基于用户访问事务文法的序列关联规则发现*

王 实, 高 文, 李锦涛

(中国科学院 计算技术研究所,北京 100080)

E-mail: shiwang@ict.ac.cn

http://www.ict.ac.cn

摘要: 在Web挖掘中,应用关联规则发现方法可以发现Web页面之间用户访问的关联度.由于Web站点内含丰富的页面结构信息,也由于用户的访问总是要遵循一定的访问顺序,因此提出一种新的可以发现用户访问序列之间关联度的方法——序列关联规则发现方法.该方法首先得到用户访问事务,然后根据正则文法,定义了一种新的用户访问事务文法,用于从用户访问事务中得到用户序列访问事务;最后应用关联规则发现算法进而发现序列关联规则.为了进一步评价所发现的序列关联规则,引入了互信息的概念.发现的序列关联规则可以帮助Web站点的设计者更好地理解用户的访问,以用于调整Web站点的结构.

关键词: Web数据挖掘;文法;序列关联规则

中图法分类号: TP311 **文献标识码**: A

当前由于World Wide Web正在飞速地发展,Web世界产生了大量的数据.因此,应用数据挖掘方法研究这些数据,即Web数据挖掘,成为当前一个新兴的重要的研究领域.Web世界的数据主要包括Web页面数据、Web结构数据以及用户访问日志.因此Web数据挖掘也分为Web内容挖掘、Web结构挖掘和Web访问信息挖掘.

Web页面之间存在着丰富的结构信息,这种结构信息和用户的访问方式是紧密相连的.当一个用户访问一个Web站点时,会留下他对该站点的访问日志.一个中等规模的站点,一天也会留下几兆的日志.日积月累的日志更是Web数据挖掘的良好对象.通过对这些日志的挖掘,可以发现用户的访问模式,用于改进Web站点的设计.这种访问模式反映着设计者所设计的Web页面所关联的优劣,可以被用来改进Web站点的结构,更好地帮助用户访问.

用户的访问模式主要分为:(1) 用户的访问序列模式[1];(2) 用户所访问页面之间的关联规则[2].前者反映的是用户的总的迁移序列,着重于序列的发现;后者反映的是群体用户所访问的页面之间的关联度,着重于独立页面之间的关联.

由于Web站点的页面存在着网状结构信息,也由于用户的访问存在着序列性,在本文中,我们将提出一种新的发现序列关联规则的方法.这种方法基于用户事务访问文法(user access transaction grammar),据此得出一个用户访问事务内的各个访问序列,然后应用关联规则发现方法和互信息的思想求得并进一步解释所发现的序列关联规则.这种序列关联规则可以更好地帮助Web站点的设计者理解用户的访问模式.

* **收稿日期**: 2000-03-06; **修改日期**: 2000-05-30

基金项目: 国家重点基础研究发展规划973资助项目(G1998030405);国家863高科技发展计划资助项目(863-306-JD06-03-4)

作者简介: 王实(1971—),男,陕西西安人,博士,主要研究领域为数据挖掘;高文(1956—),男,山东牟平人,博士,教授,博士生导师,主要研究领域为多媒体数据压缩、图像处理,计算机视觉,多模式接口,人工智能,虚拟现实;李锦涛(1962—),男,湖南华容人,博士,研究员,主要研究领域为智能化家庭信息中心平台,数字化家电应用.

现有的一些可利用的商业分析工具[3]用于分析 Log,但这些工具仅能产生一些简单的统计结果,如页面的访问频度等.

文献[1]应用 Hypertext probabilistic grammar 发现用户迁移模式,并用 grammar 的熵值评估挖掘到的模式.但它不能发现不同页面集之间的关联关系.文献[2]首次给出 Web 挖掘的定义,并且给出一个关于 Web 访问信息挖掘的系统 WEBMINER.文献中提到的思路是通过对 Web 站点的日志进行处理,将数据组织成传统的数据挖掘方法能够处理的事务数据形式,然后利用传统的数据挖掘方法(如关联规则发现算法[4])进行处理,所得的挖掘结果也是传统的数据挖掘结果,没有考虑 Web 站点的结构和用户访问的序列特性,因而不能用于发现序列关联规则.

