基于客户行为模式的Web文档预送*

朱培栋 卢锡城 周兴铭

(国防科学技术大学计算机学院 长沙 410073)

E-mail: pdzhu@nudt.edu.cn

摘要 预送作为主动cache,是cache机制由时间局部性向空间局部性的拓展.文章提出服务器主动预送的两种模式.基于单个URL的模式利用客户请求的马尔可夫链特征获取文档的时序相关模型,可进行多级预送.基于会话的模式包括基于文档属性和会话整体语义的途径,重点研究基于文档属性的途径,给出基本的聚集算法,探讨了文档兴趣的定量表达,提出反映访问时序的属性向量距离算法.对于预送性能的度量,给出请求命中率、会话命中率、预送效率和预送代价等度量方法.同时,完成大量实验,对客户行为分析的这两种模式进行比较.文章提出的由服务器访问记录提取客户行为模式的方法,不但适用于文档预送,对于服务器站点设计和ISP(internet service provider)的服务规划也有重要价值.

关键词 WWW响应性能,预送,请求相关性,马尔可夫链,会话聚集,文档属性,预送性能.

中图法分类号 TP393

单纯的cache技术只是利用了WWW访问模式的时间局部性,对于未曾访问过的内容无法缓冲,响应性能依然得不到改善,这一点在客户发现一个新的热点服务器或服务器的页面经常更新时,感觉尤其明显.另外,如果客户机器或本地代理服务器用于WWW内容缓冲的空间不大,曾经访问过的内容被覆盖,单纯的cache机制也不会产生好的响应性能.

预测客户将来可能发出的访问请求,在客户浏览当前文档时,服务器将预测的内容发送出去,当真正要访问这些文档时,只需由本地cache获取.这种预送技术(presending)利用文档的空间局部性,是主动cache,是cache机制由时间局部性向空间局部性的扩展.

分析文献[1]中的服务器访问记录,文档平均发送时间为1.37s,平均请求间隔为44.59s,可见服务器有足够的时间预送,能够保证预送的时效.35 290个访问请求获取的文档平均为8 241字节,多为小文件;WWW应用交互性强,文件命中率比字节命中率重要,所以预送的带宽损失不大.平均一次会话有15个请求,最多包含224个请求,请求次数超过30个的会话占13.4%,具有良好的预测基础和预送价值.

预送分为主动预送和请求预送.请求预送由浏览器预测并发出预取请求,由于浏览器对服务器内容和文档组织缺乏了解,只能根据简单的单层嵌入关系进行.而主动预送,由于服务器拥有大量的访问记录,对自己的文档熟悉,了解各类客户的访问模式,因此,可以更加有效地预测和预送.

1 请求相关性和预送集合构造

1.1 预送集合构造

预送过程包括学习、匹配和发送这3个阶段.学习是分析历史行为或其他客户行为,提取客户一般情况下或

---

* 本文研究得到国家863高科技项目基金资助.作者朱培栋,1971年生,博士,主要研究领域为高性能协议.卢锡城,1946年生,教授,博士生导师,主要研究领域为高性能计算机,先进网络,并行与分布处理技术.周兴铭,1938年生,教授,博士生导师,中国科学院院士,主要研究领域为高性能计算机,光互连技术,数据库.

本文通讯联系人:朱培栋,长沙 410073,国防科学技术大学计算机学院

本文1998-07-14收到原稿,1998-12-07收到修改稿

所属类别的行为模型;匹配是确定正在进行的客户访问所属的行为模型,根据共同访问模式来推断将来的请求;发送是指,将预测内容根据预送策略传输至客户,具体预送策略包括预送的层次、数目、体积、概率阈值和发送速率的选择.

学习和匹配属于预测过程,预测基于类推原则和连贯原则.类推是一种空间上的扩展,是指新的客户行为模式和现有的多数客户相似;连贯原则实现时域拓展是指客户的未来访问请求与过去的访问模式相似.本文只讨论基于服务器访问记录的学习方法.

1.2 请求相关性

HTTP是无状态应用协议,客户和服务器的一次会话在服务器看来是持续一定时间(例如3小时)的客户访问序列,一次会话内客户的访问行为相对确定.服务器的访问记录Log表示为会话序列,$Log=\langle S_1,S_2,\dots\rangle$.会话$S$为客户访问请求序列,$S=\langle Client,\langle Req_1,Req_2,\dots\rangle\rangle$.

