计算机科学2002Vol.29№.4

Web日志挖掘预处理中的用户识别技术

User Identification in the Preprocessing of Web Log Mining

吴 强 梁继民 杨万海

(西安电子科技大学电子工程学院 西安710071)

Abstract The tasks of the Web Log Mining preprocessing are analyzed and a functional model of it is presented. A user identification method based on cookie technology and extending Web Log attributes is proposed. The method can distinguish effectively the multiple users using the same one proxy server and overtake the difficulties caused by the erasion of cookies stored on the user's file system.

Keywords Web Log Mining, Preprocessing, User identification, Cookie

1 引言

互联网技术和应用的迅速发展使得可以从因特网获取的信息量日益剧增,因此迫切需要一种新的技术从这些信息中快速、及时地发现有用的知识,提高信息的利用率。作为数据挖掘技术[1]研究的一个重要领域,Web日志挖掘(Web Log Mining)是从服务器日志文件内大量的用户访问记录中抽取有用信息的过程。通过对Web日志的分析,可以构造出用户的行为模式,对于分析改进网络性能、优化网站的设计和拓扑结构以及改善企业的市场营销决策等会有极大的帮助[2,3]。

当前Web日志挖掘领域的研究已取得了很大的进展,但是目前的研究重点大都集中于挖掘算法的设计、分析和改进,对日志文件预处理方法的研究相对较少,然而正确有效地对Web日志文件进行预处理,不仅有利于随后的挖掘算法分析,而且对于最终形成准确可靠的用户行为模式也是极为重要的。

本文对Web日志挖掘预处理所要完成的任务进行了分析,提出了一个Web日志预处理功能模型;分析了现有Web日志预处理方法中的用户识别技术,提出了一种基于cookie技术和扩充日志属性的用户识别方法,这种方法不仅可以有效地识别通过同一代理服务器访问的不同用户,而且较好地解决了由用户删除本机cookit而产生的同一用户多次被标示的问题。

2 Web日志挖掘预处理

一般的Web日志文件中记录的是每个访问用户的信息,不同服务器的Web日志记录是不同的,但其中都包含有访问用户的基本信息。

表1显示的是四条Windows2000服务器的Web日志记录,其中包括:访问日期、时间、用户IP地址、用户名、服务器IP地址、方法、所请求URL资源、服务器响应状态、用户代理、发送字节数等。

表1 Windows2000服务器的Web日志

date	time	c-ip	Cs-username	s-ip	s-port	Cs-method
2001-03-30	02:22:54	202.119.80.20	-	202.117.121.7	80	GET
2001-03-30	04:47:08	159.226.65.62	-	202.117.121.7	80	GET
2001-03-30	06:53:00	202.98.83.11	-	202.117.121.7	80	GET
2001-03-30	07:07:41	211.69.197.17	-	202.117.121.7	80	GET

cs-uri-stem	sc-status	cs(User-Agent)
/papers.htm	200	Mozilla/4.0+(compatible;+MSIE+5.0;+Windows+98;+DigExt)
/index.htm	200	Internet+Explorer+4.01
/java/cont.css	200	Mozilla/4.0+(compatible;+MSIE+5.01;+Windows+NT+5.0)
/index.htm	200	Internet+Explorer+4.01

Web日志文件记录中存储的是用户访问站点信息的原始记录,直接在这些数据上面进行挖掘是比较困难的,在使用算法或工具对其分析之前,必须进行预处理。

Web日志挖掘的预处理阶段主要分为三步[4]:数据清洗、用户识别和会话识别,其中最重要的是用户识别。一般的处理流程如图1所示。首先进行数据清洗,目的在于去除日志中不相关和无效的记录,通常有几种情况:(1)一般情况下用户不会显示请求站点中的图形文件和页面样式文件,这些文件通常是站点根据请求页面中的连接自动下载的,所以只要cs-uri-stem项是以jpg、jpeg、JPG、JPEG、gif、GIF和css、js等结尾的记录都可以删除;(2)用户请求访问失败的记录,这类访问的返回代码为404(没有找到)、301(永久删除)或500(内部服务器错误)等;(3)用户请求方法中不是GET的记录也可以删除。通过数据清洗后得到净化的日志。预处理的第二步是