"Footprints"[5]的思想是:访问者在访问一个 Web 站点时,会留下"足迹",经过一段时间,最频繁访问的区域会形成路径,于是新的访问者会依据这些路径进行访问."足迹"被自动地留下,并且访问者不需要提供自己的任何信息.WUM[6]是对"Footprints"方法的一种补充,定义 g-sequences 用于挖掘迁移模式,并给出一种挖掘语言 MINT.这些方法发现的是一些局部的信息,不能发现跨不同页面集之间的关系.而本文的方法本质上是一种发现不同页面集关联关系的方法.

本文第 1 节描述了需要挖掘的对象.第 2 节定义了用户访问事务文法以及如何应用该文法得到用户序列访问事务,并给出了相应的算法.第 3 节定义了序列关联规则,并引入互信息概念以评价发现到的关联规则.第 4 节给出实验过程,并与 Cooley 的方法进行了比较.

1 挖掘对象

挖掘对象 Log 存在 Web 服务器上,其日志格式遵循 W3C 标准[7].在进行挖掘时,首先要将一段时间用户的访问日志组织成用户访问事务数据.设 L 为用户访问日志,其中的一个项 $l\in L$ 包括用户的 IP 地址 $l.ip$,用户的标识符 $l.uid$,被存取页的 URL 地址 $l.url$,以及存取访问的时间 $l.time$.

定义 1(用户访问事务). 用户访问事务被定义为

$$t=\langle ip_t,uid_t,\{(l'_1.url,l'_1.tim),\ldots,(l'_m.url,l'_m.tim)\}\rangle$$

where, for $1\leqslant k\leqslant m$, $l'_k\in L$, $l'_k.ip=ip_t$, $l'_k.uid=uid_t$, $l'_k.tim-l'_{k-1}.tim\leqslant C$.

表示一个用户对一个 Web 站点的一次访问.这里,C 是一个固定的时间窗.设一个 Web 站点有 n 个 Web 页面,每一个页面可以被记为 $a_i,i=1,\ldots,n$;那么 $A=\{a_1,\ldots,a_n\}$ 表示页面的集合,则 t 可以被简记为 $t=\langle a'_1,a'_2,\ldots,a'_m\rangle$,其中 $a'_i=l'_i.url$ 且 $a'_i\in A$.

对 Log 进行处理,找到每一个事务.寻找访问事务的算法为:(1) 对日志进行预处理;(2) 根据每一个访问者的 IP,划分日志,即在 Log 中找到每一个访问者的访问记录集;(3) 对每一个访问者的访问记录集,根据 C 进行分割,找到每一个访问者的每一次访问记录集,这时,每一个访问者的每一次访问记录集就构成了一个访问事务;(4) 在一个访问事务内每一个被访问的页面按照被访问的时间排序;(5) 最终按时间排序的所有访问事务构成我们进行挖掘的基础.处理完日志后我们就得到一个用户访问事物集.

2 用户访问事务文法(user access transaction grammar)

定义用户访问事务文法的目的是要得到用户访问事务中用户访问的有序性.一个 Web 站点的节点表是一个有限非空的符号集合,如 $A=\{a_1,\ldots,a_n\}$. A^* 表示在 A 上所有的有限长度序列的集合,它包括空序列 ε,表示用户访问该站点时可能访问的路径. A^+ 表示集合 $A^*-\{\varepsilon\}$.一个在 A 上

的序列集合 L 是 A^* 的任何子集,一个用户访问事务文法是一种能够从一个用户的访问事务中得到所有访问序列的设备.

定义 2(用户访问事务文法(user access transaction grammar)).

在一个用户访问事务 t 中,其用户访问文法是一个四元组 $G=(V,\Sigma,S,P)$,其中:

(1) V 是一个有限的序列集合:$V=\{S,A_1,\ldots,A_n\}$.

(2) Σ 是 t 事务中所访问到的页面的集合:$\Sigma=\{a_1,\ldots,a_m\}$;$V\cap\Sigma=\varnothing$. 被访问的页面 a_1 到 a_m 按被访问的时间排序,并给予相应的下标.

(3) $S\in V$ 是一个惟一的起始符.

