基于Web日志的隐私保护关联规则挖掘方法

鲍 钰 黄国兴

(华东师范大学软件学院 上海 200062)

摘 要 电子商务网站用户的每次购物会话信息会被记录在Web服务器的日志中,分析这些日志并挖掘出购物篮商品间的强关联规则,可以主动为Web终端用户提供商品推荐,优化网站服务质量。鉴于原始用户会话信息及挖掘结果的隐私保护问题,提出了一种新的数据随机干扰处理方法,即结合列置换的伪列随机化回答方法,先对原始日志信息进行变化和隐藏,然后以此为基础,给出了一种基于位逻辑与操作的高效频繁项集生成算法,进而实现了原始信息及挖掘结果均获得隐私保护的网上购物篮问题的关联规则挖掘。实验结果表明,本方法具有很好的隐私保护性、高效准确性以及适用推广性。

关键词 Web日志,隐私保护,关联规则,随机化回答

Privacy Preserving Association Rule Mining Method Based on Web Logs

BAO Yu HUANG Guo-xing

(Software Engineering Institute, East China Normal University, Shanghai 200062, China)

Abstract Each visitor's shopping session of the E-Business Web site is recorded in the Web server log files. Analyzing the log files and exploring the strong regularities in the commodities of the shopping cart, can provide the recommended goods for Web users, and improve the performance of the Web service. In order to improve the privacy preservation of the original visitor's shopping information and mining result, an effective method for privacy preserving association rule mining was presented. First, a new data preprocessing approach, Fake Column's Randomized Response with Column Replacement (FCRRCR) was proposed to transform and hide the original data. Then, an effective privacy preserving association rule mining algorithm based on bit AND operation was presented. As shown in the experimental results, the algorithm can achieve significant improvements in terms of privacy, accuracy, efficiency and applicability.

Keywords Web logs, Privacy preservation, Association rule, Randomized response

1 引言

数据挖掘对当今海量信息的分析和知识发现发挥了积极的作用,但也带来了隐私保护方面的诸多问题。例如,通过对网上购物篮的历史会话数据进行挖掘,可以发现各种商品间的强关联规则,但在一般的挖掘过程中,会暴露购物者的原始数据,从而泄露用户隐私。还有医院病例数据的挖掘、股票客户交易行为的挖掘,都会涉及到公民非常注重的个人隐私。所以,如何在数据挖掘过程中解决好隐私保护的问题,目前已经成为数据挖掘界的一个研究热点[1]。目前,隐私保护的数据挖掘方法按照基本策略主要有数据干扰法MASK[2]和查询限制法[3,4]。数据干扰法主要是利用统计学中的沃纳模型,但由于变换后的所有数据均与真实的原始数据直接相关,使得对隐私数据的保护程度并不理想,而且随机化参数的选择也都受到限制,必须偏离0.5[5]。在查询限制的策略方面,文献[3]提出了针对特定的敏感规则对原始数据进行隐藏,降低敏感规则支持度,使其不被发现的方法。但通常情况下,具体的规则在挖掘结果出来以前都是未知的。此外,上述所有方法均没有考虑到对挖掘结果的保护,即数据使用者获得了和数据提供者一样的挖掘结果。

本文提出了一种新的数据随机干扰处理方法,先通过斐波那契数列生成伪列干扰项,并通过随机化回答方法产生干扰项中信息,然后通过随机映射对其进行列置换,再以此为基础,提供给数据使用者进行频繁项集以及强关联规则的发现。对于数据使用者返回的挖掘结果,数据提供者需先剔除伪列干扰项,并进行列逆向置换,从而得到真实保密的网上购物篮商品间的强关联规则。

2 数据预处理

原始Web日志虽然包含了大量访问者的浏览信息,但有些日志是随镶嵌在网页中的图片和脚本自动产生的。网络搜索引擎使用的Web蜘蛛、软机器人或智能代理也会产生访问日志,这些日志对分析用户的访问情况都是无用的,应该过滤掉。另外,所有用户的访问日志会零星掺插在一起,所以还必须对日志按照用户进行分组,这便是用户识别。同一用户对Web站点会有很多次的访问,因此又需要在同一用户的所有