吴 强 硕士研究生,主要研究方向为数据挖掘、网络安全等。梁继民 博士,副教授,主要研究方向为信息融合、模式识别等。杨万海 教授,博士生导师,主要研究方向为信号处理、系统仿真等。

在净化日志中识别用户，由于用户机器中的 cache、代理服务器和防火墙的使用，这一步工作比较复杂，同时也是极为关键的，其主要内容包括区分用户类型、识别不同用户等，所采用的主要技术有基于 cookie 技术和基于网络拓扑结构的路径分析等。Web 日志预处理最后一步会话识别是在第二步得到的用户已识别的日志文件中，将每个用户访问站点的内容提出，得到用户访问站点的时间、页面顺序以及其它信息，并根据已有的站点拓扑结构补充用户访问该站点的部分路径，最终形成一个完整的用户会话。经过以上处理后得到的用户会话日志就可以用于进一步的挖掘算法分析。

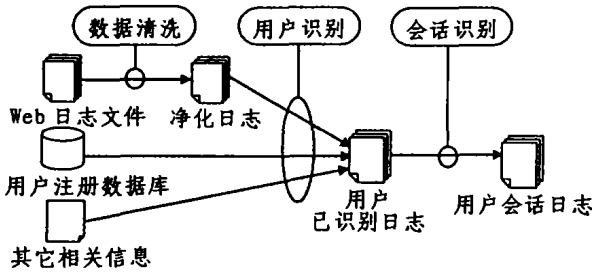

图1 Web 日志预处理模型

3 Web 日志预处理中的用户识别技术

Web 日志挖掘预处理阶段的首要任务是有效识别出不同用户，得到正确的用户会话，为进一步的挖掘算法应用提供净化数据。通常访问站点的用户可以分为四种类型[5]：

1)未识别用户，站点对于此类用户的相关信息一无所知。由于网络协议中规定站点响应用户访问请求的同时要求用户至少提供机器名，因此此类用户的信息中至少包含用户的机器名。一般情况下，这类用户可能是通过匿名代理上网的，此时站点日志保存的用户信息就是代理服务器的机器名。

2)会话用户，指可以通过用户机器里的 cookie 或是别的相关信息推断出来的用户。站点的访问者通常都是这类用户，只要他们访问过站点，就可以通过用户机器中保存的 cookie 得到他们的信息。如果用户禁止使用 cookie，站点也可以通过用户请求时使用的机器名、浏览软件、操作系统和在此之前的页面请求等信息来识别用户，但是这种识别方法不是绝对准确和可靠，尤其是在用户使用代理服务器访问站点，或者多个用户共用一台计算机的时候。

3)可跟踪用户，指在访问站点的众多用户中可以唯一、可靠地识别的用户。在网络应用的早期，站点通常是在用户请求的 URL 中加入识别信息，然后使用 CGI 脚本将随后的请求归类形成用户会话。这种跟踪方法不仅大大增加了服务器的负担，而且影响页面缓存方法的使用，降低了用户的访问速度。目前常用的跟踪方法是在用户机器中的写入 cookie 标示，从而在用户的下一次访问时将其识别出来，但这种方法会由于用户禁止 cookie 的写入或是将 cookie 文件删除而失效。

4)已识别用户，指附带有更多识别信息的可跟踪用户。此类用户的标示更多，在用户的每个请求中都有附加信息，而且可以通过使用已有的域注册数据库统计出用户相关信息，最终识别出用户。这类数据库的使用需要明确地知道代理服务器的连接，即每个使用代理服务器的用户都是可知的，而且要求在服务器上填入真实可靠的信息，在用户识别的过程中就可以逐级往下查询用户的真实身份。

Web 日志挖掘中用户识别的过程是从“未识别用户”到“已识别用户”的一个逐步深入的确认过程。根据是否更改现有的日志记录，可以将用户识别技术分为两类：一类方法不改变 Web 日志文件的结构，而是结合站点的拓扑结构分析日志中的用户请求，构造用户的访问路径，进而通过一些启发式规则来识别出用户[5~8]，比如路径分析技术[7]。另一类方法改变现有 Web 日志文件的结构，添加更多的附加相关信息，如用户机器名、内部 IP 地址等[5]。

