基于 WAMP 的网站流量数据分析①

何佩佩, 韩汉光, 谢颖华

(东华大学 信息科学与技术学院, 上海 201620)

摘 要: 传统的网站只是在页面上使用计数器进行简单的访问者数量的统计. 随着互联网技术的快速发展, 对于 Web 网站服务运营商来说, 及时、高效、准确地了解网站运营情况非常重要, 只是简单的计数统计并不能真正对网站运营状况做出全面的评估. 基于 WAMP 的网站流量数据分析系统正是对传统技术的加强, 其中 WAMP 指的是由 Windows、Apache、MySQL 和 PHP 共同组成的一个强大的 Web 应用程序平台, 本文介绍一个在 WAMP 的基础上建立的系统, 利用 Smarty 技术、Jpgraph 技术以及 Javascript 技术, 实现对网站访客信息的收集、整理、存储以及最后的图形展现, 为运营商提供有效、直观的统计信息. 通过这些量化的信息, 运营商可以借此针对性地提高服务质量、制定运营战略, 从而增强市场竞争力.

关键词: 网站流量数据分析; 数据采集; 流量统计; PHP; Jpgraph

Method of the Website Data Analysis Based on WAMP

HE Pei-Pei, HAN Han-Guang, XIE Ying-Hua

(College of Information Science and Technology, Donghua University, Shanghai 201620, China)

Abstract: The traditional website just used a counter simply to calculate the number of visitors. With the rapid development of Internet technology, web sites have increased dramatically. For web site service operators, timely, efficient and accurate understanding of website traffic is very important. Simple quantitative statistics on the website can not really make a comprehensive assessment of situation. WAMP-Based website data analysis system is to strengthen the traditional techniques and WAMP means which a powerful Web application platform is forms by Windows, Apache, MySQL and PHP. The website data analysis system is used Smarty, Jpgraph and Javascript to analyze the data traffic for operators and provide them effective and intuitive statistics information. Through these quantitative information, operators can improve service quality and develop operational strategies base on statistic data.

Key words: website traffic data analysis; data collection; traffic statistics; PHP; Jpgraph

在互联网发展到现在, 几乎大多数企业和组织都建立起官方网站, 而且这类网站正在成为各个企业组织对外宣传不可替代的公共关系手段. 分析网站流量除了获取到网站被访问量, 还可以获取访问者的相关信息. 通过分析这些信息, 网站运营者可以分析得出许多极具价值的信息, 诸如网站的受欢迎程度、访客的背景信息、有怎样的喜好, 还可以分析出网站所采用的哪些推广手段效果更好; 除此之外, 网站管理者甚至可以知道网站访问者都来自哪个区域、使用的什么类型的操作系统、用哪种类型的浏览器. 管理者可以根据这些访客信息的统计情况改善营销策略[1].

传统的简单的计数器难以帮助网站运营获得上述的参考信息, 致使为数不少的运营者由于缺乏对网站数据进行统计和分析的方法, 大量有用的数据没有被有效的利用, 导致对网站运营的情况分析只停留在数据与信息的简单汇总和流水帐式的通报, 缺乏全面和深入分析. 因此, 网站运营者迫切需要一种能够统计和分析各项数据的网站流量分析工具.

① 收稿时间:2013-11-29;收到修改稿时间:2014-01-23

网站流量数据分析系统可以使网站运营者时刻得到网站的运营情况，帮助企业制定优化网络的营销策略，减少盲目的营销运作. 同时，通过对相关数据进行统计，网站运营者可以从网站流量中发现访客访问网站的规律或趋势，并将这些规律与营销策略相结合，从而找出现有的营销策略中所存在的问题，为修正和改善网站的营销策略提供可靠的数据依据. 帮助个人或企业更好地制定营销策略，协助网站稳步发展.

1 网站流量数据的关键评价指标

网站营销过程中对各种网站流量数据进行分析，已经成为后台管理中必备的一环，通过这些分析结果，可以帮助运营者准确地了解用户行为模式和网站的运营情况. 但是,如何定量地去衡量网站运营情况，哪些又是网站运营中的关键因素,困惑着多数网站运营者.

