Web 浏览预测的 Markov 模型综述*)

林文龙　刘业政　姜元春

(合肥工业大学管理学院电子商务研究所　合肥 230009)

摘　要　Web 访问模式挖掘研究的一个重要议题是 Web 浏览预测，Markov 模型是一种经典的 Web 浏览预测模型。本文首先介绍了基本 Markov 浏览预测模型，包括基本 Markov 浏览行为模型，模型的学习训练及其在 Web 浏览预测问题中的应用；然后重点分析了扩展的 Markov 浏览预测模型，包括一序组合预测模型、高序模型、混合模型、隐 Markov 模型、连续时间 Markov 模型等，综述了各种扩展模型所考虑的浏览预测问题的本质出发点、模型的学习方法及预测方法，最后分析了 Markov 浏览预测模型有待进一步研究的问题。

关键词　访问模式挖掘，浏览预测，Markov 模型

Web Navigation Prediction Based on Markov Model—A Survey

LIN Wen-Long　LIU Ye-Zheng　JIANG Yuan-Chun

(Institute of E-Business, School of Management, Hefei University of Technology, Hefei 230009)

Abstract　Web navigation prediction is one of the most important topics for discussion in research area of Web navigation pattern mining. Markov is one kind of traditional Web navigation prediction model. This paper first introduces basic Markov Web navigation prediction models, which include basic Markov Model of navigation behaviors, its training method and its application in Web navigation prediction problem. Then several extended Markov Web navigation prediction models are introduced, which include one-order combined prediction models, higher order models, mixture Markov models, hidden Markov models, continuous-time Markov models and so on. Then the essences of each kind of extended models and their learning and predicting methods are summarized. Finally some problems of Markov Web navigation prediction models are pointed out for further research.

Keywords　Navigation pattern mining, Navigation prediction, Markov model

1　引言

随着 WWW 的快速增长，如何准确理解用户在 WWW 数字环境下的行为模式，给 Web 挖掘研究人员提出了新的挑战。Web 访问模式挖掘[1~3]是从 Web 使用数据中分析、理解、获取用户的访问行为模式，用以改进文档预取系统[4]、预送系统[5]、Web 缓存性能[6,7]，提高个性化推荐系统的推荐命中率[8]，优化站点结构[9]等。Web 访问模式挖掘研究的一个重要议题是 Web 浏览预测。按照采取的方法不同，浏览预测可以分为确定性方法和不确定性方法；按照采用的数据源对象的不同，则可以分为基于内容的方法和基于协同过滤的方法，基于协同过滤的方法又可以分为基于内存的和基于模型的。其中基于模型的方法是目前应用得比较多的方法。由于 Web 信息空间的复杂性、实时性、动态性，访问行为的有序性、异质性，访问数据的海量性，如何建立一个准确有效的用户浏览预测模型是众多 Web 挖掘研究人员的重要研究课题。已有多种方法可以对 Web 用户的浏览预测问题进行建模分析，如 OLAP[10]、神经网络[11]、贝叶斯网络[12]、聚类[13]、关联规则[14]、Markov 等。

用于 Web 浏览预测问题的 Markov 模型是一种经典的概率统计模型，在语音识别[15]、基因序列分析[16]等领域广泛应用。最早采用 Markov 模型分析 Web 用户访问行为的是 Bestavros，采用一序隐 Markov 模型预测用户的链接选择[17]。尽管有一些研究和证据表明，Web 用户的访问行为是非 Markov 性的[18]，但是 Markov 考虑了 Web 访问行为的动态性与有序性。Jespersen 等则以超文本概率文法(Hypertext Probabilistic Grammar)这种典型的基于访问行为 Markov 性假设的方法为代表，系统研究了 Markov 性假设下进行 Web 访问模式挖掘所得到的模式质量问题，证明 Markov 模型是一种适合对 Web 浏览预测问题进行建模的工具[19]。

本文将重点分析面向 Web 浏览预测问题的 Markov 模型，内容安排如下：第 2 节介绍基本 Markov 浏览预测模型，包括基本 Markov 浏览行为模型、模型的学习训练以及模型在 Web 浏览预测问题中的应用；第 3 节分析扩展 Markov 浏览预测模型，包括一序组合预测模型、高序模型、混合模型、隐 Markov 模型、连续时间 Markov 模型等；第 4 节分析 Markov 浏览预测模型有待进一步研究的问题；最后给出总结。

2　基本 Markov 浏览预测模型

2.1　基本 Markov 浏览行为模型

定义 1(Markov 链)　设随机变量序列 $\{X_t, t=0,1,2,\cdots\}$ 的状态空间 $X=\{x_1,x_2,\cdots,x_n\}$ 是离散的，如果对于 $\forall t\geqslant 0$ 及任意的状态 $i,j,i_0,\cdots,i_{t-1}\in X$ 都有

$$P\{X_{t+1}=j\mid X_t=i,X_{t-1}=i_{t-1},\cdots X_0=i_0\}=P\{X_{t+1}=j\mid$$

*)本文受国家自然科学基金项目(70672097)和国家自然科学基金重点项目(70631003)资助。**林文龙**　博士研究生，主要研究领域：Web 数据挖掘；**刘业政**　教授，博导，主要研究领域：电子商务、数据挖掘与 GDSS；**姜元春**　博士研究生，主要研究领域：数据挖掘与 GDSS。

$X_t=i\}$ (1)

恒成立，则称$\{X_t,t=0,1,2,\cdots\}$所构成的随机过程为一个Markov链。记$p_{ij}(t)=P\{X_{t+1}=j|X_t=i\}$，如果$p_{ij}(t)$与时刻$t$无关，即对于任意不同的时刻$t_1,t_2$都有$p_{ij}(t_1)=p_{ij}(t_2)$，则称此Markov链为齐次Markov链[20]。

