●易　明　邓卫华

点击流信息资源研究

摘　要　点击流信息资源,这里特指通过间接方式获取的反映站点用户点击活动的各种网络信息资源。点击流信息源主要有站点服务器、代理服务器和客户机。开发方法有:站点数据预处理、站点点击流数据挖掘。参考文献 8。

关键词　网络　信息资源　点击流　网络站点

分类号　G250.73

ABSTRACT　"Click streams" are various network information resources indirectly obtained and reflecting user's clicking activities. Information sources of "click streams" include website servers, agents and clients. In this paper, the authors also discusses some methods for the exploitation of such information resources. 8 refs.

KEY WORDS　Network.　Information resources.　Click stream.　Website.

CLASS NUMBER　G250.73

与传统的基于内联网络的信息系统相比,基于因特网站点信息系统的一个重要特点就是用户的高度自治性。站点只有开发与利用反映用户行为的点击流信息资源,进而与站点用户建立良好关系,才能使得站点提供的服务或产品更有针对性。虽然站点前台工具产生的操作失败通常会导致此次站点与用户的交互迅速结束,但是利用点击流信息资源来管理站点与其用户的关系将为站点带来持久的、系统性的竞争优势。

1　点击流信息资源的概念

目前,对信息资源的理解有广义与狭义之分。广义的信息资源是人类社会信息活动中积累起来的信息、信息生产者、信息技术等信息活动要素的集合,狭义的信息资源是人类社会经济活动中经过加工处理有序化大量积累起来的有用信息的集合[1]。本文所探讨的点击流信息资源是一种狭义的信息资源。

1.1　点击流信息资源的含义

"点击流"也有广义与狭义之分。前者是用户在因特网上的一系列点击活动,后者则是用户访问某一站点的一系列点击活动。本文研究的是一种狭义的点击流,特指站点点击流。伴随着用户点击活动的产生,与之相关的点击流信息也随之产生。站点的管理者往往不能现场观察用户的各种点击活动,无法直接获取站点用户的各种点击流信息。虽如此,但是可以通过间接方式来获取用户点击流的相关信息,即通过配置站点服务器、代理服务器和客户机来产生反映用户点击活动的记录。本文所探讨的点击流信息资源就是特指这些通过间接方式获取的反映站点用户点击活动的各种网络信息资源。

1.2　点击流信息资源的特征

点击流信息资源的本质是信息,具有信息的一般属性;它作为一种信息资源,一种网络信息资源,又具有信息资源或网络信息资源的一般属性。但作为一种特殊的网络信息资源,它也有其独有特征。

点击流信息是通过配置站点服务器、代理服务器和客户机而产生反映用户点击活动的记录信息,因此这种信息是对站点用户点击活动的客观反映,并没有添加任何主观的因素。而且它是一种间接反映用户点击活动的信息资源。

正是因为点击流信息资源是一种间接反映用户点击活动的信息,所以不能保证这些信息能够全面、准确地反映用户所有的点击活动。对点击流信息资源的收集、开发与利用要有主观思路,要运用已有的知识进行分析和判断。

点击流信息是对站点用户点击活动的反映,是一种属于个人隐私范畴的信息,不能随意公开与共享;对站点管理者而言,这些点击流信息往往还涉及一些商业秘密,同样也不能对外公开与共享。

点击流信息资源重在反映站点用户历史的点击活动,无论是在传统数据库中还是在数据仓库中,系统都积累了大量的历史点击活动信息,记录了从过去某一时间到目前各个阶段的点击流信息。通过开发这些历史点击活动信息,就可以实现点击流信息资源的有效利用,如站点决策支持、站点系统优化、站点结构优化等等。当然,有时也需要利用用户当前的点击活动信息,如站点个性化以及站点系统优化中的部分功能。

2　点击流信息源的类型

虽然因特网提供了到目标服务器的途径,但是用户和因特网之间还需要架起一座桥梁,这座桥的实体通常就是网络服务提供商(ISP)。ISP 拥有大批代理服务器以及连接常用的媒介,用户可以使用电话线拨号、DSL 或无线协议来获取 ISP 提供的服务,ISP 将用户请求发送给相应的 Web 服务器。点击流信息源主要有站点服务器、代理服务器和客户机。

2.1　站点服务器

在站点服务器端,点击流信息源主要包括 Web 服务器日志文件、内容服务器日志文件和网络监视器日志文件。Web 服务器日志文件是获取点击流信息的核心信息源,它记录了用户访问站点的数据,每当站点上的网页被访问一

