个性化 Web 推荐服务研究

韩晓莉[1,2] 李秉智[3]

(重庆邮电学院计算院 重庆 400065)[1] (中国联通重庆分公司 重庆 400042)[2]

(重庆邮电学院移通学院 重庆 400065)[3]

摘 要 本文主要论述了个性化 Web 推荐构成,提出了基于 Web 挖掘的个性化推荐服务研究中的用户聚类、Web 页面聚类、n 元预测模型及页面加权算法。利用这些算法得到的个性化信息可以准确把握用户兴趣模式并为用户提供"一对一"的具备自适应性的智能个性化服务。

关键词 个性化,Web 挖掘,Web 推荐

Personalization Research for Web Recommendation

HAN Xiao-LI[1,2] LI Bing-Zhi[3]

(Computer Science Institute of Chongqing Post and Telecommunications University, Chongqing 400065)[1]

(China Unicom Chongqing Branch Chongqing, Chongqing 400042)[2]

(Mobile Telecommunications Institute of Chongqing Post and Telecommunications University, Chongqing 400065)[3]

Abstract This paper mainly discusses the structure of Web recommendation based on personalization and provides many algorithms used in personalization service such as customer clustering, Web pages clustering, n-gram prediction model et al. With these algorithms we can exactly hold user interests model and improve the efficiency of the network information supply dramatically.

Keywords Personalization, Web mining, Web recommendation

1 前言

随着网络技术的发展及机器学习、模式识别等知识发现新技术的出现,电子商务竞争已使得信息服务方式从传统的"一对多"发展到"一对一"的个性化服务方式。随着电子商务中引入个性化用户服务方式,企业需要对 Web 环境下的客户资料数据进行深入的统计与分析,找出不同用户兴趣所在,透视隐藏在这些数据之后的更重要的用户兴趣模式信息以及关于这些数据的整体特征的描述并预测其发展趋势等。

本文介绍的个性化 Web 推荐服务是利用个性化技术将传统的数据挖掘(data mining)对象同 Web 访问信息结合起来,利用 Web 挖掘[1,2]的方法抽取用户感兴趣的潜在有用模式与信息[3],然后基于这些模式和信息为用户提供"一对一"的具备自适应性的智能个性化推荐服务[4]。这些智能个性化推荐服务可大大缩短用户在网络上的访问延迟,使得提供给用户的网络信息服务质量得到最大程度的提高。

2 基于 Web 挖掘的个性化推荐

Web 挖掘是从 WWW 上抽取知识的过程。它是从与 WWW 相关的资源和行为中抽取感兴趣的有用的模式和隐含信息[5,6],也是将数据挖掘技术和理论应用于 WWW 资源进行挖掘的一个新兴的研究领域。Web 挖掘一般可分为三个部分[7]:Web 内容挖掘、Web 结构挖掘、Web 使用挖掘。

WWW 上每一个提供信息资源的服务器上都有一个结构比较好的记录集,即 Web 访问日志。每当有获取资源的请求到来时,Web 服务器都将记录和积累这些关于用户交互作用的数据。利用 Web 挖掘方法分析不同的 Web 站点和 Web 访问日志可帮助人们根据用户访问的 Web 页面内容及用户群访问的相似性,进行页面和用户聚类分析,进而为用户提供个性化的服务[8]。系统通过挖掘 Web 服务器用户日志文件获取的用户兴趣爱好、Web 访问模式等个性化信息可以为用户提供感兴趣的站点、网页与链接,甚至直接对用户进行页面内容过滤,传送。

个性化推荐系统关键技术可概括为:通过分析服务器日志文件,挖掘用户兴趣爱好,为其提供感兴趣的站点、网页与链接,甚至可以直接对用户进行页面内容过滤、传送等;或者根据页面内容及用户群访问的相似性进行页面和用户分类聚类,根据类间的相似性进行 Web 推送。

Web 个性化推荐服务主要分为两个阶段:首先是根据 Web 个性化数据获取用户属性、分析用户兴趣特征、聚类分析、频繁访问路径发现等;然后利用推荐系统将 Web 页面内容与用户兴趣访问模式相结合,通过用户群的相似性进行 Web 页面访问预测以及内容推荐等。

