基于用户访问路径分析的网页预取模型*

许欢庆+, 王永成

(上海交通大学 计算机科学与技术系,上海 200030)

A Web Pre-Fetching Model Based on Analyzing User Access Pattern

XU Huan-Qing+, WANG Yong-Cheng

(Department of Computer Science and Technology, Shanghai Jiaotong University, Shanghai 200030, China)

+ Corresponding author: Phn: 86-21-62932933, E-mail: xuhuanqing@sjtu.edu.cn

http://www.sjtu.edu.cn

Received 2002-07-04; Accepted 2002-12-10

Xu HQ, Wang YC. A Web pre-fetching model based on analyzing user access pattern. *Journal of Software*, 2003,14(6):1142~1147.

http://www.jos.org.cn/1000-9825/14/1142.htm

Abstract: With the enormous growth of information on the Web, Internet has become one of the most important information sources. However, limited by the network bandwidth, users always suffer from long time waiting. Web pre-fetching is one of the most popular strategies, which are proposed for reducing the perceived access delay and improving the service quality of Web server (QoS). A semantics-based pre-fetching model is presented in this paper. This model predicts future requests based on latent intention that the user's current access path implies in semantics, rather than on the temporal relationships between URL accesses, which overcomes the limitation of previous pre-fetching approaches. The hidden Markov model (HMM) is employed for mining actual intention from access patterns. Experimental results show that the proposed pre-fetching model has better general performance.

Key words: Web pre-fetching; latent requirement concept; hidden Markov model

摘 要: 随着网络信息的飞速增长,互联网已成为人们获取信息的重要来源.但是,受限于网络带宽,用户往往需要忍受较长的访问延时.为了缓解这种情况,人们提出了网页预取技术,用于降低用户的访问延迟,提高 Web 服务器的服务质量.提出一种基于用户访问路径分析的服务器端网页预取模型.模型通过对用户访问序列进行语义分析,提取路径中蕴含的信息需求,依此进行网页预取决策.为了实现用户访问序列中潜在意图的挖掘,模型还引入了隐马尔可夫模型.性能测试实验的结果表明,该模型具有较好的整体性能.

关键词: 网页预取;潜在需求概念;隐马尔可夫模型

中图法分类号: TP393 **文献标识码**: A

互联网技术的飞速发展使信息的共享和发布跨越了时空的限制,网络成为人们获取信息的重要来源.但是,受网络带宽的限制,用户在访问网页之前要忍受较长的等待时间.近年来,随着网络带宽的逐步升级,延时问题

* Supported by the National Natural Science Foundation of China under Grant No.60082003 (国家自然科学基金)

第一作者简介: 许欢庆(1973－),男,江西临川人,博士生,主要研究领域为智能信息检索,Web 挖掘,信息过滤与推荐.

得到一定程度的缓解,但对网络服务质量的影响依然存在,尤其是在一些特定领域.比如低速 Modem 上网、无线网络等.为了解决这些问题,人们提出了多种技术方案,其中最主要的有缓存(caching)和预取技术(pre-fetching)两种.缓存技术已在网络节点的不同位置得到广泛应用.但随着网络资源更新频率的增加,缓存带来的性能改善不再显著[1].网页预取又称为主动缓存技术.区别于被动缓存,网页预取技术通过分析用户访问历史记录,主动预测用户可能浏览的页面,预先取出并存放在缓冲区中,以备用户的访问,从而减少用户的访问延时.实际应用证明,网页预取技术配以一定的流量平滑技术,能够大幅度减少用户的访问延时,从而提高了 WWW 服务质量[2].

现有的网页预取技术按照实施的位置,可分为以下 3 种:服务器端、代理服务器端和客户端.其中,基于服务器端的预取方法多数建立在用户访问模式中时序关系的基础上.也就是说,如果用户的访问模式记录表明,页面 *A* 随着页面 *B* 被访问也被访问的概率很大,一旦用户访问页面 *B*,则可预取页面 *A*.这类方法在预取决策时,未分析用户访问模式中蕴含的语义关系.

Azer[3]提出基于概率模型的预取方法.根据服务器 Log 数据,服务器计算出在一定时间间隔内,网页间被连续访问的概率,并建立条件概率矩阵.以此,服务器预测用户的访问请求.Sarukkai[4]运用马尔可夫链进行访问路径分析和链接预测.在此模型中,用户访问的网页集作为状态集.根据用户访问记录,计算出网页间的转移概率,作为预测依据.这两种方法在预取决策时仅考虑用户当前访问请求,而没有从整个访问序列加以考虑.

