网络用户浏览行为的分析①

张　亮[1], 赵　娜[2]

[1](中国石油大学(华东) 网络及教育技术中心, 青岛 266500)
[2](山东省青岛市黄岛区建筑工程质量监督站, 青岛 266500)

摘　要: 近年来, Web 使用挖掘成为数据挖掘领域中一个新的研究热点, Web 使用挖掘是从记录了大量网络用户行为信息的 Web 日志中发现用户访问行为特征和潜在规律. 本文结合某高校主页的真实运行数据, 通过 Web 使用挖掘对于网站的运行日志文件进行全面的挖掘分析, 分析用户对信息内容的兴趣度, 并通过用户对网页的访问数据推算出各个页面受众的兴趣度高低, 借此改良网站的内容和布局.

关键词: 用户行为; 数据挖掘; 离线分析; 在线分析; 权重

Internet Users' Browsing Behaviors Analysis

ZHANG Liang[1], ZHAO Na[2]

[1](Network Information Center, China University of Petroleum (East China), Qingdao 266500, China)
[2](Construction Quality Supervision Station, Huangdao District, Qingdao 266500, China)

Abstract: In recent years, web usage mining has become a new hotspot in the field of data mining. From the web logs which record information of a large number of network user's behavior, web usage mining discovers the characteristics and potential user access law. This paper uses many real running dates of college homepage. Aiming at running log files, we carry out a comprehensive analysis by using the web mining. Analyzing the interest measure of user to the information content. By using the user access to the page, the system can calculate the user data level of interest on each page, and thereby improving the content and layout of the site.

Key words: user behavior; data mining; off-line analysis; on-line analysis; weight

信息技术的发展, 使得互联网中积累了海量的无序的、繁杂的数据, 虽然有这些海量的数据, 但是只有极少部分是有用的. 基于 Web 使用挖掘的网络用户行为分析是指收集用户访问网站时的基本数据, 对这些用户行为数据进行统计、分析和研究, 从中发现不同用户的行为规律, 发现用户的行为模式, 了解用户的兴趣, 从而为用户提供更有效的服务.

1　用户行为数据源的获取

针对一个网站, 数据挖掘的关键步骤之一就是要采集用户兴趣的数据集. 按照服务器记录信息的不同, 数据挖掘对象来源[1,2]于客户端数据、代理端数据、服务器端数据三类数据. 在 Web 使用过程中, 从不同数据源收集而来的数据反映了用户行为的不同.

(1)客户端数据

客户端数据可以比较全面和准确的收集(利用远程 Agent)到用户数据. 所谓"客户端远程 Agent"就是运用 Applet 技术在客户端获取用户浏览行为.

(2)代理端数据

代理端可以揭示来自访问多个服务器多用户的实际 http 请求, 代理端的缓存可以减低客户端访问对网络的装载时间, 降低对 web 服务器的访问, 减少服务器端的工作负载.

(3)服务器端数据

服务器端的数据, 记录了网站用户的访问该站点时每个页面的请求信息. Web 服务器上存放的日志文

① 收稿时间:2015-10-16;收到修改稿时间:2015-11-11　[doi: 10.15888/j.cnki.csa.005180]

件时采用 ECLF(扩展型日志格式). 其格式如表 1 所示.

表 1 Web 日志属性描述

功能描述	字段名	中文含义
请求页面的日期	date	日期
请求页面的时间	time	时间
Cookie 标识	Cookie	con(cookie)
远程主机的 IP	ct-ip	IP 地址(客户端)
用户的标识	con-username	用户名
客户端连接的端口号	Ser-port	服务器端口
客户端所访问该站点的 Internet 服务	ser-name	服务名
生成日志项的服务器的 IP 地址	ser-ip	IP 地址(服务器)
生成日志项的服务器名称	ser-name	服务器名
客户端试图执行的操作	con-method	方法
访问的资源	con-uri-stem	URI 资源
客户端尝试执行的结果	con-uri-query	URI 查询
服务器响应情况	sc-status	状态
用 Windows® 使用术语表示的操作的状态	sc-win32-status	Win32 状态
服务器发送的字节	sc-bytes	发送的字节
服务器接收的字节	con-bytes	接收的字节
服务器 IP 或域名	Server	服务器
URL 请求资源和 URL 请求方法	Request	请求
估计完成浏览需要的时间	time-taken	预计时间
传输用的协议版本	con-version	传输协议版本
显示主机的内容	con-host	主机
上级页面	Superior	反向链接

本文对于用户的行为进行分析研究, 仅仅需要考虑 Web 服务器上的日志文件(Log file)即可. 在关系数据库中建立一个表 SourceLog, 用于数据源的获取, 其存放形式如表 1 所示, 表中的相应字段对应于合并后 Log 中的一个属性项, 并且可以对 Database 中的原始数据进行 SQL 操作.