次,Web服务器就在日志文件中增加一条相应的记录。这些记录数据反映了多个用户(可能同时)对单一Web站点的访问行为。

Web服务器是距离用户最远的点击流信息源,通过这种信息源所获取的点击流信息并不完全可靠。比如,由于客户各种页面请求的发送需要时间,而且在Web环境中存在多级别的缓存(如用户本地缓存、代理服务器缓存),甚至站点服务器为了提高系统的性能也会启用缓存功能,这样客户端用户的点击活动在数量与时间上并没有被Web服务器准确跟踪。另外,如果用户请求是通过POST方法传送的,那么参数在Web服务器日志中就不可见,而且防火墙(代理服务器)将使得不同的用户请求在Web服务器的日志中记录的都是防火墙(代理服务器)的IP地址。Web服务器日志是从URL的角度,而不是从用户访问目的的角度描述用户的行为。要得到用户点击行为的全面视图,必须整合应用服务器的点击流数据。另外,还可以利用应用服务器上的应用程序(如CGI)来记录用户的个人信息和以自定义的格式动态记录用户的浏览信息(需用户许可)。该收集方法在用户确定方面的准确性较服务器级高,但大量的应用程序会降低系统的效率。

网络监视器是一个可替换Web服务器的点击流信息源,一般直接放在Web服务器外,监视用户向Web服务器的请求。网络监视器直接从TCP/IP包抽取功能数据进行分析,可以检测HTTP头之外的信息,可扩展性比较好。它还可以直接获取用户通过POST方式来传送的参数,弥补了Web服务器的不足。

2.2 代理服务器

ISP提供的计算机就是所谓的代理服务器。代理服务器端日志记录了多个用户访问多个站点的点击活动,此时只收集与特定站点服务器相关的点击流。代理服务器相当于一个在客户端浏览器和Web服务器之间提供了缓存功能的中介服务器,它使用户和因特网间接相连,主要用于减少用户下载网页的时间和服务器与客户机之间的网络流量。当然,如果代理访问站点页面是通过Web应用程序动态生成的,对于用户的每次请求,代理服务器需要从Web服务器取得数据。同样,这种信息源也不能准确地识别浏览用户,对访问页面的采集不够全面,采集时间不准确。

2.3 客户机

从理论上讲,客户机是最为重要的点击流信息源,因为客户端点击流信息的收集是建立在用户行为源上,可以准确反映用户点击活动。但是,这种收集方法需要得到用户的许可。客户端的点击流信息收集需要用到远程代理(如JavaScript或Java Applets)、Plug-In、网页跟踪帧或者修改已有浏览器(如IE,NetScape,Mosaic,Mozilla)的源程序代码来增强浏览器软件的信息收集能力。

3 点击流信息资源的开发方法

3.1 站点数据预处理

3.1.1 站点结构数据预处理

站点结构是由所有页面之间的超文本链接形成的,也就是各个页面之间的链接结构。它可以用有向图G=(V,E)来表示,集合$V=\{v_1,v_2,\cdots\cdots,v_n\}$表示站点所有页面,也就是有限的非空顶点(vertex)集,集合$E=\{e_1,e_2,\cdots\cdots,e_m\}$表示站点页面之间的链接,也就是用非空顶点对表示的边(edge)集。

针对本文站点点击流数据预处理的需要,站点结构数据预处理还需要分析站点页面类型。根据页面不同的功能,站点页面可以分为站点首页、内容页面、导航页面、搜索页面和数据输入页面[2]。站点首页是站点文件结构的根,包含了大量的链接到站点其他页面的链接。内容页面是反映站点所提供的主要信息的页面,包含了大量的图片和文本,也就是用户点击活动的目标页面,反映在站点结构中就是叶子节点。导航页面的功能是引导用户访问内容页面(目标页面),这些页面中只包含少量的图片和文本。搜索页面有大量的链接到其他页面的链接,其功能类似于导航页面。数据输入页面主要是用来搜集用户数据,并和用户交互。有时,这些页面的功能可以集中在一个页面上。站点页面的分类可以由站点管理者手工来设置,或者是通过监督的学习技术自动完成,如决策树学习算法C4.5。当然,也可以利用XML将分类标签加到每个页面。