(4) P 是一个具有导出形式 $A_i\rightarrow a_t$ 或 $A_i\in a_tA_j$ 的有限导出规则集,其中 $A_i,A_j\in V$,$a_t\in(\Sigma\cup\varepsilon)$,序列 A_j 中所有页面的下标都要大于 t,并且序列 A_j 的第 1 个页面的下标等于 $t+1$.

在一个用户访问事务文法 G 中,一个一步导出 $d:s_1\Rightarrow s_2$ 是指,当应用一个导出规则从 s_1 序列得到 s_2 时,称为从 s_1 一步导出 s_2. 一个导出 $d:s_1\Rightarrow^* s_n$ 是指,有限步地应用一步导出规则从 s_1 得到 s_n. 一个序列格式是从惟一的起始符 S 导出的任意导出形式. 由一个用户访问事务文法 G 产生的序列集合是所有的仅由终止符构成的序列格式的集合,$L(G)=\{s\in\Sigma*\mid S\Rightarrow^* s\}$. 如果有一个序列 $s\in L(G)$,s 至少有两个从起始符 S 的不同导出,则称 G 为歧义的,否则称 G 是非歧义的.

在一个用户访问事务文法中,一个句子格式至少有一个非终止符. 因此,形如 $A_i\rightarrow a$ 的导出被称为最终导出,因为其终止了导出过程;形如 $A_i\rightarrow a_jA_j$ 的导出被称为转换导出. 一个导出的长度 D 被定义为在该序列被导出的过程中导出规则被应用的次数,在用户访问事务文法中相应为所产生的序列的长度,序列长度为 m 的序列我们称为 m 序列. m 序列的集合成为 m 序列集.

定义 3(用户序列访问事务 st). 一个用户访问事务经过用户访问事务文法处理所得到的序列的集合. 用户访问事务集 ST 为用户序列访问事务 st 的集合,见表 1.

Table 1 The conversion from the user access transactions to the user sequence access transactions

表 1 从用户访问事务到序列用户访问事务的转换

User access transactions①	User sequence access transactions②			
	1-Sequence set③	2-Sequence set	3-Sequence set	4-Sequence set
A_1,A_2,A_3	A_1,A_2,A_3	A_1A_2,A_2A_3	$A_1A_2A_3$	
A_4,A_2,A_3,A_5	A_4,A_2,A_3,A_5	A_4A_2,A_2A_3,A_3A_5	$A_4A_2A_3,A_2A_3A_5$	$A_4A_2A_3A_5$
A_2,A_3,A_5	A_2,A_3,A_5	A_2A_3,A_3A_5	$A_2A_3A_5$	
A_3,A_5,A_7	A_3,A_5,A_7	A_3A_5,A_5A_7	$A_3A_5A_7$	

①用户访问事务,②用户序列访问事务,③序列集.

从用户访问事务中生成用户序列访问事务的算法如下:

算法. GUSATG /* Generating User Sequence Access Transaction Grammars */

输入:$t=\langle t_1,\ldots,t_m\rangle$

```
Begin:
  k:=1;
  S_k:={t_1,...,t_m};                    /* S_k 为长度为 k 的序列的集合 */
  While k≤m
      For each s∈S^k
        p:=position(t,s);                /* 在串 t 中求得串 s 的位置 */
        If (p+k+1)≤m then
          s:=merge(s,t_{p+k+1});         /* 将 t_{p+k+1}添加到序列 s 的尾部 */
          S^{k+1}:=S^{k+1}∪{s};
        End If;
```

End For;
$k:=k+1$;
End While;
End.
输出：$S^k, k=1,\ldots,m$

对每一个用户访问事务应用该算法就可以生成一个用户序列访问事务，最终形成用户序列访问事务集.

3 序列关联规则发现

关联规则发现[4]主要用于在事物数据库中发现大项集之间的关联度.为了在用户序列访问事务集中发现序列关联规则，我们给出如下定义：

定义4(序列的支持度 *Support*(s)). 给定一个序列 s，在用户序列访问事务集 ST 中，含有序列 s 的事务的个数.

