基于隐马尔科夫模型的浏览兴趣预测

孙秀娟[1],金民锁[2],陈孝国[1]

1. 黑龙江科技学院数力系,哈尔滨 150027
2. 黑龙江科技学院信息网络中心,哈尔滨 150027

摘要 Web上的信息量正以惊人的速度增加,人们迫切需要能自动地从Web上发现、抽取和过滤信息的工具,即如何从数以亿计的页面中发现需要的内容、如何从大量的访问中发现固有的模式和关联。马尔科夫模型的网页浏览预测,仅仅从用户的浏览网页本身出发,预测用户的下一步链接,并不能捕获到用户的真正兴趣。本文提出基于隐马尔科夫模型的网页浏览路径预测,并将其与基于马尔科夫模型的方法进行对比。根据已知的浏览序列判断用户的类别,当浏览序列长度很短时,本文方法的预测准确性比马尔科夫模型低。这是由于序列长度过短,系统获取判断的信息少,增加了对用户错误分类的可能性。随着浏览序列长度逐渐增加,系统捕获的用户浏览信息越来越多,进而能够折射出用户的兴趣所在,预测准确率也逐步增加。当浏览序列长度大于或等于8时,预测准确率已经到达80%,提高了浏览兴趣预测的准确率。

关键词 马尔科夫模型;浏览预测;Web使用挖掘;聚类

中图分类号 TP393.1 **文献标识码** A **文章编号** 1000-7857(2009)18-0075-03

Prediction Based on Hidden Markov Model

SUN Xiujuan[1], JIN Minsuo[2], CHEN Xiaoguo[1]

1. *Department of Mathematics and Mechanics, Heilongjiang Institute of Science & Technology, Harbin 150027, China*
2. *Information Network Center, Heilongjiang Institute of Science & Technology, Harbin 150027, China*

Abstract The amount of information on web is increasing at an alarming rate. It is an urgent need to find tools to automatically obtain, extract and filter information from web, from hundreds of millions of pages to find the content in need, to find related patterns and associations. Markov model predicts the user's next link, only from the user's browser start page, which does not involve the real interest of the user. In this paper, HMM-based prediction of the web browser path is presented. First of all, according to known sequences to determine the browser type of user. As can be seen for the browser with a very short sequence length, the accuracy of the forecasts is lower than the Markov model. This is due to the short sequence length, the system can access only limited information to make judgement, with more classification of user errors as more likely. However, with a gradual increase in sequence length, the system may capture more and more user's browsing information to reflect user's interest and to increase the accuracy of prediction. When the sequence length is greater than or equal to 8, the forecasting accuracy rate reaches 80%.

Keywords Markov model; research prediction; web usage mining; cluster

0 引言

网页浏览预测主要是解决CPU空闲周期的应用、Web页面预先缓存、以及用户访问行为预测的问题。它在导航模式的发现中显得尤为重要,分为两种类型:记录用户导航行为的确定性方式和运用已访问过的Web页面预测接下来访问的随机性方式。Borges和Levene[1]提出从用户会话中识别序列模式的随机性方法。他们使用超文本概率文法表示用户会话。最典型的随机方法是马尔科夫(Markov)模型,在链接预测的序列模式发现中经常被运用,这主要是因为Markov模型非常适合序列化过程。

用户在Web空间的浏览过程是一个受浏览目的、文化背景、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,各个用户的浏览过程也就表现出不同的个性化特点;然而观察大量用户的浏览过程可以发现,某些用户的浏览过

收稿日期:2009-05-04

基金项目:黑龙江省教育厅科学技术面上项目(11541316);黑龙江科技学院青年教师基金项目(07-18)

作者简介:孙秀娟,讲师,研究方向为网络安全与计算,电子信箱:ruixuan992006@126.com;金民锁(通信作者),工程师,研究方向为网络安全与计算,电子信箱:jms9802@126.com

程表现出相同或相近的特点，如他们浏览的网页基本相同、浏览各个网页的顺序相似等[2]。这一现象引发了基于用户访问的页面对其进行聚类的算法，并取得较好的结果，从而证明了对 Web 用户分类的可行性。对用户进行分类，同一类别的用户由于具有相同或相近的浏览特征，可用同一个模型描述；而不同类别的用户，其浏览过程差别较大，应用不同的模型描述其特征。这样就克服了单 Markov 链模型中用一个 Markov 链描述所有用户的浏览特征而带来的不准确性，从而能更精确地描述用户的浏览特征，并有可能得到更高的预测准确率。基于这种思想，本文提出先用隐马尔科夫模型(Hidden Markov Model，HMM)对用户进行分类，然后针对不同类别的用户提炼出不同的预测模型。