3.1.2 站点内容数据预处理

站点内容数据预处理包括将文本、图片、脚本和其他多媒体文件转变为对站点点击流信息资源开发与利用有用的格式。而且,站点内容数据可以用来过滤站点点击流数据挖掘输入和输出的结果。比如,基于站点内容分类算法的结果可以用来控制发现的模式仅限于一个特点的主题。当然,在利用数据挖掘方法之前,站点内容数据预处理需要产生一系列的能够表示站点内容特征的特征词条。集合$F=\{f_1,f_2,\cdots,f_Q\}$表示站点内容数据预处理后产生的特征词条集合,Q为站点特征词条总数。集合$P=\{p_1,p_2,\cdots,p_n\}$表示站点所有页面,每一个页面都由相应的URL唯一标识。其中,$pi=\langle fw_1(f_1,p_i),fw_1(f_2,p_i),\cdots fw_1(f_j,p_i),\cdots,fw_1(f_Q,pi)\rangle$,$i\in\{1,2,\cdots,n\}$,$j\in\{1,2,\cdots,Q\}$,其中$fw_1(f_j,p_i)$就是第j个特征词条在页面$p_i\in P$中的权重。站点页面集合P可以看做是一个近似封闭的集合,特征词条权重的计算可以直接采用TF-IDF方法,计算公式为[3]:

$$fw_1(f_j,p_i)=tf_j\times\log(N/n_j)$$

其中,tf_j是第j个特征词条在页面pi中的出现频率,它描述了第j个特征词条在页面p_i中的重要程度,N是站点页面总数,n_j是第j个特征词条在所有页面中是否出现的布尔值的累加和,$\log(N/n_j)$描述了第j个特征词条在页面集合P中的重要程度。

3.1.3 站点点击流数据预处理

其步骤包括:数据过滤、用户识别、会话识别和路径补充。整个过程需要结合站点结构数据和站点内容数据。

数据过滤是指根据站点数据挖掘的需要,对站点点击流数据信息进行处理,包括删除无关紧要的数据,合并某些记录,对用户页面请求是否发生错误进行处理。当用户请求一个网页时,与这个页面有关的图片、音频、视频文件都会自动下载,并会出现在点击流数据信息中。如果挖掘的是用户访问模式,这些文件就没有什么作用了,可以删掉这些数据。但是,当挖掘目的是为了进行网络流量分析或为页面缓冲和

预取提供依据时，这些信息又会显得格外重要。又如，由于网络机器人对站点的访问行为与一般用户的访问行为是不一样的，所以通常将网络机器人的这些请求过滤掉。许多网络机器人的代理值与通常的浏览器不一样，可通过检查日志代理清除这些记录，或者是通过对网站的定时重复请求来标注出网络机器人，有时还需要利用启发式算法。

用户识别是站点点击流数据预处理最关键的一步，其重要任务就是将那些属于同一用户的点击流数据归类。由于大多数用户都是匿名访问站点，所以在用户识别的过程中，面临着很多问题[4]。比如：ISP 往往有很多代理服务器用于提供网络接入服务，这样同一个代理服务器下将有多个用户访问同一站点，甚至是在同一时间；一些 ISP 或者是私人工具会随机给每个用户 Request 分配不同的 IP 地址。在这种情况下，同一次用户点击流将拥有多个 IP 地址；一个用户可以用多台机器多个 IP 地址来访问同一站点，这样就无法识别同一用户的回头率；一个用户同时打开多个浏览器而且同时访问同一站点的不同内容；多个用户甚至可以用同一台机器来访问站点。

对于第一种情形，首先可以通过 IP 地址、用户代理的方法来识别，这样同一 IP 地址而用户代理相同的服务器访问可以分开。如果使用的是同一 IP 地址，而且用户代理也相同的话，则需要采用启发式算法。对于第二种情形，多个 IP 地址之间的前面 3 位应该是相同的，可以考虑先将那些 IP 前面部分相同的记录集合在一起，再通过启发式算法进一步区分。其余三种情形出现得比较少，一般没有深入考虑。

用户会话是指用户对站点服务器的一次有效访问。站点点击流数据中不同用户访问的页面当然属于不同的会话，当某个用户的页面请求在时间跨度比较大时，就有可能是该用户多次访问同一个网站，此时可以将用户的访问记录分成多个会话来处理。最简单的方法就是设置一个 timeout 值，如果用户访问页面的时间差超过了这个值，则认为用户开始了一个新的会话。许多商业产品都采用 30 分钟作为缺省的 timeout 值，但是 L. Catledge 和 J. Pitkow 通过实验认为，把 timeout 值设为 25.5 分钟更好一点[5]。

