Web 日志挖掘在电子商务中的应用研究

Research on the Application of Web Log Mining in E_Commerce

张新香 （中南财经政法大学信息学院 武汉 430064）

摘要：本文介绍了 Web 日志挖掘的概念和流程，提出了客户频繁访问路径和页面兴趣度挖掘算法，并给出了个性化推荐系统的构建思路，旨在为电子商务网站经营者改善网站结构提供帮助。

关键词：Web 日志挖掘 频繁访问路径 页面兴趣度 个性化推荐

1 引言

随着 internet 的日益普及，人们的购物方式发生了巨大的变化，已经由传统到商店购买直接转到 internet 上购买，这样也改变了销售商和客户之间的关系，客户所追求的不再是购买场所是否方便，而关心的是商品的价值，当然客户选择商品还有他自己的偏好，这样电子销售商必须了解客户的网上行为、价值取向、兴趣爱好，从而提升自己产品的市场竞争力。Web 数据挖掘能从 Web 服务器上大量的数据中提取出人们事先不知道但有用的信息和规律。利用 Web 数据挖掘可以发现顾客的购买偏好，发现什么样的客户是忠实的客户，为他们提供个性化服务，延长客户的驻留时间；发现潜在客户，为他们提供个性化页面，变潜在客户为忠实客户，扩大市场占有率；分析客户未来可能发生的行为，进行有针对性的营销活动，提高广告的投资汇报率；当然利用 Web 挖掘还可以实现信用评估，欺诈检测，投资组合管理，商店选址等多方面的应用。

2 Web 日志挖掘概念

一般的 Web 数据挖掘可以分为 Web 内容挖掘、Web 结构挖掘、Web 日志挖掘三类。Web 日志挖掘的主要目标是从 Web 的访问日志记录中抽取感兴趣的模式。WWW 中每个服务器保留了访问日志，记录关于用户访问和交互的信息。分析这些数据可以帮助理解用户的行为从而改进站点的结构，或为用户提供个性化的服务。

3 Web 日志挖掘流程

Web 日志挖掘流程如图(1)所示。

3.1 Web 日志记录内容

Web 服务器日志记录用户访问该站点时每个页面的请求信息，其主要结构如表 1 所示。

3.2 数据预处理

在数据预处理阶段中，根据数据挖掘的目的，对原始 Web 日志文件中的数据进行提取，分解合并，最后转化为适合进行数据挖掘的数据格式，并保存到关系数据库表或数据仓库中，等待进一步处理。数据预处理是 Web 日志挖掘整个过程的基础和有效实施的前提。数据预处理包括四个阶段：数据净化、识别用户、识别用户会话和识

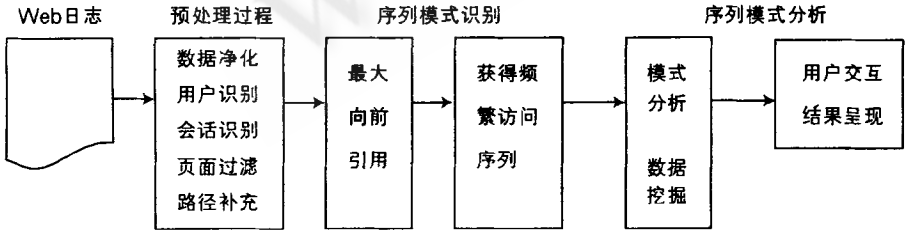

图 1 Web 日志挖掘流程

别片段(页面过滤和路径补充)。

(1) 数据净化。数据净化根据需求对日志文件进行处理,包括删除无关紧要的数据,合并某些记录,对用户请求页面时发生错误的记录进行适当的处理等等。

表 1 Web 日志记录的主要信息

域	描述
日期	用户请求页面的日期
时间	用户请求页面具体时间
客户 IP 地址	客户端主机的 IP 地址或 DNS 入口
用户名	客户端的用户名
服务器名	服务器名称
服务器 IP 地址	服务器的 IP 地址
服务器端口	服务器的端口号
方法	用户请求的方法
URL 资源	用户所请求的页面
URL 查询	用户预进行的查询
协议状态	返回 HTTP 的状态标识
发送字节数	服务器发送的字节数
接收字节数	服务器接收的字节数
所花时间	完成浏览所花的时间
协议版本	传输所用的协议版本
主机	服务器的操作系统
用户代理	服务提供者
Cookie	Cookie 标识号
参照	用户浏览的上一页

