第10卷 第4期 2010年2月
1671－1815(2010)4-0928-05

科 学 技 术 与 工 程
Science Technology and Engineering

Vol. 10 No. 4 Feb. 2010

网络广告中反 CPC 点击作弊研究

常 艳 汤小春

(西北工业大学 计算机学院,西安 710072)

摘 要 随着互联网日新月异的发展与网民数量的骤增,网络已经成为了一个更加适合传播广告并获得良好展示效果的平台。各种网络广告交易平台也如雨后春笋般出现,每次点击付费(CPC)的网络广告模式也逐渐被大家接受。可是随之而来的问题是点击作弊的预防与检测迫在眉捷,因为这将直接关系到这种广告模式能否长久。分析了 CPC 点击作弊的常见手段及特征,以及已有的反作弊手段及措施,并对通过监测 IP 来防止作弊的传统的防作弊手段,提出了一些改进措施,来有效的区分出多人共用 IP,爬虫程序 IP 以及机器人作弊程序 IP;提出了一些新的防作弊措施,如通过分析点击时间来防止作弊,以及通过自适应阈值控制来防止作弊等。

关键词 网络广告 每点击成本(Cost Per Click,CPC) 点击作弊

中图法分类号 TP393.07; **文献标志码** A

1994 年 10 月,美国《Wired》杂志网络版首次出现了 AT&T 公司等 14 家客户的旗标广告,开创了网络广告的先河[1]。随着互联网的发展,网络广告的优势不断凸现,广告效益愈发显现,逐渐成为了传媒行业的新宠。

网络广告的最大优点之一,就在于广告访问量的可统计性,点击率是衡量网络广告效果的合适标准和主要依据之一。CPC 是一种点击付费广告,根据广告被点击的次数收费,目前很多网络广告交易平台都采取这种付费方式。

随着 CPC 网络广告的流行,点击作弊的现象也越来越严重,这不仅仅是损害了广告主的利益,从长远看,如果广告主在投放网络广告后没有获得所期望的回报,那么,广告主会毫不留情地放弃网络广告的投放,最终受到伤害的还是网络广告本身,其最直接的后果就是大大降低了互联网广告的价值。

2009 年 11 月 9 日收到

第一作者简介:常 艳,E-mail:changyan_nwpu@126.com。

1 点击作弊的常见手段及特征

1.1 代理服务器

通过使用代理服务器更快的、随机的更换自己的 IP 地址等信息,使得作弊点击不易被发现。

1.2 鼓励点击

鼓励点击的方式有很多,比如广告无限靠近下载链接,或者使用诱惑性的图片误导用户;有更高级一些的,比如点击广告成为注册帐户或下载软件的条件、点击广告可以增加积分等等。

1.3 影响展示或匹配

通过关键字堆彻,增加虚假关键词,设置隐形页面,隐形文本来干扰广告的匹配。加入“大量链接机制”和别的网站交换流量,或大量的访问广告发布商的网站,但不做任何点击,这都会导致网站流量增加,影响广告的点显比(点击次数/显示次数)[4]。

1.4 雇佣人力点击广告

作弊者开设“点击工厂”,雇佣苦力进行大量点击,苦力的工资一月不超过 1 千,一分钟可贡献有效点击 5 次左右(注意规则的前提下,如更换 IP)。

1.5　作弊联盟

一些开设相同或类似广告业务的站长结成联盟,互相点击其它成员的广告,以提高作弊点击的隐蔽性,一般以 10 人左右为一个小组。

1.6　机器,程序模拟点击

使用木马和肉鸡来点击广告,通过控制点击的客户端分布和时间来实现点击正常化,让收入报告和正常情况差不多。这是比较难发现的作弊手段之一[2]。

1.7　无意的作弊点击

搜索引擎的爬虫或链接搜索程序(以下简称爬虫程序)造成的无意的广告点击[3]。

2　点击作弊的通常迹象

2.1　关键词的不正常表现

如果一个平时表现并不怎么样的关键词的赞助广告,一下子变成点击很大的关键词,就可以怀疑是点击作弊。

2.2　大量的同一 IP 点击量(或访问量)

虽然这是最简单的作弊方法,但仍有大部分的作弊案例在使用,这是最普遍的一种迹象。

2.3　突然的投资转化率下降

广告主的投资转化率在一个时间突然下降,很有可能是对手在恶意点击他们的广告。

2.4　大量访问者快速离站

雇佣苦力来点击作弊的作弊者,会有同一个问题,广告网站刚被点开,没有点击过任何链接就迅速被关闭了。

2.5　突然的点显比下降(或上升)