HMM 最早由 Baum 提出，在许多领域得到应用，尤其在语音识别中得到了广泛的应用。本文采用的是离散化输出，一阶 HMM 模型。模型可以表示为

$$\lambda=(A, B, \pi)$$

其中，输出符号集 $V=\{v_1, v_2, \cdots, v_M\}$；状态转移概率分布 $A=\{a_{ij}|1\leqslant i, j\leqslant N\}$，$N$ 为状态个数，$a_{ij}=P[q_t=j|q_{t-1}=i]$，$q_t$ 为 t 时刻的状态；输出符号的概率分布 $B=\{b_j(k)|1\leqslant k\leqslant M, 1\leqslant j\leqslant N\}$，$M$ 为离散输出符号个数，$b_j(k)=P[o_t=v_k|q_t=j]$，初试状态概率 $\pi=\{\pi_i|1\leqslant i\leqslant N\}$，$\pi_i=P[q_0=i]$。

给定 N, M, A, B, π 的 HMM 可以产生序列 $O=(o_1, o_2, \cdots, o_n)$，$o_i$ 为状态输出符号。假设状态序列 $q=(q_1, q_2, \cdots, q_n)$已知，观察序列 O 的概率计算为

$$P(O|q, \lambda)=\prod_{t=1}^{n} a_{q_{t-1}q_t} b_{q_t}(o_t) \tag{1}$$

1 隐马尔科夫模型的建立

Web 站点的节点为 HMM 的状态节点 q，给定一个虚拟的初试状态 q_i。存在概念集 $\Sigma=\sigma_1, \sigma_2, \cdots, \sigma_T$。$q_j$ 节点包含 Σ 的一个子集 $\Sigma^*=(\sigma_1^j, \sigma_2^j, \cdots, \sigma_t^j)$。两个节点 q_i，q_j 之间存在一个转移概率 $P(q_i\rightarrow q_j)$；在事务集 T 中计算任意两个节点 q_i，q_j 的概率 $P(q_i\rightarrow q_j)$，即

$$P(q_i\rightarrow q_j)=\frac{count(q_i\rightarrow q_j)}{count(q_i)} \tag{2}$$

其中 $count(q_i\rightarrow q_j)$为在事务集 T 中，q_i，q_j 同时出现且 q_j 紧随 q_i 的事务个数，$count(q_i)$为 T 中含有 q_i 的事务个数，与单 Markov 链模型中参数的计算公式相同。

在一个节点 q_j 上，群体用户对该节点上的概念集$(\sigma_1^j, \sigma_2^j, \cdots, \sigma_t^j)$ 存在一个概率分布 $P(\sigma_1^j|q_j), P(\sigma_2^j|q_j), \cdots, P(\sigma_t^j|q_j)$，即为 HMM 中状态节点的观测概率。每一个 $P(\sigma_i^j|q_j)$表示群体用户通过 q_j 的浏览访问次数，这些访问一共进行了 Σ^* 概念子集(可以重复)，则其中所含 σ_i^j 的比率近似为 $P(\sigma|q_j)$(σ_i^j 简记为 σ)。形式化如下所述。

设有 n 个事务集 $T=\{t_1, t_2, \cdots, t_n\}$，$m$ 个状态集 $Q=\{q_1, q_2, \cdots, q_m\}$，$q_j$ 状态上的物品集$(\sigma_1^j, \sigma_2^j, \cdots, \sigma_t^j)$。$T$ 事务集中第 i 个事务为

$$t_i=<t_i[1], t_i[2], \cdots, t_i[f]> \qquad t_i[f']\in Q, f'=1, 2, \cdots, f \tag{3}$$

t_i' 表示 t_i 事务中的每个分量，即每个访问结点的集合：

$$t_i'=<t_i[1], t_i[2], \cdots, t_i[f]> \qquad t_i[f']\in Q, f'=1, 2, \cdots, f \tag{4}$$

则 S_{ij} 表示 t_i 事务中 q_j 之后所有被访问的节点集合(包括 q_j)：

$$S_{ij}=\begin{cases}\{t_i[k+l]|t_i[k]=q_j, l=0, 1, 2, \cdots, (f-k)\} & q_j\in t_i'\\ \text{null} & q_j\notin t_i'\end{cases} \tag{5}$$

