## 1.独立IP数(即同一个ip一天无论访问多少次，均只计算一次)、总流量数。

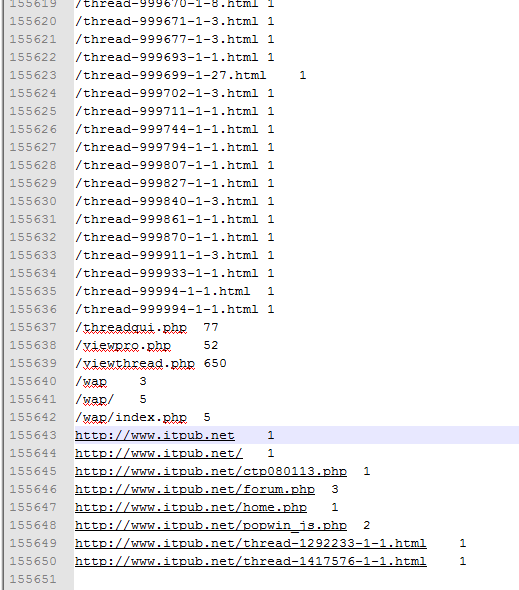
经分析，这两个指标可以放在一起求解。过滤掉无效的行，结果如下所示，其中判断该条记录是否是爬虫是分析记录中是否有spider和googlebot，代码详见附录。

|  |  |  |
| --- | --- | --- |
| 总计 | 爬虫 | 普通用户 |
|  |  |  |

其中，总流量恰好等于爬虫产生的流量加普通用户产生的 流量，但是独立IP却相等，分析可能是有些ip即作为了普通用户，又作为了爬虫访问了网站，因此导致重复计算。

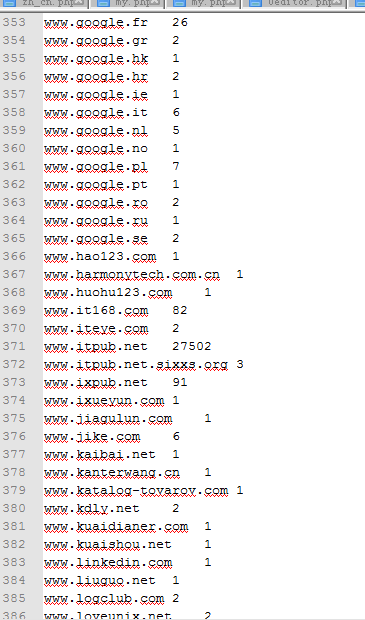
## 2.PV

首先过滤掉爬虫产生的数据，然后针对forum.php?mod=\*\*\*\*\*\*这样的访问，截取“？”号之前的部分，其他则保留全部，得到的数据如下所示，经统计后的结果仍然很多。



## 3.来源网站

首先过滤掉爬虫产生的数据，得到的结果如下所示：



对结果进行进一步统计，得到前10的来源

|  |  |  |
| --- | --- | --- |
| 域名 | 网站 | 带来的独立IP |
| [www.itpub.net](http://www.itpub.net/) | ITPUB | 27502 |
| - | 直接访问 | 17693 |
| www.baidu.com | 百度 | 10950 |
| www.google.com.hk | Google香港 | 3436 |
| www.google.com | Google | 877 |
| m.baidu.com | 手机百度 | 692 |
| www.google.com.tw | Google台湾 | 666 |
| cio.it168.com | IT168 | 445 |
| space.itpub.net | ITPUB空间 | 356 |
| www.sogou.com | 搜狗 | 351 |

## 4.浏览器统计

由于记录里面对浏览器的描述比较复杂，比如

199.255.44.5 - - [04/Jan/2012:00:00:03 +0800] "GET /thread-1497581-1-300.html HTTP/1.1" 200 14653 "-" "Mozilla/5.0 (Windows NT 6.0; WOW64) AppleWebKit/535.7 (KHTML, like Gecko) Chrome/16.0.912.63 Safari/535.7"

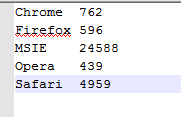
这到底是Chrome还是Safari？

所以我自己定了一个规则，如下：

1 如果用空格分隔后最后一个元素是Safari，就算是Safari浏览器，因为我观察日志中，Safari都出现在最后。

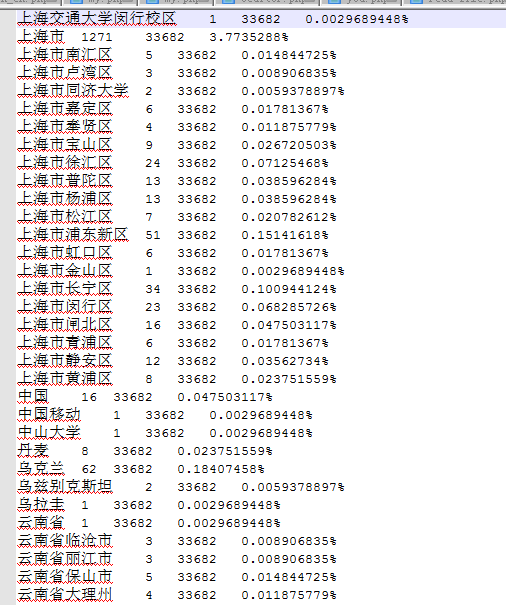
2 否则，从“来源网站”（上例是"-"）之后算起，以此在每个分隔后的元素中匹配以下浏览器（Chrome, Firefox, MSIE, Opera），匹配到哪一个就算做该记录由该浏览器访问。