自己对广告的大量点击或者对手的恶意攻击都有可能导致点显比上升。而对手大量访问网站不点击广告会造成点显比突然下降。

2.6　大量的与销售区域无关的点击来源

频繁更换代理服务器来达到避免重复 IP 的手法,很可能会导致另一个问题,即产生大量与销售区域无关的点击来源。

3　常见的防作弊手段及措施

3.1　IP 防止作弊

一般计费方式是按照 24 h 内唯一 IP,将每个 IP 记入数据库,当下一个访问 IP 与数据库已存 在的 IP 相同时,则不计费。现在上网一般是动态 IP,作弊者通过拨号器上下线来实现改变 IP 地址,可以通过 C 段 IP 来辨别,如大量出现 218.175.11.x 这种相同 C 段的 IP 号,则可能作弊。

3.2　COOKIES 防止作弊

当用户点击广告时,用 COOKIES 记录其相关参数,利用 COOKIES 可以判断同一个用户是否重复点击。这种方式的缺陷是很容易改变物理信息进行作弊,比如通过 INTERNET 选项清空 COOKIES。

3.3　点击率防止作弊

点击率 = 点击次数/共浏览的次数。可以设置点击率上限,如果某一 Adlink(广告链接)的点击率超过了这个上限,就可以认为该 Adlink 的点击存在作弊。平均点击率已从 1999 年的 5% 下降到了 1% 以内,当然还需要考虑广告面向对象与页面的访问者的交叉率,越高则表示该页面与广告的关联度越大,点击率越高。目前 富媒体广告的点击率在 2% 至 5%,普通图片点击在 0.1% 至 1%,与图片的创意有关,可以设置当点击率超过一定的百分率提示可能作弊行为(易特联盟设置在 8%)。

3.4　时间顺差防止作弊

当你打开一个有广告的页面时,一般情况下不可能瞬间点击广告,因为每个广告都是由印象转变为关注,再转变成点击行动的,所以可以设置当访问者打开网站页面几秒内点击广告为作弊行为(易特联盟设置为 3 s 内提供参考)。当你打开一个广告,在几秒内立即关闭(易特联盟设置为 2 s 提供参考),也可以判断为无效点击。

3.5　来源统计防止作弊

记录广告页面的来路,每个站点的搜索引擎来路总是占据很大的一个比例,如果该页面没有来路,可以判断这个页面的流量非连接流量,可以通过两

种途径获得,一、浏览器直接访问或者收藏夹访问;二、弹窗流量没有来路统计,也有可能是目前流行的流氓插件弹窗。该种方式也可以查询到有些站长将广告代码放置在 IFRAME 里的最终页面。

3.6 加强页面内容的智能判断

对于关键字堆彻,增加虚假关键词和设置隐性页面的作弊行为,应对放置广告的网页加强智能判断。例如:从页面的 < body > </body > 部分提取,而非 < title > </title > 或者 < meta > </meta > 部分;判断关键字是否仅在某一段落内重复,来去除恶意重复现象。对于重定向这种行为,应对页面内增加重定向分析。使用刷新标记进行重定向可以在页面内容中进行标记判断;使用 JavaScript 来进行重定向可以对页面中的 JavaScript 代码进行判断[4]。

3.7 鼠标值防止作弊

显示屏幕上的每个点都具有一个坐标值,当你在某个点按下鼠标时,都会有一个坐标值,当采用机器人点击时,为同一个鼠标值,可以只记一次点击。每次点击都会产生鼠标的 KEYUP 和 KEYDOWN 的行为,如果未能捕获到这个值,可能是模拟数据提交。

3.8 通过对广告的地域性限制防止作弊

广告的投放一般都是受地域性限制,如果产生了大量与销售区域无关的点击,则很可能是通过代理服务器在作弊。可以通过反向监查 IP 的来源是否是带有 Proxy 功能的服务器来防作弊。

3.9 通过唯一参数防止作弊

网卡 MAC 物理地址、硬盘序列号,通过该类软硬件信息生成机器码。这种方式的缺点是很难在 WEB 上应用,适合软件营销的防作弊方式。

4 对防作弊措施的改进

4.1 对 IP 防止作弊的改进

由于国内的小区宽带,校园网,局域网等,都会有很多台电脑使用同一个公共 IP 的情况,所以 24 h 内 IP 唯一的反作弊方法不够精确。

如果 24 h 内有大量点击来自同一 IP,则有 3 种可能性:① 是爬虫程序的 IP,该 IP 的点击无效,但不属于作弊行为;② 是机器人作弊程序的 IP,属于恶意点击作弊行为;③ 是多人共用 IP,该 IP 的点击属于有效点击。

