基于 Markov 链和关联规则的 Web 访问预测模型

林惠珍, 杨晨晖*, 李翠华, 陈希友
(厦门大学信息科学与技术学院, 福建 厦门 361005)

摘要: 用户访问预测是根据用户的历史访问信息和当前的访问路径预测用户下一步或将来可能访问的页面. 因此可以利用预测结果提高服务器的性能, 提高缓存的利用率和为用户提供个性化服务. 提出了基于 Markov 链和关联规则的预测模型 MAPM(Markov chain and association rule prediction model), 该模型首先使用二阶 Markov 链找到用户下一步或将来可能访问的页面集, 生成预测候选集; 然后再使用二项关联规则从正向和反向两个角度修正 Markov 的预测结果, 从而生成最后的预测页面.

关键词: Web 日志挖掘; Markov; 关联规则; 访问预测
中图分类号: TP 311 文献标识码: A 文章编号: 0438-0479(2010)04-0476-06

进入 21 世纪以来, Internet 爆炸式的增长, 使得人们真正体验到信息时代的优越性. 但是作为 Internet 的管理和研究者, 必须利用好 Internet 迅速增长带来的巨大数据资源, 并从中挖掘出有意义的知识来指导 Internet 的建设, 从而开启一个更人性化、更智能化的新 Internet 时代[1].

Web 日志挖掘是 Web 挖掘领域的一个重要研究方向, 它将数据挖掘技术成功地应用于 Web 环境的知识发现. 挖掘 Web 日志可以让 Web 站点的管理者更好地理解站点的访问情况、可以分析用户的频繁访问路径和行为规律、理解用户的行为意图、调整 Web 结构、提高 Web 应用的效率以及为用户提供个性化服务[2]. 这些功能和服务的核心是用户访问预测, 即根据用户的历史访问信息和当前访问路径, 预测该用户下一步或将来可能访问的页面. 我们可以把预测的结果提前发送给用户[3], 当用户访问这些页面时, 只要从本机或本地缓存读取即可, 这样就提高了用户的访问效率和通信资源的利用率[4]. 我们可以根据用户的爱好, 把这些预测页面部署在当前页面的醒目位置, 动态调整页面的结构, 为用户提供个性化服务, 改善用户查阅信息的效率[5]. 此外, 还可以利用预测的页面找到用户的频繁访问路径, 从而动态调整 Web 的结构, 使得 Web 结构更符合用户的访问习惯[6].

1 当前研究现状

用户访问预测的研究主要可以分为 3 类:

1) 基于关联规则. Gery 等[7] 介绍了如何利用关联规则预测用户的下一个访问页面; 在该文中, 作者还介绍了频繁序列的挖掘, 利用 N-Gram 找出所有 N 长度序列, 再计算各个序列的支持度和可信度, 从而找到支持度和可信度大于阈值的频繁序列. 无论是关联规则还是 N-Gram 都必须多轮扫描数据, 但是这些数据往往是十分庞大的, 因此该方法实时性很难保证.

2) 基于 Markov 预测. Zuckerman 等[8] 最先推出了基于 Markov 模型的用户访问预测, Markov 模型是一种简单而经典的预测模型, 但高阶 Markov 模型所覆盖的状态空间十分庞大, 导致计算复杂度过高; 而低阶 Markov 模型预测准确率较低. 邢永康等[9] 提出并建立了一种基于用户分类的新模型——多 Markov 链预测模型. 该模型提高了预测准确率, 但是时间复杂度是 $O(n^2)$, n 是 Web 站点的页面数量. 因此该预测模型的时间复杂度较高.

3) 基于路径相似的预测. Gündüz 等[10] 提出了基于点击流树的预测算法, 详细介绍了点击流树的构造算法和基于点击流树的预测算法. 预测过程首先在点击流树上查找与当前访问路径相似的历史访问路径, 然后再根据历史访问路径做出预测. 该模型的主要问题是必须在内存中维护一颗点击流树, 而点击流树是根据历史的访问记录建立的, 但是一个网站的历史访

收稿日期: 2009-11-16
基金项目: 国家自然科学基金(40627001)
* 通讯作者: chyang@xmu.edu.cn

问量会随着时间的推移而递增, 导致数据量庞大, 因此它的空间复杂度太大.