定义5(序列关联规则). s,s' 代表在用户序列访问事务集中的两个序列，其支持度大于一个给定的支持度阈值，则定义可信度 *confidence* 为

$$confidence(s,s')=\frac{support(s,s')}{support(s)}. \tag{1}$$

如果 $confidence(s,s')\geqslant\theta,\theta$ 为给定的一个阈值(例如5%)，那么 $s\rightarrow s'$ 构成一条序列关联规则.注：定义5中的 $support(s,s')$ 表示在一个 ST 中，同时含有 s 和 s' 的用户序列访问事务的个数.我们所采用的发现算法为文献[4]所述的 AprioriHybrid 关联规则发现算法.

为了进一步评价所发现的序列关联规则，我们引入互信息[8]的概念，即如果两个序列频繁出现在同一个用户序列访问事务中，那么它们之间有较高的关联度：

$$MI(s,s')=\log\frac{P(s,s')}{P(s)P(s')}. \tag{2}$$

由于这种方式没有考虑到在一个用户序列访问事务中，s 和 s' 其中一个或两个都未出现所造成的影响，因此我们进一步引入平均互信息[8]概念来处理两个序列缺席的情况，即如果两个序列总是同时出现或者同时不出现，那么它们之间有较高的关联度：

$$EMI(s,s')=\sum_{S,\overline{S}\in S}\sum_{S',\overline{S'}\in B}P(S,S')\log\frac{P(S,S')}{P(S)P(S')}. \tag{3}$$

通过序列关联规则以及对所发现的规则进一步求出互信息和平均互信息，就可以发现由用户访问事物文法而得到的用户序列访问事务集中的更加有用的信息.

与 Cooley[2]所用到的方法相比较，Cooley 等人的方法处理的是用户访问事务集，而没有考虑到一个事务内部被访问页面的被访问相关性.本文的方法所产生的关联规则集包含 Cooley 的方法所产生的关联规则集.在本文的方法中，由用户访问事物文法得出这种被访问页面的序列之间的相关性，可以更好地为站点的设计者服务，以发现的序列关联规则被用于改进 Web 站点的结构设计为例(如图1和图2所示).

通过对用户序列访问事务集的挖掘，如果在1-序列集中我们可以发现 $(B,A)\rightarrow(D,C)$ 这样一条关联规则，其可信度为80%(等价与 Cooley 的方法)，其解释为访问了 B,A 这两个节点的用户有80%也访问了 D,C 这两个节点，那么通过对2-序列集进行挖掘，我们针对该规则可以进一步知道 $AB\rightarrow CD$，可信度为70%，这样一条序列关联规则的解释为访问了 A 紧接着又访问 B 的用户有70%访问 C 后又紧接着访问了 D.由此，我们可以在 B 和 C 之间加入一个从 B 指向 C 的链接以方

便用户.也就是说,通过对序列关联规则的发现,我们可以更好地理解我们所发现的知识,以用于改进Web站点的结构设计.

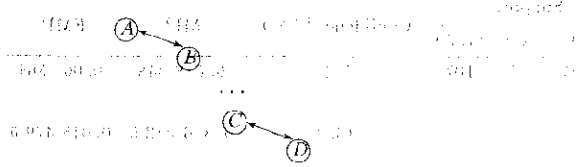


Fig. 1 There is not direct hyperlink between A, B and C, D
图1 A, B和C, D之间没有直接的链接

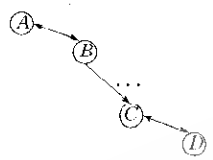


Fig. 2 Adding the direct hyperlink between B and C
图2 B和C之间加入直接的链接

4 实 验

我们选取了中国科学院计算技术研究所Web服务器(www.ict.ac.cn)上的日志作为实验对象,实验数据包括从1998年11月到1999年11月用户对计算所Web站点一年的访问数据.整个站点包括352个html页面.用户访问日志的总量为147M,包括1 749 934项.经过事务识别算法,共识别出10 399个用户访问事务,平均访问事务长度为8.8,即用户平均每次访问8.8个页面.实验环境为PentiumIII 450,64兆内存,6G硬盘,Windows NT操作系统.

我们实验的主要目的是:

(1) 给出我们的方法和Cooley的方法的比较.

求1-序列的关联规则(等价于Cooley的方法),我们得到一条关联规则,见表2.