请求的属性非常多[2],与预送有关的属性考虑请求时刻T_r和请求的文档Doc,那么请求可以表示为$Req=\langle S,T_r,Doc\rangle$,其中$S$为该请求所属的会话.对请求相关性的分析需要研究请求所属会话的相关性、请求时序和文档关联,用R_R表示请求相关性,那么$R_R(Req_1,Req_2)=\langle R_S(S_1,S_2),R_T(T_{r1},T_{r2}),R_D(Doc_1,Doc_2)\rangle$.$R_S$称为类属相关性,$R_T$为时序相关性,$R_D$为结构相关性.

2 基于URL的访问模式

2.1 基于马尔可夫链的预测

$\{\zeta(n),n=0,1,2,\dots\}$是状态空间为$I$、参数为非负整数的随机过程.若

$$P\{\zeta(n+1)=j|\zeta(0)=i_0,\zeta(1)=i_1,\dots,\zeta(n-1)=i_{n-1},\zeta(n)=i\}=P\{\zeta(n+1)=j|\zeta(n)=i\},\tag{1}$$

则称$\zeta(n)$为马尔可夫链.条件概率$P\{\zeta(k+1)=j|\zeta(k)=i\}$为在时刻$k$时的一步转移概率,记为$P_{ij}(k)$.若$P\{\zeta(k+1)=j|\zeta(k)=i\}=P_{ij}$,即状态$i$转移到$j$的概率与$k$无关,则称该过程为齐次马尔可夫链.转移矩阵$P=[p_{ij}]$.

将客户请求看做离散空间的一个事件,请求过程具有齐次马尔可夫链特征.客户在t时刻已经对文档D_i发出访问请求,设在时间$(t,t+T_W]$内对另一个文档D_j访问的概率为P_{ij},T_W称为预送窗口,P_{ij}称为预送概率,所有$P_{ij}(0\leqslant i,j<N)$构成预测方阵P,N为服务器备有的文档数目.

利用m步转移概率$P_{ij}^{(m)}$,得到文档的多级相关矩阵$P^{(m)}$,用于多级预送.

$$P_{ij}^{(m)}(n)=P\{\zeta(n+m)=j|\zeta(n)=i\},\tag{2}$$

$$P^{(m)}=P^{(m-1)}P^{(1)}.\tag{3}$$

2.2 预送性能

设客户A发出的请求req_i要求访问文档doc_i,这次请求引发预送,实际预送的所有有效文档构成的集合为$P(doc_i)$.若采用基于URL的预送模式,$P(doc_i)$在cache中的有效期至客户发出req_{i+1}为止,那么,请求req_{i+1}的命中率为

$$h_r(req_{i+1})=\sum_{d\in P(doc_i)}hit(d,req_{i+1}),\tag{4}$$

即req_{i+1}所请求的文档属于$P(doc_i)$的概率.预送文档数目越多,命中率越大,即若预送集合$P_1(doc_i)\subset P_2(doc_i)$,相应的命中率必定是$h_{r1}\leqslant h_{r2}$;命中率越高,客户从发出访问请求到文档送回的这一段等待时间$T_{access}=h_r*T_{cache}+(1-h_r)(T_{cache}+T_{remote})$和客户上机时间$T_{client}=T_{read}+T_{access}$越小,工作效率可以进一步提高,其中$T_{cache}$为cache访问时间,$T_{remote}$为从服务器远程获取文档的时间,$T_{read}$为文档阅读时间.设客户$A$在一次会话$s$期间从浏览器上发出$M$个请求,若有$hit_num$个命中,那么会话$s$的命中率表示为$h_s(s)=hit_num(s)/M$.

客户对文档doc_i进行访问所引发的预送,预送效率$u_r(doc_i)=h_r(req_{i+1})/\#[P(doc_i)]$,预送代价$q_r(doc_i)=S_{p(doc_i)}$,其中$S_{p(doc_i)}$为预送集合$P(doc_i)$中所有文档的大小之和.如果客户对文档doc的请求没有传送到服务器,而是在本地cache得到响应,我们设服务器可以得知这一信息,那么在客户阅读doc时,服务器同样可以进行基

于 doc 的预送. 会话 s 的预送效率 $u_s(s)=hit\_num(s)/\sum_{i=1}^{M}\#[F(doc_i)]$,表示在一个会话中预送一个文档命中客户请求的平均概率,可以近似看做式(4)中 $hit(d,req_{i+1})$ 相应的后验概率. 会话 s 的预送代价 $q_s(s)=\sum_{i=1}^{M}S_{p(doc_i)}$.