到稿日期:2008-11-10 返修日期:2008-12-25 本文受国家重点基础研究发展规划(973)项目(2005CB321904)资助。

鲍 钰(1977-),男,博士研究生,讲师,研究方向为知识发现与数据挖掘,E-mail:ybao@sei.ecnu.edu.cn;黄国兴(1946-),男,博士生导师,研究方向为智能信息系统、知识发现与数据挖掘。

访问日志上进行会话分组，这些都是数据预处理要做的工作。

2.1 数据净化

数据净化是指删除 Web 服务器日志中与挖掘算法无关的数据，包括消除一些通过 Web 蜘蛛、软机器人或智能代理产生的访问日志，过滤掉一些含有声音、图像文件的日志。用户一般不会显示请求页面上的图形文件，它们是根据 HTML 的超文本引用标志自动下载的。这些日志与分析用户的行为模式没有任何关系，所以通过检查 URL 的后缀，将日志中文件的后缀名为 GIF，JPEG，JPG，gif，jpeg，jpg 和 map 的项删除。另外，后缀名为 cgi，js 的脚本文件也应被删除。具体到实际的系统，可以使用一个缺省的后缀名列表帮助删除文件。列表可以根据正在分析的站点类型进行修改。除了删除上面列出的无关项外，还应当删除露宿者数据。所谓的露宿者是偶尔访问站点且逗留时间极短的用户，这些数据对挖掘过程没有什么贡献，所以在数据净化阶段也要删除。

2.2 用户识别

要识别出每一个用户，这一任务由于本地缓存、公司防火墙和代理服务器的存在变得很复杂。依赖用户的合作是最好的解决办法，但是由于涉及到隐私，这种解决办法往往难以进行。目前常用的用户识别方法有 IP 地址识别、嵌入 SessionID 等。IP 地址识别法假定每一个 IP 地址对应一个用户，是最简单易行的，不过误差也最大，因为存在多个用户共有一台机器和通过代理 IP 上网的情况。嵌入 SessionID 技术一般在电子商务记录用户购物篮内物品时最常用，使用动态的方法产生 ID 号，嵌入在用户访问请求中，也就是把一段时间内同一用户的请求都标记上相同的 SessionID 号。但是，嵌入 SessionID 只在动态网站上适用，而且是以时间间隔来判别当前 SessionID 是否有效，因此没有考虑短时间内重复访问的情况。目前以嵌入 SessionID 技术为主，本文也采用这种方法。

2.3 会话识别

在跨越时间区段较大的 Web 服务器日志中，用户有可能多次访问了该站点。会话识别的目的就是将用户的访问记录分为单个的会话(Session)。最简单的方法是利用超时。如果两个页面间请求时间的差值超过一定的界限，就认为用户开始了一个新的会话。许多商业产品将缺省超时值确定为 30min，超时的界限可以根据站点的使用统计反馈的结果进行调节，直到可以准确地识别会话。为方便起见，在本文中也使用 30min 作为会话的超时界限。通过如下算法 1 的用户访问路径会话集发现，算法可以将杂乱无章的 Web 日志信息归结为一条条用户访问路径会话记录，产生表 1 用户访问路径会话集 UVPSD(User Visit Path Session Dataset)。表 1 中 IPID =10000001，SID=10001 的记录，表示某 IPID=10000001 单个用户在这次会话中访问了 P1，P2，P3 这 3 个购物页面，即购物者购买了 P1，P2，P3 这 3 个页面上对应的商品。

表 1 用户访问路径会话集 UVPSD

IPID	SID	PAGES
10000001	10001	P1，P2，P3
10000001	10002	P1
10000001	10003	P3，P4，P5
10000002	10001	P1，P2，P3
10000002	10002	P1，P2，P3，P5