本文针对预测算法的特点, 在前人研究的基础上, 提出了基于 Markov 链和关联规则的用户访问预测模型, 该模型改进了 Markov 链二阶转移矩阵的计算方法: 首先用二阶 Markov 链生成预测候选集, 再利用二项关联规则修正预测结果. 该模型在保证预测准确率的同时, 能在线性时间内完成在线预测.

2　预测的概率模型

用 $P=\{p_1, p_2, \cdots, p_n\}$ 表示 Web 站点全部页面的集合; 用 $S=\{s_1, s_2, \cdots, s_m\}$ 表示所有的会话集合, 其中 $s_l=\{\cdots, p_i, \cdots, p_j, \cdots\}$, $1\leqslant l\leqslant m$, $1\leqslant i, j\leqslant n$; 用 $V=\{v_1, v_2, \cdots, v_k\}$ 表示用户当前的访问序列, 则下一访问页面 v_{k+1} 满足

$$v_{k+1}=\arg\max_{p\in P}\text{prob}(v_{k+1}=p\mid V)=\arg\max_{p\in P}\text{prob}(v_{k+1}=p\mid v_k, \cdots, v_2, v_1)=\arg\max_{p\in P}\text{prob}(p, v_k, \cdots, v_2, v_1)/\text{prob}(v_k\mid v_{k-1}, \cdots, v_2, v_1)\cdots\text{prob}(v_2\mid v_1)\,\text{prob}(v_1). \quad (1)$$

式(1)的值可以通过集合 P 和 S 计算, 并且 k 值越大, 结果越准确; 但是如果 k 和 S 太大, 会导致计算复杂度过高[11]. 因此式(1)理论上可行, 实际应用受到很大限制. 为了解决这个问题, 很多学者假设用户访问站点的过程是一个 Markov 过程.

3　Markov 过程

很多确定性现象都遵循这样的演变规则, 即由时刻 t_i 某过程所处的状态, 可以确定该过程在 $t>t_i$ 所处的状态, 而无需借助 t_i 以前所处的状态. 可以把上述规则延伸到随机过程中, 满足此规则的随机过程, 称为 Markov 过程[11].

定义 1　设随机过程 $\{X(t), t\in T\}$ 的状态空间为 I, 其中 T 表示时间维度. 如果对时间 t 的任意 n 个数值 $t_1<t<\cdots<t_n$, $n\geqslant 3$, $t_i\in T$ 在条件 $X(t_i)=x_i$, $x_i\in I$, $I=1, 2, \cdots, n-1$ 下, $X(t_n)$ 的条件分布函数等于在条件 $X(t_{n-1})=x_{n-1}$ 下 $X(t_n)$ 的条件分布函数, 即

$$p\{X(t_n)=x_n\mid X(t_{n-1})=x_{n-1}, \cdots, X(t_2)=x_2, X(t_1)=x_1\}=p\{X(t_n)=x_n\mid X(t_{n-1})=x_{n-1}, \cdots, X(t_{n-k})=x_{n-k}\}, (t_i\in T, x_i\in I). \quad (2)$$

则称随机过程 $\{X(t), t\in T\}$ 是 k 阶 Markov 过程, 记为 k-Markov[11].

4　基于 Markov 链和关联规则的预测模型

在第 2 节中, 分析了预测的概率计算模型, 由于式(1)的计算复杂度过高. Zukerman 等[8] 提出了 Markov 链用户访问预测模型, 它将用户的访问过程抽象为一个特殊的随机过程——齐次离散 Markov 链, 用转移概率描述用户的浏览特征, 并基于此预测用户的访问.

4.1　二阶 Markov 链模型

假设 **1**(Markov 性假设)　假设用户在 Web 上的浏览过程是一个特殊的随机过程——齐次离散 Markov 链. 即设离散随机变量 X 的值域为 Web 空间中的所有网页的集合, 则一个用户在 Web 中的浏览过程就构成一个随机变量 X 的取值序列, 并且该序列满足 Markov 性[9].