(2) 识别用户。在实际应用中,识别用户是非常重要的,由于缓存和代理服务器(包括网吧、局域网等环境)和防火墙的使用,使得识别用户变得很复杂。目前通过用户注册、Cookie 和内嵌用户 ID、客户端软件 Agent,分析用户 IP、站点拓扑的方式来确定用户身份。

(3) 识别用户会话。用户会话是指用户对服务器的一次有效访问,通过其连续请求的页面。我们可以获得他在网站中的访问行为和浏览兴趣。日志文件中不同的页面当然属于不同的会话。当某个用户的页面请求在时间上跨度较大时,就有可能是该用户多次访问同一个网站。我们可以将用户会话记录分成多个会话来处理,最简单的方法就是设置一个 tiemout 值,如果用户访问页面的时间差超过该值,则认为开始一个新的会话。

(4) 识别片段。识别片段就是找出用户会话中有意义的访问路径。由于存在客户端缓存,用户在浏览网页时,通过按下浏览器上的"后退"按钮得到的页面是从本地缓冲区中得到的,在日志文件是没有记录的,从而导致该页与用户上一次请求的页面之间没有超连接信息,在这种情况下,可以根据网站的拓扑结构,把用户的访问路径填充完整。

3.3 模式识别

模式识别是运用各种算法和技术对预处理之后的数据进行挖掘,生成模式。这些技术包括人工智能、数据挖掘、统计理论、信息论等领域的成熟技术,。数据挖掘中常用技术有路径分析,关联规则、序列模式以及分类聚类等。

3.4 模式分析

该阶段实现用户访问模式的分析,基本作用是排除模式识别中没有价值的规则或模式,从而将有价值的模式提取出来。

4 电子商务中的日志挖掘

4.1 具体实施步骤

(1) 数据准备。对 Web 日志内容进行预处理,删除无用数据,识别用户和用户会话,完善访问路径。

(2) 模式识别。采用相应的挖掘算法,算出用户频繁访问路径和页面兴趣度。

(3) 推荐。根据以上结果,构建个性化推荐系统模型。

本文主要介绍(2),(3)两块核心步骤。

在电子商务站点中,对 Web 服务器访问日志进行挖掘,能够揭示用户的频繁访问路径和兴趣页面,发现用户的浏览模式,电子商务经营者对用户的访问模式进行分析,了解用户的浏览习惯,调整电子商务站点的结构,方便用户的页面浏览和在线购买,还可以在用户频繁浏览页面上放置广告,有效提高站点的服务效率,增加站点的经济效益。

4.2 用户访问路径分析

用户在访问电子商务站点时,会在 Web 服务器日志中留下访问的"足迹",经过一段时间,最频繁访问的页面会形成路径,利用算法挖掘出用户频繁访问路径,电子商务站点的经营者能在频繁访问页面上提供超连接,页面的预取,将广告放置在用户的频繁访问页面上,商品的在线推荐等。

用户浏览网站的操作分为两类:一类是根据页面

上的超连接向前浏览新的页面，另一类是通过点击浏览器上的"BACK"按钮回退到浏览过的页面，绝大多数用户浏览已经访问过的页面不是因为页面的内容，而是由于页面之间的超连接结构，用户的回退浏览对于挖掘用户的所有访问过的页面作用不大，用户访问路径的分析关心的是用户后退操作前的所有访问过的页面，即最大向前路径。

获取最大向前路径的算法如下：

假设经过数据预处理后，已经得到用户会话，并将用户访问页面按时间排序，形成如下的会话 $X=\{X_1, X_2\cdots\cdots X_N\}$，Y 表示形成的最大向前引用事务，D 表示存储最大向前事务的数据库。

```
INPUT：用户会话集合{X}
Begin
  For each user ∈ userset do
  Y = null
  y1 = x1; j = 2; i = 2
  f = 1 /* f = 1 表明当前遍历方向是向前，f = 0 是向后 */
  While i < = n do
    Found = 0 /* 设置发现 X 和 Y 中相同页面标志 */
    k = 1
    While (k > = 1 and k < j) and found = 0 do
      If xi = yk then
        Found = 1
      Endif
      k = k + 1
    Endwhile
      If Found = 1 then
        If f = 1 then
          输出{y1, y2……yj - 1}到 DF
          j = k; i = i + 1
          f = 0
        Endif
      Else
        yj = xi; j = j + 1; i = i + 1
        f = 1
      Endif
  Endwhile
    if f = 1 then
    输出{y1, y2……yj - 1}到 DF
    Endif
  Endfor
End
Output：最大向前事务的数据库 DF
```

