基于隐马尔可夫模型的在线零售站点的自适应*

王 实, 高 文, 黄铁军, 马继勇, 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

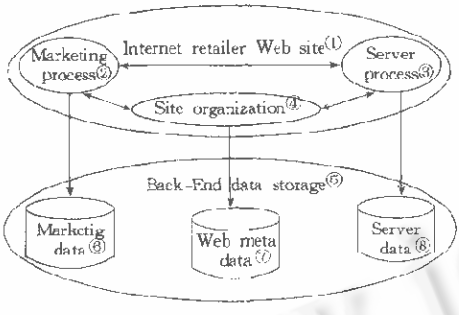
摘要: 开展在线零售业务存在的问题是,群体用户必须浏览许多无关的页面,才能最终找到自己所需要的商品.解决该问题的一个思路是:建立一个隐马尔可夫模型,通过关联规则发现算法发现关联购买集合;然后通过Viterbi算法求出从首页到一个关联购买集合中心的具有最大被购买概率的一些路径;在这些路径上标注关联购买集合;当处理完所有的关联购买集合之后,通过竞争来决定出现在导航页面上的物品集,最终将导航页合理地变成导航购买页.即站点可以自动根据群体用户的访问购买情况进行自适应.此外,该方法也是一种很好的通过建立隐马尔可夫模型来分析购买访问路径的方法,可以被广泛地用于Web站点的路径分析、广告和人工重构中.

关键词: Web数据挖掘;隐马尔可夫模型;关联规则;自适应

中图法分类号: TP311 **文献标识码**: A

在电子商务环境下,一个在线零售电子商务的业务模型如图1[1]所示.

一个标准的在线零售网站的分类结构如图2所示.其中一个节点表示一个页面.N节点表示导航页或分类页;C节点表示内容页或购买页.开展在线零售业务存在如下问题:



①零售Web站点界面处理层,②销售过程,③服务过程,④站点组织,⑤后台数据存储层,⑥市场数据,⑦Web结构数据,⑧服务数据.

Fig.1 Online retail model
图1 在线零售模型

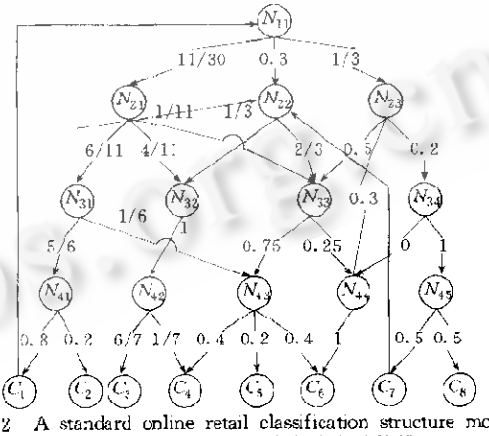


Fig.2 A standard online retail classification structure model
图2 在线零售站点Web站点分类结构模型

* 收稿日期: 1999-11-24; 修改日期: 2000-02-20

基金项目: 国家863高科技发展计划资助项目(863-306-JD05-03-4)

作者简介: 王实(1971-),男,陕西西安人,博士生,主要研究领域为数据挖掘;高文(1956-),男,山东牟平人,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;黄铁军(1964-),男,河北大名人,博士,主要研究领域为虚拟现实;马继勇(1964-),男,黑龙江哈尔滨人,博士,主要研究领域为手语识别,语音识别,基于多通道信息融合的人的身份识别研究;李锦涛(1962-),男,湖南华容人,研究员,主要研究领域为智能化家庭信息中心平台研究,数字化家电应用研究.