定义2(游历序列) 用户在Web站点信息空间上游历时，同一会话期间内*url*的请求序列称为用户的游历序列。

定义3(游历序列事务数据库*NSDB*) 群体用户的游历序列构成Web站点的游历序列事务数据库。即$NSDB=\{x^1,x^2,\cdots,x^m\}$，其中向量$x^i(i=1,\cdots,m)$表示用户的游历序列。

假设1(浏览行为的Markov性假设) 假设所有用户的游历序列是一个齐次离散Markov链，则游历序列可以看作是一离散随机变量X的取值序列，该序列满足Markov性。

定义4(基本Markov浏览行为模型) Web用户浏览行为可以表示为一个齐次离散Markov链$M=\langle X,A\rangle$，其中X是一个离散随机变量，值域为$\{x_1,x_2,\cdots,x_n\}$，每个x_i对应一个网页，称为模型的一个状态；A为转移概率矩阵，$A=(p_{ij})=P(X_t=x_j|X_{t-1}=x_i)$表示由状态$x_i$转移到状态$x_j$的转移概率，即$A=(p_{ij})_{n\times n},(i,j\in\{1,\cdots,n\})$。

转移概率矩阵A满足以下两个条件：

$$p_{ij}\geqslant 0 \text{ for all } i,j\in\{1,\cdots,n\} \quad (2)$$

$$\sum_{j=1}^{n}p_{ij}=1 \text{ for all } i\in\{1,\cdots,n\} \quad (3)$$

2.2 基本Markov浏览行为模型的训练

模型的训练主要是通过*NSDB*，对转移矩阵A做参数估计，一般可以采用极大似然估计法或Bayes参数估计法。设转移矩阵A中的待估参数集为p，采用极大似然估计法，首先定义似然函数：

$$l(p)=P(NSDB|p)=\prod_{i=1}^{m}P(x^i|p) \quad (4)$$

令S_{ij}表示在*NSDB*中状态对(x_i,x_j)的出现次数，考虑到条件(2)、(3)，可以得到p的极大似然估计值为

$$p_{ij}^{ML}=\frac{S_{ij}}{\sum_{j=1}^{n}S_{ij}} \quad (5)$$

采用极大似然估计法存在*NSDB*的问题是在数据库中，有相当一部分S_{ij}值为0。另外，当模型的学习数据不充分时，无法对参数做出准确的估值，此时可以采用Bayes参数估计法。考虑转移矩阵第i行的参数集$\{p_{i1},\cdots,p_{in}\}$，给定其先验分布为Dirichlet分布，即

$$P(\{p_{i1},\cdots,p_{in}\})=C\prod_{j=1}^{n}p_{ij}^{\alpha(q_{ij}-1)} \quad (6)$$

式中$\alpha,q_{ij}>0,\sum_{j=1}^{n}q_{ij}=1,C$是标准常量。可以求得参数的后验估计值为

$$p_{ij}^{MAP}=\frac{S_{ij}+\alpha q_{ij}}{\sum_{j=1}^{n}S_{ij}+\alpha} \quad (7)$$

模型的空间复杂度为$O(n^2)$。由于n的值一般都比较大，导致模型训练时的空间开销很大。在浏览预测问题中，当对用户的多步行为进行预测时，需要计算多步转移概率矩阵，此时空间开销显得更加显著。同时由于用户的访问行为受站点链接结构的限制，导致了转移矩阵稀疏性问题。一般采取两种策略降低算法的空间复杂度。一种比较有效的策略是状态聚类法，即在建立模型前首先对站点页面进行聚类，聚类的依据可以是基于页面内容[21]、基于页面功能[22]或是基于站点的目录结构[23]等信息。Zhu Jianhan等则利用矩阵降维压缩技术[24]提高了转移稀疏矩阵的密度。通过将转移相似度小于预定阈值的页面聚合，压缩了转移矩阵的大小[25]，其本质上也是一种状态聚类策略。状态聚类法通过把多个相近状态聚为单一状态，有效缩减了状态空间。采用状态聚类法存在的一个问题是：在对Web浏览进行预测时，预测的粒度较粗(预测的结果是聚集后的页面类别)，是一种折衷的办法。第二种策略是采用一些新颖的数据结构，如采用转移矩阵的三元组表示法或十字链表表示法，对零元素不分配存储空间来压缩稀疏矩阵的存储空间[26]。一种更合适的结构是Markov树[27~29]，Markov树将具有相同访问前缀的用户访问记录存储在从根节点出发的同一条路径中，可以避免矩阵存储结构随n值的增大产生爆炸性的存储开销。

2.3 基本Markov浏览预测模型

Web浏览预测问题[30]是Web访问模式挖掘研究的重要议题。建立有效的用户浏览预测模型，对用户的访问行为做出准确的预测，是各种浏览导航工具、文档预送系统及电子商务推荐系统的关键。

定义5(Web浏览预测) 基于*NSDB*的Web浏览预测问题可以定义如下：

$$p(x,NSDB)\rightarrow url_1(\omega_1),url_2(\omega_2),\cdots,url_l(\omega_l)\in URL$$

其中x是当前用户的游历序列，*URL*是站点页面集合，$\omega_i(i=1,\cdots,l)$是预测权重。

设向量x_t表示用户在时刻t的状态，如果此时用户处于状态x_i，则x_t的第i维等于1，其余各维都为0。根据一序Markov假设与基本Markov浏览行为模型，由下式对用户在时刻t的状态做出预测：

$$x_t=x_{t-1}\times A \quad (8)$$

在向量x_t中，概率值最大的那一维所对应的状态，就是用户在时刻t最可能的状态。一般取概率值最大的*top-N*个状态的集合或是取大于规定概率阈值的状态集作为模型的预测结果。采用这种简单的预测方法，Sarukkai在EPA服务器日志文件[31]上的实验表明，基本Markov浏览预测模型的准确率可以达到70%左右[32]。Davison则在Music Machines[33,34]等多个服务器日志数据集上对模型的预测效果做了进一步的验证，在*top-N*值较大(20)时也获得了很高的预测精度(60%～80%)[35]。