2.3 实 验

逐行分析服务器的访问记录(log file),登记每一个遇到的新文档,设服务器共有 N 个不同的文档被客户访问. 对曾经访问过的每个文档 D_i,设该文档被所有客户访问的次数为 M_i. 由于相继请求之间的时间间隔相差很大,从几秒到几千秒不等[1],所以不用时间区间作预送窗口,而是选用文档数目作为 T_w,预送概率 P_{ij} 定义为紧跟 D_i 被同一客户访问的 T_w 个文档中出现 D_j 的概率. 按时间顺序分析访问记录,设 T_w 内以 D_i 开始的各个客户的访问序列中文档 D_j 出现的总次数为 Q_j,那么,预送概率可以通过 $P_{ij}=Q_j/M_i$ 近似计算. 对所有 D_i,$i\in[1,N]$ 进行类似分析,从而得到相关矩阵 P.

文献[1]中的访问记录可以区分为 2 286 个会话,我们使用 1 000 个会话的记录用于预送表的构造,余下 1 286个会话的记录用于验证. 文献[1]只是 24 小时的访问记录,实际构造预送表需要几天或数十天的会话记录,才能足够准确地把握客户行为模式.

设预送窗口 $T_w=5$,可以构造类似于表 1 的预送表,图 1 是相应的文档相关图,表 1 和图 1 中标识的小数为预送概率. 在实际实验中,我们设预送阈值 $k_p=35\%$,将预送概率大于 k_p 的文档纳入预送备选集合. 这样,如果客户对服务器访问记录中出现的所有文档分别发出访问请求,经统计可知,各个文档的预送备选集合平均含有 3 个文档,最多为 12 个. 构造好预送表之后,当实际预送时,可以只从预送备选集合中选择预送概率最大的文档,也可选择一定数目的文档,如果在 T_w 内把预送概率大于 k_p 且小于 250KB 的文档全部预送. 则平均会话命中率 $h_s=42.5\%$,h_s 超过 85% 的会话占会话总数的 18.6%. 各会话平均预送效率 $u_s=13.2\%$,平均预送代价 $q_s=410$KB.

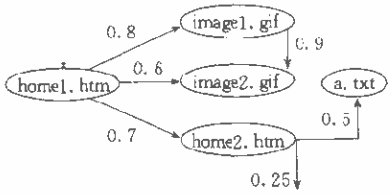

图1 文档相关图

表1 文档预送表

+	home1. htm	
—	image1. gif	0.8
	image2. gif	0.6
—	home2. htm	0.7
+	home2. htm	
—	a. txt	0.5

3 基于会话的访问模式

上述对基于 URL 的访问模式的分析,对所有客户访问 URL 的情况所进行的统计、用于预测后继文档. 这种方法简单、直接,但是没有表达会话特征和会话之间的关系,不利于语义信息的提取和应用. 而且,对每个客户请求都预送一定数目的文档,会使服务器负载加重,带宽损失比较大.

这种模式对预送表的构造和预送策略的制定是基于这样一个假设:预送内容在客户发出实际的请求之后过期. 这样,一方面,由于客户实际发出请求的时刻无法预测,使成功预测的内容因不能及时到达浏览器而浪费;另一方面,还会使已经到达浏览器的未命中文档无效,如果后继文档的预送集合包括这一文档,仍然会再次预送给客户,尽管可以在服务器方增加状态信息或由浏览器告知这一信息,但这不是基于 URL 的模式所固有的机制.

基于会话访问模式的分析和预送可以避免基于 URL 方式的上述不足. 提取会话特征并将服务器会话集合分类,判断正在进行的客户访问所属会话类别,根据该类会话的共同特征预测该客户在会话的剩余时间将要访问的文档集合,进行一次性预送. 这种方法可实现高层次的行为抽象,分析的结果易于理解,对服务器站点设计和 ISP(internet service provider)的服务规划也有一定的价值.

基于会话的访问模式包括基于会话整体语义和基于各个文档属性的方法. 基于会话整体语义的方法需要对服务器各个文档作摘要描述,提取客户在会话中已访问文档的关键词,以此作为客户兴趣的标识,并作为关键词

自动搜索相同兴趣的其他文档,将搜索概要通知客户,用于导航或将查找结果按重要程度预送给客户.文献[3]中有类似作法,用于浏览器自动对客户感兴趣站点的跟踪.由于预测的内容未蕴含在访问记录之中,此方法比较适合于客户从多个服务器搜集感兴趣的主题,所以,本文重点研究是基于各个文档属性的方法.