随着公司内部网(Intranet)的广泛应用，许多用户是通过代理服务器访问因特网的，因此站点 Web 日志文件中记录的通常是代理服务器的 IP 地址，这给正确分析用户行为带来了很多麻烦，如果仅以机器名来识别用户，会导致错误地将多个用户作为同一用户处理，可能得出错误的用户行为模式。通过使用 cookie 来标示用户的方法简单易行，能有效识别多个用户通过同一代理服务器访问同一站点的情况，因而在现有网络中得到广泛应用。但是，如果用户删除了计算机中的 cookie 文件，服务器会在用户再次访问时将 cookie 重新写入用户计算机中，从而导致将同一用户不同时间的访问当作两个用户访问，增加了后继用户行为模式分析的复杂度。

为了解决以上问题，本文提出了一种基于 cookie 技术和扩充日志属性的用户识别方法，首先在服务器端加入对用户访问站点的 cookie 读写操作，在用户访问站点时服务器不仅在用户本机内写入唯一标示此用户的 cookie，而且将这些 cookie 写入服务器的 Web 日志的扩充属性中，即在用户访问站点的每条访问记录中都有可唯一标示此用户的 cookie 属性项存在。当用户再次访问站点时，如果用户本机中的 cookie 已经存在而且有效，就继续使用该 cookie 作为用户标示；如果 cookie 已被删除或者已经失效，服务器将再次把 cookie 写入用户本机中，将用户记录为新的用户。使用以上方法，可以通过 Web 日志中的 cookie 属性项较好地识别出用户会话过程。对于由于 cookie 被删除而产生的新的用户会话，可以通过与以往日志记录相比较，找出类似的用户标示，从而将这两个用户标示合并为一个。

以上方法不仅解决了由于用户删除本机中的 cookie 给用户识别带来的困难，而且通过用户标示合并大大降低了后继挖掘算法的复杂度和计算量。下面对本文提出的这种用户识别方法在 Windows2000 Server 下的具体实现加以说明。

首先将服务器端的 Web 日志属性扩充为全部，在服务器端和站点主页、索引页添加对用户 cookie 的处理并写入 Web 日志。Windows2000日志所有的扩充属性包括：日期(date)、时间(time)、客户 IP 地址(c-ip)、用户名(cs-username)、服务名(s-sitename)、服务器名(s-computername)、服务器 IP 地址(s-ip)、服务器端口(s-port)、方法(cs-method)、URL 资源(cs-uri-stem)、URL 查询(cs-uri-query)、协议状态(sc-status)、Win32状态(sc-win32-status)、发送字节数(sc-bytes)、接收字节数(cs-bytes)、所花时间(time-taken)、协议版本(cs-version)、主机(cs-host)、用户代理(User-Agent)、cookie、参照 cs(Referer)等。表2所示为一条用户访问实例。其中 userid=2001050713571400001项为扩充的 cookie 属性项。cookie 的发送是通过 http 头来实现的，它早于文件的传递。cookie 中包含 name，value，expires，path，domain，secure 等内容。其中 name 为 cookie 名，name 不能使用分号和逗号，有多个 name 值时用分号分隔；value 为此 cookie 值；expires 为 cookie 的有效期限，即此 cookie 的生存期；path 为 cookie 支持的路径，如果 path 是一个路径，则 cookie 对这个目录下的所有

文件及子目录生效，如果 path 是一个文件，则 cookie 指对这个文件生效；domain 为对 cookie 生效的域名；如果给出 secure 标志，表示 cookie 只能通过 SSL 协议的 https 服务器来传递。

表2 Windows2000 Server 用户访问记录实例

date	time	c-ip	cs-username	s-sitename	s-computername
2001-05-07	05:57:16	192.117.127.1	-	W3SVC1	PIII3

s-ip	s-port	cs-method	cs-uri-stem	cs-uri-query	sc-status
192.117.127.33	80	GET	/test/top.htm	—	200

Sc-win32-status	sc-bytes	cs-bytes	time-taken	cs-version	cs-host
0	3691	277	0	HTTP/1.	piii3

User-Agent	Cookie	cs(Referer)
Mozilla/4.0+(compatible;+MSIE+5.01;+Windows+NT+5.0)	userid=2001050713571400001	http://piii3/test/default.htm

