Web 用户访问日志数据挖掘研究

张　娥　　冯耕中　郑斐峰

(西安交通大学管理学院　西安　710049)

摘　要　简要介绍了 Web 用户访问日志数据挖掘研究内容,综述了 Web 用户访问日志数据挖掘研究的基础,包括常用术语含义、用户访问 Web 的几种习惯和用户访问日志的分布情况,论述了如何识别用户访问的 Web 服务器的会话期间,指出了 Web 用户访问日志数据挖掘研究的难点所在。

关键词　Web 用户访问日志　数据挖掘　Web 数据挖掘

1　Web 用户访问日志研究介绍

1.1　*问题提出背景*　据国际互联网组织 W3C 统计,目前全世界有 1/10 的人接触过互联网。而从中国 CNNIC 的统计数据来看,中国上网的人数已经达到了 2000 万,平均每人每星期在网上浏览的时间为 2.1 小时。各种基于 Internet 网络的应用业务也如雨后春笋般地发展起来,例如网上商店、网上银行、远程教育、远程医疗等。Web 为商业、教育等各个行业开辟了新的发展契机,特别是方便、快捷、高效的电子商务,据统计,美国在 1998 年到 2002 年里,电子商务发展速度超过了 30%。到 2002 年,全球消费者的电子贸易额将达到数千亿美元的规模。毫无疑问,随着互联网的普及,整个商业生态将会改变,形成真正以网络为基础的互联网经济。在这个经济架构中,所有的新兴和传统企业都将依赖全局性的电子商务而运作。强大的互联网服务将成为整个商业运作的发动机。如果说互联网是一张有形的网,那么互联网背后用户使用互联网的方式就是一张无形的网,有形网是由知识连接编织而成,无形的网是由用户使用互连网的行为习惯组成的,其中包含的信息具有不可估量的价值。由于互联网传输协议的无状态性,对企业来讲,如何发现并利用人们在互联网上的行为习惯非常困难。于是出现了 Web 用户访问日志数据挖掘(Web Usage Mining)。

1.2　*研究意义和本文内容*　互联网作为第四媒体,在人们生活、企业商业活动中的角色越来越重要。研究人们从 Web 上获取信息的模式、获取信息的类型,从而可以得到用户兴趣偏好等方面的信息,以客户为中心已经成为新的经营理念之一,从日志中提取模式,以便掌握用户习惯,这不仅可以指导企业发掘潜在客户、潜在需求,也方便用户制定战略决策。Web 用户访问日志研究也将成为未来各互联网企业决胜千里的利器之一。Web 用户访问日志数据挖掘就是利用数据挖掘的技术挖掘分析用户访问留下的日志文件,挖掘用户访问模式,为网站经营管理和结构调整提供决策支持;为企业发现新市场机会,进行市场决策;提高通过网站施行的营销效果,以及为企业进行战略决策提供有价值的潜在的信息。

Web 用户访问日志数据挖掘是信息检索与数据挖掘两个研究领域的交叉,是新兴的研究学科。目前还有很多问题有待探索。本文将从用户使用 Web 的方式入手,就 Web 用户访问日志数据挖掘研究中的一些主要基础性问题做深入探讨。