算法 1 用户访问路径会话集发现算法

函数 END_TIME(Sk)返回会话 Sk 中最后一个访问页面产生的日志记录时间。

函数 CLOSE_SESSION(Sk)从 Open_Sessions_Set 中移除 Sk。

函数 WRITE_SESSION(Sk)将 Sk 写入用户访问路径会话集 UVPSD 中。

函数 OPEN_SESSION(Si)向 Open_Sessions_Set 中添加 Sk。

```
L: The set of input logs
L = {L1,L2,L3 ,…,Li,…,L|L|} , ∀i ≤ |L|
|L|: the number of input logs
Li = {IPi,TIMEi,METHODi,URLi,PROTi,CODEi,BYTESi }
S: The set of sessions
S = { S1,S2,S3,…,Si,…,S|S| } , ∀i ≤ |S|
|S|: The number of sessions
Si = { IPi,PAGESi }
PAGESi={ PAGE1,PAGE2,…,PAGEm,…,PAGE|PGSi| }, ∀m ≤ |
        PGSi|
|PGSi|: The number of PAGESi in Session_i
//Input: L, |L|, Δt
//Output: S, |S|
Function DSUSKSD (|L|, L, Δt)
  For each Li of L
    If Methodi is 'GET' AND Urli is 'WEBPAGE'
       If ∃Sk ∈ Open_Sessions_Set with IPk = IPi then
          If ((Timei-END_TIME(Sk)) < Δt) then
             //Δt: 会话超时界限设为 30 分钟
             Sk = (IPk ,PAGESk ∪ PAGESi)
             //属同一会话，合并访问页面
          Else
             CLOSE_SESSION(Sk)
             WRITE_SESSION(Sk)
             Si = { IPi,PAGESi }
             OPEN_SESSION(Si)
       End if
    Else
       Si = { IPi,PAGESi }
       OPEN_SESSION(Si)
    End if
  End if
  End For
  For each Si of Open_Sessions_Set
    CLOSE_SESSION(Si)
    WRITE_SESSION(Si)
  End For
End Function
```

3 事务数据库的布尔矩阵表示

将每种商品购买页面作为一个项，并拥有一个确定的顺序号。客户的一次购物作为一个事务，并用一个长度为商品种类总数的 BOOL 序列来表示：购买了相应序号的商品，该位就取值为 1，否则就取值为 0。表 1 的用户访问路径会话集 UVPSD 用布尔矩阵存储，形式如表 2 所列。

表 2 购物者会话集的布尔矩阵表示 M

TID	IPID	SID	P1 P2 P3 P4 P5
1	10000001	10001	1 1 1 0 0
2	10000001	10002	1 0 0 0 0
3	10000001	10003	0 0 1 1 1
4	10000002	10001	1 1 1 0 0
5	10000002	10002	1 1 1 0 1

表 1 的事务数据库经一次扫描后，就可映射成表 2 的布尔矩阵，故对事务数据库的挖掘问题就可转化为对其布尔矩阵的分析。表 2 中增加了一个新字段 TID，表示事务标号，而 IPID，SID 两个字段可以在后续的工作中剔除。

4 结合列置换的伪列随机化回答方法

在结合列置换的伪列随机化回答方法 FCRRCR(Fake Column's Randomized Response with Column Replacement)中，伪列标题的添加可以采用随机数的方式生成，也可根据需要通过一些著名的数列公式产生。本文采用斐波那契数列生成如下伪列 F1，F2，F3，F5，并按照序号关系添加在相应的真实列标题 P1，P2，P3，P5 之后。伪列 F1，F2，F3，F5 中的具体数据填充，采用随机化回答方法：

给定随机化参数 $0\leqslant x,y,z\leqslant 1$，且 $x+y+z=1$。设 $F(i,j)$代表伪列 Fi 在事务编号 $TID=j$ 的取值；$P(i,j)$代表原始列 Pi 在事务编号 $TID=j$ 的取值。对于项值 $b\in\{0,1\}$，随机化函数 $R(b)$以 x 的概率选择取值为 0，以 y 的概率取值不变即 b，以 z 的概率取值为 1。采用公式 $F(i,j)=R(P(i,j))$，即可产生伪列中的干扰填充数据。

