Web 日志挖掘中的用户识别算法①

肖 慧[1,2]，王立华[2]

[1](上海海洋大学 信息学院，上海 201306)
[2](中国水产科学研究院 渔业工程研究所，北京 100141)

摘 要：介绍了现有的用户识别算法，针对用户识别目前存在的问题提出了 IASR(IP, Agent, Session and Referrer)用户识别算法。该算法采用重写 URL 的用户跟踪技术，引入会话(Session)来识别用户，能够高效准确地识别访问同一代理服务器的不同用户，很好地解决同一用户直接从浏览器地址输入 URL 信息访问站点造成的“多用户问题”。最后，对用户识别算法的发展趋势进行了展望。

关键词：用户识别；重写 URL；会话机制；Web 日志挖掘

User Identification Algorithm in Web Log Mining

XIAO Hui[1,2], WANG Li-Hua[2]

[1](Information College, Shanghai Ocean University, Shanghai 201306, China)
[2](Institute of Fisheries Engineering, Chinese Academy of Fishery Sciences, Beijing 100141, China)

Abstract: The paper introduces some existing user identification algorithms, proposes IASR (IP, Agent, Session and Referrer) user identification algorithm to solve existing problems on user identification. The proposed algorithm overwrite URL in order to track users, efficiently and accurately identifies different users accessing the same proxy, and satisfactorily solves “Multi-User Problem” due to accessing Web via directly inputting URL in browser’s address bar. At last, the paper prospects future development of user identification algorithm.

Keywords: user identification; overwrite URL; session mechanism; Web log mining

Web 日志挖掘是指通过分析挖掘 Web 服务器日志记录发现用户访问 Web 页面的模式。Web 日志挖掘一般分为三个步骤：数据预处理、模式挖掘和模式分析。数据预处理是指根据挖掘的目的，对原始的 Web 日志文件进行有效地分析，提取、合并，最后形成适合进行模式挖掘的数据文件。它是 Web 日志挖掘的基础和实施有效挖掘算法的前提，在整个 Web 日志挖掘过程中起着非常重要的作用。数据预处理包括数据清理、用户识别和会话识别三个部分，其中用户识别是关键环节，它直接影响了会话识别效果的好坏，间接影响了整个 Web 日志挖掘的成败。

用户识别是指从大量的 Web 日志记录中找出访问 Web 站点的具体用户，包括用户的 IP、操作系统、浏览器等用户信息。但是由于本地缓存、防火墙、代理服务器的存在，使得用户识别比较困难。多个用户通过代理服务器的请求在日志中具有相同的标识，防火墙的设置使得不同的用户在日志记录中的 IP 都是防火墙的 IP，本地缓存使得请求记录无法被服务器记录，这些都给用户识别造成了困扰。针对上述难题，许多学者纷纷提出了有效的用户识别算法。通过总结这些算法的特点，得出了用户识别算法通用的三条启发式原则[1,2]：

(1) 如果 IP 地址不同，则认为是不同的用户；

(2) 如果 IP 地址相同，但操作系统或浏览器软件(用户代理域)不同，则认为是不同的用户[3]；

(3) 如果 IP 地址相同，用户使用的操作系统以及

① 基金项目:国家基础条件平台建设项目渔业科学数据平台建设项目(2005DKA31800-03)
收稿时间:2010-09-13;收到修改稿时间:2010-10-23

浏览器软件也相同，那么根据网站拓扑结构对用户进行识别，如果用户请求的页面不能从已访问的任何页面到达，则认为这是一个新用户[4,5]。

然而，这些规则并不能非常准确的识别每一个用户。例如，当一个用户使用多种浏览器或直接在地址栏输入 URL，这时会被认为是多个用户，即“多用户问题”；具有相同 IP 地址的用户使用同种操作系统和同种浏览器网站，并且浏览的页面集合相同，则会被认为是同一个用户。

文献[6]中提出了一种基于 cookie 技术和扩充日志属性的用户识别方法，这种方法可以有效地识别通过同一代理服务器访问的不同用户，但是它要求服务器端和客户端都要支持 cookie，具有一定的局限性。文献[7]提出了基于日志引用页的用户识别算法，该方法简单实用，但是缺乏有效性和准确性。