1.1 网站流量(PV, Passengers Volume)

网站流量统计常用来衡量网站的整体效果，主要的统计指标包括:

①页面浏览数(Page View)指在一天(00:00-24:00)内，访问站点的页面数总和. 用户对统一页面的多次访问，将被累计记录.

②独立访客数(Unique Visitors)指将每台独立的上网电脑(以 cookie 为依据)视为一个用户，在一天(00:00-24:00)内，访问站点的不重复用户数.

③独立 IP 数指在一天(00:00-24:00)内，访问站点的所以用户使用的不重复 IP 数. 相同的 IP 地址只被统计 1 次.

1.2 用户行为(VB, Visitors' Behavior)

用户行为主要反映用户如何访问站点、在网站上的停留时间、访问了哪些页面等，主要统计指标包括:

①访问深度: 一次完整的站点访问中，用户浏览的页面数. 浏览页面越多，深度越深.

②访客时长: 用户访问站点的持续时间.

③来源分析: 分析站点用户的来源类型. 来源类型分为搜索引擎、其他网站、直接输入网址和标签和站内跳转.

1.3 用户访问方式(VAM, Visitors Access Mode)

用户访问方式主要反映用户的地理位置、设备、浏览器名称和版本、操作系统等，主要的统计指标包括:

①地理位置: 网站的访客来源于哪个省、市、自治区或国外

②浏览器: 网站的访客所使用的浏览器类型

③屏幕分辨率: 访客所使用的各种屏幕分辨率

④操作系统: 网站的访客所使用的操作系统

2 系统的设计方案

2.1 系统的结构设计

系统采用标准 B/S 三层结构开发，三层结构分别为: 用户界面层(UI)、业务逻辑层(BLL)和数据层(DB)[2].

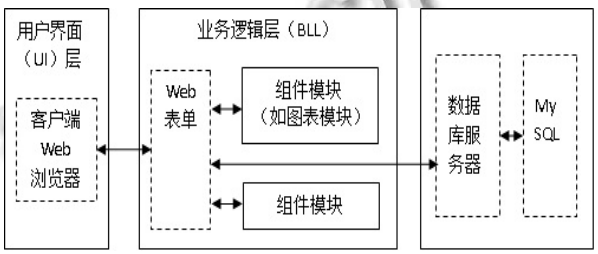

图 1 系统结构图

①用户界面层(UI). 通过浏览器，用户界面层与用户进行交互、接收用户的输入信息并将服务器端传来的数据呈现给客户

②业务逻辑层(BLL). 主要由 Web 表单和组件服务组成，负责接收浏览器传来的请求，并将其发送给数据层，同时将请求处理结果发送给浏览器. 其中 Web 表单是向客户呈现数据和信息的基础，也是响应和处理客户信息和数据，以及显示 Web 表单交互生成的信息和数据的基础.

③数据层(DB). 利用 PHP 的 MySQL 扩展库来操纵数据，为业务逻辑层提供数据服务，如对数据表进行 Insert、Delete、Select、Update 操作，存储数据处理结果、返回检索数据的结果等[3].

2.2 系统的模块设计

2.2.1 web 数据采集模块

目前，在网站运营过程中,Web 服务器端获取流量数据的方法大致分为三类.

①监听网络数据包

数据包是网络传输中的基本单元，主要由发送端 IP 地址、传输数据和接收端 IP 地址组成，里面包含了发送端用户所有数据信息. 通过监听网络数据包获取流量数据,需要在客户端和 Web 服务器之间加一个基于软件或者硬件的包嗅探器，才能对经过的所有数据包进行监听，从中提取出有用的信息，比如用户的 IP

地址、请求访问的 URL 资源以及浏览器类型等, 接收到数据包的时间还可以作为用户访问的时间.