(1) 全体用户对某些物品的兴趣要远远大于另一些物品,但页面的分类层次设计必须严格遵循商品的分类结构,否则一般用户就无法访问.这种矛盾将导致大量用户必须浏览许多无关页面和进入过多层次才能最终完成交易.解决该问题的思路是将图2上的导航页(N 页)变成导航内容页(NC 页).这样,用户就可以在 N 页上直接购买自己需要的商品.现有的做法是根据人工经验或简单的统计方法来生成导航内容页.而由于用户的购买兴趣经常发生变化,也因为页面结构的复杂性,这两种方法都存在不足,所以我们要做的就是如何自动地、有效地、合理地、更加智能地解决这个问题.

(2) 许多用户购买的物品类似于啤酒和尿布这样的物品,即在页面结构分类上两者相距很远,但很多顾客会同时购买.于是,这些用户就不得反复进退多个 Web 页.我们要做的就是如何自动地发现这些关联物品集,自动建立包含其在内的导航内容页,以帮助用户访问.

所以,需要建立一个模型和相应的算法,从而把 N 页自动变成 NC 页,而又能保证原有的 N 页之间的导航关系不被破坏.NC 页将满足大部分用户的需求,使他们不需访问过多的层次,或尽量不需绕路就能进行交易.

文献[1]将数据挖掘的技术应用于电子商务的环境下,以发现市场智能.其方法局限于传统的挖掘手段.本文不仅应用传统的挖掘手段(如关联规则发现算法[2])而且还应用隐马尔可夫模型进行 Web 站点结构的有限智能调整.文献[3]给出 Web 挖掘的定义,本文在此基础上,把挖掘的结果用于改进在线零售站点的组织结构,以利于群体用户的访问.WebLogMiner[4]方法用 OLAP 技术来实现对 Web 日志数据的预测、分类和时间序列分析.而本文所述方法主要用于站点的自动重新设计,并且不会破坏原有的分类结构,即自适应.在文献[5,6]中,应用这些方法的目的是自动定制不同的用户访问界面,对比来说,我们的方法(1) 是一种优化方法;(2) 周期性、离线地进行挖掘;(3) 挖掘的对象是全体用户共同的访问购买兴趣,挖掘结果面向全体用户;(4) 不需要特定的某一个或某一类用户的信息.文献[7]用 PageGather 聚类方法的结果:索引页,以实现 Web 站点对外部访问的自适应.它潜在地使整个 Web 站点的结构平面化,在处理大量 Web 页面的时候,索引页的数量也一样会很多;而我们的方法不会形成附加的索引页,被提升的内容自然而然地出现在它们应该出现的地方.

本文第1节简述所要用到的隐马尔可夫模型.第2节给出在线零售站点的隐马尔可夫模型以及解决上述问题的思路.第3节讨论如何初始化模型.第4节给出实验比较过程.

1 隐马尔可夫模型

隐马尔可夫模型(hidden Markov model,简称 HMM)[8]被广泛地用于语音识别之中.本文所采用的是离散化输出,一阶隐马尔可夫模型:

(1) 一个状态集合 Q,具有指定的初始状态 q_I 和最终状态 q_F.

(2) 一个状态转移集,每个元素为 $(q \to q')$.

(3) 一个离散的输出符号集:$\Sigma=\sigma_1,\sigma_2,\ldots,\sigma_M$.

从初始状态开始,转移到一个新的状态,观测到一个输出符号,如此反复,直到最终状态,于是就产生一个符号串:$X=x_1,x_2,\ldots,x_t$.每一个转移存在着一个转移概率 $P(q \to q')$.在一个状态观测到一个特殊符号的概率为 $P(\sigma|q)$.那么,在一个隐马尔可夫模型 M 上,一个串 X 被观测的概率为在所有可能路径上求概率之和:

$$P(X|M)=\sum_{q_1,\cdots,q_l\in Q^l}\prod_{k=1}^{l+1}P(q_{k-1}\to q_k)P(x_k|q_k). \tag{1}$$

这里,q_0 和 q_{l+1} 为初始状态 q_I 和最终状态 q_F. x_{l+1} 为中止符号.