除了对用户的下一步浏览行为做出预测，Zhu Jianhan等进一步预测用户经过L次点击行为后的状态[25]：

$$x_{t+L}=x_t\times A^L \quad (9)$$

Sarukkai则提出路径生成Markov模型(Tour generation Using Markov Models，TUMM)，可以对用户的访问路径进行预测[32]：

$$P(x_{t+L},x_{t+L-1},\cdots,x_{t+1})=P(x_t)\prod_{i=t}^{t+L-1}P(x_{i+1}|x_i) \quad (10)$$

但是Sarukkai没有给出量化评估生成路径的质量问题，也没有一些实际站点对该方法的应用效果进行实证分析，因此路径生成模型在实际应用中是否有效还是个问题。

3 扩展Markov浏览预测模型

3.1 一序组合预测模型与高序模型

(8)式的简单预测模型有待改进的一点是模型没有太多太细致的考虑用户的访问历史，这样就不能很好地区分不同用户的访问行为模式。为了得到更好的预测结果，一种可以充分利用用户访问历史的预测形式是采用对一序多步Mark-

ov模型进行组合预测，合成的方法可以采用取加权平滑值，也可以采用取最大值，即

$$x_t = a_1 x_{t-1} \times A^1 + a_2 x_{t-2} \times A^2 + \cdots + a_h x_{t-h} \times A^h \quad (11)$$

$$x_t = \max\{a_1 x_{t-1} \times A^1, a_2 x_{t-2} \times A^2, \cdots, a_h x_{t-h} \times A^h\} \quad (12)$$

其中，A^h 表示 Markov 链的 h 步转移矩阵。$a_i(i=1\cdots h)$ 是权值，满足 $\sum_{i=1}^{h} a_i = 1$。权值一般采用经验平滑值，对较近的访问记录取较大的权值，即 $1 > a_1 > a_2 > \cdots > a_h > 0$。

考虑用户访问历史的另一种办法是使用高序 *Markov* 模型，设 k 为高序模型的序数，则用户在 t 时刻的状态 x_t 取决于 $x_{t-1}x_{t-2}, \cdots x_{t-k}$，即

$$P(x_t \mid x_{t-1}, \cdots, x_1) = P(x_t \mid x_{t-1}, \cdots, x_{t-k}) \quad (13)$$

Deshpand 等及 Pirolli 等分别就不同序的模型在预测准确率方面的性能做了比较实验，证明在 *top-N* 值较小时，高序模型比低序模型的预测效果要显著得多[36,37]。与一序多步组合预测相对应，对不同序的高序模型也可以进行加权组合的预测，权值的确定一般也是采用经验平滑值，给较高序模型的预测结果以较大的权值。Xing Dongshan 等提出一种准确性投票的权值确定方法，设 k 序模型的预测准确率为 $Accuracy_k$，则定义 k 序模型预测结果的权值为 $PW_k = Accuracy_k / \sum_i Accuracy_i$。他们的实验表明，采用这种方法在各序的预测结果都比基本模型的准确性要高[38]。

但是，高序模型在进行预测时，常常会出现匹配的序列过少，导致较低的预测覆盖率以及过高的状态空间复杂性等问题。在不至于影响模型预测效果的情况下，可以采用一些有效的状态空间剪枝方法。Deshpand 等提出支持度剪枝、置信度剪枝与错误剪枝三种剪枝文法[36]。考虑到训练集中支持度较低的状态在测试集中的支持度值也较小，去除此类状态并不会影响模型的预测精度，这就是支持度剪枝文法；将训练集中状态看作是规则的前件，预测的结果页面看作是规则的后件，则可以删除置信度小于预定阈值的规则集，即为置信度剪枝文法；错误剪枝文法则是预先设定一个验证集，删除在验证集中预测错误的状态集。他们的实验表明，这三种剪枝文法在显著减小模型的状态空间复杂性的同时，还获得了预测准确率的少量提高。对于高序模型的低覆盖率问题，Pitkow 等提出全 k 序 Markov 模型(All-Kth-Order Markov Models)的解决方法[39]。他们的方法基于这样的一种策略：首先使用用户当前的游历序列与模型中最长的序列进行匹配来预测用户的请求，如果没有找到长度为 k 的序列，就依次查找长度为 $k-1, k-2, \cdots$ 的序列。在最坏的情形下，当前用户的访问行为可以通过一序模型得到匹配。这种首先在高序模型中"尝试"匹配的策略在一些情形下会增加无谓的计算开销，因此并不是一种很好的解决方法。

3.2 混合 Markov 模型(Mixture Markov Models)

用户在 Web 信息空间的游历行为是一个受浏览目的、背景知识、兴趣爱好等多种因素影响的复杂过程。由于这些因素的差异，各个用户的浏览行为也就表现出不同的个性化特点，基本 Markov 浏览预测模型采用一个 Markov 链描述所有用户的浏览特征，明显过于简单，其预测结果必然存在较大的误差。基于各个用户的浏览特点，可以将所有用户分为多个类别，使同一类别的用户具有相同或相似的浏览特征，然后分别为每个类别的用户设立单独的 Markov 模型，用以描述该类别用户的浏览特征，这样就克服了基本模型中只用一个 Markov 链描述所有用户的浏览特征所带来的不准确性，并能得到更高的预测准确率。与基本模型相对照，XING Yong-Kang 等将这种采用多个 Markov 链来描述多个类别用户的浏览特征的模型称为多 Markov 链模型，而把基本模型称为单 Markov 链模型[40,41]，Cadez 等[21]，Sen 等[42]及 Eren Manavoglu 等[43]则称之为混合 Markov 链模型，本文简单称之为混合模型。