找出最大向前路径后，可以借用改进的关联规则算法来挖掘出用户频繁访问路径。思路如下：首先简单统计所有含一个页面项目集的出现频率，并找出那些不小于最小支持度的页面项目集，即 1 维最大项目集，然后开始循环处理直到没有最大页面项目集生成，在第 K 次循环中，根据第 K - 1 步生成的(K - 1)维最大页面项目集产生 K 维候选页面项目集，对事务数据库进行搜索，得到候选项目集的支持度，找出大于最小支持度的 K 维最大项目集。

求出满足最小支持度的所有频繁路径的算法如下：

符号说明：

$K-itemset$：K 维页面项目集

L_k：具有最小支持度的最大 K - itemset

C_k：候选的 K - itemset，潜在的最大页面项目集

```
INPUT 事务数据库 DF
BEGIN
  C1 = {candidate 1 - itemset}
  L1 = {c ∈ C1 | c. support > = minsupport}
K = 2
While L(k-1) < > null do
  Ck = cand_gen(L(k-1))
For all candidate c ∈ Ck do
  c. support = c. support + 1
Endfor
  Lk = {c ∈ Ck | c. support > = minsupport}
k = k + 1
Endwhile
End
OUTPUT：长度为 K 的高频页面集 L
Function cand_gen(lk - 1)
  Begin
    Insert into Ck
    Select p. item1, p. item2, ……p. itemk - 1
```

```
From lk-1 p, lk-1 q
Where p.item2 = q.item1 and p.item3 = q.item2
……p.itemk-1 = q.itemk-2 /* 体现了页面连续性 */
For all itemset c ∈ Ck do
 For all (k-1)维 c 的子集 s
  If s 不属于 lk-1 then
     Delete c from Ck
 endif
endfor
  endfor
return ck
```

4.3 用户浏览页面兴趣度的测量

用户在变化访问路径时，也存在页面滞留时间的改变，用户在不喜欢的页面访问时间较短，在喜欢的页面停留的时间较长，定义 user 对网页 page 的兴趣度如下：

$$f(user,page) = \frac{user\ 浏览\ page\ 所用的时间}{user\ 的总浏览时间} \times \frac{page\ 字节数}{路径中总字节数}$$

算出用户的页面兴趣度，并将结果从大到小排列，可以为站点优化和个性化推荐提供依据，算法如下：

```
INPUT：用户浏览信息
OUTPUT：用户浏览页面的兴趣度
Function pageinterest
Begin
  for logx=1 to n
  k=locate(browtime,page_logx)
  browtime(k) = browtime(k) + browtime
 endfor
  for j=1 to pagenum
  totaltime=totaltime+browtime(j)
  totallength=totalength+length(j)
 endfor
 for j=1 to pagenum
 interest(j) = (browtime(j)/totaltme) * (length(j)/totallength)
 endfor
End
```

4.4 个性化推荐系统模型

有了用户频繁访问路径和页面兴趣度，可以给出个性化推荐系统模型，个性化推荐系统有两部分组成：脱机部分和在线部分。脱机部分由数据准备和特定的挖掘任务组成，数据准备将用户的 Web 服务器日志信息进行处理，生成事务数据库和用户数据库，挖掘任务完成用户频繁路径和页面兴趣度生成。在线部分主要由推荐引擎、Web 服务器和用户浏览器构成，主要完成在线的页面推荐工作。推荐引擎首先分析当前用户和用户的当前会话，分析出其频繁访问路径和兴趣页面集合，将该用户的请求发给 Web 服务器，当返回页面后，推荐引擎将相应的推荐集附加在返回页底部，方便用户点击浏览。

5 结束语

利用 Web 日志挖掘可以生成用户频繁路径和页面兴趣度，能为电子商务网站经营者改善网站结构提供理性依据，有较强的现实意义。当然日志挖掘中还可以从用户聚类、主题兴趣度、关键词兴趣度计算等方面来优化个性化推荐系统，从而更好的为用户服务，提高商务站点的经济效益。

参考文献

1 王曰芬等，电子商务网站设计与管理[M]，北京大学出版社，2002。

2 张健沛、刘建东、杨静，基于 Web 的日志挖掘数据预处理方法的研究[J]，计算机工程与应用，2003。