下面的算法通过分析一个 IP 在 24 h 内点击的数据,来区分这三种 IP。

算法依赖于如下参数:该 IP 在 24 h 内的点击总数(nums_day),点击的 Adlink 总数(link_nums_day),点击中出现的不同的 HTTP_REFERER 总数(page_nums_day),点击中出现的不同的 HTTP_USER_AGENT 总数(agent_nums_day),点击不同的广告发布商的总数(user_nums_day),点击最多的发布商的点击次数(max_user_nums)。

4.1.1 多人共用 IP VS 爬虫 IP 和机器人 IP

由于不同用户其浏览器版本,操作系统版本,以及其浏览网页的习惯等都不尽相同,所以对于多人共用 IP, agent_rate(agent_nums_day/nums_day)和 page_rate(page_nums_day/nums_day)都会偏大。机器人 IP 和爬虫 IP 一般来自于少数几个固定的电脑或服务器,所以其 agent_nums_day 会非常小,一般是 2 - 3 个,page_rate 也会偏小,甚至根本取不到其 HTTP_REFERER 值。

4.1.2 机器人 IP VS 爬虫 IP

机器人程序以增加点击量为主要目标,其点击一般会集中在少数几个 Adlink 上,或少数几个发布商所发布的 Adlink 上,其 link_rate(links_nums_day/nums_day)和 user_rate (user_nums_day/*nums_day*)一般会偏小。而爬虫 IP 意在获取网站感兴趣的信息,每天都会爬取很大的页面量,虽然随着其对各个网站感兴趣的程度不同,其对各个网站的访问量也会有很大差异,但是爬虫 IP 的点击一般会比较均匀的落在多个 Adlink 上,所以其 link_rate 和 user_rate 会很大,前者一般在 90% 以上。

4.1.3 考虑 nums_day 比较小的情况

当 nums_day 偏小时,link_rate 和 user_rate 会相对偏大,从而影响算法的准确性。需要再根据(max_user_rate) max_user_nums/nums_day 进行判断,对于爬虫 IP,这个比例会偏小;对于机器人 IP,这个比

例则会偏大。Max_user_rate 的阈值应该比 link_rate 和 user_rate 的阈值小很多。

4.2　阈值自适应防止作弊

在防作弊系统中,一般会设置多个参数,对每个参数设置阈值,通过查看用户的数据中有多少参数超过了相应的阈值,判断用户是否作弊。

挖掘这些参数之间的关系,使这些参数根据它们之间的关系来自适应,从而更准确的检测和预防作弊。

例如,点击率上限 VS 无效点击率,一个广告的无效点击率越高,说明这个广告的点击数据越不正常,这时点击率上限应该跟着下降。

4.3　点击时间防止作弊

现在有很多网络广告交易平台允许普通用户把广告贴到自己的博客,个人主页里来赚取广告费,这一类广告有其特殊的规律。图 1 和图 2 是广告被投放到博客上以后的点击情况。

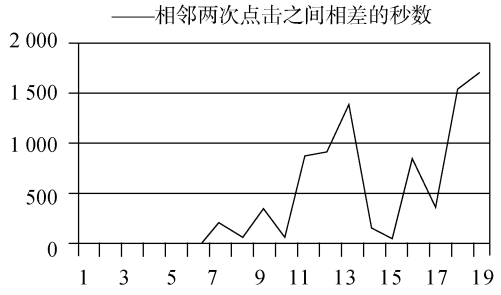

图 1　正常点击的情况

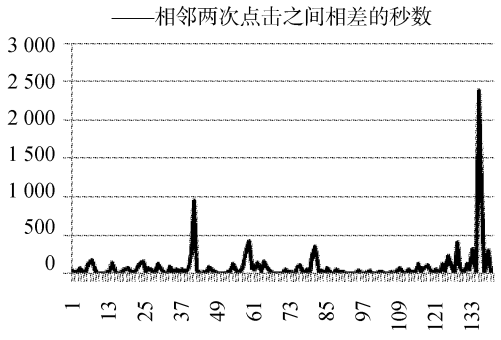

图 2　作弊点击的情况

从图 1 和图 2 中可以看出,当广告刚被投放到博客上以后,短时间内点击会比较多,但是时间越长,点击便会越少,即点击之间的时间间隔呈增大的趋势。为了便于分析,采用下面的算法来转换这些数据:

(1) 定义一个二维数组 a, 对于 $i=0,1,2\cdots$, $a[i][0]$存放广告被点击的时间(Unix 时间戳形式), $a[i][1]$存放在 $a[i][0]$时刻广告被点击的总次数。