2 用户行为分析

不同的用户对网页的兴趣度也不同, 如何满足不同用户的各种需求是网站管理员最挂心的事. 上面已经采集 Web 服务器日志, 并将其归集起来, 接着对用户行为进行分析, 利用上面归集起来的数据. 用户行为分析[3,4]分为离线分析和在线分析两部分.

2.1 离线分析

离线分析就是对 Web 日志进行预处理、分析、挖掘, 为在线分析准备频繁序列模式.

2.1.1 日志文件的预处理

预处理就是将采集到的用户原始的行为数据进行分析, 消除错误的、冗余的、不完整的数据信息, 获得一组可以挖掘、适宜分析的对象. 数据预处理阶段包括以下几个过程.

(1) 数据清洗

数据清洗是指从多个服务其中读取并合并有关日志数据, 然后删除 Web 日志文件中与数据挖掘无关的数据, 这些无关的数据主要包括: 一些非 HTML 文件(如图片和音频文件)、样式文件和脚本文件、用户访问失败的记录、不是 GET 的数据记录, 弹出式广告的记录等. 比如分析者可能只希望分析某一时间段(2015 年 5 月 1 日-2015 年 5 月 31 日)用户行为规律, 可以通下列语句来实现.

```
Select *
From SourceLog
Where time Between 01/May/2015:12:00:00 And 31/May/2015:12:00:00
```

代理发出的请求还大量的存在日志中, 将会不影响挖掘结果, 必须对此进行处理. 因此, 从日志中识别代理或网络爬虫的访问时必需的. 数据清洗的流程图如图 1 所示.

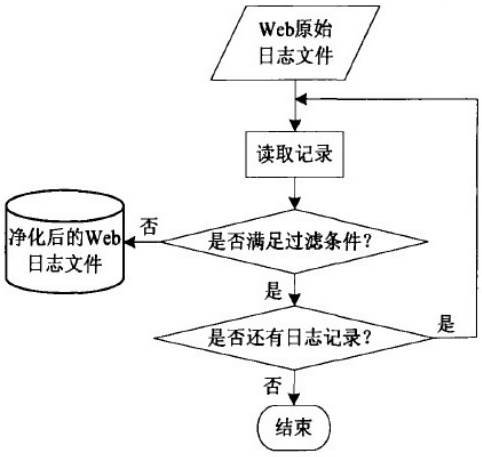

图 1 数据清洗的流程图

数据清洗的最后一步就是规范统一资源定位器地址(即 URL 地址), 分析公共网关接口数据(即 CGI 数据).

(2) 用户和会话识别

用户会话[5]是一个用户一次访问一个 Web 网站时所浏览的所有网页的集合, 通过连续请求的页面, 可以获得其在网站中的浏览兴趣行为和访问行为. 为了

网络用户行为进行研究分析, 必须将不同的用户区分出来, 一般将 IP 地址、代理类型 Agent 结合起来去辨识一个用户, 辨识出用户后其访问记录就必须划分为会话. 会话识别中, 常常设置一个超时时限, 若请求的来源的网页文件超过设定的时限限制, 则认为它来自一个新的会话. 用户识别和会话识别的流程图如图 2 和图 3 所示.

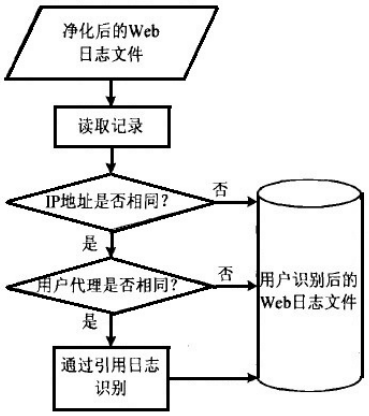

图 2　用户识别的流程图

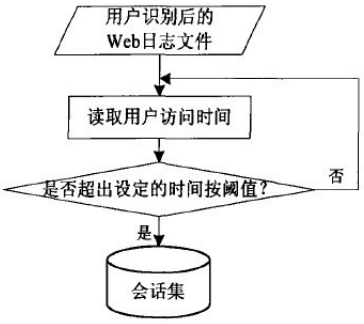

图 3　会话识别的流程图

一般采用基于页面访问时间的启发式方法, 其算法如算法 1 所示. 在同一用户访问的页面序列中, 如果两个相邻页面的时间差 $Time(T_{i+1})$- $Time(T_i)$超过给定的时间限值, 则认为从页面 $i+1$ 开始用户的另一次会话, 一般使用 30min 作为缺省阀值. 另外, 为了消除偶然用户的访问出现对模式的识别, 将会话长度小于 2 的会话在数据库中删除.