假设 2(用户分类假设) 根据用户游历序列事务数据库 $NSDB$，可以将用户分为 K 类，使得同一类用户之间的浏览行为特征相似性最大，不同类之间的用户浏览行为特征相似性最小。设 $C=\{c_1, c_2, \cdots, c_k\}$ 表示用户的类别，任意一个用户属于类别 $c_k(k \in \{1, \cdots, K\})$ 的概率为 $P(C=c_k)$，则有 $\sum_{k=1}^{K} P(C=c_k) = 1$。

假设 3(类 Markov 链假设) 假设同一类用户的浏览过程是一个特殊的随机过程——齐次离散 Markov 链。

定义 6(混合模型) 混合模型可以表示为一个四元组 $M=\langle X, K, P(C), MC\rangle$，其中，$X$ 是一个离散随机变量，值域为 $\{x_1, x_2, \cdots, x_n\}$，每个 x_i 对应一个网页，称为模型的一个状态，K 表示模型包含的用户类别数目，$C=\{c_1, c_2, \cdots, c_k\}$ 表示用户的类别，其分布函数 $P(C)$ 表示不同类别用户的概率分布，$MC=\{mc_1, mc_2, \cdots, mc_k\}$ 为类 Markov 链的集合，每一个元素 mc_k 是描述类别为 c_k 的用户浏览特征的 Markov 链，称为类 Markov 链，它的转移矩阵可以表示为

$$A_k = (p_{kij})_{n \times n}, (i, j \in \{1, \cdots, n\}, k \in \{1, \cdots, K\})$$

对混合模型的学习主要是确定以下几个参数：

1) 用户类别数 K；

2) 任意一个用户属于类别 c_k 的概率 $P(C=c_k)$；

3) 类 Markov 链的转移矩阵。

混合模型的训练其实是用户聚类与类 Markov 链的学习过程。XING Yong-Kang 等采用聚类的方法对混合模型进行学习[40,41]。设有学习数据 $NSDB=\{x^1, x^2, \cdots, x^m\}$，第 k 类所包含的用户游历序列的数目为 m_k，将初始状态看作是一种特殊聚类结果，即每个用户的游历序列为一个类别，此时 $K=m, m_k=1$ 且有 $P(C=c_k)=\frac{m_k}{m}$，采用 Bayes 估计计算转移概率矩阵和初始分布中的每一项可以得到初始的 m 个 Markov 链，对于任意的两个 Markov 链 mc_k 与 mc_l，及其对应的转移矩阵 A_k、A_l，它们的近似程度可以用交叉熵距离定义为

$$sim(A_k, A_l) = \sum_{i=1}^{n}\sum_{j=1}^{n} p_{kij} \log \frac{p_{kij}}{p_{lij}} / n \quad (14)$$

于是可以定义 mc_k 与 mc_l 的相似程度为 $sim(mc_k, mc_l) = 2/(sim(A_k, A_l) + sim(A_l, A_k))$，利用该式计算 m 个 Markov 链两两之间的相似度大小，定义评价聚类结果质量的准测函数为模型对于学习数据的后验概率最大，通过尝试合并具有较大相似度的 Markov 链的迭代算法可以找出使准则函数取极值的聚类结果，即最好的聚类结果，完成对多 Markov 链模型的学习。该聚类学习算法的时间复杂度为 $o(m^5 n^2)$。

Cadez 等[21]，Sen 等[42]及 Eren Manavoglu 等[43]则都采用了 EM 算法对混合模型进行学习。设混合模型中的待估参数集为 θ，首先给 θ 一个先验概率分布，然后采用 EM 迭代算法，用 $p(c_k \mid \theta)$ 表示参数集 θ 下混合模型中第 k 个类 Markov 模型的边际概率，即 $\sum_{k=1}^{K} p(c_k \mid \theta) = 1$，$p_k(x \mid c_k, \theta)$，表示参数集 θ 下第 k 个类 Markov 模型中具有访问行为 x 的用户概率，$p(c_k \mid x, \theta)$ 表示参数集 θ 下具有访问行为 x 的用户属于第 k 个类

Markov 模型的概率，算法的 E 步是在当前给定的 θ 下，将具有访问行为 x^i 的用户分类到类别 c_k，即

$$P_{i,k}(\theta)=p(c_k|x^i,\theta)=\frac{p(c_k|\theta)p_k(x^i|c_k,\theta)}{\sum_{j=1}^{K}p(c_j|\theta)p_j(x^j|c_j,\theta)} \quad (15)$$

算法的 M 步是采用极大似然估计法，用学习数据 $NSDB$ 修正参数 θ 的估计，使之满足：

$$\theta^{MAP}=\arg\max_\theta p(\theta|NSDB)=\arg\max_\theta p(NSDB|\theta)/p(NSDB) \quad (16)$$

实际应用时，通常采用如下的目标函数 Q 来代替上式：

$$Q(\theta,\theta_{old})=\sum_{i=1}^{m}\sum_{k=1}^{K}P_{i,k}(\theta_{old})\log[p_k(c_k|\theta)p_k(x^i|c_k,\theta)]+\log p(\theta) \quad (17)$$

当算法两次迭代值差小于 0.01%，让算法收敛，由此完成混合模型的学习。

采用混合模型进行浏览预测的方法可以用公式描述如下：

$$p(x_{t+1}|x_t,\cdots,x_1)=\sum_{k=1}^{K}p(x_{t+1},c_k|x_t,\cdots,x_1)=\sum_{k=1}^{K}p(x_{t+1}|x_t,\cdots,x_1,c_k)p(c_k|x_t,\cdots,x_1)=\sum_{k=1}^{K}p(x_{t+1}|x_t,c_k)p(c_k|x_t,\cdots,x_1) \quad (18)$$