建立 HMM 的一个普遍目的是找到一个状态序列 $V(X|M)$,它具有观察序列的最大概率:

$$V(X|M)=\arg\max_{q_1\cdots q_l\in Q^l}\prod_{k=1}^{l+1}P(q_{k-1}\to q_k)p(x_k|q_k), \tag{2}$$

可以采用 Viterbi 算法求得 $V(X|M)$.

2 在线零售站点的隐马尔可夫模型

2.1 定义模型

(1) N,C 节点为 HMM 的状态节点 q.

(2) 存在物品集 $\Sigma=\sigma_1,\sigma_2,\ldots,\sigma_M$.

(3) C 节点包含 Σ 的一个子集$(\sigma'_1,\ldots,\sigma'_m)$.

(4) 直接相联的节点 q,q' 之间存在着一个转移概率 $P(q\to q')$.

(5) 在每一个节点 q 可能发生的购买行为,即群体用户经过 q 购买 σ 的概率为 $P(\sigma|q)$,也即标准 HMM 中状态节点的观测概率.

2.2 关联规则发现

设定支持度和可信度,通过对交易事务数据采用关联规则发现算法,并在规则的基础上进行聚类,就可以得到一些关联购买的物品集.每个物品集中被购买最多的某一个物品为其聚类中心.只取其中具有很大支持度的关联物品集,这样可以防止运算量过大.

2.3 发现路径

(1) 对每一个关联物品集 $x_i=(\sigma'_1,\ldots,\sigma'_m)$,通过以首页和中心购买页这两个节点为起始点和终止点,以该物品集作为每一次观察结果,形成一个序列,我们就可以通过 Viterbi 算法求得从首页到中心购买页之间的具有最大购买该物品集可能的 j 条状态路径(例如,j 可以取所有购买该物品集的路径的 50%).

(2) 取 $X=x_1,x_2,\ldots,x_l$ 且 $x_1=x_2=\ldots=x_l=(\sigma'_1,\ldots,\sigma'_m)$.

(3)

$$V(X|M)=\arg\max_{q_1\cdots q_l\in Q^l}{}_{1\ldots j}\prod_{k=1}^{l-1}P(q_{k-1}\to q_k)p(x_k|q_k), \tag{3}$$

于是,这些路径就是用户最有可能购买该关联物品集的路径.

(4) 在这些路径上将这个物品集及其被购买的概率(在 N_{11} 节点的被购买概率,即物品在根节点的被购买概率)标注在路径上的每一个 N 节点上.

(5) 对所有的购买聚类集进行步骤(1)~(4)的操作.

(6) 最终,由于中间路径上的每一个导航节点的内容部分容量有限,所以,只找出每个节点上的前 m 个在该节点上具有最大购买概率的物品集作为最终的内容部分.

3 模型的初始化

3.1 挖掘对象

挖掘的对象存在于如图1所示的后台数据存储层之中,分为:

(1) 用户的访问日志(在如图1所示的服务数据中).服务器上的日志格式遵循W3C标准[9],见表1.

Table 1 The log format

表1 用户访问日志格式

Field①	Description②
Date	Date, time, and timezone of request③
Client IP	Remote host IP and/or DNS entry④
User name	Remote log name of the user⑤
Bytes	Bytes transferred (sent and received)⑥
Server	Server name, IP address and port⑦
Request	URI query and stem⑧
Status	http status code returned to the client⑨
...	...

①域,②描述,③请求日期,时间,时区,④客户端IP地址或DNS入口,⑤客户端用户名,⑥收发字节数,⑦服务器名,IP地址和端口号,⑧URL请求,以及详细地址,⑨返回给客户端的状态码.

(2) 用户的交易记录即传统的交易事务数据(在如图1所示的市场数据中),见表2.

Table 2 The purchasing transaction

表2 用户交易事务

Field①	Description②
TransactionID	Transaction identification③
Commodities	The commodities that the user purchases in the transaction④

①域,②描述,③事务标识符,④在该用户访问事务中,用户所购买的物品集.