定义 2　二阶 Markov 链预测模型可以表示为一个四元组 $\{(x_i, x_j), \boldsymbol{p}(1), \boldsymbol{p}(2), x\}$, 其中 (x_i, x_j) 表示用户最近访问的两个页面, $\boldsymbol{p}(1)$ 表示一阶概率转移矩阵, $\boldsymbol{p}(2)$ 表示二阶转移矩阵, x 表示预测页面. $p_{ij}(1)$ 表示访问页面 x_i 之后访问 x_j 的概率; $p_{ij}(2)$ 表示访问页面 x_i 之后, 访问 $(*, x_j)$ 的概率, 其中"*"代表任意网页. 一阶和二阶转移矩阵的学习算法如下:

算法 1　一阶概率转移模型 $\boldsymbol{p}(1)$ 的学习算法

输入: 用户会话集 S.

输出: 概率转移模型 $\boldsymbol{p}(1)$.

```
while(s_i ∈ S)
{while((x_u, x_v) ⊆ s_i)
{p_{x_u x_v}++; //计数递增
p_{x_u}++;
x_u = x_u.next; //访问 s_i 中的下一个页面
x_v = x_v.next;
}
s_i = s_i.next; //访问 S 中的下一个会话
}
foreach(x_u, x_v)
{p_{x_u x_v} = p_{x_u x_v}/p_{x_u}}; //计算转移概率
```

二阶概率转移模型 $\boldsymbol{p}(2)$ 的学习算法和算法 1 类似, 只要把第 2 行用 while($(x_u, *, x_v)\subseteq s_i$) 替代即可, * 号表示任意页面.

分析算法 1 可知, 它的时间复杂度是 $O(\lambda|S|)$, 空间复杂度是 $O(n^2)$, 其中 $|S|$ 代表用户会话集(训练数

据集)的大小, n 代表 Web 站点页面的个数.

4.2 二阶 Markov 链预测算法(MPA)

用 n 维向量 $\boldsymbol{v}(t)=(v_1, v_2, \cdots, v_n)$ 表示用户 t 时刻的状态,如果某用户 t 时刻的访问页面为 x_j,则让 $\boldsymbol{v}(t)$ 的第 j 维为 1,其余维都为 0. 用 n 维向量 $\boldsymbol{x}(t)=(x_1, x_2, \cdots, x_n)$ 表示 t 时刻用户访问各页面的概率,即 $\boldsymbol{x}(t)=(p(X_t=x_1), p(X_t=x_2), \cdots, p(X_t=x_n))$[10]. 假设某用户的 t 时刻的页面访问序列为 $(\cdots, x_i, x_j)$,则当前状态 $\boldsymbol{v}(t)=(v_1=0, v_2=0, \cdots, v_j=1, v_{j+1}=0, \cdots, v_n=0)$,前一状态 $\boldsymbol{v}(t-1)=(v_1=0, v_2=0, \cdots, v_i=1, v_{i+1}=0, \cdots, v_n=0)$,则下一时刻用户访问各个页面的概率为

$$\boldsymbol{x}(t+1)=\alpha_1\boldsymbol{v}(t)\boldsymbol{p}(1)+\alpha_2\boldsymbol{v}(t-1)\boldsymbol{p}(2), \quad (3)$$

其中 α_1 为一阶转移矩阵的权值, α_2 为二阶转移矩阵的权值,令 $\alpha_1+\alpha_2=1$,它们的取值在实验分析部分做详细介绍. 最后从 $\boldsymbol{x}(t+1)$ 中选出概率最大的页面作为预测结果.

算法 **2** MPA

输入:用户的当前访问序列 $(\cdots, x_i, x_j)$.

输出:预测页面或预测页面集.

根据当前访问序列构造 $\boldsymbol{v}(t)$ 和 $\boldsymbol{v}(t-1)$.

$\boldsymbol{x}(t+1)=\alpha_1\boldsymbol{v}(t)\boldsymbol{p}(1)+\alpha_2\boldsymbol{v}(t-1)\boldsymbol{p}(2)$.

返回向量 $\boldsymbol{x}(t+1)$ 中值最大的维对应的页面(或 $\boldsymbol{x}(t+1)$ 中值大于某阈值的维对应的页面集).

其中算法 2 的时间复杂度是 $O(n)$,其中 n 是 Web 站点页面的数量.