$S_{ij,\sigma}$ 表示在 S_{ij} 集合中含有 σ 节点的集合：$S_{ij,\sigma}=\{q|q\in S_{ij}, q$ 中含有 $\sigma\}$，则在 q_j 节点上，群体用户对 σ 感兴趣的概率为

$$P(\sigma|q_j)=\sum_{i=1}^{n}\|S_{i,j,\sigma}\| \Big/ \sum_{\sigma'=1}^{T}\sum_{i=1}^{n}\|S_{i,j,\sigma'}\| \tag{6}$$

通过对 log 文件的分析，可以建立这样的 HMM 模型：假设浏览序列 $S^k=(q_1, q_2, \cdots, q_n)$，那么观察序列为 $O=\left(\underbrace{\sigma, \sigma, \cdots, \sigma}_{n}\right)(\sigma\in\Sigma)$，其概率为

$$\begin{aligned}P(O|q, \lambda)&=a_{q_1q_2}b_{q_2}(\sigma)a_{q_2q_3}b_{q_3}(\sigma)\cdots a_{q_{n-1}q_n}b_{q_n}(\sigma)\\&=(P(q_1\rightarrow q_2)P(\sigma|q_2))\cdot(P(q_2\rightarrow q_3)P(\sigma|q_3))\cdots\\&\quad(P(q_{n-1}\rightarrow q_n)P(\sigma|q_n))\end{aligned} \tag{7}$$

同理，给定一浏览序列 $W=(w_1, w_2, \cdots, w_n)$，通过隐马尔科夫模型可以获得用户对概念 σ 感兴趣程度，它可以用 $R(\sigma|W)$ 表达：

$$R(\sigma|W)=P(O|q, \lambda)=(P(w_1\rightarrow w_2)P(\sigma|w_2))\cdots(P(w_{n-1}\rightarrow w_n)P(\sigma|w_n)) \tag{8}$$

其中，$P(w_{i-1}\rightarrow w_i)$为状态 w_{i-1} 网页转换到状态 w_i 的概率，$P(\sigma|w_i)$ 为页面 w_i 输出概念为 σ 的观察概率。

如果 $P(\sigma|w_i)$表述的是概念 σ 在浏览序列第 i 步受用户关注的概率，则观察序列 $O=\left(\underbrace{\sigma, \sigma, \cdots, \sigma}_{n}\right)$的概率表示概念 σ 是浏览序列潜在需求概念的概率。

2 隐马尔科夫模型的网页浏览预测方法

Web 使用数据的预处理过程共包括 3 个主要步骤和 1 个可选步骤[3]，这一预处理过程负责将初始化的 Web 网站访问使用数据转换为服务会话。因为隐 Markov 模型通过网页概念集合反映用户兴趣，这些概念集合可从网页和网站上的分类信息中提炼出来。聚类、分类、Markov 模型、N-gram 模型仅仅从分析用户的网页浏览路径获得用户兴趣，而 HMM 是从用户浏览过的网页内容本身出发进一步预测用户的真正兴趣所在。当然，网页内容本身的分析是很困难的，它涉及到自然语言理解和处理，而自然语言理解和处理本身就是一大难题。

基于 HMM 的网页浏览预测首先计算概念集合中每个概念的 $R(\sigma|S^k)$值，选出 $R(\sigma|S^k)$值最大的概念，即用户最感兴趣的概念[4]。然后，系统为每一个概念集训练出一个分类器或预测模型（本文选择 Markov 模型）。为用户选择完对应的概念后，运用针对此概念的马尔科夫模型进一步预测用户浏览路

径,给用户推荐网页。系统的结构见图 1。

图 1　预测系统结构

Fig. 1　Predict systematic structure

Web 使用挖掘研究者已经提出了基于聚类/分类[5]、多马尔科夫模型[6]预测用户浏览路径,其本质思想是首先对用户进行分类,然后再有针对性地进行预测。但是它们在第一步的预测中并不能非常准确地预测用户的兴趣所在,这样就对后面的预测和推荐产生了一定的影响。本文提出的基于 HMM 的用户兴趣分类和下一步的预测,其思想在本质上和先前的研究方式是一样的,首先从用户浏览的内容出发对用户兴趣分类,即识别群体用户的兴趣,克服了用户兴趣长短变化的影响,且能周期性、离线地挖掘。这样挖掘的是群体用户兴趣的偏好和行为,不需要某一个特定类别的用户信息,同时能自动跨越不同页面的分类集,为下一步的预测提供了良好的保证,整体上减少了样本的干扰,全面、客观地评价用户和服务用户。