(3) 路径补充

由于网络机制等原因, 识别出的会话序列可能并不完善, 日志记录中可能还遗漏了一些用户访问的过程, 所以还需要对用户会话进行路径补全. 路径补全是根据网络的拓扑机制, 如果用户使用了代理服务器, 网页浏览跟踪便无法从客户端获取, 便需要通过路径补全, 推断出一些缓存网页的浏览情况, 对遗漏的请求补全到会话当中.

算法:*Web*日志中生成用户事物数据库D
输入:清洗过的*Web*日志
输出:用户事物数据库D
order all records by IP/Agent and Time increasingly
For each IP/Agent
　Create a new user session in D
　For i=1 to the number of records of this IP/Agent
　 If(Time(T_{i+1})-Time(T_i))<t Then
　　Insert this record into user session
　 Else
　　Create a new user session in D
　 End If
　End For
End For

路径补全的流程图如 4 所示.

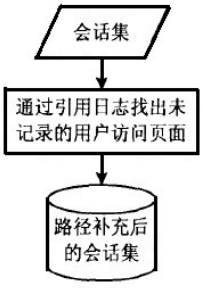

图 4　路径补全的流程图

用 *UIP* 代表用户 *IP* 地址, *UID* 代表用户 *ID* 号, URL_i 代表该第 *i* 号页面,TIM_i 代表第 *i* 号页面访问时间, HIT_i 代表第 i 号页面访问次数, 通过预处理的过程便可以形成一个用户访问矢量:
$<UIP,UID<URL_1,TIM_1,HIT_1>,<URL_2,TIM_2,HIT_2>..<URL_n,TIM_n,HIT_n>>$

给推断出的浏览网页赋予一个时间是路径补全的另一个任务.

2.1.2 模式发现和分析

由于网站包含的页面数量巨大, 用户真正感兴趣的网页很难完全列举出来, 因此, 将页面的访问次数和页面的平均访问时间结合起来, 来反映页面兴趣度, 即页面权重. 根据下面提供的页面权重计算公式,

$$W_p = \lambda W_f + (1-\lambda)W_a \quad (1)$$

其中, W_p 代表页面权重, W_f 代表按照访问频率计算页

面权重, W_a 代表按照平均访问时间计算页面权重, λ(0<λ<1)是 W_f和 W_a所占比重. 当 W_f>W_a时, $\lambda\in(0.5,1)$, W_f<W_a时, $\lambda\in(0,0.5)$.

根据用户的访问页面顺序的组成特点, 访问序列的权重可以进行定义如下: 给定一个用户访问序列 $S=\{P_1,P_2,\cdots,P_m\}$, 其中 $P_i(1\le i\le m)\in$ 项集 I, 则该用户访问序列的权重计算公式如式 2 所示.

$$W(S)=(W(P_1)+W(P_2+\cdots+W(P_m))/m=\sum_{k=1}^{m}W(P_k)\Big/m \quad (2)$$

通过文献[6]-[10]我们可以知道验证一个访问序列是否是频繁的主要依据是 $\sup port(S)\ge \min\_\sup$, 当用户访问序列 S 满足 $\sup port(S)<\min\_\sup$, 访问序列 S 为非频繁访问序列. 实际上, 每个 Web 页面都由各自的特点, 不是完全相同的, 在计算最小支持度($\min\_\sup$)时, 我们应该考虑考虑页面之间的差异. 因此下面, 我们对以上算法进行改进增加权重, 使其更加合理. 具体步骤如下所示.

(1)扫描数据库经过数据清理的数据, 计算出普通用户访问各页面的权重 W_p 以及各页面的访问序列权重 *W(S)*;

(2)普通用户访问各页面的权重 W_p 以及各页面的访问序列权重 *W(S)*作为标准权重.

通过连接操作, 生成候选访问序列, 根据用具判断是否为频繁序列的条件, 若标准权重低于所要判断的项集, 则作为频繁项集保留.

下面以某高校主页为例, 针对 2015 年 5 月 1 日至 2015 年 5 月 31 日这段时间, 各页面的访问浏览次数和平均访问时间进行有效的统计计算, 选择所计算出来的有效权重中权重最高的 15 个页面作为研究对象.