2.1 个性化推荐系统技术分类

个性化推荐技术能充分提高站点的服务质量和访问效率,从而吸引更多的访问者。目前存在着许多个性化推荐服务系统,它们提出了各种思路以实现个性化推荐服务。个性化推荐服务系统根据其所采用的推荐技术可以分为两种。

2.1.1 基于规则的技术 企业 Web 站点管理员根据用户统计数、静态个性文件或用户会话(User Session)记录制定一系列规则并利用这些规则为特定用户提供特定服务。规则可以通过用户个人输入静态信息来获取,也可以基于用户动

韩晓莉 工程师。李秉智 教授。

态浏览信息来建立。系统在信息推荐时根据当前用户浏览历史的感兴趣内容，判断新的用户偏好内容，并按其重要程度排序后推荐给用户。基于规则的推荐系统结构如图 1 所示：关键词层主要提供用户的静态属性，以关键词的定义静态规则；动态属性层根据用户的浏览偏好动态定义个性化规则；用户接口层则按下两层提供的规则提供个性化推荐服务。

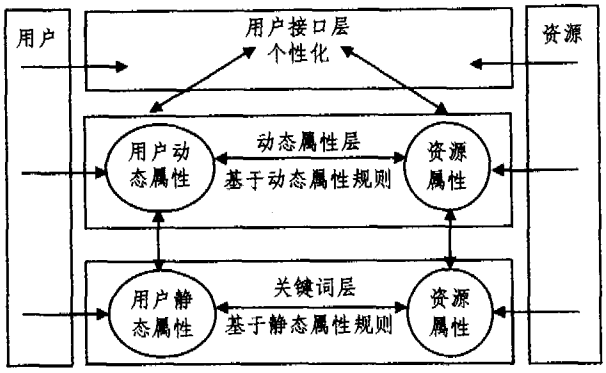

图 1　基于规则的个性化推荐

2.1.2　信息过滤技术　可分为基于内容过滤的技术和协作过滤技术。基于内容的过滤技术(Content-based Filtering)，直接利用 Web 信息，通过用户历史访问内容挖掘用户访问模式并将该模式需求同 URL 结合以满足用户个性化需求。基于内容的过滤技术的关键是相似度计算，但从用户历史访问内容中挖掘用户兴趣属性也非常重要，它直接关系到推荐内容是否与户兴趣度相关；协作过滤技术(Collaborative Filtering)推荐关键在于用户聚类，它是利用用户的访问信息，通过用户群的相似性进行内容推荐，所以可以为用户推荐新的感兴趣内容。图 2 为基于内容过滤技术和协作过滤技术的个性化推荐系统结构。

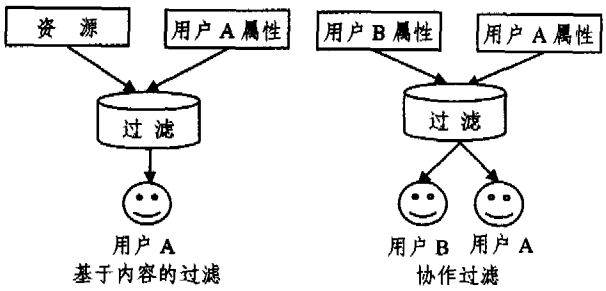

图 2　基于信息过滤技术的个性化推荐

1)基于内容的过滤。是通过分析对象的内容来形成一个对访问者兴趣的表达。通常，这种分析识别每个对象的一组关键字属性，然后填写属性值。

2)协作过滤。使用显式或隐式评价，收集一组对象访问者的意见，来形成具有相似意向的同等组，然后研究同等组，从而预测特定的访问者对于某项的兴趣。与基于内容的过滤寻找与访问者过去喜欢的对象类似的商品不同的是，协作过滤寻找具有类似兴趣的访问者来生成推荐。

2.2　Web 推荐系统原型

2.2.1　基于内容过滤的系统　Letizia 系统[9]是由 MIT 开发的，具有智能导航功能。它采用了一种基于行为的用户兴趣建模方法，即通过跟踪用户的浏览行为推测用户兴趣，建立用户兴趣模型。例如该系统可自动从用户当前页面出发，对所有超链接指向的链宿页面进行宽度优先搜索。在分析页面内容后与用户兴趣模型比较，进而找出用户可能感兴趣的页面，在单独的窗口中显示推荐给用户的 URL 列表。