4.3 二项关联规则

页面的二项关联规则可以体现出页面之间被组合访问的关联度. 本文使用矩阵表示页面二项关联规则,矩阵的列和行分别是 Web 站点的页面,矩阵中存取的是页面被组合访问的支持度. 只要扫描一轮会话集即可建立二项关联矩阵,并根据新会话增量更新矩阵.

4.4 基于 Markov 链和关联规则的预测模型(MAPM)

k 阶 Markov 链预测模型忽略了较早的历史访问知识,假设下一访问页面只与最新的 k 个页面有关系,简化了预测模型,减少了计算时间,但是预测准确度也随之下降. 针对这个问题,本文提出了 MAPM 算法. 首先用 Markov 链预测模型返回下一步有可能会访问的页面集 m 和这些页面对应的 Markov 预测概率;再利用二项关联规则从正向和反向两个角度来修正预测结果. 假设当前用户已访问的页面序列为 $v=\{p_1, p_2, \cdots, p_s\}$,其中 s 代表一次会话的长度, Markov 预测的结果集 $m=\{r_1, r_2, \cdots, r_t\}$,其中 t 代表预测算法推荐的页面数,且对应的 Markov 预测概率 $mp=\{mp(r_1), mp(r_2), \cdots, mp(r_t)\}$. 以下是反向和正向修正过程的介绍:

反向修正过程:对于任意的 $r_i\in m$,分别计算 v 中所有页面与 r_i 的可信度,如果所有的可信度都小于阈值 min_rule,则无论 $mp(r_i)$ 的值多大,都从 m 中删除页面 r_i,故反向修正过程具有否决权. 由于所有的可信度都小于阈值 min_rule,我们可以认为页面 r_i 与 v 中的所有页面不存在超链接关系,故用户下一步访问该页面的可能性很小.

正向修正过程:对于任意的 $r_i\in m$,分别计算 v 中所有页面与 r_i 的可信度,如果存在页面 $p_j\in v$,且 p_j 与 r_i 的可信度($\mathrm{conf}(p_j\rightarrow r_i)$)大于阈值 max_rule,则 r_i 的预测概率为

$$\mathrm{preditProb}(r_i)=\lambda_1\cdot mp(r_i)+\lambda_2\cdot\sum_j w_j\cdot \mathrm{conf}(p_j\rightarrow r_i),$$

其中 $w_j=j/|v|$, $\lambda_1+\lambda_2=1$, λ_1 代表 Markov 的预测权值, λ_2 代表关联规则的预测权值,它们的取值在实验分析中做详细讨论. w_j 代表规则 $\mathrm{conf}(p_j\rightarrow r_i)$ 的权值,由于在访问序列 v 中, p_j 离当前时间越近,这条规则的参考价值越大,反之参考价值越小.

算法 **3** MAPM

输入:用户的当前访问序列 $(v=p_1, p_2, \cdots\cdots, p_s)$,阈值 min_rule 和 max_rule.

输出:预测页面

利用二阶 Markov 预测算法计算出用户有可能访问的页面集 m 和对应的 Markov 预测概率 mp;

```
foreach(r_i ∈ m)
{preditProb(r_i) = mp(r_i); // 初始化预测概率
foreach (p_j ∈ v)
{conf(p_j → r_i) = R_{p_j r_i}; } // 从关联矩阵 R 中查找相应规则的可信度
// 逆向修正
foreach(p_j ∈ v)
{ if ( foreach p_j ∈ v, (conf(p_j → r_i)) < min_rule
{remove r_i from m; } // 从 m 中删除 r_i
}
foreach(p_j ∈ v)
{ if ( conf (p_j → r_i)) > max_rule)
{preditProb(r_i) + = λ· w_j · conf(p_j → r_i); } // 计算预测概率
}
}
返回 predictProb 中值最大的页面
```

算法 3 的时间复杂度是 $O(n)$,其中 n 是 Web 站点页面的数量.

证明　算法的第 1 行的时间复杂度是 $O(n)$, 第 2 行的循环执行 $|m|$ 次, 第 4, 6, 8 行的循环都是执行 $|v|$ 次. m 是 Web 站点页面的子集, 它的大小由 Markov 预测算法调控, 一般都大大小于 n. $|v|$ 是用户访问的长度, 是服从泊松分布的随机变量, 故它的数学期望是 λ 所以算法的时间复杂度为 $O(n+\lambda|m|)$, 通常情况下, $\lambda|m|$ 大大小于 n, 所以 $O(n+\lambda|m|)$ 可以近似为 $O(n)$.