Schechter 等人[5]构造用户访问路径树,采用最长匹配方法,寻找与当前用户访问路径匹配的历史路径,以此预测用户接下来的访问请求.路径树的构造与匹配需要的时间和空间复杂度较高.

徐宝文等人[6]提出一种基于数据挖掘的预取模型.模型利用客户端浏览器缓冲区数据,挖掘其中蕴含的兴趣关联规则,以此预测用户可能选择的链接.在此模型中,用户兴趣表现为对词条的兴趣,兴趣关联规则表示从一个词条转向其他词条的可能性.利用兴趣关联规则,结合用户当前访问的页面的轨迹和用户访问的当前页面,预测用户可能访问的链接.此模型在预取决策时,仅考虑了词条间转移的可能性,未考虑用户对超链指向页面的感兴趣程度.同时,在挖掘兴趣关联规则时也没有考虑词条间兴趣转移具有的传递性.

XU Cheng-Zhong 等人[7]引入神经网络实现基于语义的网页预取.通过抽取网页超链描述文字信息中的关键词作为神经网络的输入,神经网络输出结果作为预取依据.用户浏览路径途径的页面作为训练样本反馈给神经网络进行学习.由于关键词的多义性会影响预取的准确性,模型对预取网页的范围采取分类处理,不同类别构造不同的预取器.虽然这在一定程度上提高了模型的预取准确性,但却限制了模型的实用性.

朱培栋等人[8]提出提炼用户会话特征,按语义对用户会话进行分类.在回应用户访问请求时,服务器计算当前用户访问路径与各类别中心的距离,确定会话所属的类别.根据会话所属类别的共同特征,预测用户可能访问的文档,一次性地预送到客户端.这种方法在计算会话与类别中心的向量距离时,过于严格遵循访问模式的时序关系.

本文中,我们提出一种基于用户访问路径分析的服务器端预取模型.区别于现有的基于访问时序关系的预取方法,模型通过分析用户的访问路径,挖掘其中蕴含的用户信息需求,据此预测用户的下一步访问请求.为了实现用户访问序列中潜在意图的挖掘,模型还引入了隐马尔可夫模型(hidden Markov model,简称 HMM).

1 预取模型

用户在互联网冲浪时,通常具有明显的目的性.在特定需求的驱动下,用户连续选择网页阅读.因此,访问路径可以理解成用户为了满足信息需求提出的请求序列.显然,如果我们能够从用户访问路径中理解出用户的信息需求,则据此可以较为准确地预测用户的请求.我们将这种用户信息需求表示成对特定概念的需求.假设用户浏览路径为 $Page_a \rightarrow Page_b \rightarrow Page_c$,当前页面 $Page_c$ 包含指向页面 $Page_d$,$Page_e$ 和 $Page_f$ 的超链.通过分析用户访问路径可知,潜在需求概念为 *c*.评估页面 $Page_d$,$Page_e$ 和 $Page_f$,如果页面 $Page_d$ 与概念 *c* 最相关,则 $Page_d$ 是用户下一步最有可能访问的页面.

由上述分析可知,如何挖掘用户访问路径所蕴涵的信息需求概念,是进行实现网页预取的关键.在本模型中,我们运用隐马尔可夫模型(HMM)进行用户访问路径分析.其中,访问路径途径的 URL 对应于 HMM 的状态,访问页面涉及的概念构成输出符号集,模型的输出观察序列转变成概念输出序列.我们计算出概念观察序列的

输出概率,将其作为评价依据,判断此概念为访问路径潜在需求概念的可能性.显然,输出概率越大,表明此概念越有可能是访问路径所蕴涵的信息需求概念.最能满足这些概念的候选网页,被选定为预取目标.

1.1 服务器Log预处理

Web 服务器日志记录了每个用户访问请求的如下属性:访问时间、用户 IP 地址、访问资源的文件名或脚本、参数域.我们将在一段时间内用户连续提交的请求序列定义为用户会话.通过预处理,服务器 Log 文件可以整理成服务器会话集.预处理步骤为:(1) 剔除访问多媒体文件、脚本文件的用户请求;(2) 按用户的 IP 地址,将 Log 文件分割成独立的访问记录集;(3) 将每个访问记录集的请求按时间排序,设立时间窗口阈值 tw,分割访问记录集,时间间隔小于 tw 的相邻访问请求同属于一个用户会话,然后,所有的会话组成用户会话集;(4) 所有用户会话集组成服务器会话集.