LIRA 系统[10]是由 Stanford 开发的，具有主动服务功能的系统。在用户网络浏览过程中选择与用户兴趣模型相似度高的页面提交给用户，并要求用户给出明确的评估值，然后根据用户提供的相关反馈结果修改搜索和选择启发值，调整用户兴趣模型。该系统特点在于利用了启发式搜索算法，对搜索规模进行了限制，从而兼顾了效率。

Personal Web Watch[11]是由 CMU 开发的。在用户浏览 Web 时，该系统根据超文本链接标记文本，预测叶节点页面的兴趣度，并对用户感兴趣页面中的超链做标记，建议用户访问。在用户离线后 PWW 根据记录的访问页面及 URL 地址分析页面内容，更新用户的兴趣模型，进行页面推荐，从而指导下一次浏览。因为 PWW 根据构造的用户兴趣模型进行页面推荐从而减少了用户为查找所需信息耗费的时间，同时也提高了浏览质量，并且由于 PWW 具有自学习的功能，用户兴趣模型也可不断地加以更新及调整。

2.2.2　协作过滤系统　Webwatcher[12]是一个非常著名的导航器，它使用一个称为信息查找助理的 Agent，导航用户在网上的浏览过程。该系统通过对用户选择“链路”或站点进行跟踪学习，学习产生哪一超链是可能到达目标信息的知识，通过采用这些知识来帮助用户定位希望的信息，改善导航质量。

ProFusion Personal Assistant[13]是一个信息过滤工具，它使用用户明确的相关度反馈决定用户感兴趣的领域，以此为据，将元搜索引擎 ProFusion 返回的结果进行过滤，决定哪些结果提交给用户，哪些抛弃。

3　Web 推荐系统中的用户聚类及兴趣模型构建

个性化推荐服务中的用户聚类主要是指通过分析 WWW 服务器的日志文件获取 Web 用户行为模式，并将其量化，然后利用一定的算法进行用户聚类的过程。

3.1　基于神经网络的用户聚类算法

利用 Web 使用挖掘中，通过对日志文件进行数据清洗可获得数据挖掘源。对该数据源进行扫描建立用户会话(user session)，即每遇到一个新的 IP 地址就为其创建一个用户会话，以后将隶属于该 IP 地址发出的连续请求都加入该会话中(连续请求指两个请求间的时间间隔不超过预先设定的阈值)。在同一次用户会话中，若用户访问了网站中的 n 个页面，则该会话可用一个 n 维向量表示，其 i 维向量值为用户对第 i 个页面的兴趣度即权重。如此，根据日志文件提取的用户访问信息就可以用模式向量形式表达出来。

考虑到神经网络具有良好的聚类特性，可利用神经网络对用户访问模式向量进行聚类分析。

Kohonen[14]神经网络是基于无监督方式学习方法进行训练的神经网络。当向量进入 Kohonen 神经网络后，权值向量与输入向量具有最小欧氏范数距离的神经元作为神经元竞争中的获胜者。这样在网络训练稳定后，每一领域的所有节点对某种输入具有类似的输出，其获胜神经元的权重按以下方法训练：$W_{ij}(t+1)=W_{ij}(t)+\eta(t)[x_i(t)-W_{ij}(t)]$，其中 $\eta(t)$ 为衰减因子，i,j 分别表示输入层和输出层神经元的序号，权值向量是被随机初始化的。训练结果将权值向量逐渐靠近输入向量，经过一定数量的训练后输出的获胜神经元就能表示输入的不同用户的模式向量所属聚类。

以 Kohonen 神经网络的扩展 SOFM(自组织特征映射)模型为例[15](SOFM 模型是在 Kohonen 模型的基础上,在输出层神经元之间增加了侧向连接权值,从而在输出层引入侧反馈机制,同时将欧氏范数距离函数改为墨西哥草帽函数,在逐步缩小的邻域内以侧反馈的方式调节网络权值)。算法流程可描述为:

输入向量 X 表示为 $X=[x_1,x_2,\cdots,x_p]^T$,与输出层神经元 j 对应的权值向量 W_j 为 $W_j=[W_{j1},W_{j2},\cdots,W_{jp}]^T$,$j=1,2,\cdots,n$。选择权值向量 W_j 与输入向量 X 最为匹配的输出层神经元即为获胜的输出层神经元。我们选择权值向量距离输入向量有最小欧氏范数值的输出层神经元作为获胜神经元。如果用 $i(x)$来指定获胜神经元,有 $i(x)=k$

当 $\|W_k-X\|<\|W_j-X\|$,$j=1,2,\cdots,n$

获胜神经元学习过程可表示为

$$W_j(n+1)=\begin{cases}W_j(n)+\eta(n)[X-W_j(n)], & j\in\Lambda_{i(x)}(n)\\ W_j(n), & \text{其他}\end{cases}$$

其中 $\Lambda_{i(x)}(n)$表示第 n 次迭代时的邻域函数

经过该学习训练过程,输出的获胜神经元即能表示不同用户模式向量所属聚类

3.2 用户兴趣模型的自动评价方法

自动评价方法的思想主要基于:若用户长时间或高频率的访问某一个站点或某一具体页面,表明其对该站点或页面的兴趣度高。而访问时间及频度恰恰可作为兴趣度量的权重。

自动评价的算法如下:

首先,定义用户访问一个 IP 地址次数为 number(user,IP),考虑到数据清洗阶段要去除用户偶尔访问的站点,此时可设定在固定时间段 time 中 number(user,IP)≥2

用户访问一个 IP 地址的时间为:

$$Time(user,IP)=\frac{AllTime(user,IP)}{\max_{(visitedIP)}(AllTime(user,IP))/Size(IP)}$$

用户访问一个 IP 地址的新鲜度为

$$newtime(user,IP)=\frac{\sum_i^{number(user,IP)} visit_i(user,IP)-Visit(start)}{numer(user,IP)}$$

其中 $Visit_i(user,IP)$表示 user 在固定时间段 time 中第 i 次访问 IP 的时间,$Visit(start)$为开始记录记录日志的时刻

此时得到用户对某个 IP 地址自动兴趣评价值为:

$$Value(user,IP)=(\log_2 number(user,IP))\times(1+Time(user,IP+Newtime(user,IP)))$$

算法如下:

在对 Web 日志进行必要的预处理后,根据用户的 ID 号划分对应的访问记录集并统计用户访问一个 IP 地址次数 number(user,IP),剔除访问次数小于 2 的 IP。然后计算用户对其访问 IP 地址的评价值。以用户 ID 为行,IP 为列构建用户兴趣矩阵,然后利用基于近邻的方法[16,17]即可为用户提供个性化的推荐服务。

4 个性化服务中的 Web 页面分类

4.1 基于模糊聚类算法的 Web 页面聚类

模糊集理论[18]是 Zadeh 于 1965 年提出的,其定义如下:设 $U=\{u_1,u_2,\cdots,u_n\}$为论域,若集合 R 是其上的一个模糊集,则有 $R=\{(u_1,f_R(u_1,)),(u_2,f_R(u_2,)),\cdots(u_i,f_R(u_i,)),\cdots(u_n,f_R(u_n,))\}$。$f_R:U\to[0,1]$是模糊集 R 的隶属函数,$f_R(u_i,)$为 u_i 的隶属度。在两个模糊集 A 与 B 的上的运算有:

$\wedge:f_{A\wedge B}(u_i)=\mathrm{Min}(f_A(u_i),f_B(u_i)),\forall u_i\in U$;

$\cup:f_{A\cup B}(u_i)=\mathrm{Max}(f_A(u_i),f_B(u_i)),\forall u_i\in U$。

应用模糊算法进行 Web 页面聚类时,主要就是构造页面间的模糊相似矩阵。定义 Web 访问用户集合 $C=(C_1,C_2,\cdots,C_N)$,某一站点所有 URL 集合 URL=$\{u_1,u_2,\cdots,u_n\}$中 URLu_i 可用用户访问情况表示为:$I_j^u=\{(c_i,f_{I_j^u}(c_i))\mid c_i\in C\}$,其中 $f_{I_j^u}(c_i)=hits(c_i)/\sum_{j=1}^n hits(c_j),\in[0,1]$,$n$ 表示用户数量。