为了打破 Cookie 造成的局限性，同时高效地识别访问同一代理服务器的不同用户，解决同一用户直接在地址栏输入 URL 造成的“多用户问题”，本文提出了重写 URL 的用户跟踪方法，利用用户 IP、用户代理(Agent)、用户会话(Session)和引用页(Referrer)来识别用户，该方法简称为 IASR 用户识别算法。

1　IASR用户识别算法

1.1 重写 URL

Web 服务器跟踪用户的状态通常有 4 中方法[8]：(1)在 HTML 表单中加入隐藏字段 HIDDEN，它包含用于跟踪用户状态的数据；(2)重写 URL，是它包含用于跟踪用户状态的数据；(3)用 Cookie 来传送用于跟踪用户状态的数据；(4)使用会话(Session)机制。目前许多用户识别算法都依赖 Cookie，然而由于 Cookie 具有一定的风险性，有可能泄露用户的私人信息，因此用户采用禁止 Cookie 的模式访问网站，这给依赖 Cookie 的用户识别算法的可行性带来了考验。会话(Session)机制也是一种常用的用户跟踪方法。会话是指用户对某个 Web 站点连续请求的集合，其具体实现的方法有两种：Cookie 和重写 URL。由前面的分析可知，如果客户端不支持 Cookie，依赖 Cookie 的会话机制也无法实现。因此，本文提出重写 URL 的会话跟踪机制，其原理如下：

(1) 判断当前 Web 组件是否支持会话，如果不支持会话，则 Web 应用中的所有链接地址保持原值；否则，则进入(2)；

(2) 判断浏览器是否支持 Cookie，如果支持 Cookie，则 Web 应用中的所有链接地址保持原值；否则，在原来链接地址后面加上 sessionid 字段，字段值必须为具有唯一性的字符串。

采用重写 URL 的用户跟踪方法，即使用户禁止使用 Cookie，服务器也能够通过会话机制跟踪用户。会话 ID 分别记录到日志记录的 Cookie 字段和请求 URL 字段中，这样几乎每一条日志记录都包含会话标识。

1.2 IASR 用户识别算法

目前几乎所有的用户识别算法都遵循了经典的三条启发式原则，IASR 用户识别算法也不例外。IASR 用户识别算法利用重写 URL 的用户跟踪算法，使得每一条日志记录都包含会话标识，然后引入会话判断，扩展三条启发式原则。会话判断原则为：在 IP 地址和用户代理均相同的情况下，如果会话相同，则认为是相同用户。下面对本文提出的这种用户识别方法在 Windows XP 下的具体实现加以说明。

首先设计重写 URL 的功能。根据重写 URL 的会话跟踪机制原理，要求当前的 Web 组件支持会话，并且在服务端编写 URL 重写函数，确保在不支持 Cookie 情况下的访问记录包含会话标记，从而方便进行用户跟踪。该步骤为 IASR 用户识别算法搭建了可行性环境。

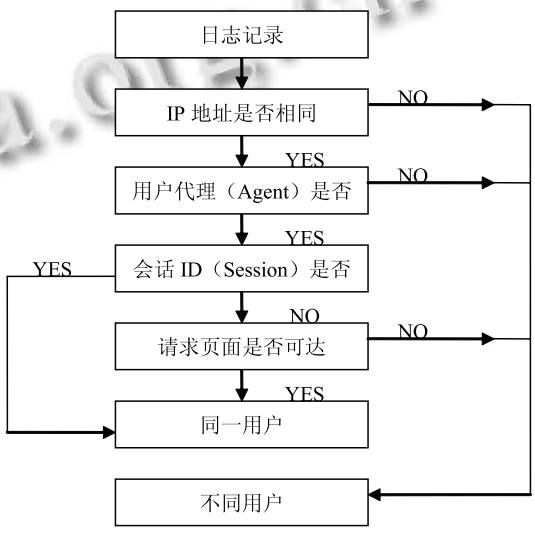

图 1　IASR 算法流程图

接下来提取会话标识，进行用户识别。通过访问实现了重写 URL 会话跟踪机制的 Web 应用，每一条

日志记录都包含会话标识。对于支持 Cookie 的访问记录，会话标识位于 cs(Cookie)字段，对于不支持 Cookie 的访问记录，会话标识位于 cs-uri-stem 字段。编写相应的代码可以将会话标记提取出来进行用户识别。如果两条记录具有相同的 IP 地址和用户代理域，而且它们的会话标识也相同，则认为这两条访问记录属于同一用户，并且属于同一会话，具体实现流程如图 1。