1.2 信息提取

信息提取过程抽取出反映网页主题的特征词,计算特征词在页面中的权重,用特征词向量表示网页.经过预处理(分词、剔除停用词)之后,网页 p 可表示成特征词集 $Term=\{t_1,t_2,\ldots,t_m\}$,$t_i$ 为特征词.特征词在文档中的出现频率在一定程度上反映了文章主题.但在多数情况下,用户仅通过简单浏览即可确定页面内容或作出超链选择,文章中一些敏感性较高的词给予用户强提示作用.这些敏感性高的词的出现频率并不一定很高,而是出现在网页标签中,如 title,anchor text,url,key 等.因此,我们调整了特征词的频率计算.设特征词 t_i 在正文、标题、关键词、超链、超链描述中出现的频率分别为 $tf_{\text{other}}(t_i)$,$tf_{\text{title}}(t_i)$,$tf_{\text{key}}(t_i)$,$tf_{\text{url}}(t_i)$和 $tf_{\text{Anchore}}(t_i)$.特征词频率计算公式如下:

$$tf(t_i)=tf_{\text{other}}(t_i)+A_1\cdot tf_{\text{title}}(t_i)+A_2\cdot tf_{\text{key}}(t_i)+A_3\cdot tf_{\text{url}}(t_i)+A_4\cdot tf_{\text{anchor}}(t_i), \tag{1}$$

其中 A_1,A_2,A_3 和 A_4 是调整系数.特征词的权重计算公式如下:

$$w_p(t_i)=\frac{\left(\frac{tf(t_i)}{tf_{\max}}\right)}{\sqrt{\sum_{j=1}^{m}\left(\frac{tf(t_j)}{tf_{\max}}\right)^2}}, \tag{2}$$

$tf_{\max}$ 是特征词出现的最大频率值.权值越大的特征词反映文档主题概念的能力越强.权重超过一定阈值的特征词,组成文档特征词向量.规范化向量中特征词的权重,文档表示为

$$p=\{(t_i,w_p(t_i))\mid t_i\in T\}. \tag{3}$$

1.3 访问路径分析

Web 服务器在回应访问请求的同时,按照访问请求到达的时间顺序,将请求存入到对应的用户缓冲区中.在预定的时间阈值 tw 内,如果没有新的请求进入用户缓冲区,则缓冲区自动刷新.在用户缓冲区清空之前,保存在缓冲区里的请求序列就是当前用户的访问路径.设路径 $Path$,$Path=\{p_1,p_2,\ldots,p_n\}$,$p_n$ 为用户当前浏览页面.概念集 $C=\{c_1,c_2,\ldots,c_m\}$.我们运用离散隐马尔可夫模型分析访问路径.

隐马尔可夫模型最早由 Baum 提出,在许多领域得到运用.模型可以表示为 $\lambda=(A,B,\pi)$.

- N 为状态个数,t 时刻状态表示为 q_t;
- M 为离散输出符号个数,输出符号集 $V=\{v_1,v_2,\ldots,v_m\}$;
- 状态转移概率分布 $A=\{a_{ij}\mid 1\le i,j\le N\}$,

其中

$$a_{ij}=P[q_{t+1}=j\mid q_t=i]. \tag{4}$$

- 输出符号的观察概率分布 $B=\{b_j(k)\mid 1\le k\le M,1\le j\le N\}$,其中

$$b_j(k)=P[o_t=v_k\mid q_t=j]. \tag{5}$$

- 初始状态概率 $\pi=\{\pi_i\mid 1\le i\le N\}$,其中

$$\pi_i=P[q_1=i]. \tag{6}$$

给定 N,M,A,B,和 π,隐马尔可夫模型可以产生观察序列 $O=(o_1o_2\ldots o_n)$,o_i 为状态的输出符号.假设状态转移序列 $q=(q_1q_2\ldots q_n)$已知,则观察序列 O 的概率计算如下:

$$P(o\mid q,\lambda)=\prod_{t=1}^{n}P(o_t\mid q_t,\lambda)=b_{q_1}(o_1)\cdot b_{q_2}(o_2)\cdot\ldots\cdot b_{q_n}(o_n)\ . \tag{7}$$