3.1 基本聚集(clustering)算法

聚集就是根据数据子集的相似性对数据集合分类,是一种无教练(unsupervised)的机器学习.客户和服务器的一次会话可用属性向量 $v=(a_1,a_2,\ldots,a_N)$ 表示,元素 a_i 表示客户对服务器文档 i 的兴趣程度,N 为服务器备有的文档数目.会话 p 和 q 分别用其属性向量表示:$p=(p_1,p_2,\ldots,p_N)$,$q=(q_1,q_2,\ldots,q_N)$.每个会话作为向量空间的一个点,p 和 q 之间的相似性可用欧几里得距离(Euclidian distance)d 来度量,

$$d(p,q)=\sqrt{\sum_{i=1}^{N}(p_i-q_i)^2}. \tag{5}$$

判断会话 p 是否属于聚集 c,需要计算 p 与 c 的质心 m 的距离.在不引起混淆的前提下,可用 $d(p,c)$ 代表 $d(p,m)$.若 $c=\{v_1,v_2,\ldots,v_k\}$,则 m 的属性 m_i 为

$$m_i=\frac{v_{1i}+v_{2i}+\ldots+v_{ki}}{k}. \tag{6}$$

聚集分为层次方法和非层次方法,我们选用 Leader 层次算法[4].

算法 1. 会话聚集的 Leader 算法.

输入:一组向量 V

输出:一组聚集 $C=\{C_1,C_2,\ldots\}$,$C_i\subset V$.

过程:

(0) C 初值为空.

(1) 对 V 中每个向量 v:

寻找聚集 c 使 v 和 c 的质心距离最短,记为 d_{min};

如果 d_{min} 小于距离阈值 *DistanceThresh*,

则将 v 加入 c,

否则,将 $\{v\}$ 加入 C.

3.2 基于各个文档属性的聚集

向量 v 的属性项为服务器上的文档,属性值为客户对该文档感兴趣的程度 *val*.如果会话 v_1 和 v_2 访问的文档很少有不相同的,对相同文档的兴趣差异也小,并且访问顺序相一致,则 v_1 和 v_2 就应纳入同一聚集.

3.2.1 访问兴趣

• 文献[5]将文档的访问次数作属性值 *val*,但是,即使客户在一次会话中对同一文档兴趣非常浓厚,多次阅读,在服务器记录中也无法反映这一点.这是由于:在一次会话中,客户一般不会从服务器多次下载同一文档;浏览器的 cache 验证选项,多数情况下不是设置为每次访问都验证;即使每次验证,如果 cache 使用过期验证模型(expiration model)[2],也不与源服务器通信.

但是,可以由服务器统计出文档 doc 所引出的对其他文档的访问总次数 R_{num}.由超文本结构看,如果对 doc 的直接链接的下一层(physically linked)文档访问的次数多,表明客户对 doc 所含内容兴趣浓厚.

• 文档的大小不适于作 *val* 值,由于同一文档在不同会话中大小不变.但是,如果不考虑客户阅读速度的差异,文档的阅读时间可以度量客户的兴趣程度.同一文档的阅读时间 T_{read} 变化很大,文献[1]中不同客户在对某个 4KB 文档阅读时,以 10s 为最小区分单位,T_{read} 就有 6s～1297s 之间 43 种可能,可用来划分客户兴趣群.请求间隔和文档传输时间分别记为 $T_{inter\_request}$ 和 T_{on},有 $T_{read}=T_{inter\_request}-T_{on}$.由于客户在阅读时,可能被其他事情中断,所以设 T_{read} 最大为 1 800s,若计算出的 $T_{read}>1\ 800$s,就记为 1 800s.文献[1]中平均 $T_{inter_request}=44.59$s,平均 $T_{on}=1.37$s,平均 $T_{read}=43.22$s.

根据上述分析,客户对文档感兴趣的程度 $val=f(R_{num},T_{read})$.为了权衡表征兴趣的两个参数,令 $f(R_{num},T_{read})=R_{num}\sqrt{T_{read}/10}$.例如,文档 doc_1 的 $R_{num1}=3$,$T_{read1}=160$s,则 $val_1=7$;文档 doc_2 的 $R_{num2}=6$,$T_{read2}=10$s,则 $val_2=7$.这种定义符合客户的主观判断.

3.2.2 访问顺序

以式(5)计算向量距离，忽略了访问顺序. 例如，客户 A 的访问序列为(8,19,4,...)，客户 B 的访问序列为(4,8,19,...). 在进行预送时，对于 A，访问文档 8 时可预送 19 与 4；而对于 B，8 的后继就不包括 4. 以式(5)计算会话相似性，这种差异无法表达，而预送必须反映访问的时序关系.