表 2 15 个权重最高的页面

页面	URL	浏览页面次数	访问权重(%)	平均访问时间	时间权重(%)	页面权重(%)
Page1	/index	8945	0.026	151	0.326	0.176
Page2	/news.upc.edu.cn/	6777	0.020	109	0.326	0.173
Page3	/library.upc.edu.cn/	6534	0.019	198	0.326	0.172
Page4	/bbs.upc.edu.cn/forum.php	37419	0.110	759	0.062	0.086
Page5	/youth.upc.edu.cn/	8734	0.026	311	0.109	0.068
Page6	/cupnews/js/xwxc.html	43210	0.127	97	0.007	0.067
Page7	/news.upc.edu.cn/tzgg/	17317	0.051	455	0.081	0.066
Page8	/ special/bcjt/	34789	0.102	305	0.027	0.065
Page9	/newsupc/news_sdxz/	23551	0.069	321	0.042	0.056
Page10	/newsupc/wymb/	9874	0.029	261	0.081	0.055
Page11	/special/bcjt/xltj/20110608/050716.html	12976	0.038	289	0.068	0.053
Page12	/special/bcjt/tsyfx/rlts/20090528/110318.html	17560	0.052	271	0.047	0.050
Page13	/special/bcjt/bjjt/ls/20110315/042741.html	27991	0.082	121	0.013	0.048
Page14	/newsupc/news_yw/20120320/092810.html	9986	0.029	207	0.064	0.047
Page15	/tech/index.html	15279	0.045	211	0.042	0.044

由表 2 可以看出, 在改进网站结构时, 由于 page1、page2 和 page3 页面的访问量较大, 可以将其提升到首页.

2.2 在线分析

本在线分析是为了访问者作即时推荐, 依据用户访问的页面, 推测用户即将访问的页面, 并将其加入到推荐页面中, 供访问者选择. 主要包括以下三个步骤:

(1)明确系统参数, 确定合适的滑动窗口时一个反复调节的过程;

(2)利用最长匹配算法与频繁访问模式集合中的项集进行匹配, 获得访问序列中对应的滑动窗口数据, 找到匹配序列模式;

(3)页面推荐, 如果按上一步找到匹配序列模式, 则可以加入推荐页面的集合, 如果没有找到与当前用户匹配的序列模式, 则不加入推荐页面的集合.

在线处理模块如算法 2 所示

一般而言, 取一个经验值作为滑动窗口的长度 w 的取值. 根据研究统计分析[11]: 滑动窗口的宽度一般设置为 3 或 4, 这是因为大部分用户在浏览网页的时候, 习惯点击“后退”或者“前进”按钮. 通过的在线分析, 找到相应的推荐集, 以链接的形式给出推荐集中的页

面, 并显示在当前的网页中, 以起到动态提供相关链接的目的.

算法2:在线处理模块执行过程的算法

输入：用户访问序列P和滑动窗口的长度w

输出：推荐页面的集合R_p

获得用户访问序列P中最后的w个页面构成的页面序列S;

获得频繁访问模式集合中长度w+1的项构成的集合F_{w+1};

For F_{w+1}中任意的项F

{ if（F去掉最后一个页面后与当前的页面序列S相同）

把F中的最后一个页面加入推荐页面的集合R_p;

}

最后输出推荐页面的集合R_p;

3 结语

本文根据某高校主页的统计信息, 通过对访问数据的量化分析了解用户行为, 通过 Web 使用挖掘发现网络用户对高校主页信息的兴趣度, 再依据数据分类等方式, 调整基础数据优化网站布局, 从而提高高校主页的用户满意度和点击率. 下一步研究是如何提高算法效率, 以实现根据浏览器信息个性化推荐网页.

参考文献

1 杨玉梅.Web 日志挖掘中的数据预处理技术研究.科技视界,2014,(12):20,24−25.

2 于升峰,蓝洁.基于用户行为挖掘和 RSS 技术的知识服务模式研究.情报探索,2011,(8):93−95.

3 Liu B. Web 数据挖掘.北京:清华大学出版社,2013:384−422.

4 McCarty JA, Hastak M. Segmentation approaches in data-mining: A comparision of RFM, CHAID, and logistic regression. Journal of Business Research, 2006, 60(6): 656−662.

5 于飞,丁华福,姜伦.Web 日志挖掘中数据预处理技术的研究.计算机技术与发展,2010,5(20):47−50.

6 Xing DS, Shen JY. Efficient data mining for Web navigation patterns. Information and Software Technology, 2004, (46): 55−63.

7 王小姣.聚类分析及其在 Web 日志挖掘中的应用研究[学位论文].济南:山东师范大学,2011:29−37.

8 唐伟,周倩.网络用户信息浏览路径挖掘研究的发展.情报理论与实践,2013,36(6):125−128.

9 刘贵平.基于浏览行为的用户价值细分研究.内蒙古大学学报(自然科学版),2014,45(6):623−627.

10 Sungjune P, et al. Sequence-based clustering for Web usage mining:a new experimental framework and ANN- enhanced K-means algorithm. Data & Knowledge Engineering, 2008, 65(3): 512−543.

11 Buchner A, Mulvenna M. Discovering internet marking intelligence though online analytical web usage. SIGMOD Record, 1999, 27(4): 23−38.