站点点击流数据预处理的最后一步就是推测缓存引用页面。浏览器主要提供了 3 种方法来支持用户对缓存页面的访问。最常用的方法（基于 GVU 调查的结果[6]）利用"back"按钮。第 2 种方法是通过链接那些已经浏览访问过的页面。第 3 种方法就是直接从历史记录中浏览。因为无法识别用户是采用哪种方法，所以在路径补充环节就默认用户是采用第 1 种方法。值得强调的是，此环节的完成需要借助启发式算法。同时，路径补充除了需要补充由于缓存而导致的页面丢失，还需要注意补充由于 POST 方法而导致无法被 Web 服务器日志记录的参数，此时需要借用网络探测器。

站点点击流数据预处理最终会产生 k 个交易事务形成的交易事务集 $T=\{t_1,t_2,\cdots,t_k\}$ 和 n 个属于交易事务的页面记录集 $P=\{p_1,p_2,\cdots,p_n\}$，每一个页面记录都由相应的 URL 唯一标识。

3.2 站点点击流数据挖掘

就分析和建立模型的技术和算法而言，站点点击流数据挖掘和传统的数据挖掘差别并不是特别大，很多方法和分析思想都可以运用，所不同的是站点的数据格式和传统的数据库格式有一定的区别。这一点在数据预处理阶段就表现出来了。点击流信息资源的开发除了主要采用数据挖掘方法，还可以采用一般的统计方法和在线分析处理方法。

3.2.1 聚类分析

在点击流数据挖掘中常用的聚类策略有两种：交易事务聚类和站点页面聚类。如果是交易事务聚类，则需要赋予每个交易事务中各个页面一定的权重。这个权重可以通过各种方法进行计算，如可以用二进制来代表在交易事务中这个页面有或无，或者用用户在每个页面的停留时间来表示，或者用这个页面所在的特定领域的重要性来衡量，当然也以采用停留时间与页面长度的比值来表示。如果是站点页面聚类，则需要首先将站点页面表示成站点特征词条的形式，然后依据各个特征词条在各个页面中的权重再来聚类。常用的聚类算法有 k-means 方法、ARHP 方法等等。

3.2.2 关联规则分析

关联规则分析主要是针对交易事务而言的，它既能捕获在交易事务中同时出现的页面之间的关联度（不考虑页面出现的顺序），也能分析同时出现在交易事务中的特征词条之间的关联度。目前，主要采用的是 Apriori 算法，这种算法可以找到经常出现在交易事务中的频繁页面集和频繁特征词条集。

3.2.3 序列模式分析

序列模式就是那些在很多交易事务中经常出现的序列。序列模式分析能够捕获经常被用户访问的页面轨迹。一般的序列 $<s_1,s_2,\cdots,s_l>$ 必须满足以下条件：对于交易事务 $t=<p_1,p_2,\cdots,p_r>(l\leqslant r)$ 中，存在 1 个正整数 $1\leqslant a_1<a_2<\cdots<a_l\leqslant r$，对于所有 i 值都有 $s_i=P_{ai}(i\in\{1,2,\cdots,1\})$。如果存在整数 $0\leqslant b\leqslant r-l$，对于所有的 i 值都有 $CS_i=p_{b+i}$，则序列 $<cs_1,cs_2,\cdots,cs_n>$ 为连续序列。在连续序列模式中，每一个邻近的元素 s_i 和 s_{i+1}，都必须连续出现在交易事务 t 中，而一般的序列模式可以是非连续地出现在交易事务中。连续序列模式反映的是用户频繁导航路径，而序列模式代表了更为一般的导航模式。关联规则分析中的 Apriori 算法也能够用于发现序列模式和连续序列模式，马尔柯夫模型也是一种重要的发现连续序列模式方法[7]。

4 点击流信息资源的利用策略

4.1 站点系统优化

站点系统优化可以采用统计学的方法，对点击流数据信息进行多种分析和统计，包括频繁访问页面、单位时间访问频度、访问量的时间分布等，从而帮助站点指定页面缓存、网络传输或者是数据分布。还可以通过对点击流数据信息进行序列模式分析，建立时间和空间位置的预测模型，以此决定预先提取和缓存策略以减少用于产生页面的延迟。甚至还可以利用点击流数据挖掘得出网络侵入、欺骗以及试图闯入等影响网络安全的行为模式。

4.2 站点决策支持

通过对用户点击活动与站点提供的服务或产品之间关系的挖掘，更好理解用户的意图，发现用户需求特征与趋势，识别潜在用户，以此进行相关站点决策。（转第 92 页）

参考文献

1,3 崔雷,郑华川. 关于从MEDLINE数据库中进行知识抽取和挖掘的研究进展. 情报学报,2003 (4)