此时可建立页面间的模糊相似矩阵 $M_{n\times n}^f$,矩阵中的元素值为:

$$m_{i,j}^f=\sum_{k=1}^n f_{I_j^u}(c_k)\wedge f_{I_j^u}(c_k)/\sum_{k=1}^n f_{I_j^u}((c_k)\vee f_{I_j^u}(c_k)$$

因该矩阵为对称矩阵,所以在计算相似度时只取一半数据,以给定的阈值构造相似类。由于模糊矩阵 $M_{n\times n}^f$不满足传递性,故只能得到含有公共元素的相似类而非等价类。具体而言:对于每一个 $URL_i\in URL$,根据给定的阈值 σ 构造相似类$[URL_i]_S=\{URL_j/URL_j\in URL,m_{i,j}^f\geqslant\sigma\}$会具有相同的元素。如$[URL_i]_S=\{URL_i,URL_m\}$;$[URL_j]_S=\{URL_j,URL_m\}$ 即$[URL_i]_S\cap[URL_j]_S\neq\phi$。此时将具有公共元素的相似类归并得到对应的等价类即为 *Web* 页面聚类的结果(如合并相似类$[URL_i]_S$、$[URL_j]_S$ 有$[URL_i]_S=\{URL_i,URL_m,URL_j\}$)。

同理将用户 C_i 用浏览子图的 URL 序列表示为:

$G_i^c=\{(URL_j,f_{G_i^c}(URL_j),URL_j\in URL)\}$

建立客户相似矩阵:

$$m_{i,j}^f=\sum_{k=1}^n f_{G_i^c}(URL_k)\wedge f_{G_i^c}(URL_k)/\sum_{k=1}^n f_{G_i^c}(URL_k)\vee f_{G_i^c}(URL_k)$$

按页面聚类相同方法即可进行用户聚类。

5 基于 *n* 元预测模型的 Web 页面推荐

n 元预测模型的原型来自于自然语言的 n 元语言模型[19]。利用该方法无须用户介入向系统反馈明确的喜好信息,仅需分析 Web 日志文件。用户不需要增加任何的额外负担。

5.1 *n* 元预测模型的构造

n 元预测模型的算法是利用对 Web 申请频率的统计建立一个 n 元的预测模型。由于每个用户会话是由一系列的 Web 申请构成,对其长度可定义为 n 元。以 $n-1$ 元项为索引建立一张查找表,表中记录训练集合的会话过程中出现在 $n-1$ 项后的 m 个不同的 Web 申请出现的次数并求其条件概率。算法的精髓就在于构建这样的一张 hash 查找表 Y:

设 R 是截止当前页面用户浏览的页面数目,n 是预测所需最少访问路径长度。

若 $R<n$,则不能利用本模型预测。如果 $R\geqslant n$ 有:

先以 hash 表 Q 存储在 n 长度的访问路径后发生的不同的 Web 页面请求。hash 表 Y 存储模型预测结果;R 是截止目前用户提出的页面请求个数;D 是在当前页面请求后不同的 m 个请求集合,其元素表示为 D_i。

对每一个 D_i 如果满足 $P(D_i/R)>\varepsilon$ 且 $P(D_i/R)>Y_{n-m}(R)._R$(即在 R 个页面请求发生后出现 D_i 请求的概率大于预定的阈值且大于 hash 表 $Y_{n-m}(R)$预测结果的概率 $Y_{n-m}(R)._R$)

则将 $Y_{n-m}(R)$ 赋值为 D_i；$Y_{n-m}(D_i)._R$ 赋值为 $P(D_i/R)$

整个过程需将 D 集合遍历一次。

5.2 基于 n 元预测模型的访问页面请求预测

页面请求预测的算法简单描述如下：

```
Begin
  For I= max(Length(R),max Length of prediction model) down to
      n
      If Index of R∈ hash table(Y_{i-m}) then
        Return(Y_{i-m})
      Endif
  End For
  Return()
End
```

实际操作中，以日志文件中提取的一个会话序列进行分割将其中一部分作为训练样本，进行预测模型的训练，而利用另一部分作为预测集。以 R^+ 表示预测正确次数，R^- 表示预测错误次数。R 是用户 Web 请求总数。此时有：