3.3 隐 Markov 模型(Hidden Markov Models, HMM)

当用户在 Web 站点信息空间游历时，实际上是带有某种目的的，也就是说用户对某种东西是感兴趣的。基本模型只是简单的考虑用户请求页面之间的动态时序性，没有考虑蕴涵在用户不同的访问路径中的用户兴趣概念。与此不同，HMM 通过定义用户感兴趣的事物或概念为用户访问概念集，群体用户在每一个访问状态节点都存在对访问概念集的一个概率分布，由此可以发现用户带有兴趣的迁移模式[44,45]。这种带有某种兴趣的迁移模式实质上是一种关联规则，反映的是用户的偏好访问。

定义 7(带有用户访问兴趣的 HMM 模型)

1) 视 Web 站点的节点为 HMM 的状态节点 x，

2) 存在概念集 $\Sigma=\sigma_1,\sigma_2,\cdots,\sigma_T$，

3) 对于任意一个节点 x_j 包含 Σ 的一个子集 $(\sigma_1^j,\sigma_2^j,\cdots\sigma_t^j)$，

4) 任意两个节点 x_i,x_j 之间存在一个转移概率 $p_{ij}=P(x_i\rightarrow x_j)$，

5) 在一个节点 x_j，群体用户对该节点的概念集 $(\sigma_1^j,\sigma_2^j,\cdots,\sigma_t^j)$ 存在一个概率分布 $P(\sigma_1^j),\cdots,P(\sigma_t^j)$ 即为标准 HMM 中状态节点的观测概率。每一个 $P(\sigma_t^j|x_j)$ 意义为(σ_t^j 简记为 σ)：群体用户通过 x_j 共访问了 Σ'概念(允许重复)，则其中所含 σ 的比率即近似为 $P(\sigma_t^j|x_j)$。

形式化如下：

设有学习数据 $NSDB=\{x^1,x^2,\cdots,x^m\}$，状态集 $X=\{x_1,x_2,\cdots,x_n\}$，x_j 状态上的概念集为 $(\sigma_1^j,\sigma_2^j,\cdots,\sigma_t^j)$，$NSDB$ 中第个用户游历序列 x^i 可以表示为

$$x^i=\langle x_1^i,x_2^i,\cdots,x_f^i\rangle,x_{f'}^i\in X,f'=1,\cdots,f \quad (19)$$

用 $x^{i'}$ 表示 x^i 中的每个分量，即每个被访问节点的集合：

$$x^{i'}=\{x_1^i,x_2^i,\cdots,x_f^i\},x_{f'}^i\in X,f'=1,\cdots,f \quad (20)$$

用 $SS_{i,j}$ 表示游历序列 x^i 中 x_j 之后的所有被访问节点的集合(包括 x_j)，即

$$SS_{ij}=\begin{cases}\{x_{k+l}^i|x_k^i=x_j,l=0,\cdots,(f-k)\}, & x_j\in x^{i'}\\ NULL, & x_j\notin x^{i'}\end{cases} \quad (21)$$

用 $SS_{i,j,\sigma}$ 表示 $SS_{i,j}$ 在集合中含有 σ 的节点的集合，即

$$SS_{i,j,\sigma}=\{x|x\in SS_{i,j},x\text{ 中含有 }\sigma\} \quad (22)$$

则在 x_j 节点上，群体用户对 σ 感兴趣的概率为

$$P(\sigma|x_j)\approx\frac{\sum_{i=1}^{m}\|SS_{i,j,\sigma}\|}{\sum_{\sigma'=1}^{m}\|SS_{i,j,\sigma'}\|} \quad (23)$$

通过对日志文件的分析，我们可以建立这样一个 HMM 模型。

定义 8(兴趣关联规则 $IR(\sigma|x)$) 已知一个用户游历序列 $x=(x_1,x_2,\cdots,x_t)$ 和用户访问兴趣 σ，那么兴趣关联规则 $IR(\sigma|x)$ 为

$$IR(\sigma|x)=(P(x_1\rightarrow x_2)\times P(\sigma|x_2))\times(P(x_2\rightarrow x_3)|P(\sigma|x_3))\times\cdots\times(P(x_{t-1}\rightarrow x_t)\times P(\sigma|x_t))$$

并且 $IR(\sigma|x)\geqslant\varepsilon$($\varepsilon$ 为一个给定的可信度阈值)。

由兴趣关联规则 $IR(\sigma|x)$，可以将 Web 信息空间中包含有概念子集 σ 的节点作为具有访问序列 x 的用户的浏览预测页面。

3.4 连续时间 Markov 模型(Continuous-Time Markov Models)

前述所有模型的一个共同特点是：建立的用户浏览行为模型都是离散时间模型。在浏览预测问题中，离散时间模型虽然可以对用户的访问页面做出预测，但不能预测用户在何时请求该页面。这种"何时"问题，在一些情形下是有效用的，如在文档预送系统中，可以综合考虑网络繁忙情况与用户对页面的请求时刻给出文档更合理的预送时刻。Qiming 等将用户的游历序列看作是一个连续时间 Markov 过程，给出了该问题的一个求解方法[46]。