5　实验结果及分析

5.1　实验数据介绍

实验数据来源于 1998 年 2 月微软服务器 "www.microsoft.com" 用户一周的访问日志(从 http://kdd.ics.uci.edu/databases/msweb/msweb.html 上下载). 实验训练数据和测试数据的页面数和会话数如表 1.

表 1　实验数据

Tab. 1　Experimental data

数据分类	页面数	会话数	长度 ≥2 的会话数
训练数据	294	32711	26168
测试数据	294	5000	3452

训练数据中各个页面被访问的频度分布情况如图 1.

测试数据中, 会话长度(访问路径的长度)大于等于 $x(x \geqslant 2)$ 的会话数分布如图 2.

5.2　预测准确率定义

预测算法最重要的两个属性是时间复杂度和预测准确率; 预测准确率很大程度上取决于 Web 访问日志是否规律. 结合实际的应用价值, 本文给出了两种预测准确率的计算方法, 定义如下:

定义 **3**　1-预测准确率指的是预测的页面在下一步被访问的比率.

定义 **4**　x-预测准确率指的是预测的页面在将来(从当前至会话结束)被访问的比率.

5.3　基于 Markov 链和关联规则的预测分析

在以往的 Markov 预测中, k 阶转移矩阵是通过定理 $p(k)=p^k(1)$[11] 计算的(记为方法 1). 而本文的 k 阶转移矩阵是通过学习算法建立的(记为方法 2).

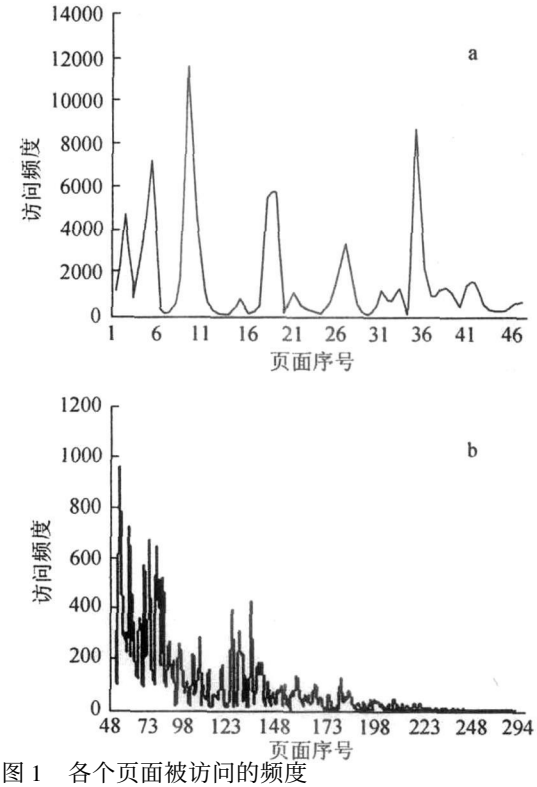

图 1　各个页面被访问的频度

a. 1~ 47 页; b. 48~ 294 页

Fig. 1　The access frequency of each page

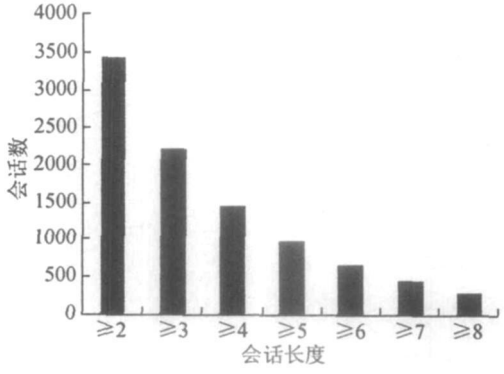

图 2　会话数分布图

Fig. 2　Session number distributing graph

在算法 2 中 α_1 代表一阶转移概率的权值, $\alpha_2(\alpha_2=1-\alpha_1)$ 代表二阶转移概率的权值. 在实验过程中, 让 α_1 从 1.0 逐步等分下降到 0.1, 计算 10 次预测准确率. 其中, 在方法 1 中, α_1 为 0.7 时, 预测准确率最高; 在方法 2 中, α_1 为 0.5 时, 预测准确率最高. 此外, 方法 2 的平均 1-预测准确率和平均 x 预测准确率分别比方法 1 高 4.25% 和 5.68%. 经实验分析和观察可知, 当 α_1(1 阶转移概率的权值)取 0.5 时, MPA 算法的预测准确率最好. 因此在 MAPM 算法中, α_1 为 0.5; 经过反复的实验, 当 min_rule 取 10%、max_rule 取