2　Web 用户访问过程

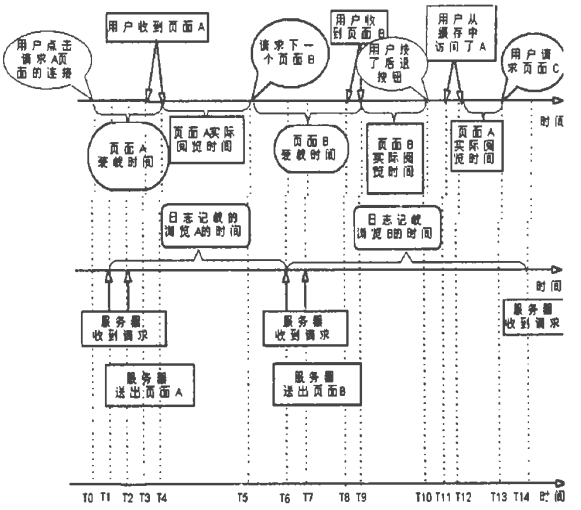

图 1　用户访问服务器处理过程图

用户访问过程示意如图 1 所示。T_0 时刻用户提出对 A 页面的请求,服务器在 T_1 时刻收到请求,在时刻 T_2 处理请求,在服务器上找到 A 页面,并在 T_3 时刻发出页面内容,浏览器收到页面经过解释于 T_4 时刻呈现在客户端。T_5 时刻用户浏览完 A 页面发出对下一个页面 B 的请求,服务器在 T_6 时刻收到请求,在时刻 T_7 处理请求,在服务器上找到 B 页面,并在 T_8 时刻发出页面内容,浏览器收到页面经过解释于 T_9 时刻呈现在客户端。如此重复。用户访问过程中有可能要从缓存中读取页面,如图 1 中第二次访问 A 页面。总体上讲,用户访问过程就是用户提出服务请求,服务器响应请求的过程。

3　Web 用户访问日志分布及特点

Web 用户访问日志分别记录在三个地方:客户端、代理服务器和 Web 服务器端。客户端浏览器记录了单用户访问多个网站的情况;Web 服务器的日志则记录了多个用户访问一个网站的情况;代理服务器日志跟踪记录了多个用户访问多个网站的情况。三个日志数据集记载了用户使用网络资源的不同模式。客户端日志数据记录了单用户访问多服务器的模式,Web 服务器端日志数据记载的是多用户访问单服务器的模式,而代理服务器端日志记载的是多用户访问多服务器的访问模式。

这几种数据集合记录的数据类型、所反映用户浏览行为的信息和获取相应信息的方法差异也很大。代理服务器和 Web 服务器日志数据的收集是由服务器自动记录的,客户端日志数据则需要有专门的程序收集,比如客户端的代理软件或者经过修改的浏览器等。相对而言,服务器端日志格式标准化程度是最高的。W3C 组织规定了服务器日志的两种格式:CLF (Common Log Format) 和 ECLF (Extended Common Log Format), ECLF 格式数据最基本的数据域包括客户端 IP、服务器 IP/名、用户名称/ID 号(User)、请求时间(Time)、请求(Request)、状态(status)、服务器接受到的比特大小(BytesRecvd)、服务器响应用户请求发送出的比特大小(BytesSent)、处理时间(Processtime)、参引(referrer)页面、代理(agent),表 1 就是某网站的访问日志片段。只有当请求文件需要认证时,用户一栏才有数据。时间记录了服务器响应用户请求发出文件的时间。请求域记录了用户请求的方法(Method)、URL 和使用的协议。其中方法(Method)一栏一般有 GET、POST 或者 HEAD 三种。GET 指从 WEB 服务器请求了一个对象;POST 表示向服务器发送信息;HEAD 指只取一个对象的头;URL 记录了本地文件系统的一个页面或者响应请求的可执行文件。状态是由服务器记录的,以表示对一个请求的响应情况。200－299 一般表示成功,300－399 表示页面重新定向,400－499 表示处理一个请求时失败了,500－599 表示 WEB 服务器有问题。其中最常见的错误是 404,它表示请求的文件没有找到。BytesRecvd 记录用户发出请求向服务器端送出了多少字节,(BytesSent)比特大小页面记录了处理请求服务器送出了多少字节。参引页面记载了发出请求的 URL。当然,用户输入地址或者通过书签进行的访问其参引页面值为空。最后,代理字段记录用户使用的操作系统和浏览器软件类型。

用户在访问网站过程中留下的足迹也不是随机的,用户需求不同访问请求的页面也不同。网站设计人员在设计网站时,通常是按照某种主题分类分层组织网站结构的,有大量重复的导航页面。因此,网站链接结构可以表示为一个图。用户要利用导航页面提供的访问链接在自己感兴趣的页面之间跳转,因为这些 URL 很少有人知道、被记住,所以几乎没有人愿意自己输入网站深层的文件 URL 地址来访问页面。所以,用户访问网站过程中都要访问一些共同的导航页面。

表 1　某网站日志片段

IP	ServerIP	Time	Request	Status	Bytes Recved	Bytes Sernd	Process time	Method	Refer
137. 132. 124. 30	202. 117. 29. 25	Jun 19 1999 09:08AM	/default. asp	200	253	312	2356	GET	null
129. 125. 23. 152	202. 117. 29. 24	Jun 19 1999 09:09AM	e－intro- duction. htm	200	120531	373	1E＋05	GET	http://www. xjtu. edu. cn/u- nit/graduate/e－index. htm
202. 102. 144. 211	202. 117. 29. 24	Jun 19 1999 09:11AM	/robots. txt	404	63	202	623	GET	null
137. 132. 5. 50	202. 117. 29. 24	Jun 19 1999 09:42AM	/default. asp	200	312	429	4040	GET	null