IASR 用户识别算法采用重写 URL 的用户跟踪技术，消除了一些用户识别算法对 Cookie 的依赖，加强了算法的通用性；通过引入会话识别用户，可以快速准确地识别相同用户，从整体上提高了算法的效率和准确性。

2 实验与结果

本文在 Windows XP 下对用户识别通用方法和 IASR 用户识别算法分别进行了实验。实验分为两个步骤：(1)从平台服务器上大量的日志记录中随机选取部分日志记录，进行数据清理；(2)进行用户识别操作。本实验抽取了 2009 年 2 月份的日志记录，经过数据清理之后，消除了不完整的数据记录和与挖掘无关的字段信息。从清理过的日志记录随机抽取 100 条日志记录作为实验数据，实验结果如表 1 所示。

表 1 实验结果表

方法	总记录（条）	耗时（ms）	用户数（个）
IASR	100	16	45
通用方法	100	31	58

实验结果分析：实验总记录数均为 100 条，通用方法耗时 31ms，识别的用户数为 58 个，而 IASR 算法耗时 16ms，识别的用户数为 45 个，IASR 用户识别算法的速度几乎是通用算法的 2 倍，识别用户的准确性也高于通用算法。

通用方法由于没有引入 Session 来识别用户，很多 IP 地址相同、用户代理相同、会话 ID 相同的记录，经过路径分析而得出多个用户的结果。如表 2 所示，IP 地址为 172.31.4.87 的日志记录被识别出包含 2 个用户，而且这两个用户拥有相同的会话 ID。然而，会话 ID 具有唯一性，一个会话只能代表一个用户，这两个用户实际上是同一个用户。因此，通用的三条启发式规则无法对直接从浏览器地址栏输入 URL 信息的用户进行准确有效的识别。

表 2 具有相同 Session 的用户表

ID	IP	Agent	Sessions	URLs
1	172.31.4.87	Mozilla/4.0+(compatible; +MSIE+7.0; +Windows+NT+5.1)	lstat_bc_1136402=13584688982071541668;+ASPSESSIONIDSSRRRQRD=OJLHJLJCEDBBNNKAJGFDHHII	http://fishery.agridata.cn/index.asp http://fishery.agridata.cn/detail.asp?db=FishData&table=A040374&id=22 http://fishery.agridata.cn/document/A040374/渔业生态环境监测规范 4 报告编制.doc http://fishery.agridata.cn/grade2.asp?cataid=104 http://fishery.agridata.cn/grade3.asp?st=llsj&id=A040367 http://fishery.agridata.cn/grade3.asp?st=llsj&id=A040374
2	172.31.4.87	Mozilla/4.0+(compatible; +MSIE+7.0; +Windows+NT+5.1)	lstat_bc_1136402=13584688982071541668;+ASPSESSIONIDSSRRRQRD=OJLHJLJCEDBBNNKAJGFDHHII	http://fishery.agridata.cn/ http://fishery.agridata.cn/grade2.asp?cataid=104 http://fishery.agridata.cn/grade2.asp?cataid=55 http://fishery.agridata.cn/grade3.asp?st=llsj&id=A040320 http://fishery.agridata.cn/grade3.asp?st=llsj&id=A040367

IASR 用户识别算法借助 Session 来识别用户，其主要作用是快速识别出同一用户。如果两条日志记录具有相同的会话 ID，则肯定是同一用户。该方法不仅能够准确有效地识别直接从浏览器地址栏输入 URL 信息进行 Web 访问的用户，而且能够迅速地对具有相同会话 ID 的用户进行识别，省去路径分析过程，提高了用户识别效率。

3 总结与讨论

用户识别算法是 Web 日志挖掘领域的一个研究热点，与会话识别和模式挖掘相比，用户识别的重要性显而易见。本文提出的 IARS 用户识别算法是对三条

启发式规则的扩展和改进，其特点和优势主要体现在两个方面：

(1) 服务器端应用程序的改进。重写 URL 涉及到服务器端应用程序的修改，在服务器端 Web 应用支持 Cookie 的情况下，采用重写 URL 的用户跟踪方法，即使用户浏览器禁止 Cookie 的使用，服务器端仍然能够准确的识别每个用户。这是 IASR 用户识别算法的前提，只有每条日志记录都包含会话 ID，才能准确的实现 IASR 用户识别操作。