20%、λ(代表 Markov 的预测权值)取 0.9 时,λ(关联规则的预测权值,λ= 1- λ)为 0.1,算法 MAPM 的 1-预测准确率最好. 图 3 是 MAPM 和 MPA 的最大 1-预测准确率的比较图.

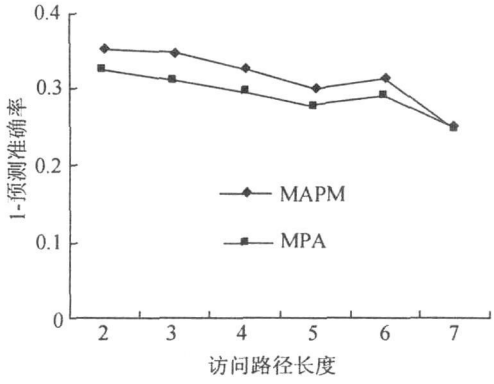

图 3 MPA 和 MAPM 算法最大 1-预测准确率比较图

Fig. 3 Comparison of maximum 1-prediction accuracy between MPA and MAPM

在计算 x-预测准确率的过程中,λ 和 λ 分别为 0.5 时(其余参数未变),MAPM 算法的 x-预测准确率最好. 图 4 是 MPA 和 MAPM 算法最大 x-预测准确率比较图.

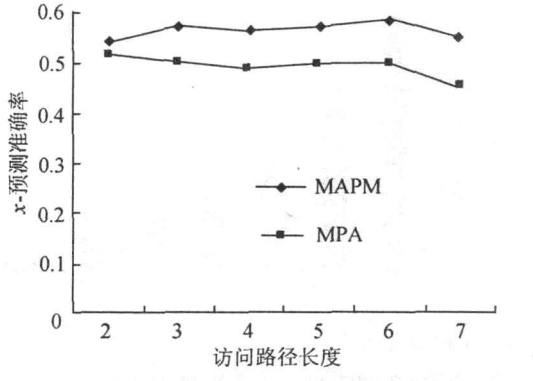

图 4 MPA 和 MAPM 算法最大 x-预测准确率比较图

Fig. 4 Comparison of maximum x-prediction accuracy between MPA and MAPM

在 MPA 算法中,无论是 1-预测准确率还是 x-预测准确率,α 都是取 0.5 时(代表一阶转移概率和 2 阶转移概率的权值相等)最好,也就是说当前访问页面和上一时刻的访问页面在 Markov 预测过程中的作用是等价的. 然而,在 MAPM 算法中,λ 为 0.1 时(λ 为 0.9),1-预测准确率最好,而 λ 为 0.5 时(λ 为 0.5) x-预测准确率最好. λ 为 0.1 说明关联规则修正对结果的影响较小;而 λ 为 0.5 说明关联规则修正对结果影响较大. 根据经验可知,在一步转移过程中,当前已访问页面和下一访问页面之间 Markov 性更强一些;然而,在已访问页面和将来的访问页面之间关联关系更密切.

5.4 实验结果和点击流树(CST)对比

CST 是基于路径相似预测的最典型算法,该算法首先根据 Web 日志构建 CST,然后再利用该树进行预测[8]. 为了方便对比,让 MAPM 推荐预测概率前 3 名的 3 个页面,然后计算平均 1-预测准确率.

从表 2 可以看出,MAPM 的预测准确率略高于 CST. CST 的实验环境:奔腾 4CPU(2.4 GHz)+ 552 MB 内存;本文的实验环境:奔腾 4CPU(频率 2.93 GHz)+ 552 MB 内存. 因此折算 CPU 频率的差别后,MAPM 预测效率比 CST 高 2.6 倍.