假设 4(浏览行为的连续时间 Markov 性假设) 用户在 Web 信息空间的游历行为可以看作是一个连续时间 Markov 链。

设 $\{X(t),t\geqslant0\}$ 是代表用户游历序列的连续时间 Markov 链，转移矩阵为 $P(t)=e^{Rt}$，$R=\{r_{ij}\}$，为其相应的转移速率矩阵，v_i 是用户在状态 i 时的转移速率，p_{ij} 为状态 i,j 的转移概率，则有

$$r_{ij}=v_i\times p_{ij}\text{ if }i\neq j \quad (24)$$

$$r_{ij}=-v_i\text{ if }i=j \quad (25)$$

假设 5(浏览时间的指数分布假设) 设 t_{ij} 是群体用户离开 i 访问 j 之前在状态 i 的停留时间，则 t_{ij} 服从指数分布。

对所有可能的状态对 (x_i,x_j) 按照 t_{ij} 的大小排序为 t_{ij} ($t_{i1}<t_{i2}<\cdots t_{il}$)，设在时刻 t_{im} 群体用户对状态 i 的访问次数记为 N_{im}，注意到 N_i 代表的是最后一个用户离开状态 i 时群体用户对状态 i 的访问次数，因此在时刻 t_{im} 状态 i 的离开概率为 N_{im}/N_i，可以用一个指数分布函数来表示状态 i 的离开概率：

$$F(t)=1-e^{\lambda t},t>0\text{ and }F(t)=0,t<0 \quad (26)$$

式中，λ 等于模型中的转移速率 v_i，于是有

$$v_{im}=-(\ln(1-N_{im}/N_i))/t_{im}\ m=1,2,\cdots,n-1 \quad (27)$$

$$v_i=(v_{i1}+v_{i2}+\cdots v_{i(n-l)})/(n-l) \quad (28)$$

根据(24)、(25)式求得转移速率矩阵，由此建立用户浏览行为的连续时间 Markov 模型。采用带有时间参数的转移矩阵 $P(t)=e^{Rt}$ 可以对用户请求的页面及其请求的时刻做出预测。

3.5 其它一些扩展模型

除了考虑访问行为的时序因素，一些研究者还考虑了其它的一些因素，用以改进基本模型。通过考虑请求页面之间的时序性、引用页与请求页的相关性，考虑请求页面间的时序

性及引用页与请求页的相关性，Zukerman 等提出了四种不同的 Markov 模型[47]：时间模型(Time Markov Models)，仅基于用户最后的请求页面来预测下一步的链接；二序模型(Second-order Time Markov model)，基于用户最后的两次请求页对访问行为进行预测，时间模型与二序模型其本质上为基本模型与二序的高阶模型；空间模型(Space Markov model)，基于引用页推测用户的下一个请求；链接空间—时间模型(Linked Space-Time Markov model)，同时使用引用页和最后请求页来预测用户的下一个请求。Albrecht 等进一步将这四种不同的 Markov 模型用于文档预送系统，并采用了一个决策理论框架用以确定具有最大收益率的预送文档，获得了较好的效果[48]。

Zukerman 的四种不同的 Markov 模型都是基于 Web 日志的不同数据建立的。在 Web 浏览预测问题中，一个可供利用的数据是站点结构，Eirinaki 等在基本的 Markov 模型中集成链接分析的方法，采用 PageRank 算法分析页面重要度作为浏览预测时的初始概率分布。他们的实验表明，这种结合访问数据与页面重要度的综合预测方法能取得比仅基于用户访问记录的预测方法更好的效果[49]。Anderson 等则充分考虑了站点页面的不同主题与概化层次结构，将站点页面划分为主题页面集，对每个主题页面集定义一组具有概化层次性质的变量，如一个电子商务站点的可以描述为：{MainEntryPage()，ProductPage(Product，StockLevel)，CheckoutPage()}。此时 Markov 模型中的状态不再是站点页面，而是某个主题页面集，并且是该主题页面集的概化层次变量的函数。这种扩展的模型更适合实际 Web 站点的情形，Anderson 等将其称为关系 Markov 模型(Relational Markov Models)，RMM 适合对具有异质状态空间与稀疏数据的大型站点进行建模分析[50]。

4 进一步研究的问题

尽管现有的 Markov 浏览预测模型在预测准确率、覆盖率方面已取得较满意的成果，但浏览预测问题的实际应用背景中的一些特殊要求使得这一领域仍存在一些需要进一步研究的问题。这些问题包括：

1)冷开始问题及与其它方法的组合预测问题。Markov 浏览预测模型是基于用户访问历史的，如果用户一开始并没有访问任何页面，则模型没有办法对该用户的访问行为做出预测，此时可以考虑采用集成基于内容过滤的预测方法或是基于用户特征分析的方法[51]。另一方面，在实际浏览预测问题中，Markov 的随机统计方法与其它方法(如反映用户访问行为学特征的信息搜寻理论)结合的组合预测模型应该能获得较高的预测准确率。

2)模型的动态性问题。Web 是个动态性很强的环境。用户兴趣、浏览目的、站点结构、页面内容以及日志数据都是动态变化的，不能实时反映 Web 动态特征的静态 Markov 模型在实际应用中不能从根本上解决浏览预测问题。要求模型要在以下几个方面实现动态性，一是研究一些增量式的学习算法[52]，使得模型可以随日志数据量的增长进行动态学习；二是在模型的预测结果没有实际命中时，模型能根据用户实际的访问行为进行动态自适应学习；三是模型能随用户兴趣、浏览目的变化进行动态预测；四是在站点结构、页面内容变化时，模型能做出同步反馈，调整预测的结果页面。

3)复杂度与预测精度的矛盾问题。一般来说，高序模型、混合模型等能取得比基本模型更高的预测精度，但在获得高预测精度的同时，也带来了更高的模型复杂度问题。在实际应用时，如何在模型的复杂度与预测精度的矛盾中取得一个合适的"平衡点"，是个有待解决的问题。一般来说，在实际应用中对各种预测模型的选择取决于对预测精度，模型复杂度与实时性等各方面的综合考虑。

4)模型的评价问题。在实验系统中，一般采用模型对测试集的预测精度与覆盖率作为模型的评价尺度。而在实际应用中，应当结合浏览预测问题的应用背景采取相应的评价尺度，如在文档预送系统中，能在多大程度上减轻用户访问的延时问题；在推荐系统中，能在多大程度上提高推荐的满意度。目前尚没有证据表明实验系统中的高预测准确率模型，就一定能在实际应用系统中取得好的应用效果。