这里,我们将访问路径作为隐马尔可夫模型的状态转移序列,页面中概念集作为状态输出符号集.观察序列为 $O=\left(\underbrace{c'c'\ldots c'}_{n}\right)(c'\in C)$的概率计算公式如下:

$$P(o\mid q,\lambda)=b_{p_1}(c')\cdot b_{p_2}(c')\cdot\ldots\cdot b_{p_n}(c')\ , \tag{8}$$

其中 $b_{p_j}(c')$ 表示页面 p_j 输出概念为 c'的观察概率.显然,如果 $b_{p_j}(c')$ 表述的是概念 c'在访问路径第 j 步受用户关注的概率,则观察序列 O 的概率可用来表示概念 c'是路径 Path 潜在需求概念的概率.由于在预取决策过程中,用户访问路径固定不变,问题转化成求取固定状态转移的观察序列输出概率 $P(o\mid q,\lambda)$.

分析访问路径的特性,路径中页面旨在细化概念,引导用户到达最能满足概念的页面.因此,我们计算出页面对某个概念的引导能力,作为概念的观察概率.在预处理后的服务器 Log 中,用户对应的会话集为 $S=\{s_1,s_2,\ldots,s_m\}$,会话 s_j 表示如下:

$$s_j=\{s_j[1],s_j[2],\ldots,s_j[t]\},\ 1\le j\le m, \tag{9}$$

其中,t 为会话 s_j 的长度.会话中跟随在页面 p_i 之后的页面组成集合 $T_j(p_i)$:

$$T_j(p_i)=\{s_j[k+l]\mid s_j[k]=p_i,l=1,\ldots,t-k\}\ . \tag{10}$$

考虑整个用户会话集,可以得到

$$T(p_i)=\bigcup_{s_j\in S}T_j(p_i)\ . \tag{11}$$

观察概率 $b_{p_j}(c')$ 可以计算如下:

$$b_{p_i}(c')=\frac{\sum_{p_j\in T(p_i)}w_{p_j}(c')}{\sum_{p_j\in T(p_i)}\sum_{c''\in C}w_{p_j}(c'')}\ . \tag{12}$$

将上述公式代入式(8),计算每个概念对应观察序列的概率 $P(o\mid q,\lambda)$.

1.4 网页预取

预取模型对用户当前访问路径进行分析,依此进行预取决策.为了便于用户浏览,网页作者在设计页面时,会使用简洁、精炼的文字,概括超链指向网页的内容.这些文本被称为超链描述信息(即 Anchor Text).用户阅读页面时,仅根据超链描述信息就可以判断出链接页面是否满足信息需求.因此,从预测用户下一步可能选择的链接的角度考虑,我们在分析访问路径时,只需考虑超链描述文本中的概念即可.

设页面 p_n 中包含的超链集为 $L=\{l_1,l_2,\ldots,l_m\}$,$Text(l_i)$表示超链 l_i 的描述文本.建立伪文档 Pseudo-Doc,有

$$\text{Pseudo-Doc}=\bigcup_{l_i\in L}Text(l_i)\ . \tag{13}$$

依据下述步骤, 提取 Pseudo-Doc 中包含的概念集 C:(1) 剔除 Pseudo-Doc 包含的 Html 标签;(2) 如果 Pseudo-Doc 为中文文本,进行分词;(3) 去除文本中的泛滥词;(4) 如果 Pseudo-Doc 为英文文本,进行去词根处理.文本中包含的词构成概念集 C.计算概念集 C 中每一个概念作为用户潜在需求概念的概率.评价当前页面 p_n 中包含的超链,计算超链预取的优先权值,公式如下:

$$pw(l_i)=\sum_{c'\in Text(l_i)}P\big(O(c')\mid q,\lambda\big)\ . \tag{14}$$

$O(c')$表示观察序列为 $O=\left(\underbrace{c'c'\ldots c'}_{n}\right)$,$n$ 为路径长度.显然,权值 $pw(l_i)$愈大,超链 l_i 愈有可能被用户访问.根据设定的预取阈值ξ,系统预取前ξ个超链链接页面,以备用户访问.