Table 2 The 1-sequence association rule

表2 1-序列的关联规则

(1-Sequence) to (1-Sequence)①, support②≥10, confidence③=25%
(/cjc/cjccw.html,/cjc/cjcc.html,/cjc/introc.html,/cjc/cjcw2.html)→(/cjc/centc.html,/cjc/abstc.html)

①(1-序列)对(1-序列)的关联规则,②支持度,③可信度.

根据我们的方法,相对于这条关联规则,算法发现的结果见表3.

Table 3 Some discovered sequence association rules

表3 一些发现的关联规则

The sequence association rule①, support②≥5, confidence③≥1%①	Confidence(%)
/cjc/cjccw.html→/cjc/cjcc.html/cjc/introc.html→/cjc/cjcw2.html	98
/cjc/cjccw.html/cjc/cjcc.html/cjc/introc.html/cjc/cjcw2.html→/cjc/centc.html/cjc/abstc.html	15
/cjc/cjccw.html/cjc/cjcc.html/cjc/cjcw2.html/cjc/introc.html→/cjc/centc.html/cjc/abstc.html	7
/cjc/cjccw.html/cjc/cjcc.html/cjc/introc.html/cjc/cjcw2.html→/cjc/abstc.html/cjc/centc.html	2
/cjc/cjccw.html/cjc/cjcc.html/cjc/cjcw2.html/cjc/introc.html→/cjc/abstc.html/cjc/centc.html	1

①序列关联规则,②支持度,③可信度.

显然,与Cooley的方法相比,我们的方法对发现的规则给出了更好的解释.

(2) 互信息的引入.引入互信息可以更好地对所发现的序列关联规则给予解释,见表4.

对比式(1)和式(2),在引入$P(s')$参数后,可以更好地说明s和s'的相关性.即如果$P(s)$和$P(ss')$固定不变,那么随着$P(s')$的增加,两个序列互信息的值会缩小.较大的$P(s')$会降低互信息的值.互信息的值更好地解释了所发现的规则.如表4所示的第2和第3条规则:第2条规则相对于第3条规则而言有较低的可信度,但却有较高的互信息和平均互信息值.即说明/cjc/cjccw.html/cjc/absc.html/cjc/abstc.html比/cjc/cjccw.html/cjc/othersc.html/cjc/ly.html更有可能是一条用户访问的路径.

Table 4 The associations between different sequences

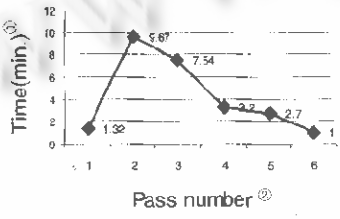
表4 不同序列之间的关联度

No①	Sequence association rules②: $s \rightarrow s'$	Support③			Confidence④(%)	MI⑤	EMI⑥
		(s)	(s')	(s,s')			
1	/cjc/cjccw.html→/cjc/contc.html /cjc/cont98c.html	2 047	125	107	5.2	0.638 348	0.005 364
2	/cjc/cjccw.html→/cjc/absc.html /cjc/abstc.html	2 047	312	281	13.7	0.366 042 5	0.015 479 0
3	/cjc/cjccw.html→/cjc/otherstc.html /cjc/ly.html	2 047	407	309	15	0.583 238	0.013 204
4	/cjc/cjccw.html→/cjc/abstc.html /cjc/contc.html	2 047	550	533	26	0.692 238	0.035 739

①编号,②序列关联规则,③支持度,④可信度,⑤互信息,⑥平均互信息.

(3) 算法执行的性能.

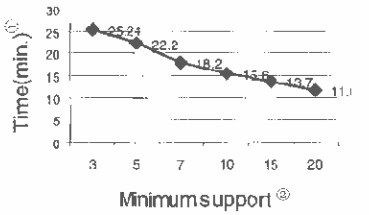
我们采用 AprioriHybrid 来求得序列关联规则. 图3给出算法执行时每一遍所用的时间. 图4则给出最小支持度和算法的运行时间的关系.



①时间(分),②遍数.

Fig. 3 The execution time of each pass (the minimum support is 3)

图3 算法每一遍所需的时间(最小支持度为3)



①时间(分),②最小支持度.