(2) 对于 $i=0,1,2\cdots$, $a[i][0]=a[i][0]-a[0][0]$。

(3) 定义数组 b, $b[0][0]=a[0][0]$, $b[0][1]=a[0][1]$。

(4) 对于 $i=1,2,\cdots$, $j=0,1,2\cdots$

```
{
如果(a[i][0]>b[0])
{
 j++;
 b[j+1][0]=a[i][0];
 b[j+1][1]=a[i][1];
 }
 i++;
 }
```

清除数组 a。

(5) 对于 $i=0,1,2\cdots$, $a[i][0]=b[i][0]/3600$, 清除数组 b。

(6) 重复执行(3)-(4), 如果数组 b 的个数等于 3,转向(10)。

(7) 对于 $i=0,12,\cdots$, $a[i][0]=b[i][0]/3$, 清除数组 b。

(8) 执行(3)-(4), 如果数组 b 的个数等于 3,转向(10)。

(9) 转向(7)。

(10) 算法结束。

转换后的数据如下图所示,横轴是 $b[i][0]$的值,竖轴是 $b[i][1]$的值,$i=0,1,2$。可以通过比较两天线段之间的斜率来判断是否作弊,正常点击情况下,第二条线段的斜率会高于第一条线段的斜率,如图 3 所示,而作弊点击的情况则正好相反,如图 4 所示。

5　结束语

虽然以上防作弊手段和措施可以有效地预防和

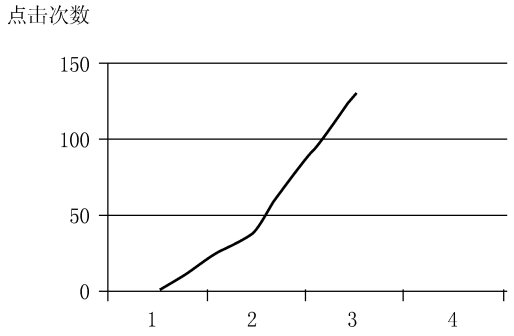

图3 正常点击的情况(数据转换后)

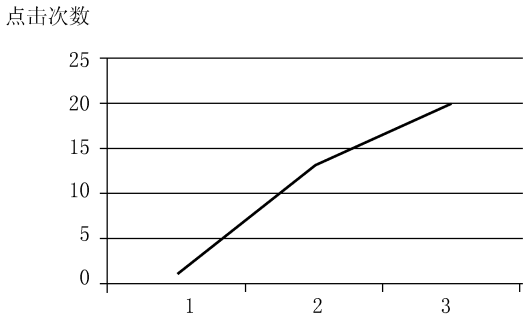

图4 作弊点击的情况(数据转换后)

检测网络广告中的点击作弊现象。但是点击作弊的现象依然层出不穷,而且作弊手段和花样也一直在不断翻新,所以反作弊不是一天两天能解决的,甚至是一个只能"接近完满解决"的问题。

只有彻底杜绝点击作弊和其他广告欺骗的现象,才能保护广告主和网络广告交易平台的正当利益,从而保护网络广告本身,促进网络广告的健康发展。可以肯定的是,各大网络交易平台以后在防作弊上投入将会只多不少,点击作弊将会越来越困难。

参 考 文 献

1 Devore J L. Probability and statistics for engineering and the science (4th ed.). Duxbury Press,1995

2 张 佶. 网络广告点击作弊的常见来源与迹象. http://www.zhangji.net/. 2007

3 Fetterly D, Manasse M, Najork M. Spam, Damn Spam, and statistics: using statistical analysis to locate spam web pages, In: Proceedings of WebDB, 2004

4 王利刚,赵政文,赵鑫鑫. 搜索引擎中的反 SEO 作弊研究. 成都,计算机应用与研究,2009

Anti CPC Click-cheating on Network Advertisement

CHANG Yan, TANG Xiao-chun

(Computer Collage of Northwestern Polytechnical University, Xi'an 710072, P. R. China)

[**Abstract**] With the rapid development of Internet and the double-quick increment of the number of netizens, the network has become more suitable platform for dissemination and high quality exhibition of advertisement. A variety of online advertising trading platforms are springing up like mushrooms, and CPC (cost-per-click) model of online advertising is gradually being accepted. However, the problem appears with this phenomenon is that click fraud prevention and detection of pressing, because it will have a direct bearing on whether CPC can become a long-term advertising model. About the characteristics of the CPC click-cheating means hich analyzes the existent methods of preventing and detecting the CPC click-cheating, and puts forward some improvement measures on the traditional anti-cheating by monitoring IP address, effectively distinguishing the IP shared by many people, reptiles IP, as well as IP of robot cheat program; and proposed some new anti-cheating measures, such as anti-cheating by analyzing click-time, anti-cheating by adaptive threshold controlling, and so on.

[**Key words**] network advertisement CPC click-cheating