设项的总数为 k，则对于用 0－1 序列表示的事务 $M=(m_1,m_2,\cdots,m_k)$，干扰后的事务 $N=(n_1,n_2,\cdots,n_k)$可以通过上述伪列随机化回答方法填充得到。

设 i 是一个项，Spi 表示列 Pi 基于 M 的支持度，Sfi 表示列 Fi 基于 N 的支持度。设 M 中的列 Pi 经过处理，得到 N 中的列 Fi，则 Pi 和 Fi 的取值和对应概率如表 3 所列，分析得到式(1)：

$$Sfi=Spi*(y+z)+(1-Spi)*z=y*Spi+z \qquad (1)$$

表 3 FCRRCR 方法的数据映射概率

NO	Pi Fi	Probability
1	0 0	x+y
2	0 1	z
3	1 0	x
4	1 1	y+z

根据 M 中的支持度阈值，可以利用式(1)来为每个伪列选择不同的随机化参数 x，y，z，这会对挖掘效率和干扰效果起到非常大的影响。当 Sfi 过大时，会使得所有的伪列都成为频繁 1 项集，这时干扰强度最大，但会加大运算量；当 Sfi 过小时，大部分伪列会在求解频繁 1 项集时被剔除，这将提高后续的挖掘效率，但干扰强度最小。

假设 M 中的支持度阈值为 0.4。这里为了简化，每个伪列的随机化参数 x，y，z 选择相同的数值，x＝0.1，y＝0.5，z＝0.4，可以得到表 4 中数据。

表 4 经过伪列填充后的购物者会话集 N

TID	*P1 F1 P2 F2 P3 F3 P4 P5 F5*
1	1 0 1 1 1 1 1 0 0 1
2	1 1 0 1 0 1 0 0 1
3	0 1 0 1 1 1 1 1 0
4	1 1 1 0 1 0 0 0 1
5	1 1 1 1 1 1 0 1 0

为了对原始日志进一步干扰，可以采用随机数的方式生成如下的列置换映射关系，如表 5 所列。

表 5 列置换映射关系表

原始列名	置换后列名
P1	*N8*
F1	*N1*
P2	*N3*
F2	*N9*
P3	*N7*
F3	*N6*
P4	*N2*
P5	*N4*
F5	*N5*

经过列置换处理后，最终得到表 6：可提供给数据使用者进行挖掘的伪造事务集 S′。

表 6 伪造事务集 S′

TID	*N1 N2 N3 N4 N5 N6 N7 N8 N9*
1	0 0 1 0 1 1 1 1 1
2	1 0 0 0 1 1 0 1 1
3	1 1 0 1 0 1 1 0 1
4	1 0 1 0 1 0 1 1 0
5	1 0 1 1 0 1 1 1 1

5 基于位逻辑与操作的高效频繁项集生成算法

在伪造的事务集 S' 上，数据使用者可采用目前国际上众多学者基于 Apriori 算法提出的一些改进或扩展方法，如 DHP 方法[6]、Partition 法[7]、频繁闭项集法[8]、FP2Growth 算法[9]、闭包项集格[10]、TBAR 算法[11]、动态剪枝[12]等，进行关联规则的挖掘。这里提出一个基于位逻辑与操作的高效频繁项集生成算法。

引入路径深度 DEPTH 和决策阈值 COUNT。DEPTH＝所需要求解购物篮中关联商品种类的最大个数，COUNT＝支持度阈值＊伪造的事务集 S' 的事务行数。

算法 2 基于位逻辑与操作的高效频繁项集生成算法

1)SET C1＝{(SUBSETi，SUPi)/0＜i＜n－1，i∈N}，FOR EACH (SUBSETi，SUPi)∈C1，满足 SUBSETi 取值为 S′所有单个商品列名(SUCH AS：N1－N9)。SUPi 为 SUBSETi 中单个节点在 S′中出现的次数(即对应列名下 BOOL 数值为 1 的个数)。