读写 cookie 是在服务器和客户两端进行的，具体实现流程如图2所示。客户端发送页面请求，浏览器检测本机是否有相应的 cookie 存在，如果有就将 cookie 附在 http 头部，发送给服务器；服务器端响应客户端的请求，读出 cookie 值，写入 Web 日志。其中基于 cookie 的用户识别方法关键是 cookie 值的设置。在实际使用中因为 cookie 值主要是用来标识用户（用 userid 表示），所以必须使用唯一标识，用户标识可以用多个属性值合并表示，例如：userid＝日期＋时间＋访问站点序号。同时要将 cookie 的有效期设置为一个固定时间段，超出这个时间段后的用户请求可以认为是一个新的用户请求，这是为了区分一台计算机多个用户使用的情况，而且对最终用户行为模式的形成不会有太大影响。通过使用 cookie，所有用户将由其 cookie 值唯一确定，可以有效地识别出使用代理服务器的用户，对于代理服务器中使用 cache 的情况也能较好地区分。

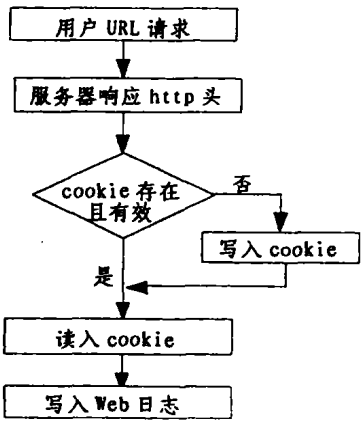

图2 cookie 的读写流程

本文在 Windows2000 Server 上对以上基于 cookie 技术和扩充日志属性的用户识别方法进行了实验，实验结果表明，该方法可以准确地识别出多个使用同一代理服务器访问站点的用户。

结论 正确有效地对 Web 日志文件进行预处理，不仅可以大大降低日志挖掘算法的数据处理量，而且有助于准确可靠地分析出用户的行为模式，为优化网站设计提供可靠的参考信息。本文首先给出了一个 Web 日志文件预处理的功能模型，并提出了一种基于 cookie 技术和扩充日志属性的用户识别方法，这种方法不仅可以有效地识别通过同一代理服务器访问站点的不同用户，而且通过用户标示合并，较好地解决了由于用户删除本机 cookie 文件、服务器将 cookie 再次写入而产生的同一用户多次标示的问题。

基于 cookie 的用户识别方法的局限性在于不是所有的客户端浏览器都支持或允许使用 cookie。在不能使用 cookie 的情况下，必须采用别的方法，比如路径分析方法，该方法在所得到的用户访问站点子图的基础上，根据网站的拓扑结构补充形成完整的用户访问路径，即用户访问站点时间序列，然后进行用户识别。实际应用中可以将这两类技术结合使用，一方面使用 cookie 标示每个用户，另一方面，使用路径分析技术分析用户访问站点的时间序列，从而准确可靠地识别每个用户，为进一步的日志挖掘做准备。

参考文献

1 Leung Y, Leung K S. An Intelligent Expert Systems Shell for Knowledge-Based Geographic Information Systems. International Journal of Geographic Information Systems, 1993(7): 351～355

2 Zaiane O R, Xin M, Han J. Discovering Web Access Patterns and Trends by Applying OLAP and Data Mining Technology on Web Logs. In: Proc. Advances in Digital Libraries Conf. (ADL'98), Santa Barbara, CA, April 1998. 19～29

3 Pei J, Han J, Mortazavi-Asl B, Zhu H. Mining Access Pattern efficiently from Web logs. In: Proc. 2000 Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD'00), Kyoto, Japan, April 2000

4 Cooley R, Mobasher B, Srivastava J. Web mining: Information and Pattern discovery on the World Wide Web. In: Proc. IEEE Intl. Conf. Tools with AI, Dec. 1997

5 Pitow J. In Search of Reliable Usage Data on the WWW. In: Proc. of the 6th Intl. World Wide Web Conf. Santa Clara, CA, 1997. 451～463

6 Cooley R, Mobasher B, Srivastava J. Grouping web page references into transactions for mining world web browsing patterns: [Technical Report TR 91-021]. University of Minnesota, Dept. of Computer Science, Minneapolis, 1997

7 Pirolli P, Pitkow J, Rao R. Silk from a Sow's Ear: Extracting Usable Structures from the Web. In: Proceedings of CHI'96, Vancouver BC, ACM Press, 1996. 118～125

8 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)：66～72