3　预测实验结果分析

为了检验预测模型的性能,选择了 www.usth.edu.cn 服务器日志进行预测性能模拟实验。抽取服务器 log 中 2008 年 10 月 1 日至 2009 年 5 月 31 日的访问记录,组成实验数据集。预设时间窗口阈值 T_w=0.4 h。经过预处理后,数据集的主要特性如下:用户请求总数 24 437,用户会话集个数 76,用户会话总数 4 687,最短会话长度 2,最长会话长度 12。数据集中,60% 的会话作为训练样本集,剩余会话作为测试集。

定义了 6 个概念,Σ={学术、院系、机构、招生、就业、人事}。根据每个页面内容,在各个页面上分别标注 Σ 的子集;根据网络日志分别建立 HMM 和马尔科夫模型。实验测试的是网页浏览预测的准确率 η。

$$\eta=\frac{pre^{+}}{pre^{+}+pre^{-}} \tag{9}$$

pre^{+}, pre^{-}分别表示预测结果正确和错误的样本集。

图 2 中横坐标表示浏览序列长度,纵坐标表示预测准确率,实验选择了 Markov 模型和本文方法进行对比。本文提出的基于 HMM 的网页浏览路径预测,首先要根据已知的浏览序列判断用户的类别。由图 2 可以看出,当浏览序列长度很短时,预测的准确性比 Markov 模型低,这是由于序列长度短时,系统获取判断的信息很少,对用户错误分类的可能性就大。随着浏览序列长度逐渐增加,系统捕获的用户浏览信息越来越多,此时能够折射出用户的兴趣所在,预测准确率也逐步增加。当浏览序列长度≥8 时,预测准确率已经到达

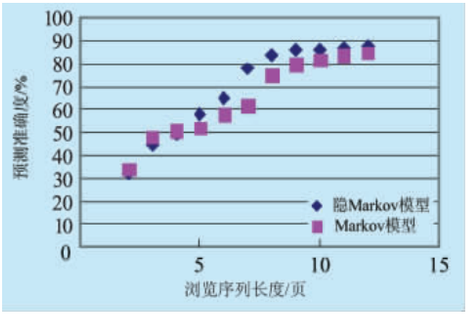

图 2　实验预测结果

Fig. 2　Experimental prediction result

80%,证明本文方法优于单 Markov 链模型。

4　结论

网络被认为是人类史上的第四次工业革命。今天 Web 已经成为信息发布、交互及获取的主要工具。Web 上的信息量正以惊人的速度增加,人们迫切需要能自动地从 Web 上发现、抽取和过滤信息的工具。如何从数以亿计的页面中发现需要的内容、从大量的访问中发现固有的模式和关联,成了人们迫切希望解决的问题。本文重点提出了基于 HMM 的网页浏览预测模型,并且从挖掘网页内容的角度出发捕获用户的真正兴趣,提出基于 HMM 的网页浏览预测,提高预测的准确率,并通过实验验证了当浏览序列长度逐渐增加时,HMM 优于单 Markov 链模型。

参考文献(References)

[1] Borges J, Levene M. A fine grained heuristic to capture web navigation patterns[J]. *SIGKDD Explorations*, 2000: 2(1): 40–50.

[2] 李颖基, 彭宏, 郑启伦. Web 日志中有趣关联规则的发现[J]. 计算机研究与发展, 2003, 40(3): 436–439.
Li Yingji, Peng Hong, Zheng Qilun. *Computer Research and Development*, 2003, 40(3): 436–439.

[3] 朱明. 数据挖掘[M]. 合肥: 中国科学技术大学出版社, 2002: 59–281.
Zhu Ming. Data mining [M]. Hefei: China University of Science and Technology Publishing House, 2002: 259–281.

[4] 周欣, 沙朝锋, 朱扬勇, 等, 兴趣度——关联规则的另一个阈值 [J]. 计算机研究与发展, 2000, 37(5): 627–633.
Zhou Xin, Sha Chaofeng, Zhu Yangyong, *et al. Computer Research and Development*, 2000, 37(5): 627–633.

[5] 王实, 高文. 路径聚类: 在站点中的知识发现 [J]. 计算机研究与发展, 2001, 38(4): 483–486.
Wang Shi, Gao Wen. *Computer Research and Development*, 2001, 38(4): 483–486.

[6] 邢永康, 马少平. 多 Markov 链用户浏览预测模型 [J]. 计算机学报, 2003, 26(11): 1511–1517.
Xing Yongkang, Ma Shaoping. *Chinese Journal of Computers*, 2003, 26 (11): 1511–1517.

(责任编辑　杨书卷)