4　用户访问 Web 的几种习惯

互联网上用户可以根据自己的兴趣爱好选择访问网站和网站上的页面,研究表明,多数用户访问具有以下几个特点:a. 每次访问只有一个主题,并且都是从辅助页面开始浏览,最终目的是为了浏览一个内容页面;b. 用户花在一个页面上的时间与该页面对用户是辅助页面还是内容页有关;c. 访问过程中只有在改变访问主题时,才会访问前面访问过的页面以跳转到另外的页面;d. 用户一次访问的时间都不会超过一个最大的限制－时间窗口。

5　确定用户访问服务器会话期间

前面讲了用户访问日志中,从用户角度来看有用户会话期间,从特定服务器角度来看有服务器会话期间。由于同一用户访问可能通过代理服务器,分布在浏览器端、代理服务器和 Web 服务器端,几乎无法确定。在目前技术水平上,确定用户访问服务器会话期间比较可行。服务器会话期间是每一用户在一段时间内访问服务器、按时间顺序请求的页面集合,而服务器要并发处理多个用户的请求,要从多个相互交织的用户访问会话期间中正确区分出所有用户的访问事务也存在一定的困难。确定用户访问会话期间三个重要的因子需要确定:a. 确定用户。前面讲了不同数据源数据记载的用户访问过程不同,而且用户访问和服务器资源不是一对一的关系。比如,服务器端日志可能记载了一个用户可以在多个客户端提交请求,多个用户也可以在一个客户端提交请求。由于缓存、防火墙和代理服务器等的存在,准确确定出每个用户很困难,研究人员设计的启发式推断用户的方法,除非通过在客户端跟踪用户的行踪得到第一手的访问资料,很难准确确定用户。b. 确定用户访问时间。用户请求页面和浏览页面的准确时间很难确定,如图 1 中记录了用户请求的三个页面。可以看出服务器记载用户浏览页面时间有较大的偏差,页面读取时间也有偏差,根据服务器端记载的用户浏览页面时间显然要比客户端实际的浏览时间长,比如在图 1 中,按照服务器记载时间计算页面 A 时间是 $T_6－T_1$,但是实际 A 页面的浏览时间是 $T_5－T_4$。实际浏览 B 页面的时间是 $T_{10}－T_9$,但是服务器记载的浏览时间是 $T_{14}－T_6$。受客户端连接处理速度、页面大小和网络拥挤程度的影响,服务器记载的用户浏览页面时间误差大小甚至可以达到几分钟,因此,几乎无法准确确

定用户访问时间。c.确定用户访问的页面。为了提高浏览的效率,一般的浏览器(IE、Netscape 等)都提供了三种基本的方法访问缓存页面:通过使用浏览器的后退键;通过点击当前访问事务已经访问过的页面链接,这个页面在不在历史堆栈中取决于前面已经访问过的路径;直接从浏览器的历史记录中访问。图 1 中第二次请求页面 A 就是从客户端的缓存中读取的。由于客户端缓存的存在,有些用户浏览请求页面信息并没有在服务器端日志中记载,因此服务器记载的日志不能完全反映用户请求的页面情况。确定用户访问 Sesion 时要对用户访问路径进行完善,推断出用户从访问缓存中访问的页面信息。识别路径不完善的方法是直接根据当前页面的参引页面是否是访问事务中当前页面的前一个页面。如果没有出现说明路径不完善,需要把参引页面插入到访问事务中当前页面之前;否则,就说明这一条路径是完善的。以图 1 中用户请求页面为例,在日志数据库中记录的页面和参引页面为<(A,null),(B,A),(C,B)>,对应的用户访问事务就是<A,B,C>,由于日志记载的 C 页面的参引页面 B 不是可以链接到达 C 的前一个页面,推断用户是从缓存中访问了 A,因此,把 A 页面加入到 C 页面之前,从而完善了用户访问路径,这时访问事务是<A,B,A,C>。至于 A 页面的访问时间以 C 页面访问时间减去常数 t 得到。