(3) Web页面的结构信息(在如图1所示的Web结构数据中).

3.2 计算转移和购买概率

在进行挖掘时,首先要将一段时间内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$l.time$.访问事务被定义为

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

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

这里,C是一个固定的时间窗.对Log进行处理,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) 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.

处理完日志后可以得到两组事务集:(1) 用户访问事务集Ta;(2) 用户交易事务集Tt(可以从市场数据中直接得出).此时,处理Ta和Tt,把用户每次发生购买行为和由此而进行的路径访问提取出来,形成Tat事务集,构成我们建立隐马尔可夫模型的基础,见表3.

Table 3 The access and purchasing transaction

表 3 存取交易事务

Field①	Description②
TransactionID	Transaction Identification③
Commodities	The Commodities that the user purchases in the transaction④
AccessPath	The user went through the web path for buying the goods⑤

①域,②描述,③事务标识符,④在该用户访问事务中,用户所购买的物品集,⑤在该事务中,用户为了购买该物品集所访问的路径.

在 Tat 中计算两个直接相联页面的单步转移概率为

$$P(q_i \rightarrow q_j) \approx \frac{count(q_i \rightarrow q_j)}{count(q_i)}, \tag{4}$$

其中 $count(q_i \rightarrow q_j)$ 表示在事务集 Tat 中 q_i 和 q_j 直接连通,且从 q_i 一步到 q_j 的访问事务的个数. $count(q_i)$ 为 Tat 中含有 q_i 的事务的个数.

在页面 q 可能购买某物品集 σ 的概率,即群体用户经过 q 购买 σ 的概率为

$$P(\sigma|q) \approx \frac{count(\sigma \wedge q)}{count(q)}, \tag{5}$$

其中 $count(\sigma \wedge q)$ 表示在 Tat 中同时出现 σ 和 q 的记录的个数. $count(q)$ 表示在 Tat 中出现 q 的事务的个数. 全部概率根据站点的拓扑结构和 Tat 计算得出.

文中所述的转移和购买概率存在一定的相关性,但网络的转移主要取决于两者的乘积,这样并不影响整个算法的运行.

4 实 验

4.1 HMM 与 MM 的比较

为何采用 HMM 而不是 MM?我们建立了一个以图 2 为背景的 8 个物品的简单站点例子,运行了一段时间以得到数据. 图上已标注了转移概率,表 4 给出在每一个节点的购买概率(设每个页只有一种物品,C_1 页有 G_1,...,C_8 页有 G_8):

Table 4 The purchased probability of each commodity in each node

表 4 在每个节点对每种物品的购买概率

	G_1	G_2	G_3	G_4	G_5	G_6	G_7	G_8
N_{11}	4/30	1/30	6/30	5/30	2/30	10/30	1/30	1/30
N_{21}	4/11	1/11	3/11	1/11	0	2/11	0	0
N_{22}	0	0	3/9	4/9	0	2/9	0	0
N_{23}	0	0	0	0	0.2	0.6	0.1	0.1
N_{31}	4/6	1/6	0	0	0	1/6	0	0
N_{32}	0	0	6/7	1/7	0	0	0	0
N_{33}	0	0	0	2/6	1/6	3/6	0	0
N_{34}	0	0	0	0	0	0	0.5	0.5
N_{41}	0.8	0.2	0	0	0	0	0	0
N_{42}	0	0	6/7	1/7	0	0	0	0
N_{43}	0	0	0	0.4	0.2	0.4	0	0
N_{44}	0	0	0	0	0	1	0	0
N_{45}	0	0	0	0	0	0	0.5	0.5

如果采用 MM 只计算一步转移概率,那么从 N_{11} 到 C_6 的路径中:

(1) Path1(N_{11},N_{23},N_{33},N_{43},C_6)从 N_{11} 经过 N_{23},N_{33},N_{43} 到 C_6 的转移概率为

$P(N_{11}\rightarrow N_{23})\times P(N_{23}\rightarrow N_{33})\times P(N_{33}\rightarrow N_{43})\times P(N_{43}\rightarrow C_6)=1/3\times 0.5\times 0.75\times 0.4=0.05$.