预测精度：$precision=\frac{R^+}{R}$

预测能力 $Application=\frac{R^+ +R^-}{R}$

在实际应用中 n 元模型的大小同页面数成正比。实验表明[20]在所有页面申请中，大部分页面被访问次数较少。在页面总点击中占很小的比例。如访问次数小于 10 次的页面占总页面空间申请的 85%，其访问总数不到 10%。根据这一特性可以对 n 元模型进行压缩，从而提高模型的实用性。

6 Web 个性化推荐中的页面加权

在应用 Web 使用挖掘实现 Web 个性化推荐时，为了提高页面个性化推荐的准确率和覆盖率，可以对 Web 推荐信息进行加权处理。

6.1 基于页面访问时间的加权

Web 日志对用户浏览行为不但记录访问的页面内容，而且还包括用户对该页面的起始访问时间，通过计算相邻页面访问时间差即可判断用户对一个页面的浏览时间。考虑到网络延时等影响因素，该时间并不能精确描述用户的页面浏览时间。这里可以使用第 4 节提到的客户端数据获取技术及时准确地获取用户页面浏览时间，并将其作为基于页面访问时间的加权计算数据源。

通常用户在一个页面的浏览时间越长，说明用户对该网页的兴趣度越高，在推荐时应该将其优先推荐给用户。所以对于一个 Web 个性化推荐系统而言，应该将用户页面浏览时间作为一个权重纳入页面推荐得分的计算中来，以提高个性化推荐系统的准确率。

设在对 Web 个性化数据经过预处理后，获取含有 n 个用户会话的事务，其中第 k 条会话含有 m 个页面的会话为：

$\{(Pid_{1,k}, time_{1,k}), \cdots, (Pid_{i,k}, time_{i,k}), (Pid_{i+1,k}, time_{i+1,k})(Pid_{m,k}, time_{m,k})\}$

其中 $Pid_{i,k}$ 表示第 K 条会话中的第 i 个访问页面，$time_{i,k}$ 是用户访问第 k 条会话中的 $Pid_{i,k}$ 页面起始时间。则用户在第 K 条会话中的第 i 个访问页面上的浏览时间为：

$\Delta time_{i,k}=time_{i+1,k}-time_{i,k}$

特殊地，对于最后一个页面的浏览时间可以用该条会话的页面访问时间的平均值代替。

第 i 个访问页面在第 k 条会话中的访问时间权重可以用下式获得。

$W_k(Pid_{i,k})=\Delta time_{i,k}/\sum_{i=1}^{m}\Delta time_{i,k}$

根据上面的计算，访问页面 $Pid_{i,k}$ 在该事务中的访问时间权重为：

$W(Pid_i)=\sum_{k=1}^{s}W_k(Pid_i)/s$

其中 s 为给定事务中出现 Pid_i 的会话个数，该权重的值介于 0～1 之间。

6.2 基于页面间距离的加权

对于 Web 站点来说，页面之间都会存在一定的最短连通距离。即便两者不连通，它们之间的距离也可以通过它们与根的距离绝对值相加或其他计算方法来取得。当推荐页面与当前页面之间的距离较远时，用户访问该推荐页面的概率就越小，此推荐页面的价值也就越高。