Fig. 4 The relation between the minimum support and the execution time

图4 最小支持度和执行时间的关系

5 结论以及将来的工作

在 Web 挖掘中,传统的一些关联规则发现方法可以发现 Web 页面之间用户访问的关联度. 由于 Web 站点内含丰富的 Web 页面结构信息,也由于用户的访问总是要遵循一定的访问顺序,使得本文提出一种新的可以发现用户访问序列之间的关联度的方法——序列关联规则发现方法. 该方法包含了 Cooley 的方法,并且进一步发现用户不同的访问序列之间的关系. 在该方法中,我们首先挖掘用户访问日志以得到用户访问事务,然后根据正则文法定义了一种新的用户访问事务文法,以用于从用户访问事务中得到用户序列访问事务,同时也给出了相应的算法. 然后应用关联规则发现算法进而发现序列关联规则. 为了进一步评价所发现的序列关联规则,本文引入互信息的概念. 发现的序列关联规则可以帮助 Web 站点的设计者更好地理解用户的访问,以用于调整 Web 站点的结构.

本文提出的方法的特点是:(1) 发现的是序列关联规则;(2) 周期性、离线地进行挖掘;(3) 挖掘的对象是全体用户的迁移行为,挖掘的是全体用户的访问兴趣,挖掘的结果面向 Web 站点的设计者;不需要特定的某一个或某一类用户的信息;(4) 方法在本质上是自动地跨不同页面分类集的,发现的两个序列不一定在 Web 站点上有直接链接.

Web 站点的设计者可以根据该方法发现页面集之间的关系,据此考虑增加两个序列之间的直

接链接,以利于用户访问.我们进一步的工作将主要应用这种方法预测用户的访问行为,进行实时个性化推荐.

References:

[1] Borges, J., Levene, M. Data mining of user navigation patterns. In: Brij, Masand, ed. The Web Usage Analysis and User Profiling Workshop. San Diego, CA: ACM Press, 1999. 31~36.

[2] Cooley, R., Mobasher, B., Srivastava, J., *et al.* Data preparation for mining world wide web browsing patterns. Knowledge and Information Systems, 1999,1(1):17~24.

[3] Stort, R. Web Site Stats: Tracking Hits and Analyzing Traffic. Osborne: McGraw-Hill, 1997.

[4] Agrawal, R., Srikant, R. Fast algorithms for mining association rules. In: Fayyad, U., ed. Proceedings of the 20th VLDB Conference. Santiago, Chile: IEEE Society Press, 1994. 487~499.

[5] Wexelblat, A., Maes, P. Footprints: history-rich web browsing. In: Jan, Pedersen, ed. Proceedings of the Conference on Computer-Assisted Information Retrieval (RIAO). New York: IEEE Society Press, 1997. 75~84.

[6] Spiliopoulou, M. The laborious way from data mining to web mining. International Journal of Computing Systems, Science and Engineering, 1999,3(2):42~47.

[7] Luotonen, A. The common log file format. 1995. http://www.w3.org/pub/WWW/.

[8] Rosenfeld, R. A maximum entropy approach to adaptive statistical language modeling. Computer, Speech, and Language, 1996,10(1):51~67.

Sequence Association Rule Discovery Based on User Access Transaction Grammar*

WANG Shi, GAO Wen, LI Jin-tao

(*Institute of Computing Technology, The Chinese Academy of Sciences, Beijing* 100080, *China*)

E-mail: shiwang@ict.ac.cn

http://www.ict.ac.cn

Abstract: In web mining, applying association rule discovery can discover the association between different web pages accessed by users. Because there is the rich structure information in the website and the access of the users conforms to some kinds of sequences, a new approach is presented in this paper to discover the association between the access sequences, which is the sequence association rule discovery. In this approach, first the Log is mined in the web server to get the user access transactions, and then according to the regular grammar, a new user access transaction grammar is defined in order to get the user sequence access transactions from the user access transactions. Subsequently, the association rule discovery algorithm is employed to discover the sequence association rules. To evaluate these rules, the mutual information is proposed. The result of this approach can help the designer of the website to understand the user access patterns better, and according to this result, the designer can adjust the structure of the web site.

Key words: web mining; grammar; sequence association rule

* Received March 6, 2000; accepted May 30, 2000

Supported by the National Grand Foundamental Research 973 Program of China under Grant No. G1998030405; the National High Technology Development 863 Program of China under Grant No. 863-306-JD06-03-4