(2) Path2$(N_{11},N_{23},N_{33},N_{44},C_6)$从 N_{11} 经过 N_{23},N_{33},N_{44} 到 C_6 的转移概率为

$P(N_{11}\rightarrow N_{23})\times P(N_{23}\rightarrow N_{33})\times P(N_{33}\rightarrow N_{44})\rightarrow P(N_{44}\rightarrow C_6)=1/3\times 0.5\times 0.25\times 1=0.125/3$.

因此就要选择 Path1 路径.而如果计算 HMM,那么:

(1) Path1$(N_{11},N_{23},N_{33},N_{43},C_6)$从 N_{11} 经过 N_{23},N_{33},N_{43} 到 C_6,在中间的每一个状态点始终购买 G_6 的概率为

$$(P(N_{11}\rightarrow N_{23})\times P(G_6|N_{23}))\times(P(N_{23}\rightarrow N_{33})\times$$
$$P(G_6|N_{33}))\times(P(N_{33}\rightarrow N_{43})\times P(G_6|N_{43}))\times(P(N_{43}\rightarrow C_6)\times P(G_6|C_6))=$$
$$(1/3\times 0.6)\times(0.5\times 3/6)\times(0.75\times 0.4)\times(0.4\times 1)=0.006.$$

(2) Path2$(N_{11},N_{23},N_{33},N_{44},C_6)$从 N_{11} 经过 N_{23},N_{33},N_{44} 到 C_6,在中间的每一个状态点始终购买 G_6 的概率为

$$(P(N_{11}\rightarrow N_{23})\times P(G_6|N_{23}))\times(P(N_{23}\rightarrow N_{33})\times P(G_6|N_{33}))\times$$
$$(P(N_{33}\rightarrow N_{44})\times P(G_6|N_{44}))\times(P(N_{44}\rightarrow C_6)\times P(G_6|C_6))=$$
$$(1/3\times 0.6)\times(0.5\times 3/6)\times(0.25\times 1)\times(1\times 1)=0.0125.$$

因此,就要选择 Path2 路径.只定义转移概率而不引入购买概率就会出现这样的问题:Path1 比 Path2 的转移概率高,但并不意味着群体用户经过 Path1 购买 G_6 物品的概率比 Path2 高,而且在有些情况下恰恰相反.

进一步经过关联规则发现算法,我们可以得出 $G_6,G_3,G_4,G_1,(G_3,G_4)$ 为大项集.$G_6\rightarrow G_4$ 为一条关联规则.N_{11} 到 C_6 的购买 G_6 的路径有很多条,根据 Viterbi 算法可以求出具有最大购买概率的 3 条路径,见表 5.

Table 5 The 3 paths that have maximal purchased probability from N_{11} to C_6 for G_6

表 5 N_{11} 到 C_6 的购买 G_6 的具有最大购买概率的 3 条路径

The path from N_{11} to $C_6$①	The probability that G_6 is purchased②
Path $(N_{11},N_{23},N_{44},C_6)$	$(1/3\times 0.3\times 1)\times(0.5\times 1\times 1)=0.06$
Path $(N_{11},N_{23},N_{33},N_{44},C_6)$	$(1/3\times 0.5\times 0.25\times 1)\times(0.6\times 0.5\times 1\times 1)=0.0125$
Path $(N_{11},N_{23},N_{33},N_{43},C_6)$	$(1/3\times 0.5\times 0.75\times 0.4)\times(0.6\times 0.5\times 0.4\times 1)=0.006$

①N_{11} 到 C_6 的路径,②购买 G_6 的概率.