设页面权值函数 $W(p,m)$ 用于计算各页面的加权值。P 代表用户当前访问页面，m 表示候选推荐页面，其值通过对两个页面间的距离进行归一化计算获取。

在得到页面访问时间的权重或页面距离权重后，各推荐页面的推荐得分可定义为：

$S=R+a*W$

其中 S 为候选推荐页面最终推荐得分，R 为候选推荐页面初始得分，a 是归一化调节因子，W 即为得到的页面访问时间的权重或页面距离权重。

推荐页面在使用基于页面距离和页面访问时间的加权技术后得到的推荐得分可以较客观地评价一个页面对用户的重要程度。

结束语 本文主要讨论了个性化 Web 推荐服务研究，给出了利用 Web 挖掘方法的个性化推荐服务研究中重要的用户聚类、Web 页面聚类、页面推荐以及 Web 个性化推荐中的页面加权算法。本文提出的算法使得 Web 信息服务提供者根据用户网络浏览行为可正确把握其兴趣所在并可动态对其兴趣改变进行跟踪，从而实现最大效率地为用户提供方便快捷且实用的个性化服务。本文介绍的个性化 Web 推荐系统通过对用户兴趣进行分析利用和学习，实现了用户浏览页面预测及预取其提供给用户，满足了广大用户迫切需要的适应自己特定兴趣的个性化服务，缩短了网络请求延迟时间，提高网络服务的准确性。从而使得网络资源利用个性化、智能化、高效化。随着电子商务、电子政务以及网络远程教育等电子服务的发展，研究 Web 环境下的个性化推荐服务具有重要且现实的意义。

参 考 文 献

1 韩家炜，孟小峰，等. Web 挖掘研究. 计算机研究与发展，2001，38(4)：405～413

2 Chen Ming-Syan, Han Jiawei, Yu P. Data Mining: an overview from a database perspective [J]. IEEE transaction on knowledge and date engineering, 1996,8(6)

3 Cooley R, et al. Data Preparation for Mining World Wide Web Browsing Patterns. Knowledge and Information Systems, 1999,1(1)

4 李勇. 基于内容的智能网络多媒体信息过滤检索. 情报理论与实践，2001，2：137～139

5 Imielinski T, Manila H. A database perspective on knowledge discovery. Communications of ACM, 1996,39(11): 58～64

6 Cooley R, Mobasher B, Srivastava J. Web Ming: Information and pattern discovery on the World Wide Web. In: Proc. of the 9th Int'l Conf. on Took with Artificial Intelligence (ICTAI'97), CA, 1997. 558～567

（下转第 141 页）

• S_2 表示已缓存到 ESP 上的内容

• S_1 表示正在从原服务器下载的内容

• S_3 表示还未下载的内容

$S_u = S_1 + S_2 + S_3$

6)流调度完毕,算法结束。

综上所述,此算法的关键在于:缓存的文件大小刚好能够满足连续码流连续播放的要求。当边缘服务器给用户流调度完毕时,缓存为空。在用户 U_1 请求未命中时,算法采用让原服务器为用户 U_1 提供普通流媒体在线播放的方式,也就是用户 U_1 在观看 p_u 时并没有享受到 SMCDN 的好处,但这样却同时满足了:a)让 U_1 在其延时容忍时间内就得到服务;b)触发了 ESP 启动原服务器"拉"下 S_{p1} 的线程;c)牺牲一用户的 QoS,但其他用户 $U_2, U_3, \cdots\cdots$ 再访问 p_u 时就可以享受到 SMCDN 带来的高质量服务。而在缓存整个文件法中,用户必须等待 ESP 将整个文件下载之后才能得到服务,这样的延时是用户不能容忍的。

4 实验结果分析

本文针对 SMCDN 提出了基于缓存部分流媒体内容的缓存算法思想,通过参与"下一代互联网内容分发网络"研究项目,在基于 IPv6 的网络环境下,将这种策略在实际系统环境中进行了应用研究,对其性能进行了测试并与其它的主流 CDN 系统进行了比较(如表 1)。

表 1 SMCDN 系统与其它系统性能比较

CDN系统 / 性能参数	采用部分缓存的 SMCDN	其它 CDN 系统的最优性能
从用户提出请求到得到反应的时间(单位:s)	<3s	<4s
客户端播放器的缓冲时间(单位:s)	<10s	<10s
播放中延迟(单位:s)	<2s	<5s
支持的网络协议	Ipv4/Ipv6	Ipv4
同时并发提供不同节目的流(单位:个/ESP)	200~300	150
内容路由并发性能(Media DNS 在每秒钟处理的客户的个数)	230~250	150~200
ESP 的命中率	90%	30%

从表 1 可以得知,采用部分缓存的 SMCDN 系统,在同时提供不同节目的流和 ESP 的命中率这两项性能指标方面提高得比较明显,从而在相同的硬件条件下可以存取更多的流媒体节目。但是,应用部分缓存策略的 SMCDN 系统,实现较为复杂,同时受实际网络情况的影响比较大,与在良好的实验环境下相比,在稳定性上有一定的下降。这方面,我们正在做进一步的研究。