②分析服务器日志

当用户浏览网站时, 便与远程计算机(即 Web 服务器)产生了 Internet 连接, 服务器端通过 TCP/IP 协议请求向客户端传送文件. TCP/IP 协议将文件数据打包分成分组, 通过 Internet 网络传到用户的计算机. 这些分组将在用户计算机中重组并显示在浏览器中. Web 服务器定位和传递用户请求的文件, 每一次请求后, 服务器便在日志文件中记录信息交换结果. 常用的 Web 服务器如 IIS、Apache 都具有一套完整灵活的日志系统, 均能在服务器端自动生成日志文件, 里面记录了里面记录了用户访问网站的各种数据, 包括 IP 地址、访问时间、访问方式、请求内容等.

③添加页面标记

页面标记法指的是在需要被监测的网页源文件中插入一小段可执行的 JavaScript 程序代码. 当该网页被下载到客户端的浏览器上时, 这段 JavaScript 代码就会被执行. 然后, 它就会如实的将采集到的访客行为信息不间断的发给所对应的服务器[4].

本系统采用添加页面标记法的方法进行 Web 数据的采集. 添加页面标记需要在被统计的网站上嵌入一段 JavaScript 代码, 该代码的作用就是向处理程序发送各种流量数据, 接着, 处理程序将实时存储和处理接收到的数据, 并使用 Web 的方式向运营者提供数据分析报告, 形成一套完整的流量分析系统的平台.

2.2.2 数据库模块

数据库模块的主要功能是接收 web 数据采集模块中 JavaScript 代码传入的参数, 分析数据并存入数据库. 在 MySQL 控制台中, 创建一个名为 counter 的数据库, 该数据库中包含的数据表及其相应的功能如表所示.

表 1 counter 数据库中包含的数据表及其功能

数据表	功能
Member	用户注册信息表
counter_detail	最近访客信息表
counter_daily	每日访问统计表
counter_month	每月访问统计表
counter_year	每年访问统计表
counter_area	地理位置统计表
counter_browser	来源网站统计表

①注册用户信息表

表 2 member 注册用户信息

字段	类型数据	备注
id	‘id’int(11)unsigned NOT NULL auto_increment	自动编号
username	‘username’varchar(20)NOT NULL default	用户名
password	‘password’varchar(20)NOT NULL default	密码
tel	‘QQ’varchar(50)default NULL	手机
QQ	‘QQ’varchar(20)default NULL	QQ
email	‘email’varchar(50)default NULL	邮箱

其中 id(自动编号)字段作为该表的主关键字(primary key), 唯一标识了一个注册用户信息.

②最近访客信息表

表 3 counter_detail 最近访客信息表

字段	类型数据	备注
id	‘id’int(8)unsigned NOT NULL auto_increment	自动编号
date	‘date’varchar(10)NOT NULL default	访问日期
time	‘time’varchar(8)NOT NULL default	访问时间
ip	‘ip’ varchar(15)NOT NULL default	访问者 ip
os	‘os’varchar(255)default NULL	访问者操作系统
brows	‘brows’varchar(255)default NULL	访问者浏览器信息
ref	‘ref’varchar(255)default NULL	来源 URL
area	‘area’varchar(255)default NULL	访问者地理位置
host	‘host’varchar(255)default NULL	来源域名

其中 id(自动编号)字段作为该表的主关键字(primary key), 唯一标识了一个访问者信息.

③每日/每月/每年访问统计表

表 4 每日/每月/每年访问统计表

字段	类型数据	备注
id	‘id’int(8)unsigned NOT NULL auto_increment	自动编号
name	‘name’varchar(100)NOT NULL default	统计标识
value	‘value’int(10) default NULL	统计值
date	‘date’varchar(10) NOT NULL default	统计时间

其中 id(自动编号)字段作为该表的主关键字(primary key), 唯一标识了一个访问记录信息. 另外由 name, date 字段组成联合唯一值字段

地理位置统计表及来源网站统计表的结构与上表相同, 通过 name 和 value 字段对需要统计的数据进行存储. 而 date 字段则存储每条记录的统计时间, 其中因为地理位置和来源页的日期不需要统计, 所以表中

就不对其日期作记录.

