计算具有时态路径约束关系的记录的个数，作为 $c'$的支持度。

## 4 TPRAR 算法性能分析

### 4.1 应用举例

设有如表1所示的样本表 S，其中 TID 为事务标识符；Itemset 为相应事务所含包的项集(每一项为一类业务，本例中共有 A、B、C、D 四类业务，分别表示 FTP、Telnet、HTTP、SMTP 业务)；Timezone 为相应事务的时间区间；Path 为相应事务的路径集合

(下转第256页)

织的管理模式，不仅提高了灵活性，而且有助于提高生产效率，优化管理模式，实现信息资源共享，降低成本，实现了生产计划、工艺、仓库、经管、生产调度、原材料供应、组装工具配套等多个部门的网上协同作业。采用了先进的 Web Service 技术为整个软件架构的可扩展性和重用性，提供了有力的保障。同时在 Web Service 的调用过程中，通过本地缓存来减少远程访问数据库的次数，提高了系统的性能。

## 参 考 文 献


1 罗鸿，王忠民. ERP 原理设计实施. 电子工业出版社，2003

2 Ching C，Holsapple C W，Whinston A B. Toward IT support fro co-ordination in network organizations. Information & Management，1996，30：179～199

3 Martin Vervijmeren，Software Component Architecture in Supply Chain Management. Computers in Industry，2004，53：165～178

4 张向东. 重庆建设工业(集团)公司车间生产管理信息系统的研制. 2002，10

5 VanderMeer D，Datta A，Dutta H，Thomas H，Ramamri K，Navathe S B. FUSION：A System Allowing Dynamic Web Service Composition and Automatic. In：Proceedings of the IEEE Intl. Conf. on E-Commerce(CEC；03)

6 Vossen G，Westerkamp P. E-Learning as a Web Service. In：Proc. of the Seventh International Database Engineering and Applications Symposium (IDEAS'03)

8 Roy J，Ramanujan A. Understanding Web Service. Perspectives，IEEE 2001

9 Wu Chun-Hsin，Su Da-Chun，Chang Justin，Wei Chia-Chen，Lin Kwei-Jay，Ho Jan-Ming. The Design and implementation of Intelligent Transportation Web Services. In：Proc. of the IEEE International Conference on E-Commerce(CEC'03)


(上接第252页)

(本例中共有 $s_1$、$s_2$两个地址，$m_1$、$m_2$两个测量点地址，$d_1$、$d_2$两个目的地址)。时间区间扩展因子 f 为 1.5，最小支持度 minsup 为2。

表1 样本表 S

| TID | Item set | 时间区间 | 路径 |
|---|---|---|---|
| 100 | A,B,C,D | [100,160] | s1,m1,m2,d1 |
| 200 | A,B | [60,120] | s1,m2,d1 |
| 300 | A,D | [105,165] | s2,m1,m2,d1 |
| 400 | B | [100,140] | s1,m2,d2 |
| 500 | A,B,D | [320,400] | s2,m2,m2,d2 |

计算支持度，按 TPRAR 算法得到大项集表 $L_1$ (如表2)，其中路径字段的比较满足完全约束关系，且合并后的路径取 $P_1$和 $P_2$的交集。

表2 $L_1$表

| 项集 | 时间区间 | 路径 |
|---|---|---|
| A | [80,110] | $m2$ |
| B | [80,110] | $m2$ |
| D | [90,180] | $m1,m2$ |

### 4.2 算法性能分析

由 TPRAR 算法可知，算法的性能由三部分决定。第一步生成候选大项集，第一遍扫描的时间复杂度为 $O(t)$，不是第一遍扫描时链接过程的时间复杂度为 $O(n^2)$，修剪过程的时间复杂度为 $O(n)$。第二步添加时间区间和路径字段，时间复杂度为 $O(n)$。第三步计算支持度，形成大项集表，时间复杂度为 $O(n)$。故整个 TPRAR 算法的性能为 $O(n^2)$。Apriori 算法的时间复杂度为 $O(k)$。增加了时态路径约束后，TPRAR 算法比 Apriori 算法时间复杂度有所下降，综合分析有以下原因：①链接过程中，Apriori 算法采用的是 Hash 表操作，时间复杂度为 $O(k)$，而 TPRAR 算法用的是数据库操作，作二次循环，时间复杂度为 $O(n^2)$；②其他处理过程中 Apriori 算法采用的是 Hash 表操作，时间复杂度为 $O(k)$，而 TPRAR 算法中采用的是数据库操作，进行循环操作，时间复杂度为 $O(n)$。

**结束语** 如今网络业务种类越来越多，业务量越来越大，严重影响了网络性能。在业务流分析中应用 TPRAR 算法具有以下好处：首先以业务为单位进行分析，对网络的分析更加宏观，属业务级的分析；其次可充分利用已测量到的大量数据，分析已有业务的行为，了解业务间的关系；然后通过对已有业务的分析，可以预测未来业务的行为。应用 TPRAR 算法可以找出满足路径约束的各种业务之间的前后因果关系，从而达到从宏观的角度来分析网络业务间的特性，为网络管理人员进行网络优化和网络控制提供强有力的支持，以利于网络业务的开展。

## 参 考 文 献


1 杨新宇，郑守淇，曾明，等. 用于业务流以设计的一种多 Agent 模型[J]. 西安交通大学学报，2001，35 (8)：834～838

2 楮立文，廖建新，陈俊亮. 自相似业务流的建模及排队性能分析[J]. 通信学报，1999，8(20)：7～12

3 杨新宇，郑守淇，等. 面向网络业务流设计的时态路径约束关联规则算法[J]. 西安交通大学学报，2001，35 (10)：1029～1033

4 王清毅，张波，蔡庆生. 目前数据挖掘算法的评价[J]. 小型微型计算机系统，2000，1(11)：75～78