表 2 MAPM 与 CST 实验结果对比

Tab. 2 Comparison between MAPM and CST

算法	数据集	平均 1-预测准确率	平均预测时间/ms
MAPM	Msweb	0.559	0.025
CST	ClarkNet	0.551	0.080

6 小 结

Web 日志挖掘是 Web 挖掘领域的一个重要研究方向,它将数据挖掘技术成功地应用于 Web 环境的知识发现,用户访问预测是 Web 日志挖掘的重要分支. 通过对 Web 日志进行挖掘,可以帮助站点管理者发现用户访问页面的行为规律,理解用户的行为意图,从而为用户提供个性化服务;挖掘 Web 日志,可以预测用户将来的访问页面,并把预测的页面提前发送给用户,从而改善用户的访问效率和提高网络资源的利用率;还可以利用预测的页面优化代理服务器或 Web 服务器的缓存置换策略,从而改善服务器的性能和设计.

本文首先介绍了用户访问预测研究的现状,分析了当前典型预测算法的优缺点和预测的概率模型,详细介绍了二阶 Markov 预测模型和基于 Markov 和关联规则的预测模型. 由于用户访问 Web 站点的行为具有 Markov 性,而访问的页面之间具有关联性,所以用 Markov 和关联规则相结合的方法能较好的与用户的行为特征相吻合. 实验结果证明了 MAPM 具有较好的预测准确率和预测效率.

参考文献:

[1] Castellano G, Fanelli A M, Torsello M A. Computational intelligence techniques for Web personalization[J]. Web Intelligence and Agent Systems, 2008, 8: 253-272.

[2] Subhash K S, Kulkarni U V. A new approach for on line recommender system in Web usage mining[C]// Proceedings of the 2008 International Conference on Advanced Computer. Washington DC, USA: IEEE Computer Society, 2008: 973-997.

[3] Albrecht D W, Zukerman I, Nicholson A E. Pre-sending documents on the WWW: a comparative study[C]// Proceedings of the 16th International Joint Conference on Artificial Intelligence, IJCAI' 99. San Franciso, CA, USA: Morgan Kaufmann Publishers Ins, 1999.

[4] 曾春, 邢春晓, 周立柱. 个性化服务技术综述[J]. 软件学报, 2002, 13C(10): 1952-1961.

[5] Sun Yang, Li Huajing, Isaac G Councill, et al. Personalized ranking for digital libraries based on log analysis[C]// Proceeding of the 10th ACM Workshop on Web Information and Data. California, USA: ACM, 2008: 133-140.

[6] Tao Yuhui, Hong Tzungpei, Su Yuming. Web usage mining with intentional browsing data[J]. Expert Systems with Applications, 2008, 4: 1893-1904.

[7] Gery M, Haddad H. Evaluation of Web usage mining approaches for user's next request prediction[C]// WIDM' 03. New Orleans, USA: ACM, 2003: 74-81.

[8] Zuckerman I, Albrecht D, Nicholson A. Predicting user's requests on the WWW[C]// Proceedings of the 7th International Conference on UserModeling. New York: Springer, 1999: 275-284.

[9] 刑永康. 马少平. 多 Markov 链用户浏览预测模型[J]. 计算机学报, 2003(11): 1510-1517.

[10] Gündüz S. Tamer M Özsu. A Web page prediction model based on click-stream tree representation of user behavior[C]// SIGKDD' 03. USA: ACM, 2003: 535-540.

[11] 盛骤, 谢式千, 潘承毅. 概率论与数理统计[M]. 北京: 高等教育出版社, 2001.

Web Access Prediction Model Based on Markov Chain and Association Rule

LIN Hui-zhen, YANG Chen-hui*, LI Cui-hua, CHEN Xi-you

(School of Information Science and Technology, Xiamen University, Xiamen 361005, China)

Abstract: User access prediction is the core of Web log mining, which predicts the next access page or the future access pages according to the history access information and the current access path. We can make use of the prediction result to improve the Web server performance, increase the cache utilization and provide users with personal service. In this paper we proposed Markov chain and association rule prediction model (MAPM). This model uses second-order Markov chain to find the pages which users may visit in next step or future, so as to generate the candidate prediction page set, and then corrects the Markov prediction result on forward and reverse perspective according to the two-items association rules, and gets the last prediction page.

Key words: Web log mining; Markov; association rule; access prediction