此时,因为没有竞争者,可以把 G_6,G_4 直接放到这些路径的沿途页面上.

4.2 与简单统计方法的比较

简单统计方法分为两种,这两种方式应用于非树形的站点结构都存在着严重的缺点:

(1) 从一个 N_i 节点开始,它所能到达的所有的内容节点中具有最大被购买概率 $P(G|N_I)$(在首节点的被购买概率,即物品在根节点的被购买概率)的物品,被作为内容部分放到该 N_i 节点上.

(2) 从一个 N_i 节点开始,它所能到达的所有的内容节点中具有最大被购买概率 $P(G|N_i)$(在该节点的被购买概率,即物品在该节点的被购买概率)的物品,被作为内容部分放到该 N_i 节点上.

第 1 种方法的缺点是:以 N_{31} 为例,在该节点应该放入 G_6,但群体用户到达该节点的主要购买目的是 G_1 而不是 G_6.第 2 种方法是对第 1 种方法的改进,可以从我们建立的 HMM 中直接得出.但如果站点的结构是复杂的网状层次结构,如图 2 所示,那么当物品在导航站点竞争时,第 2 种方法又会遇到如下问题:

设每一个导航页的内容部分只能存放一个物品,那么,根据该方法:N_{11} 节点放入 G_6 物品,N_{21} 节点放入 G_1 物品,N_{22} 节点放入 G_4 物品.但是,因为 G_3 是仅次于 G_6 的被购买的物品,N_{21} 和 N_{22} 节

点中至少有一个应该放入 G_3 物品.这就产生如下问题:当站点的结构不是树形结构时,如果一个商品经常被购买,那么网站的设计者会开辟很多到它的通路,但是这些通路会相对降低其在这些通路的节点上的被购买概率;而那些只有一条路的商品(相对来说被购买概率低),因为它们的被购买概率不被分散,往往会被排到前面.

解决这个问题的思路就是采用我们的办法来求路径.在上述例子中,只选择通过 N_{21} 节点的路径,最终 G_3 物品放置 N_{21} 节点上.根据我们的算法,N_{22} 放入 G_4,N_{21} 放入 G_3,N_{31} 放入 G_1,...,其结果明显优于第 2 种简单统计方法.如果整个站点的结构是一个严格的树状结构,那么,我们的方法等价于第 2 种简单统计方法.

4.3 具有真实背景的实验

Proxy 代理是一个理想的实验环境.通过代理,我们可以得到一些用户对某个站点的访问记录(等价于在该站点上的访问日志)和该站点的 Web 页面结构.至于购买的物品集,网上的 MP3 音乐站点是一个很好的在线零售商务站点的模拟.由于有很多用户非常喜欢下载 MP3 音乐(即相当于用户购买物品),所以,我们就可以通过 Proxy 代理服务器对某一个 MP3 音乐站点进行监控,以得到我们需要的全部数据.我们分析了在一段时间(3 个月)内,3 个 MP3 站点的访问情况,并建立了 HMM 以进行挖掘,见表 6.

Table 6 The mining result of three MP3 music sites

表 6 对 3 个 MP3 音乐站点挖掘的情况

Site NO①	Site type②	The number of the access transactions③	The number of the large items④	The total number of the paths to the large items⑤	The number of the levels⑥	Compare with the second simple statistic approach⑦
Site1	FTP	3 124	10	10	4	Same⑧
Site2	HTTP	1 027	12	31	4	There are 3 large items that are put in irrational way; HMM hasn't the problem (j is 50% of the total paths).⑨
Site3	HTTP	1 311	13	57	5	There are 4 large items that are put in irrational way; HMM hasn't the problem (j is 60% of the total paths).⑩

①站点,②类型,③访问事务个数,④大项集个数,⑤到大项集的总路径数,⑥层数,⑦与第 2 种简单统计方法的比较,⑧一致,⑨简单统计方法有 3 个大项集分布不合理;HMM 方法没有这种情况(j 为总路径的 50%),⑩简单统计方法有 4 个大项集分布不合理;HMM 方法没有这种情况(j 为总路径的 60%).