统计文献[1]的服务器访问记录，客户访问过的不同文档共有 5 104 个，平均每个会话对 13 个文档访问，最多访问 85 个文档，所以属性向量可采用紧致形式表示为序偶〈pos,val〉的序列，其中 pos 为文档编号，val 就是上文所述的属性值. 会话 v_1 和 v_2 对应的属性向量都采用链表表示，链表元素为〈pos,val〉. 计算向量距离用以下方法.

算法 2. 反映访问顺序的向量距离计算：$distance(v_1,v_2)$.

输入：$pair*v_1$，$*v_2$

输出：$sqrt(sum)$

过程：

(0) $sum=0$.

(1) 循环 1：起始：$p=v_1$，$q=v_2$；终止：p 或 q 为空

若 $p\rightarrow pos<q\rightarrow pos$，则 $sum+=(p\rightarrow val)^2$，$p++$；

若 $p\rightarrow pos>q\rightarrow pos$，则 $sum+=(q\rightarrow val)^2$，$q++$；

否则，$sum+=(p\rightarrow val-q\rightarrow val)^2$；$p++$；$q++$.

(2) 循环 2：对 v_1 的其他元素，$sum+=(p\rightarrow val)^2$.

(3) 循环 3：对 v_2 的其他元素，$sum+=(q\rightarrow val)^2$.

例如，$v_1=$(〈8,1〉〈19,1〉〈4,2〉)，$v_2=$(〈4,2〉〈8,1〉〈19,1〉)，那么，$v_3=$(〈4,2〉〈8,1〉)和 v_1 的距离 $d_{31}=3$ 和 v_2 的距离 $d_{32}=1$，很好地反映了访问的时序特征.

3.3 匹配与预测

设服务器从收到客户 A 发出的 L_1 个访问请求时开始对后继请求预测，这 L_1 个请求构成会话 s 的一部分，记为 v_1. 服务器聚集算法对已有会话分类，$C=\{C_1,C_2,...\}$，$C_i=\{v_{i1},v_{i2},...,v_{ik}\}$.

设第 i 个聚集的判断函数为 f_i，可以采用指标极小化判断标准. 如果 $\forall i\neq k$，

$$f_k(v_1)<f_i(v_1), \tag{7}$$

则 $v_1\in C_k$. 这里，$f_i(v_1)=d(v_1,C_i)$. 只预送 C_k 中兴趣值超过 l 且不在 v_1 中的内容，具体方法如下. 设 C_k 的质心为 m_k，首先滤去 m_k 中兴趣值小于 l 的文档，得到 v'_2，$v'_2=h_l(m_k)$. 其中 $h_l(p)=q$：若 p 的第 i 个属性值 $p_i\geqslant l$，则 q 相应的属性值 $q_i=p_i$；否则 $q_i=0$. 然后排除已在 v_1 中的文档，那么，预送文档集合 $P(v_1)=sub(v'_2)-sub(v_1)$，其中 $sub(p)=\{n|p_n>0\}$.

在实际会话中，客户 A 又在浏览器上发出 L_2 个访问请求，会话 s 的向量表示 $v=v_1+v_2$，则会话命中率 $h_s(s)=\#[sub(P(v_1))\cap sub(v_2)]/\#[sub(v)]$，会话预送效率 $u_s(s)=\#[sub(P(v_1))\cap sub(v_2)]/\#[sub(P(v_1)]$，预送代价 $q_s(s)=S_{p(v_1)}$.

3.4 预送实验

与基于马尔可夫链的实验相似，使用文献[1]中 1 000 个会话的记录用于会话聚集，余下 1 286 个会话记录用于验证. 实验分以下步骤进行：(1) 预处理：访问属性计算，会话向量表示；(2) 聚集；(3) 匹配；(4) 预送.

滤掉少于 4 个访问请求的会话，共有 760 个会话参与聚集，生成 68 个子集，去掉只含 1 个会话的子集，余下 37 个子集用于预测. 令 $L_1=3$，客户发出 3 个请求之后进行匹配，匹配成功后，令预送兴趣阈值为 1，结果是每个会话平均预送文档 14 个，平均会话命中率 $h_s=37.6\%$，平均预送效率 $u_s=40.3\%$，平均预送代价 $q_s=130.2$KB.

