基于兴趣的Web挖掘中用户身份的识别新方法

张春生,庄丽艳
(内蒙古民族大学 数学与计算机科学学院,内蒙古 通辽 028043)

摘 要:分析了基于用户兴趣的Web挖掘中用户身份识别的重要性,以及对下一步路径填充的影响。针对目前众多的Web挖掘用户身份识别方法的缺陷,提出了一种基于兴趣的Web挖掘用户身份识别新方法。该方法通过取得用户的CPU号或网卡号,对用户的打扰为零,实现用户的个人身份识别。该方法简单易行,对Web挖掘中用户身份识别所存在的5个难题解决4个,提高了用户身份识别的准确率,为Web挖掘的准确性提供了保障。

关键词:兴趣;Web挖掘;身份;识别
中图分类号:TP311 **文献标识码**:A **文章编号**:1673-629X(2009)05-0062-03

New Method on Identification of User in Web Mining Based on Interesting

ZHANG Chun-sheng, ZHUANG Li-yan
(School of Mathematics and Computer Sci. of Inner Mongolia Univ. for Nationalities, Tongliao 028043, China)

Abstract: Analyzes the importance of identification of user in web mining based on interesting and the effect of the subsequent path filling. A new method on identification of user in web mining based on interesting is proposed, for the limitations of several methods on identification of user in web mining in nowadays. The interference of the method to user equals to zero by obtaining CPU number and network card of user, so that realizes personal identification of user. The method is easy to use and it settles four of five problems of user identification existed in web mining, and raises the accuracy rates of identification of user, and provides guarantee for the accuracy of web mining.

Key words: interesting; Web mining; identity; identification

0 引 言

随着Internet的迅猛发展,其信息量逐渐增大,1999年的时候估计有3.5亿个网页,并且以每天100万的速度增长,Google最近宣布已经索引了30亿网页。万维网是目前最大的数据库,如何有效地存取这些数据是一项具有挑战性的任务[1]。

解决这些问题的一个有效途径就是Web挖掘。Web挖掘就是把数据挖掘的技术应用到Web数据中以发现感兴趣的有用模式和隐含信息[2]。

而事实上,数据挖掘对所处理的数据有严格的质量要求。数据挖掘的一个关键步骤就是为挖掘任务建立起合适的数据集[3],因此数据挖掘之前的数据预处理也就显得至关重要。根据统计,三分之二的数据挖掘分析家们认为在一个完整数据预处理要花费整个挖掘时间的60%左右[4]。

Web挖掘分为3类:Web内容挖掘、Web日志挖掘、Web结构挖掘[5]。对于Web内容挖掘和Web结构挖掘来说,用户的身份显得不太重要,但对Web日志挖掘,尤其是对基于用户兴趣的挖掘来说,用户身份的识别非常重要,它将直接影响路径的填充,也严重影响挖掘的结果。

但从目前所使用的处理方法来看,都围绕着服务器本身的日志做文章,都不能真正令人满意,所以,文中提出一种新的日志格式和获取方法。

1 目前Web格式和用户身份识别方法

1.1 IIS5.0中W3C扩展日志文件格式

IIS5.0中W3C扩展日志文件格式包括[6]:

(1)日期(date):用户请求页面的日期;

(2)时间(time):用户请求页面的具体时间;

(3)客户IP地址:客户端主机的IP地址或DNS;

(4)username:客户端的用户名;

(5)用户代理(UserAgent):服务的提供者;

收稿日期:2008-09-08
基金项目:内蒙古人才基金资助项目(8批);内蒙古教育科研项目资助(NJZY07140)
作者简介:张春生(1965-),男,河北乐亭人,硕士,教授,研究方向为数据库技术、数据挖掘。

(6)服务器 ip 地址(ip) :服务器的 IP 地址;

(7)服务器端口(s2port) :服务器的端口号;

(8)方法(method):用户的请求方法;

(9)URL 资源(Uri2stem) :用户的请求页面;

(10)URL 查询(Uri2query):用户欲进行的查询;

(11)协议状态(status) :返回 HTTP 的状态标识;

(12)服务器名(computer name):服务器名称;

(13)发送字节数(sc - bytes) :服务器发送的字节数;

(14)接收字节数(cs - bytes):服务器收到的字节数;

(15)所花时间(time2taken):完成浏览所花费的时间;