实验结果说明,如果一个在线零售站点在设计上越接近网状结构,Web 页面结构层次越深,页面之间的结构越复杂,则算法效果越好.该算法适用于处理大规模的、复杂的 Web 站点自适应问题.

5 结论以及将来的工作

我们的方法本质上是 Web 访问信息挖掘(Web usage mining)中的一种推荐(recommendation)方法,即根据群体用户对在线零售电子商务站点的访问,在 Web 站点上,推荐根据以前群体用户的兴趣而得到的知识,以加速当前群体用户对站点的访问.我们首次将 HMM 引入到在线零售站点的购买路径分析之中,拓展了 HMM 的应用领域,较好地解决了在线零售电子商务网站的群体自适应问题,而且该方法也可以被广泛用于 Web 站点的路径分析上.在我们的方法中,建立模型不需要复杂的训练过程,购买概率和转移概率都比较容易计算.

我们进一步的工作将不仅是 Web 访问信息挖掘中的推荐方法,而且也是预测(prediction)方

法.通过将这两种方法结合起来,我们不但能够在 Web 站点上推荐我们所发现的用户的兴趣,而且也将能够预测用户的兴趣.

References:

[1] Buchner, A. G., Mulvenna, M. D. Discovering internet marketing intelligence through online analytical web usage mining. SIGMOD Record, 1998,27(4):54~61.

[2] Agrawal, R., Srikant, R. Fast algorithms for mining association rules. In: Mehta, M., ed. Proceedings of the 20th VLDB Conference. Santiago, Chile: AAAI Press, 1994. 487~499.

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

[4] Zaiane, O. R., Xin, M., Han, J. Discovering Web access patterns and trends by applying OLAP and data mining technology on Web logs. In: Han, J., ed. Proceedings of the Advances in Digital Libraries. CA, Santa Barbara: IEEE Press, 1998. 19~29.

[5] Joachims, T., Freitag, D., Michell, T. Web-Watcher: a tour guide for the world wide web. In: Mitchell, T., ed. Proceedings of the 15th International Joint Conference on AI. Nagoya: AAAI Press, 1997. 770~775.

[6] Fink, J., Kobsa, A., Nill, A. User-Oriented adaptivity and adaptability in the AVANTI project. In: Fink, J., ed. Designing for the Web: Empirical Studies. Redmond, WA: Microsoft Press, 1996.

[7] Perkowitz, M., Etzioni, O. Adaptive Web sites: automatically synthesizing Web pages. In: Umeshwar, Dayal, ed. Proceedings of the AAAI'98. Madison, Wisconsin: AAAI Press, 1998. 727~732.

[8] Rabiner, L. R. A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE, 1989,77(2):257~286.

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

Adaptive Online Retail Web Site Based on Hidden Markov Model*

WANG Shi, GAO Wen, HUANG Tie-jun, MA Ji-yong, 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: There is a problem in online retail: the conflict between the different interests of all customers to different commodities and the commodity classification structure of Web site. This problem will make most customers access overabundant Web pages. To solve the problem, the Web page data, server data, and marketing data are mined to build a hidden Markov model. The authors use association rule discovery to get the large item set. Viterbi algorithm is used to find some paths that come from the root Web page to the Web page that the center of the large item set is in. This large item set is marked in the nodes that are in the paths. Through these steps, one can calculate all item sets and mark them in these paths. The large item sets will compete in the nodes for the limited space. Through this method the Web site will adjust itself to reduce the total access time of all users. This method can also be used in analysis of paths, advertisements, and reconstructing the Web site.

Key words: Web mining; hidden Markov model; association rule; self-adaptation

* Received November 24, 1999; accepted February 20, 2000

Supported by the National High Technology Development Program of China under Grant No. 863-306-JD06-03-4