2.2.3 分析界面展示模块

分析界面展示模块位于网站后台管理中, 采用当前流行的DIV+CSS页面设计技术, 将样式信息与网页内容分离, 更好地控制页面布局, 更好地制作展现界面. 同时, 利用 Javascript 脚本, 来实现一些特定的效果. 接下来,, 详细叙述一下各种界面的功能设计.

①登入界面. 流量分析系统的登录方式是需要对管理员账号和密码两者都进行验证, 只有两项对应匹配才能够进入分析模块. 而这些登录验证信息都是预先在数据库中进行了分配, 并存储在 member 表中.

②最近访客信息界面. 该界面是最基础的数据分析界面, 主要针对网站的一些日常访问情况的统计分析, 主要功能是显示最近 10 个访客的详细信息.

③每日/每月/每年统计信息界面. 该界面主要为运营者提供日统计、月统计和年统计访问量报表.

④历史数据查询界面. 该界面主要提供历史数据查询功能, 展现访客在访问过程中的各种信息.

⑤地理位置统计界面. 该界面这要根据访客的 IP 地址来识别访客的来源地址, 可以实现统计各地区访问网站的人数.

⑥统计图形分析界面. 该界面主要将数据转化成柱状图、折线图、饼状图等图形形式, 运营者课直观地观察到网站运营状况.

3 系统的技术解决方案

3.1 系统运行处理的流程

系统的处理流程主要是由几个步骤组成, 首先访客发送访问请求, 经过检验模块的判断, 如果该次访问已经被记录过, 则立即返回访客的请求页面; 若该条记录未被记录过, 便记录下访客的各类信息, 并将经过处理过后的数据存储于数据库中. 然后等待管理员发出查看请求, 由业务逻辑判定调用哪个统计功能, 从数据库中调用相关数据, 利用绘图工具绘制出统计图表, 最终显示在统计页面上. 基于 Web 的网站流量分析系统的流程图如图 2 所示.

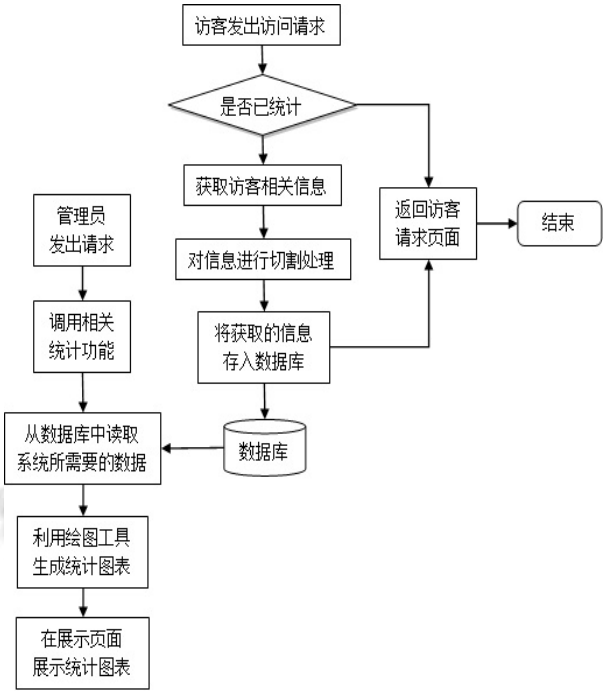

图 2 系统流程示意图

3.2 Javascript 脚本的插入

添加页面的技术是在被监测的网页源文件中增加一段程序代码, 用来采集该网页的访问量或特定的流量数据, 进而进行相关数据的统计分析. 程序代码通常是一段JavaScript的代码, 放置在网页源文件的结尾处, 即</Body>之前, 以确保是在显示完页面内容之后, 再进行相关流量数据的采集. 另外, 执行代码并不会占用服务器资源, 因此不会影响访客的浏览感受[5].

3.3 获得地理位置的公共函数

通过访客的 IP 地址可以得到访客所在的地理位置, 从而实现对网站访客来源于哪个省、市、自治区或国外进行统计分析.