(16)协议版本(version):传输用的协议版本;

(17)主机(host): 服务器的操作系统;

(18)Cookie (Cookie):Cookie 标识符;

(19)参照(Referer):用户浏览的上页。

1.2 用户身份识别困难的 5 个问题

用户身份识别困难的 5 个问题[2]:

(1)单个 IP 对多个服务器用户访问会话:ISP 利用 Proxy 代理为用户提供服务,统一 IP 访问一个 Web 站点时可能是不同的用户。

(2)多个 IP 对单个服务器用户会话:有些 ISP 对来自同一个用户的请求,会随机分配若干个 IP 中的一个给用户,这样一个用户进程会有不同的 IP。

(3)多个 IP 对单个用户:从不同机器上访问 Web 的同一个用户因为不同的进程而拥有不同的 IP ,这也使得追踪同一个用户变得复杂。

(4)多个服务器进程对单个用户:这种情况发生在用户打开多个浏览器窗口,同时对同一个站点的不同部分进行访问。

(5)单客户端对多用户:如家庭,很多人共用一台机器。

1.3 目前的识别算法

目前的用户识别方法有:cookies 方法、内嵌用户 ID、客户端软件 agent 等,由于用户可能关闭 cookies,或不愿注册等,导致算法不够理想[3,7~13]。

例如:

文献[3]利用历史访问序列解决除 ISP 用 Proxy 为用户提供服务以外的用户身份识别问题。

文献[10]利用 IP 地址、浏览器类型、网站拓扑结构来解决用户身份的识别,很显然不够理想。

文献[11]利用 cookies 方法,但不是所有的浏览器都支持 cookies,同时该方案的正确性有限。

纵观以上方案,都有一定的可取性,但也都存在着一定的缺陷。

2 基于兴趣的 Web 挖掘用户身份识别方法

要合理地识别用户的身份,仅靠现有的服务器记录的日志是不能达到要求的,必须有额外的资源开销,所以可采用以下两种方案:

(1)用户注册。

对于敏感内容,一旦确定准备进行挖掘,则采用用户注册的方式,只有注册用户才能浏览信息,这样可彻底解决用户身份的识别问题,当然,这种方法给用户带来不便,它适合于内容专一的网站。

比如:要挖掘不同用户对销售产品的兴趣偏好,可采用此办法。

(2)采用自定义日志方式。

定义 1:日志 L 是如下 4 元组的集合 $L = \{< User_1, Udatetime_1, D - url_1, S - url_1 >, \cdots, < User_n, Udatetime_n, D - url_n, S - url_n >\}$,其中 $User_i$ 是第 i 个用户的标识,标识可以是客户机的 CPU 标号或硬盘标号或网卡标号;$Udatetime_i$ 是第 i 个用户的访问时间戳;$D - url_i$ 是第 i 个用户访问的目标 URL 地址;$S - url_i$ 是第 i 个用户访问的源 URL 地址。

算法描述:

```
function networkpage_click
{
User=读取当前客户的 CPU 标号或硬盘标号或网卡标号
Datetime=getdatetime()
D-url=转向页面的 URL 地址
U-url=现页面的 URL 地址
insert L values(User,Datetime,D-url,S-url)
}
```

值得说明的是:如果网页用 ASP. net 编写,User 可用当前客户的 CPU 标号或硬盘标号或网卡标号,如果网页用 ASP 编写,限于技术,User 可用当前客户的网卡标号。

3 ASP.net 读取 CPU 标号、硬盘标号、网卡标号方法

```
public static void GetCpuInfo(out string cpuInfo)
{
//得到 cpu 信息
string cpuInfo="";
ManagementClass cimobject=new ManagementClass("Win32_Processor");
ManagementObjectCollection moc=cimobject.GetInstances();
foreach(ManagementObject mo in moc)
{
```

```
_cpuInfo=mo.Properties["ProcessorId"].Value.ToString();
}
cpuInfo=_cpuInfo;
}
//获取硬盘 ID
string _HDInfo="";
ManagementClass cimobject1 = new ManagementClass("Win32_
DiskDrive");
ManagementObjectCollection moc1=cimobject1.GetInstances();
foreach(ManagementObject mo in moc1)
{
_HDInfo=(string)mo.Properties["Model"].Value;
} HDInfo=_HDInfo;
}
public static void GetMacAddress(out string MacAddress)
{
//获取网卡硬件地址
string_MacAddress="";
ManagementClass mc = new ManagementClass("Win32_ Net-
workAdapterConfiguration");
ManagementObjectCollection moc2=mc.GetInstances();
foreach(ManagementObject mo in moc2)
{
if((bool)mo["IPEnabled"]==true)
MacAddress=mo["MacAddress"].ToString();
mo.Dispose();
}
MacAddress=_MacAddress;
}
```