5)模型的合理解释性问题。从 Lawless 等使用的聚类分析研究中可以知道用户在浏览 Web 页面时可以分为信息搜寻(information seekers)、特征探索(feature explorers)和随意超链使用者(apathetic hypertext)等三种浏览路径类型[53]。信息搜寻理论从 Web 访问的行为学特征出发，假设用户在 Web 站点信息空间的浏览行为是带有目的的，并假设用户总是倾向最大化搜寻活动的获取率，即单位费用上获得的信息量，采用该理论的浏览预测方法可以对预测结果做出合理性解释[54]。但是另一方面，用户在 Web 站点信息空间的游历行为是一个同时受浏览目的、背景知识、兴趣爱好、站点结构、网络环境等多种因素交叉影响的复杂过程，完全基于行为学的角度建立确切的用户浏览预测模型是不现实的。大量随意超链使用者对 Web 站点的访问形成了 Web 群体用户访问行为的统计学特征，Markov 模型正是基于大量用户的浏览记录建立的，是一种通过研究 Web 群体用户访问行为的数量特征来分析 Web 访问模式进行浏览预测的方法。从本质上讲，浏览预测问题的 Markov 模型是一种基于群体用户访问历史的随机概率统计模型。对于个体用户的浏览预测，与基于行为学的方法相比，Markov 模型缺乏对预测结果的合理性解释。

总结 针对 Web 浏览预测问题的特点，Web 挖掘研究工作者提出了各种不同的 Markov 浏览预测模型。对各种模型在实践中的应用价值，还有待于在大量的 Web 站点进行实证研究与广大 Web 用户的检验，进而进一步完善模型。

参 考 文 献

1 Chen Jiyang, Sun Lisheng, Osmar R Z, et al. Visualizing and discovering Web navigational patterns. In: Proceedings of the 7th International Workshop on the Web and Databases: collocated with ACM SIGMOD/PODS, Paris, France, 2004. 13～18

2 Don Zimmerman, Pat Walls. Exploring navigational patterns on the Web. In: Proceedings of IEEE Professional Communication Society International Professional Communication Conference and Proceedings of the 18th annual ACM International Conference on Computer Documentation: Technology & Teamwork, 2000. 581～591

3 Srivastava J, Cooley R, Deshpande M, et al. Web usage mining: discovery and applications of usage patterns from Web data. ACM SIGKDD Explorations Newsletter, 2000, 1(2): 12～23

4 Joseph D, Grunwald D. Prefetching Using Markov Predictors. IEEE Transactions on Computers, 1999, 48(2): 121～133

5 Li Tianyi. Building Association-Rule Based Sequential Classifiers for Web-Document Prediction. Data Mining and Knowledge Discovery, 2004, 8(3): 253～273

6 Davison B D. The Design and Evaluation of Web Prefetching and Caching Techniques: [PhD thesis]. Department of Computer Science, Rutgers University, 2002

7 Yang Qiang, Zhang Haining H. Web-Log Mining for Predictive Web Caching. IEEE Transactions on Knowledge and Data Engineering, 2003, 15(4): 1050～1053

8 Eirinaki M, Vazirgiannis M. Web mining for Web personalization. ACM Transactions on Internet Technology (TOIT), 2003, 3(1): 1～27

9 Srikant R, Yang Yinghui. Mining Web logs to improve Web site organization. In: Proceedings of the 10th International Conference on World Wide Web, Hong Kong, 2001. 430～437

10 Zaiane O, Xin M, Han J. Discovering Web access patterns and trends by applying OLAP and data mining technology on Web logs. In: Proceedings of the Advances in Digital Libraries Conference, 1999. 19～29

11 Manevitz L M, Yousef M. A Web navigation system based on a neural network user-model trained with only positive Web documents. Web Intelligence and Agent System, 2004, 2(2): 137～144

12 Chan P K. A non-invasive learning approach to building Web user profiles. In: Proceedings of 5th ACMSIGKDD International Conference, Workshop on Web Usage Analysis and User Profiling, Springer, 1999. 7～12

13 Paliouras G, Papatheodorou C, Karkaletsis V, et al. Clustering the Users of Large Web Sites into Communities. In: Proceedings of International Conference on Machine Learning (ICML), Stanford, California, 2000. 719～726

14 Mobasher B, Dai Honghua, Luo Tao, et al. Effective personalization based on association rule discovery from Web usage data. In: Proceedings of the 3rd International Workshop on Web Information and Data management, Atlanta, Georgia, USA, 2001. 9～15

15 Huang X D, Hon A H W, Lee K F. Large-vocabulary speaker-independent continuous speech recognition with semi-continuous hidden Markov models. In: Proceedings of the Workshop on Speech and Natural Language, 1989. 276～279

16 Pachter L, Alexandersson M, Cawley S. Applications of generalized pair hidden Markov models to alignment and gene finding problems. In: Proceedings of the Fifth Annual International Conference on Computational Biology, Montreal, Quebec, Canada, 2001. 241～248

17 Bestavros A. Using Speculation to Reduce Server Load and Service Time on the WWW. In: Proceedings of the 4th ACM International Conference on Information and Knowledge Management, Baltimore, MD, 1995. 403～410

18 Huberman B A, Pirolli P L T, Pitkow J E, et al. Strong Regularities in World Wide Web Surfing. Science, 1998, 280: 95～97

19 Jespersen S, Pedersen T B, Thorhauge J. Evaluating the markov assumption for Web usage mining. In: Proceedings of the 5th ACM International Workshop on Web Information and Data Management, ACM Press, 2003. 82～89

20 Kijima M. Markov Processes for Stochastic Modeling. London Chapman&Hall, 1997

21 Cadez I V, Heckerman D, Smyth P, et al. Model～based clustering and visualization of navigation patterns on a Web site. Knowledge Discovery and Data Mining, 2003