2)SET d＝1 //d 为当前路径深度。

3)选择 Cm 中 SUPi ≥ COUNT 的元组(SUBSETi，SUPi)，构成 Lm。

4)If d ＜ DEPTH Then

Goto 5

Else

Exit

End If

5)FOR EACH {SUBSETx，SUPx}，{SUBSETy，SUPy}∈Lm，x≠y，SET SUBSETz ＝ SUBSETx ∪SUBSETy。对于节点个数为 m＋1 的 SUBSETz，扫描其在 S′中出现的次数 SUPz，并将(SUBSETz，SUPz)插入到 Cm＋1 中。

6)m ＝ m＋1，Goto 3。

注意，由于 S' 对事务记录的 BOOL 矩阵式存储结构，在上述算法第 5 步骤，对于节点个数为 $m+1$ 的 SUBSETz，扫描其在 S' 中出现的次数 SUPz。这个匹配过程可以采用基于位的逻辑与操作来优化性能，具体做法是：假定 SUBSETz 为{N1，N3}，我们构造如下的二进制位序列{1，0，1，0，0，0，0，0，0}作为匹配模板，对于 S' 中每行事务用此模板和其进行基于位上的逻辑与操作，仅当逻辑与后的结果与匹配模板相等时，

SUBSETz 的出现次数增 1。

在表 6 伪造事务集 S' 上，执行基于位逻辑与操作的高效频繁项集生成算法(这里 $M=2,S=3$)，可以得到如下频繁 1 项集：{N1}，{N3}，{N5}，{N6}，{N7}，{N8}，{N9}；频繁 2 项集：{N1，N6}，{N1，N7}，{N1，N8}，{N1，N9}，{N3，N7}，{N3，N8}，{N5，N9}，{N6，N7}，{N6，N8}，{N6，N9}，{N7，N8}，{N7，N9}，{N8，N9}。

数据提供者在获得数据使用者挖掘出来的频繁项集后，通过表 5 列置换映射关系表，进行列逆向置换并剔除含有任一伪列的项集，得到频繁 1 项集：{P2}，{P3}，{P1}；频繁 2 项集：{P2，P3}，{P2，P1}，{P3，P1}。

获得真正的频繁项集后，根据如下支持度和置信度计算公式，数据提供者可以求得购物篮商品间的关联程度：

$$\text{support}(A \Rightarrow B) = \text{support\_count}(A \cup B)/\text{all\_count}$$

6 实验结果分析

在某中等规模的电子商务网站上，对其服务器上 30 日产生的 Web 日志采用本文中的方法进行了实验。结果表明，利用式(1)来为每个伪列选择不同的随机化参数 x,y,z 时，这里 $y=p$，图 1 给出了 FCRRCR 方法和 MASK 方法平均项集误差随参数 p 变化的情况。

可以看出，MASK 方法的误差变化比较大。当 p 接近 0 或 1 时，挖掘结果比较准确；但此时的隐私破坏系数接近于 1，方法对隐私的保护程度很差；在 p 从 0 或 1 逐渐接近 0.5 的过程中，隐私破坏系数会逐渐减小，隐私保护的程度在不断提高，但挖掘结果的准确性将显著下降。而本文提出的 FCRRCR 方法，误差变化相对比较平稳，随着 p 值，也就是真实数据所占的比例从 0 增加到 1，隐私破坏系数也从 0 增长到 1，方法对隐私的保护程度不断下降，而挖掘结果的准确性不断提高。

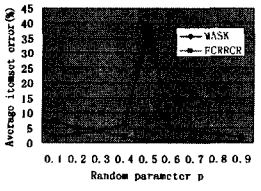

图 1 FCRRCR 方法和 MASK 方法的平均项集误差

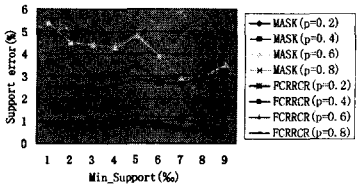

图 2 FCRRCR 方法和 MASK 方法的支持度误差

在图 2 中，详细给出了当随机化参数 p 分别取值 0.2，0.4，0.6，0.8 时 FCRRCR 方法和 MASK 方法在最小支持度阈值 $s=1‰,2‰,3‰,4‰,5‰,6‰,7‰,8‰,9‰$ 情况下的支持度误差比较。