4 结束语

采用用户注册的方法,可彻底解决用户身份的识别问题;采用自定义日志方式,可解决 5 个问题中的(1)~(4),因为问题(4)只影响路径填充,不属于用户身份识别问题,只有问题(5)无法解决。

对于问题(5),目前的解决方案也比较多,比如用根据网页拓扑结构,对用户的访问路径进行分析,达到确定用户的目的,由于存在着用户行为的随意性,问题(5)的解决方案不够理想,需进一步在日志文件上进行研究。

此方案虽然有一些资源浪费,但省去了复杂的用户识别问题,同时也减轻了数据清洗的复杂度,应是一个可取的方法。

参考文献:

[1] Kantardzic M. 数据挖掘——概念、模型、方法和算法[M]. 北京:清华大学出版社,2003.

[2] Dunham M H. 数据挖掘教程[M]. 北京:清华大学出版社,2005.

[3] 李超锋. Web 使用挖掘中数据预处理算法的设计与实现[J]. 中南民族大学学报:自然科学版,2007,25(1):56-60.

[4] Srivastava J, Cooley R, Deshpande M, et al. Web usage mining: Discovery and applications of usage patterns from web data[J]. Proc ACM SIGKDD, 2000,1(2):12-23.

[5] 刘立军. Web 使用挖掘的数据预处理[J]. 计算机科学,2007,34(5):200-204.

[6] 李烈彪. Web 日志挖掘中数据预处理方法的研究[J]. 计算机技术与发展,2007,17(7):45-48.

[7] 李 煊,庄镇泉. Web 访问挖掘预处理的用户识别算法[J]. 计算机工程与应用,2002(7):172-176.

[8] 马瑞民. Web 日志挖掘中数据预处理技术的研究[J]. 计算机工程与设计,2007,28(10):2358-2360.

[9] Tanasa D, Trousse B. Advanced data preprocessing for intersites web usage mining[J]. IEEE Intelligent Systems, 2004 (3/4):59-65.

[10] 吴 荣. Web 日志挖掘的用户识别算法研究[J]. 微型电脑应用,2007,23(9):61-62.

[11] 吴 强,梁继民,杨万海. Web 日志挖掘预处理中的用户识别技术[J]. 计算机科学,2002,29(4):64-66.

[12] 李 利,王秀峰. Web 应用中识别用户身份的一种方式[J]. 计算机与现代化,2004,23(3):101-104.

[13] 管 涛,罗建军,冯博琴. 基于粗糙集的 web 用户的识别与规则提取[J]. 计算机工程与应用,2003(25):114-117.

(上接第 61 页)

[4] Trolltech. Qt Reference Documentation Qt/ Embedded version. 2.3.10[M]. Norway: Trolltech 公司,2001.

[5] Watson N. Using QT to Develop for Embedded Linux[J]. Embedded Linux Journal,2001(3):59-61.

[6] 程 龙,刘彦明,鲁 刚,等. Qt/E 中文环境的搭建及对常用输入设备的支持[J]. 计算机工程,2007,33(15):275-276.

[7] 芮建武,吴 健,孙玉芳. 国际化标准框架下蒙文操作系统的设计[J]. 计算机研究与发展,2006,43(4):716-721.

[8] The Unicode consortium[EB/OL]. 2008-09. Unicode, Inc: http://www.Unicode.org.

[9] 张方辉,王建群. Qt/Embedded 在嵌入式 Linux 上的移植[J]. 计算机技术与发展,2006,16(10):64-66.

[10] 吴伟清. 基于 QTE 的嵌入式 Linux 中文环境解决方案[J]. 计算机工程,2005,31(2):87-88.

[11] 刘 淼. 嵌入式系统接口设计与 Linux 驱动程序开发[M]. 北京:北京航空航天大学出版社,2006.