与基于马尔可夫链的方法相比，在 h_s 相差不大的情况下(为 0.88 倍)，u_s 却高出很多(为 3.05 倍)，q_s 低很多(为 0.32 倍). 可见，基于会话的方法以较小的预送代价换取了相当程度的会话命中率.

4 相关工作

预取是并行和分布式文件系统提高存取效率的通用方法，在以文件为单位的预取中，多数通过应用提示进

行预取. 文献[6]中的预测方法是其中较有智能的一种，使用多阶上下文模型对文件系统的事件进行跟踪，匹配的前件是事件序列，不如本文提出的基于马尔可夫链的方法简单.

WWW 是分布式系统的特殊应用领域，简洁的 HTTP 协议使客户的访问行为得到准确的记录，并易于分析. 文献[7,8]对 Web 预取有简单的介绍，但是没有涉及预送代价等性能度量，没有考虑基于会话语义的途径. 文献[5]对客户访问模式分类，用于动态链的构造，但是未把握访问的时序特征，客户对文档兴趣的度量采用访问计数是不恰当的. 文献[9]通过统计文档翻页和窗口放大次数，获取客户对在线报纸某一主题的喜好程度，实现起来较为繁琐，并且不是基于服务器访问记录.

数据挖掘可以从大规模数据库中发现潜在的模式或规则[10]，本文对 WWW 访问记录的分析运用了数据挖掘的基本原理，借鉴了基本方法. 现有的服务器记录分析工具，如 WWWStat，Wusage 和 WebLog[11]等，不支持文档相关性和会话特征分析，无法用于预送和实质的客户行为预测.

5 结　语

主动预送只要求 cache 有效期为相继请求之间的间隔或一次会话持续时间，对客户尚未请求的内容可以缓冲并快速响应，同时减少浏览器的缓冲空间需求. 基于 URL 的马尔可夫链模式简单、直接. 基于会话聚集的途径，以较小的预送代价换取了相当程度的会话命中率，表达更为丰富且易于理解的语义特征. 作为一种无需客户干预的主动服务，文档预送对于改善响应性能和浏览质量、增强 ISP 竞争力具有重要的意义.

参考文献

1 ftp: //ita.ee.lbl.gov/traces/epa-http.txt.Z. Downloaded at April, 1998
2 Feilding R *et al*. Hypertext transfer protocol——HTTP/1.1. IETF RFC 2068, Jan. 1997
3 Ackerman M *et al*. Learning probabilistic user profiles. AI Magazine, 1997,18(2):48～56
4 Hartigan J. Clustering Algorithms. New York: John Wiley Press, 1975
5 Yan T W *et al*. From user access pattern to dynamic hypertext linking. Computer Networks and ISDN Systems, 1996, 28(1):1007～1014
6 Kroeger T M, Long D D. Predicting file system actions from prior events. In: Gray R ed. Proceedings of the 1996 USENIX Annual Technical Conference. Monterey, CA: Usenix Association, 1996. 319～328
7 Mroz M. A client based prefetching implementation for the world-wide-web. Technical Report, Computer Science Department, Boston University, Jan. 1996
8 Padmanabhan V P, Mogul J. Using predictive prefetching to improve world-wide-web latency. ACM Computer Communication Review, 1996,26(3):22～26
9 Sakagami H, Kamba T. Learning personal preference on online newspaper articles from user behaviors. Computer Networks and ISDN Systems, 1997,29(1):1447～1455
10 Fayyad U *et al*. From data mining to knowledge discovery in databases. AI Magazine, 1996,17(3):37～54
11 http: //awsd.com/scripts/weblog/. July 1998

Web Document Presending Based on User Behavior Patterns

ZHU Pei-dong　LU Xi-Cheng　ZHOU Xing-ming

(*School of Computer　National University of Defense Technology　Changsha*　410073)

Abstract　Presending is an active service which extends caching mechanism from temporal locality to spatial locality. Two modes of extracting user behavior patterns are proposed to predict future requests from clients for efficient presending. URL-based mode exploits the Markov-chain features of request series, and can be used for hierarchical presending. Session-based mode captures more semantics, and the authors' work emphasizes the clustering algorithm, feasible document weight definition, and attribute-vector-distance computation representing order of accesses. Their performance is evaluated using appropriate metrics such as request hit rate, session hit rate, presending efficiency and presending cost. Numerous experiments are carried out to compare the two modes. These methods are used for web presending, while they are helpful to web server design and ISP (internet service provider) service planning.

Key words　WWW responsiveness, presending, request dependency, Markov chain, session clustering, document attribute, presending performance.