22 Li S, Montgomery A, Srinivasan K, et al. Predicting online purchase conversion using Web path analysis. Graduate School of Industrial Administration, Carnegie Mellon University, Pittsburgh, PA, 2002

23 Anderson C, Domingos P, Weld D. Adaptive Web navigation for wireless devices. In: Proc. 17th Int Joint Conf. on Artificial Intelligence, San Francisco, CA, 2001. 879～884

24 Spears W M. A compression algorithm for probability transition matrices. SIAM Matrix Analysis and Applications, 1998, 20(1): 60～77

25 Zhu Jianhan, Hong Jun, Hughes J G. Using Markov Chains for Link Prediction in Adaptive Web Sites. In: Proceedings of the First International Conference on Computing in an Imperfect World, 2002. 60～73

26 Silva M. Sparse matrix storage revisited. In: Proceedings of the 2nd Conference on Computing Frontiers, Ischia, Italy, 2005. 230～235

27 Shafer G, Shenoy P P, Mellouli K. Propagating belief functions in qualitative Markov trees. International Journal of Approximate Reasoning, 1987, 1(4): 349～400

28 Laird P, Saul R. Discrete sequence prediction and its applications. Machine Learning, 1994, 15(1): 43～68

29 Fan Guoliang, Xia Xianggen. Maximum likelihood texture analysis and classification and classification using wavelet～domain hidden markov models. In: Proceedings of the 34th Asilomar Conference on Signals, Systems and Computers, Pacific Grove, CA, 2000

30 Géry M, Haddad H. Evaluation of Web usage mining approaches for user's next request prediction. Proceedings of the 5th ACM International Workshop on Web Information and Data Management, New Orleans, Louisiana, USA, 2003. 74～81

31 Laura Bottomley of Duke University. EPA～HTTP server logs. http://ita.ee.lbl.gov/html/contrib/EPA～HTTP.html

32 Sarukkai R R. Link prediction and path analysis using markov chains. In: 9th International World Wide Web Conference, 2000

33 Perkowitz M, Etzioni O. Adaptive Web sites: Automatically synthesizing Web pages. In: Proceedings of the Fifteenth National Conference on Artificial Intelligence, Madison, WI, AAAI Press, 1998

34 Perkowitz M, Etzioni O. Adaptive Web sites. Communications of the ACM, 2000, 43(8): 152～158

35 Davison B. Learning Web request patterns. In: Web Dynamics – Adapting to Change in Content, Size, Topology and Use, Springer, 2004

36 Deshpande M, Karypis G. Selective Markov Models for Predicting Web-Page Accesses. ACM Transactions on Internet Technology, 2004, 4(2): 163～184

37 Pirolli P L, Pitkow J E. Distributions of surfers' paths through the World Wide Web: Empirical characterization. World Wide Web, 1999, 2: 29～45

38 Xing Dongshan, Shen Junyi. A New Markov Model For Web Access Prediction. Computing in Science and Engg, 2002, 4: 34～39

39 Pitkow J, Pirolli P. Mining longest repeating subsequence to predict world wide Web surfing. In: 2nd USENIX Symposium on Internet Technologies and Systems, Boulder, CO, 1999

40 邢永康,马少平. 多 Markov 链用户浏览预测模型. 计算机学报, 2003, 26(11): 1510～1517

41 邢永康,马少平. 一种基于 Markov 链模型的动态聚类方法. 计算机研究与发展, 2003, 40(2): 129～135

42 Sen R, Hansen M H. Predicting a Web user's next request based on log data. Journal of Computational Graphics and Statistics, 2003, 12(1): 143～155

43 Manavoglu E. Probabilistic User Behavior Models. http://citeseer.ist.psu.edu/726257.html, 2003

44 王实,高文,李锦涛,等. 基于隐马尔可夫模型的兴趣迁移模式发现. 计算机学报, 2001, 24(2) 152～157

45 王实,高文,黄铁军,等. 基于隐马尔可夫模型的在零售站点的自适应. 软件学报, 2001, 12(4): 599～606

46 Huang Qiming, Yang Qiang, Huang J Zhexue, et al. Mining of Web-Page Visiting Patterns with Continuous-Time Markov Models. Lecture Notes in Computer Science, 2004, 3056: 549～558

47 Zukerman I, Albrecht D W, Nicholson A E. Predicting users' requests on the WWW. In: Proceedings of the Seventh International Conference on User Modeling, Banff, Canada, 1999. 275～284

48 Albrecht D, Zukerman I, Nicholson A. Pre-sending documents on the WWW: A comparative study. In: IJCAI99 – Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence, Stockholm, Sweden, 1999

49 Magdalini E, Michalis V, Dimitris K. Web path recommendations based on page ranking and Markov models. In: Proceedings of the 7th annual ACM International Workshop on Web Information and Data Management, Bremen, Germany, 2005. 2～9

50 Anderson C, Domingos P, Weld D. Relational Markov models and their application to adaptive Web navigation. In: Proc. 8th ACM SIGKDD Intl Conf on Knowledge Discovery and Data Mining, Edmonton, Canada, 2002. 143～152

51 Sergio F, Sergio G, Andrea T, et al. Mining User Preferences, Page Content and Usage to Personalize Web site Navigation. World Wide Web, 2005, 8(3): 317～345

52 Yen Show-Jane, Lee Yue-Shi, Hsieh Min-Chi. An Efficient Incremental Algorithm for Mining Web Traversal Patterns. In: Proceedings of the IEEE International Conference on e-Business Engineering, 2005. 274～281

53 Lawless K, Kulikowich J M. Domain knowledge, interest, and hypertext navigation: A study of individual differences. Journal of Educational Multimedia and Hypermedia, 1998, 7(1): 51～69

54 Pirolli P, Card S K. Information Foraging: [Technical Report]. UIR, 1999