3.3.1 格式化 IP 地址函数

d_ip 函数将传入的IP地址的死段数字地址数组参数格式化成三维字符并返回, 比如传入 127.0.0.1, 格式化输出为 127.000.000.001.

3.3.2 获得地理位置函数

从 IP 地址库中查找与之对应的地理位置, IP 地址库是 ipdata 目录下的所有文本文件, 每个文件都保存着 IP 地址对应关系数据, 并且文本文件命名也是根据 IP 地址的第一段数字开始的. 本系统采用了自定义函数 ip(), 传入 IP 地址参数, 返回对应的地理位置. 检索 IP 文本数据时, 先根据 IP 地址的第一段字, 判断打开相应的文本文件. 读取文本, 逐行判断用户传入的 IP 是否在检索的地址范围内, 如果是, 则返回, 否则继续检索下一行.

3.4 访客信息的采集(JS 脚本)

被监测的网页源文件中所添加的 JavaScript 脚本, 用来收集访客信息, 并向分析系统发送统计信息数据.

①使用 navigator.appName 获得客户端浏览器信息, 用于对用户的访问行为进行分析.

```
var soft="";
soft=navigator.appName; //获得客户端浏览器信息
if(soft != "")
//如果有信息, 则进行 escape 处理, 以保证通过 url 传输正常
{ soft=escape(soft); }
```

②使用 document.referrer 获得客户端来源页信息, 用于对用户的来源类型进行分析.

```
var fromurl="";
try{fromurl=top.document.referrer;}
//用 document.referrer 是来源页面
catch(err)  { fromurl=""; }
finally { fromurl=(fromurl=="")?
document.referrer:fromurl; }
fromurl=escape(fromurl);
```

③使用 document.URL, 获得当前 url, 用于对用户的访问深度进行分析.

```
var resource="";
resource=document.URL;
//获得当前 url, 使用 document.URL 即可
if(resource != "")
    resource=escape(resource);
```

④使用 navigator.userAgent 得到浏览器版本, 用于对用户的访问行为进行分析.

```
var a="";
a=navigator.userAgent;
//浏览器版本可以使用 navigator.userAgent 得到
if(a != "")
    a=escape(a);
```

⑤根据得到的客户信息构造传递用的查询字符串并调用接受客户端信息的 PHP 程序

```
r="?referer="+formurl+"&resolve="+resolve+"&color="+color+"&title="+title+"&resource="+resource+"&a="+a;
document.open();
//打开一个新的 HTML 文档
document.write( "<script type=\"text/javascript\" scr=\"http://127.0.0.1/counter/Smarty/test.php"+r+"\"></script>");
```

3.5 接收并存储访客信息

接收上面来自 JavaScript 传入的参数, 分析数据并存入数据库.

①取得客户端 IP 地址的函数

分别通过环境变量 HTTP_CLIENT_IP、HTTP_X_FORWARDED_FOR 和 REMOTE_ADDR 来获得客户端 IP 并返回.

②获得客户浏览器版本和客户操作系统

客户浏览器版本函数: 通过对传入的浏览器信息匹配检索, 获得具体具体的浏览器名称及版本. 客户操作系统函数: 根据操作系统特征字符串, 逐个匹配比较, 以获得客户操作系统的确切名称.

③更新数据库

编写一则更新数据库的函数, 通过该函数, 可以把获得数据插入数据库. 该函数原理是, 接收三个参数, 分别为$tablename、$typename、$today. 先以接收到的$typename 值和$today 值为条件, 更新数据表的 value 值, 使之增 1, 检索数据库并使用 mysql_affected_rows 函数来判断 SQL 执行受影响的行数, 如果大于 1, 则表明更新成功, 否则就插入一条新记录.