2,6 ,12 Qin He. Knowledge Discovery Through Co-Word Analysis. Library Trends, 48, 1999(1)

4 Ying Ding et al.. Bibliography of information retrieval research by using co-word analysis. Information Processing and Management, 2000(4)

5,7 ,11 Callon et al.. Co-Word Analysis For Basic And Technological Research. Scientmetrics, 1991 (22)

8 Coulter Neal et al.. An Evolutionary Perspective of Software Engineering Research Through Co-Word Analysis. Technical ReportCMU/SEI-95-TR-019ESC-TR-95-019

9 Law et al.. Policy and the Mapping of scientific change: A Co-Word Analysis of Research into Environment acidification. Scientometrics, 14(3-4), 1988

10 张晗,崔雷等. 生物信息学的共词分析研究. 情报学报,2003 (5)

13 Kostoff Ronald N. et al.. Database Tomography for Information Retrieval. Journal of Information Science, 23 (4)1997

冯 璐 中国科学院文献情报中心,中国科学院研究生院情报学专业2004级博士研究生。通信地址:北京市海淀区中关村北四环西路33号。邮编100080。

冷伏海 教授,中国科学院文献情报中心博士生导师、情报研究部副主任。通信地址同上。

(来稿时间:2005-04-13)

(接第75页)对于电子商务站点而言,有关客户购买行为的数据是非常重要的,这些数据可以用来识别客户关系生命周期的特定阶段,进而帮助站点制定相应的客户策略。

4.3 站点结构优化

目前,站点内容的组织方式都是从站点的角度来安排的。它往往与站点用户所期望的组织方式有所差异。站点结构优化的过程也就是消除差异的过程。这些差异可以通过分析点击流数据信息得到。比如,如果站点用户在所期望的位置找不到目标页面,往往会点击"后退"按钮或者直接点击新链接继续寻找。如果点击"后退"按钮,用户的"后退"点击流就为优化站点结构提供了一种思路,即用户点击"后退"按钮的页面所在位置就是用户针对这个目标页面的期望位置,此时站点设计可以考虑调整站点结构或者是在期望位置添加指向目标页面的链接。另外关联规则分析和序列模式分析会发现站点各个频道之间或者是站点一般页面之间的关联度,这样可以调整站点频道在页面中的位置,或者是在关联度较高的一般页面之间增加链接。

4.4 站点个性化

所谓站点个性化实质上就是为站点用户提供个性化的站点访问体验。由于传统的手工决策规则系统方法、基于内容的过滤代理系统方法、协作过滤系统方法的种种不足,点击流数据挖掘已经成为站点个性化主流方法[8]。比如可以采用聚类分析方法,在数据预处理的基础上实现基于站点使用和站点内容的交易事务聚类,然后导出站点的使用文档和内容文档,在此基础上结合当前用户会话形成基于站点使用和站点内容的个性化推荐集,最后在整合两种推荐集的基础上完成个性化推荐。关联规则分析方法和序列模式分析方法同样也可以用来完成站点的个性化推荐,其基本原理和聚类方法是一样的。

参考文献

1 马费成,李纲,查先进. 信息资源管理. 武汉:武汉大学出版社,2000

2 Peter Pirolli, Jame Pitkow, Ramana Rao. Silk from a Sow's Ear: Extracting Usable Structures from the Web. Proc. In ACM Conf. Human Factors in Computing Systems, 1996

3 D. D. Lewis, et al. Training algorithms for linear text classifiers. In Proceedings of the Nineteenth International ACM SIGIR Conference on Research and Development in Information Retrieval, 1996

4 R. Cooley, B. Mobasher, and J. Srivastava. Data preparation for mining World Wide Web browsing patterns. Journal of Knowledge and information Systems, 1999(1)

5,6 L. Catledge, J. Pitkow. Characterizing browsing behaviors on the World Wide Web. Computer Networks and ISDN Systems, 1995(6)

7 Jianhan, ZhuJun, Hong John G. Hughes. Using Markov models for web site link prediction. In Proceedings of the thirteenth ACM conference on Hypertext and hypermedia, 2002.

8 J. Srivastava et al. Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data. SIGKDD Explorations, 2000(2)

易 明 华中师范大学信息管理系讲师。通信地址:武汉。邮编430079。

邓卫华 华中农业大学经济管理学院讲师。通信地址:武汉。邮编430070。

(来稿时间:2005-07-07)