2 性能评价

2.1 实验设计

为了检验预取模型的性能,我们选择了 Http:\\naxun.sjtu.edu.cn 服务器日志进行预取性能模拟实验.该网站的主要话题是中文信息处理方面的相关技术.抽取服务器 log 中从 2001 年 11 月 1 日~2001 年 12 月 31 日的访问记录,组成实验数据集.预设时间窗口阈值 tw=0.5 小时.经过预处理后,数据集的主要特性如下:用户请求总数为 34 576,用户会话集个数为 73,用户会话总数为 5 694,最短会话长度为 1,最长会话长度为 11.数据集中,60%的会话被作为训练样本集,剩余会话用于测试集.

我们定义了请求命中率、平均会话命中率和带宽浪费率作为模型性能评价指标.

定义 1. 请求命中率表示被成功预取的用户请求数与用户提交的请求总数的比率.

定义 2. 会话命中率表示用户会话中被成功预取的请求数与用户提交的请求总数的比率.平均会话命中率即为会话命中率的平均值.

定义 3. 带宽浪费率表示预取页面中未被用户访问的页面数与总预取页面数的比率.

为了对比测试模型的预取性能,我们选用 Sarukkai[4]提及的方法作为对比实验方案.它运用马尔可夫链模型(MCM)进行链接预测,为了便于叙述,文中将我们提出的模型简称为 PAP,将对比模型称为 MCP.

2.2 测试结果分析

测试数据集中,每一个用户会话都作为当前访问路径,对访问路径中每一步进行预取仿真测试.图 1 显示了不同步进下(2~11),预取模型的请求命中率的变化情况.在实验中,预取网页阈值ξ设定为 4,图中横坐标表示用户浏览步进,纵坐标为请求命中率.由图可知,随着访问步进由 2 增加到 6,PAP 模型的请求命中率由 50.2%递增至 57.1%.当步进超过 6 后,PAP 模型的请求命中率转为下降趋势(由 57.1%降至 53.3%).PAP 模型的请求命中率发生变化的原因是,随着用户浏览步进的延伸,用户的信息需求意图变得明确,提高了挖掘的准确性.但是,当浏览路径进一步拉长时,用户的兴趣注意力发生转移的可能性增加,从而给预测引入了噪声,降低了预取准确度.

与 PAP 模型比较,访问步进对 MCP 模型的请求命中率影响不大,其值始终保持在 51.1%~51.5%.这是由于它所依赖的进行预测的网页转移概率由训练样本集计算出来后,并不随着浏览步进的递增而改变.请求命中率的曲线图还表明,在用户浏览开始之初(步进为 2 和 3),两个模型的请求命中率相近.主要原因是,当用户浏览路径较短时,路径的语义理解较为困难.此时,我们的模型与基于时序的预取方法类似.

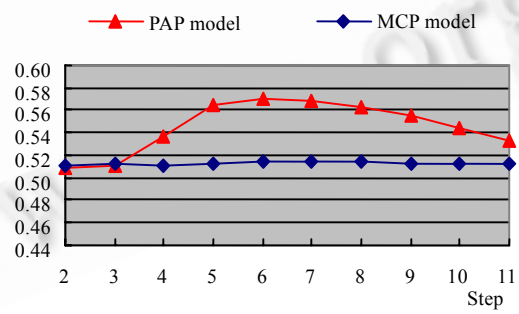

Fig.1 Request hit ratio vs number of steps

图 1 不同步进下,预取模型的请求命中率

预取阈值ξ是影响预取模型性能的主要因素.我们测试不同预取阈值ξ下(1~7),预取模型性能的变化情况.图 2 和图 3 分别显示预取模型的平均会话命中率和带宽浪费率的变化.

图 2 的测试结果数据显示,随着预取阈值ξ的提高,两种模型的平均会话命中率总体趋势是在增加.PCP 模型的平均会话命中率由 50.1%增加到 59.7%,MCP 模型的平均会话命中率由 47.3%增加到 54.1%.曲线也表明:在ξ=4 的前后段,平均会话命中率的增长速度不同.后段的增长速度明显变缓.两个预取模型的平均会话命中率指标均高于请求命中率,这是因为预取的页面在缓冲区中有一定的生存周期.在生存周期内,预取页面可以为会话

中以后的请求访问继续服务.