在服务器会话期间的基础上就数据可以进一步根据算法要求再将服务器会话期间处理成数据挖掘算法所需要的用户访问事务形式。

6　研究难点与方向

前面介绍了 Web 用户访问的基本过程、用户访问的 footprint——访问日志分布情况,客户端、代理服务器端缓存的存在,使用户访问日志分别存在于服务器、代理服务器和客户端,又因为互联网传输协议 http 的无状态性,从 Web 用户访问日志中探究用户访问规律最大的难点在于如何把分布于不同位置的访问日志经过预处理,形成一个个用户一次的访问会话期间。通常来讲,对于静态 Web 网站,服务器端的日志容易取得,客户端和代理服务器用户访问日志不容易取得。在只有服务器端的访问日志可得的情况下,就需要对用户访问情况做适当的假设。假设用户在一段时间里访问的页面有关联关系;假设用户访问在短时间内访问 Web 的目的不变,访问内容的主题一样;假设用户与用户访问网站的主机 IP 地址有某种对应关系。其次,由于一个完整的 Web 页面往往是由一个个图片和框架页面组成的,而用户访问服务器也有并发性,在确定用户访问内容时,必须从服务器日志中甄选出某个用户实际请求的页面和页面的主要内容。另外,由于目前已经有的数据挖掘算法主要是在大量交易数据基础上发展起来的,在处理海量 Web 用户访问日志中也需要重新设计算法处理结构。

由此可见,用户访问日志研究的难点集中在三个方面:如何综合利用客户端、代理服务器端和服务器端的日志,对用户行为做适当的假设,从而确定用户个体;如何对日志流进行预处理得到每个用户的访问内容;如何设计有效的算法处理这些日志,挖掘出用户访问规律,指导企业制定营销战略。

参 考 文 献

1　World Wide Web Committee Web Usage Characterization Activity. http://www.w3.org/WCA/, 2000

2　J. Pitkow and L. Catledge. Characterizing Browsing Behaviors on the World Wide Web. Computer Networks and ISDN Systems, 27(6), 1995

3　J. Pitkow. Summary of www Characterizations. In 7th International World Wide Web Conference, 1998

4　Pazzani M., & Billsus, D.. Learning and Revising User Profiles: The Identification of Interesting Web Sites, Machine Learning 27, 313－331, 1997. Cyrus Shahabi, Amir Zarkesh, Jafar Adibi, Vishal Shah, Knowledge Discovery from Users Web－Page Navigation, In Proceedings of the IEEE RIDE97 Workshop, April 1997

5　Bamshad Mobasher, R. Cooley and J. Srivastava. Grouping Web Page References into Transactions for Mining World Wide Web Browsing Patterns. Proceedings of the 1997 IEEE Knowledge and Data Engineering Exchange Workshop (KDEX－97), November 1997

(责编:京阳)

(上接第 47 页)积极参与。可以看到,开发者和使用者由于相互之间的专业背景不同,所用的名词语汇以及表达习惯不同,造成了沟通上的困难,但这只是浅层的原因。深层的原因是对于企业信息化的认识不一致,以及各自的目标不一致,相当一部分源于各自的价值观问题。所以,必须加强人员信息行为的管理,通过人们的一致行动,以改进一个组织的信息环境,营造新型的企业信息文化。

基于以上的讨论,在信息技术、组织管理、信息资源三方面的基础上,本文大胆地提出企业信息化工程的"四维"体系框架,可以说这是一种全面分析企业信息化的新思路(见图 3)。

企业信息化的"四维"体系
- 人员行为
- 信息资源
- 组织管理
- 信息技术

图 3　"四维"企业信息化工程体系

通过在这四个方面的具体努力和积极推进,以期进一步提高企业信息化的整体水平。

4　结　语

技术是企业信息化的条件,管理是企业信息化的基础,信息是企业信息化的核心,人员是企业信息化的关键。根据我国企业信息化建设的进程,本文分析了企业信息化的模型,从而形成了企业信息化工程的体系。这些对我国企业信息化建设具有参考价值和指导意义。

参 考 文 献

1　Currie, W., Galliers, B. Rethinking Management Information Systems. New York: Oxford University Press Inc., 1999

3　王众托.企业信息化与管理变革.北京:中国人民大学出版社,2001

2　仲秋雁,刘友德.管理信息系统.大连:大连理工大学出版社,1998

4　杜　栋.信息管理学教程.北京:清华大学出版社,2002

(责编:加勃)