当 p 的取值较小时，MASK 方法的误差比 FCRRCR 方法要小，准确性要高；而当 p 的取值超过 0.4 以后，FCRRCR 方法的误差就要低于 MASK 方法了。在理论分析和实验结果的基础上，权衡数据的隐私性和挖掘结果的准确性，我们建议在区间[0.4，0.6]上选取随机化参数 p 的值，使用 FCRRCR 方法进行隐私保护的关联规则挖掘。

结束语 本文提出了一种新的数据随机干扰处理方法——FCRRCR。先将购物者的原始会话对应的布尔矩阵进行列置换和伪列干扰，然后针对经过 FCRRCR 方法处理的数据，给出一种既简单又高效的基于位逻辑与操作的频繁项集生成算法，进而实现了原始会话信息及挖掘结果均获得隐私保护的网上购物篮问题的关联规则挖掘。

还在某中等规模的电子商务网站上，采用 FCRRCR 方法对其一个月的原始日志进行处理后，再提交给数据使用者进行频繁项集的挖掘。实验结果表明，本方法具有很好的隐私保护性以及适用推广性。

不过，通过购物篮的关联规则挖掘实验，发现式(1)中伪列的随机化参数对数据隐私性和挖掘效率的影响甚大。本文是根据经验及折中估算的方法来调整参数。在未来的工作中，我们希望能够引入更好的方法来解决上述问题，并进一步提高挖掘算法的运行效率。

参 考 文 献

[1] Verykios V S, Bertino E, Fovino I N, et al. State-of-the-Art in privacy preserving data mining[J]. SIGMOD Record, 2004, 33(1): 50-57

[2] Agrawal S, Krishnan V, Haritsa J R. On addressing efficiency concerns in privacy-preserving mining[C]// Lee Y J, Li J Z, Whang K Y, et al., eds. Proc. of the 9th Int'l Conf. on Database Systems for Advanced Applications. LNCS 2973. Jeju Island: Springer-Verlag, 2004: 113-124

[3] Oliveira S R M, Za-ane O R. Privacy preserving frequent itemset mining[C]// Clifton C, Estivill-Castro V, eds. Proc. of the IEEE ICDM Workshop on Privacy, Security and Data Mining. Maebashi: Australian Computer Society, 2002: 43-54

[4] Kantarcioglu M, Clifton C. Privacy-preserving distributed mining of association rules on horizontally partitioned data[J]. IEEE Trans. on Knowledge and Data Engineering, 2004, 16(9): 1026-1037

[5] 赵俊康. 统计调查中的抽样设计理论与方法[M]. 北京：中国统计出版社，2002

[6] Park J S, Chen M S, Yu P S. An effective hash based algorithm for mining association rules[A]// ACM SIGMOD International Conference Management of Data[C]. 1995: 175-186

[7] Savasere A, Omiecinski E, Navathe S. An efficient algorithm for mining association rules in large database[A]// Proc. of 21st Intl Conf. on Very Large DataBase [C]. Zurich, Swizerland, 1995: 432-443

[8] Pasquoier N, Bastide Y, TaouilL R, et al. Discovering frequent closed item sets for association rules[A]// ICDTp99[C]. Israel, 1999: 398-416

[9] Han J, Pei J, Yin Y. Mining frequent patterns without candidate generation[A]// Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data[C]. Dallas, Texas: ACM Press, 2000: 1-12

[10] Pasquoer N, Bastide Y, Taouil R. Efficient mining of association rules using closed item set lattices[J]. Information System, 1999, 24(1): 25-46

[11] Berzal F, Cubero J2C, Marin N. TBAR: An efficient method for association rule mining in relational databases[J]. Data & Knowledge Engineering, 2001, 37: 47-64

[12] 皮德常，秦小麟，王宁生. 基于动态剪枝的关联规则挖掘算法[J]. 小型微型计算机系统，2004，25(10)：1850-1852