图 3 的数据表明:带宽浪费率在预取阈值ξ小于 4 的范围内,随着预取阈值ξ的增加,呈现下降趋势.PCP 模型的带宽浪费率由 39.7%降至 30.0%,MCP 模型的平均会话命中率由 44.0%降至 36.7%.但随着预取阈值ξ的进一步增加,PCP 模型的带宽浪费率由 30.0%增加到 45.7%,MCP 模型的平均会话命中率由 36.7%增加到 47.3%.这是因为,在预取阈值增加的初期,每次预取的页面数目增加,预取准确度提高,造成带宽浪费率的下降.随着预取的页面数目进一步增加,预取准确度增长速度趋于平缓,带宽浪费率从而转为增长趋势.由此可见,预取阈值的选择应综合考虑命中率和带宽浪费.在本实验中,预取阈值选择为 4 或 5 较为合理.对比图 2 与图 3 中的曲线可知,PCP 预取模型具有更好的总体性能.

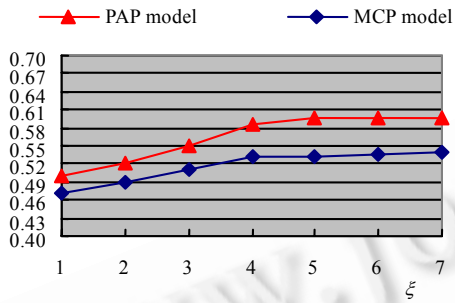

Fig.2 Avg. session hit ratio vs the pre-fetching threshold ξ

图 2 不同预取阈值ξ下,预取模型的平均会话命中率

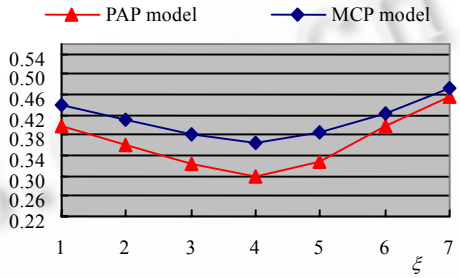

Fig.3 Effects of per-fetching threshold ξ on waste bandwidth ratio

图 3 不同预取阈值ξ下,预取模型的带宽浪费率

3 结 语

网页预取技术缓解了服务器访问延时的问题,提高了 Web 服务的质量.在许多实际系统中,预取技术都得到了较好的运用.在本文中,我们提出一种基于用户访问路径分析的服务器端预取模型.模型引入隐马尔可夫模型对用户访问路径进行语义分析,基于此进行预取决策.实验证明,我们所提出的模型具有较好的预取性能.

网页预取技术与许多研究领域具有相通性.本文提出的模型经过相应的修改,可应用于个性化信息推荐、信息过滤领域.

References:

[1] Thomas MK, Darrel DEL, Jeffrey CM. Exploring the bounds of Web latency reduction from caching and prefetching. In: Proceedings of the USENIX Symposium on Internet Technologies and Systems. California: USENIX Association, 1997. 13~22.

[2] Crovella M, Barford P. The network effects of prefetching. In: Proceedings of the IEEE Conference on Computer and Communications (INFOCOM'98). San Francisco, 1998. 1232~1240.

[3] Bestravros A. Using speculation to reduce server load and service time on the WWW. In: Proceedings of the CIKM'95. Baltimore, 1995. 403~410.

[4] Sarukkai R. Link prediction and path analysis using Markov Chains. Computer Networks, 2000,33(1-6):377~386.

[5] Schechter S, Krishnan M, Michael DS. Using path profiles to predict http requests. In: Proceedings of the 7th International World Wide Web Conference. Brisbane, 1998. 457~467.

[6] Xu BW, Zhang WF. Applying data mining to Web pre-fetching. Chinese Journal of Computers, 2001,24(4):1~7 (in Chinese with English abstract).

[7] Xu CZ, Tamer IL. Semantics-Based personalized prefetching to improve Web performance. In: Proceedings of the 20th IEEE Conference on Distributed Computing Systems. 2000. 636~643.

[8] Zhu PD, Lu XC, Zhou XM. Web document presending based on user behavior patters. Journal of Software, 1999, 10(11): 1142~1147 (in Chinese with English abstract).

附中文参考文献:

[6] 徐宝文,张卫丰.数据挖掘技术在 Web 预取中的应用研究.计算机学报,2001,24(4):1~7.

[8] 朱培栋,卢锡城,周兴铭.基于客户行为模式的 Web 文档预送.软件学报,1999,10(11):1142~1147.