(2) 引入会话（Session）识别用户。通过服务器端应用程序的改进，日志记录中均包含有会话信息。IASR 用户识别算法将三条启发式规则扩展成为四条判断规则。当两个用户具有相同的 IP 地址和用户代理时，不是直接采用路径分析方法，而是先进行会话判断，如果这两个用户的会话 ID 相同，直接断定他们为同一用户；否则，再进行路径分析。

目前大多数用户识别算法都需要 Cookie 的支持，然而，IASR 用户识别算法摆脱了对 Cookie 的依赖，提高了用户识别算法的通用性。它还引入会话识别，能够高效准确地识别通过同一代理服务器访问站点的用户和直接在浏览器地址栏输入 URL 信息访问站点的用户。尽管如此，用户识别还存在一些亟待解决的问题。比如，同一用户使用多种浏览器访问站点的情况，本地缓存造成的日志记录不完整等问题。就当前存在的这些问题，可以看出用户识别算法还具有较大的发展空间，其发展趋势包括两个方面：(1)提高用户识别算法的有效性和准确性。基于经典的三条启发式规则，增加新的识别条件或者改进服务器端应用程序，提高用户识别算法的效率，增加用户识别算法的准确性；(2)突破三条启发式规则，提出创新算法。目前的用户识别算法都被三条启发式规则所束缚，要推动 Web 日志挖掘领域的发展，必须突破现有的算法模式，从新的角度提出用户识别算法。

参考文献

1 赵伟,何丕廉,陈霞,谢振亮.Web 日志挖掘中的数据预处理技术研究.计算机应用,2003,23(5):62－65.

2 熊忠阳,周亚峰.Web 访问挖掘的预处理技术的研究.计算机技术与发展,2007,17(8):11－15.

3 Pirolli P, Pitkow J, Rao R. Silk from a Sow's Ear: Extracting Usable Structures from the Web. CHI-96, Human Factors in Computing Systems, 1996.

4 李煊,庄镇泉.Web 访问挖掘预处理的用户识别算法.计算机工程与应用,2002,38(7):173－176.

5 陆丽娜,杨怡玲,管旭东,魏恒义.Web 日志挖掘中的数据预处理的研究.计算机工程,2000,26(4):66－68.

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

7 方成效,袁可风.Web 日志挖掘的数据预处理研究.计算机与现代化,2006,(4):79－82.

8 孙卫琴.Tomcat 与 Java Web 开发技术详解.北京:电子工业出版社,2009.

（上接第 116 页）

线路连接，15 台服务器，逐条光线路、定时光线路、循环光线路自动测试和人工测试，数据都能得到及时响应。设定故障点人工测试和本软件自动测试都能及时报警并显示。系统运行了大约 3 月后，发现有一台远端交换机经反复测试无法通讯，经现场查看发现，交换机柜风扇停机，交换机无法充分散热造成交换机故障，从而将故障及时排除。但不可否认的是，由于要利用 ICMP 协议，因此要使得系统正常工作，必须使到目标主机的设备不能阻止 ICMP 数据包，一定程度上造成网络潜在的不安全性。此外，系统也没有充分利用目前基于 SNMP 协议的图形化软件界面的优点，各条线路的流量并不是按照 SNMP 协议获得[5,6]等都使得系统显得不直观，有待于进一步研究。

参考文献

1 陈东,邓鸿.MRTG 在监控网络主干状况的应用.潍坊学院学报, 2005,5(4):35-36.

2 杨龙江,高静.图形化方针 PING 命令的设计.电脑编程技巧与应用,2003,9:22-24.

3 马金虎.VisualC#创建 TraceRt 命令.电脑编程技巧与应用, 2004,10:49-51.

4 刘小明.MRTG 日志文件的分析研究.电脑学习, 2007,6: 22-23.

5 贾志强.基于 TCP/IP 的流量检测技术.中国石油大学胜利学院学报, 2007,21(1):14-15.

6 卢苇,严斌宇,郑畅.MRTG 在校园网状态检测参数中的应用.四川大学学报(自然科学版), 2007,38(2):189-191.