结论 随着用户对网络流媒体需求的不断增加,人们迫切希望在互联网上看到电视或录像机一样效果的内容,所以流媒体内容分发网络是达到此目的的选择方法之一。本文提出的部分缓存策略应用到流媒体内容分发网络,体现了流式技术与 CDN 相结合的优势,有效提高了系统的命中率和服务质量。由于 CDN 涉及内容广泛,本文仅仅从 SMCDN 中的 ESP 的缓存策略进行了初步研究,因此需要不断完善和提高。

参 考 文 献

1 胡海清,傅鹤岗,朱庆生. 基于软件 Agent 技术的内容分发网络研究[J]. 计算机应用,2004,24(6):51~53

2 Lu Jian. Reactive and Proactive Approaches to Media Streaming [C]. In: Information Technology: Coding and Computing, 2001. Proc.. Intl. Conf. on 2001. 5~9

3 Yoshimur T, Yonemoto Y, Ohya T, et al. Mobile Streaming Media CDN Enabled by Dynamic SMIL [J]. ACM, 2002. 651~661

4 Fujita N, Enomoto N, Iwata A, et al. Coarse-Grain Dynamic Replication Schemes for Scalable Content Delivery Networks [J]. IEEE, 2002. 2235~2239

5 王薇薇,李子木. 基于 CDN 的流媒体分发技术研究综述[J]. 计算机工程与应用,2004(8):121~125

6 Cronin E, Jamin S, Cheng Jin, et al. Constrained Mirror Placement on the Intemet [J]. IEEE, 2002, 20(7): 1369~1382

7 Qiu Lili, Padmanabhan V N, Voelker G M. On the placement of Web Server Replicas [J]. IEEE. 2001, 3(22~26): 1587~1596

8 Kangasharju J. Object replication strategies in content distribution networks [J]. Computer Communications, 2002. 25: 376~383

9 Sen S, Rexford J, Towsley D. Proxy Prefix Caching for Multimedia Streams [J]. IEEE, 1999, 3(21~25): 1310~1319

(上接第 138 页)

7 Srivastava J, et al. Web usage mining: Discovery and application of usage patterns from Web data. SIGKDD Explorations, 2000, 1 (2)

8 Zaiane OR, Xin M, Han J. Discovering Web access patterns and trends by applying OLAP and date mining technology on Web logs . In: Proc. of Advances in Digital Libraries Conf. (Adl'98), Santa Barbara, CA, 1998. 19~29

9 Lieberman H. Letizia: An Agent that assists Web browsing. In : Proc. of the 14th Intl. Joint Conf. on Artificial Intelligence, Montreal, IJCAI'95. Aug. 1995. 924~929

10 Balabanovic M, Shoham Y. An adaptive Agent for automated Web browing. Joural of Visual Communication and Image Representation, 1995, 6(4)

11 Mladenic D. Personal Web Watch: Design and Implementation: [Technical Report IJS-DP-7472]. Department for Intelligent System, J. Stefan Institute

12 Joachims T, Freitag D, Mitchell T. WebWatcher: a tour guide for the World Wide Web. In: Georgeff M P, Pollack E M, eds. Proc. of the Intl. Joint Conf. on Artificial Intelligence. San Francisco: Morgan Kaufmann Publishers, 1997. 770~777

13 http://profusion. ittc. ukans. edu

14 王耀南. 智能控制系统——模糊逻辑. 专家系统. 神经网络控制,湖南大学出版社,1996

15 Pandya A S, Macy R B. 神经网络模式识别及其实现. 北京:电子工业出版社,1999

16 李勇,桑艳艳. 网络文本数据分类技术与实现算法. 情报学报,2002,21(1):21~26

17 Herlocker J L, et al. An Algorithmic Framework for Performing Collaborative Filtering. SIGIR, 1999. 230~237

18 Zadeh L A. Fuzzy sets. Information and Control. 1965, 8: 338~353

19 Lee K F, Mahajan S. Automatic Speech Recognition: The Development of The SPHINX System. Dordrecht, Netheriands: Kluwer, 1989

20 苏中,马少平,等. 基于 Web-Log Mining 的 N 元预测模型. 软件学报,2002,13(1):136~141