3.6 利用 Jpgraph 实现统计数据的可视化显示

JpGraph 是 PHP 专门提供图表的类库, 它使得作图变成了一件非常简单的事情. 首先从数据库中取得统计数据, 定义标题, 图表类型, 然后通过 PHP 中的 Jpgraph 类库, 便可在统计数据的基础上完成各类统计图, 包括坐标图、柱形图、折线图等. 效果如下所示.

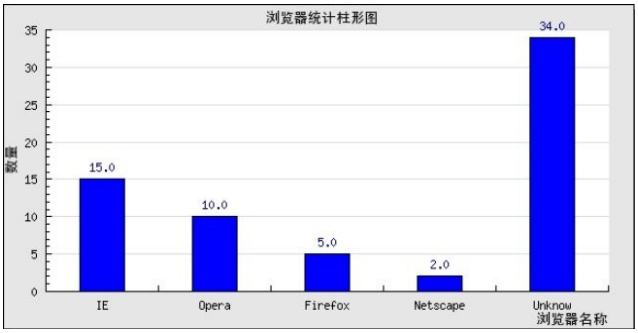

图 3 访客浏览器信息柱状图

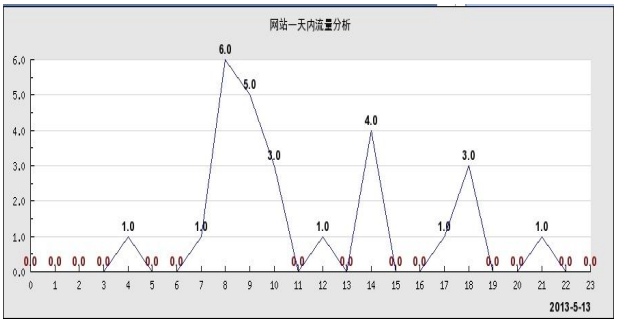

图 4 日流量统计折线图

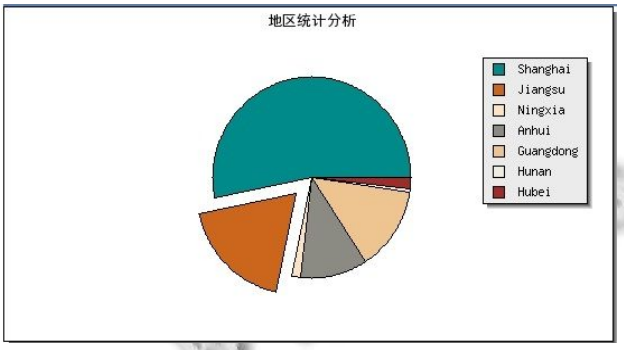

图 5 地区统计分析饼状图

4 结语

基于 WAMP 的网站流量数据分析系统主要是对网站运营过程中的各种流量信息和访客信息进行统计分析，因此，这些数据的实质可反映当前网站的运营情况. 然后，通过一定的技术手段，将这些数据生成报表或图形，反映出一定的规律，为运营者制定策略提供可靠的依据.

通过对网站流量数据分析的理论和技术的研究，开发了一个满足一定实际应用需要的网站流量数据分析系统. 在开发过程中，主要完成了以下工作. 首先，结合网站流量数据分析的常用指标和实际应用的需求，利用加添页面标记的方式实现了 Web 流量数据的采集. 然后，根据网站流量数据的特点和分析界面的不同功能，设计了数据库并实现了后台的数据存储. 最后，研究了 Web 应用程序开发的相关技术，设计并实现了系统前端的页面设计和系统后台的数据分析功能.

参考文献

1 杜晓春.基于 Web 的网站数据分析软件 Wysistat 的设计与实现[学位论文].西安:西安电子科技大学,2010.
2 王风玲.基于 PHP+MYSQL 的新闻发布系统的研究与实现.计算机应用与软件,2012,36(24):42−44.
3 宗小忠.基于 LAMP 的网站流量分析系统图表功能的设计与实现.沙洲职业工学院学报,2010,(2):5−7.
4 单哲.网站流量统计分析技术研究[学位论文].哈尔滨工程大学,2012.
5 李雯.电子商务平台中流量统计模块的设计研究.硅谷,2010,(20):87.