得到结果如下：



## 5.IP地域统计

利用纯真数据库及其读取代码统计独立IP所属地域，得到如下结果：



进一步整理，得到如下结果：

国内：

|  |  |  |
| --- | --- | --- |
| 北京 | 5416 | 18.6% |
| 广东 | 5192 | 17.8% |
| 山东 | 1851 | 6.3% |
| 河北 | 1618 | 5.5% |
| 上海 | 1515 | 5.2% |
| 浙江 | 1430 | 4.9% |
| 辽宁 | 1314 | 4.5% |
| 河南 | 1292 | 4.4% |
| 江苏 | 1062 | 3.6% |
| 天津 | 823 | 2.8% |
| 台湾 | 810 | 2.8% |
| 黑龙江 | 775 | 2.7% |
| 山西 | 632 | 2.2% |
| 香港 | 629 | 2.2% |
| 吉林 | 594 | 2.0% |
| 安徽 | 485 | 1.7% |
| 福建 | 439 | 1.5% |
| 内蒙古 | 439 | 1.5% |
| 湖北 | 393 | 1.3% |
| 陕西 | 380 | 1.3% |
| 广西 | 372 | 1.3% |
| 四川 | 346 | 1.2% |
| 湖南 | 339 | 1.2% |
| 重庆 | 297 | 1.0% |
| 江西 | 153 | 0.5% |
| 云南 | 150 | 0.5% |
| 贵州 | 88 | 0.3% |
| 新疆 | 82 | 0.3% |
| 甘肃 | 71 | 0.2% |
| 海南 | 60 | 0.2% |
| 澳门 | 48 | 0.2% |
| 青海 | 42 | 0.1% |
| 宁夏 | 30 | 0.1% |
| 西藏 | 4 | 0.0% |

国外：

|  |  |  |
| --- | --- | --- |
| 美国 | 1687 | 41.5% |
| 新加坡 | 388 | 9.5% |
| 加拿大 | 223 | 5.5% |
| 日本 | 219 | 5.4% |
| 英国 | 146 | 3.6% |
| 德国 | 134 | 3.3% |
| 马来西亚 | 118 | 2.9% |
| 澳大利亚 | 105 | 2.6% |
| 法国 | 100 | 2.5% |
| 俄罗斯 | 90 | 2.2% |
| 韩国 | 75 | 1.8% |
| 印度 | 62 | 1.5% |
| 乌克兰 | 62 | 1.5% |
| 越南 | 52 | 1.3% |
| 意大利 | 34 | 0.8% |
| 荷兰 | 33 | 0.8% |
| 泰国 | 32 | 0.8% |
| 西班牙 | 29 | 0.7% |
| 巴西 | 29 | 0.7% |
| 瑞典 | 26 | 0.6% |
| 印度尼西亚 | 20 | 0.5% |
| 土耳其 | 14 | 0.3% |
| 阿根廷 | 13 | 0.3% |
| 阿联酋 | 13 | 0.3% |
| 芬兰 | 12 | 0.3% |
| 波兰 | 12 | 0.3% |
| 墨西哥 | 12 | 0.3% |
| 罗马尼亚 | 11 | 0.3% |
| 菲律宾 | 11 | 0.3% |
| 沙特阿拉伯 | 11 | 0.3% |
| 柬埔寨 | 11 | 0.3% |
| 委内瑞拉 | 11 | 0.3% |
| 埃及 | 11 | 0.3% |
| 以色列 | 11 | 0.3% |
| 爱尔兰 | 10 | 0.2% |
| 新西兰 | 9 | 0.2% |
| 瑞士 | 8 | 0.2% |
| 挪威 | 8 | 0.2% |
| 哥伦比亚 | 8 | 0.2% |
| 伊朗 | 8 | 0.2% |
| 丹麦 | 8 | 0.2% |
| 秘鲁 | 7 | 0.2% |
| 捷克 | 7 | 0.2% |
| 哈萨克斯坦 | 7 | 0.2% |
| 比利时 | 6 | 0.1% |
| 摩尔多瓦 | 6 | 0.1% |
| 巴基斯坦 | 6 | 0.1% |
| 葡萄牙 | 5 | 0.1% |
| 欧洲 | 5 | 0.1% |
| 斯洛伐克 | 5 | 0.1% |
| 南非 | 5 | 0.1% |
| 厄瓜多尔 | 5 | 0.1% |
| 匈牙利 | 5 | 0.1% |
| 黎巴嫩 | 4 | 0.1% |
| 智利 | 4 | 0.1% |
| 拉脱维亚 | 4 | 0.1% |
| 巴拿马 | 4 | 0.1% |
| 安哥拉 | 4 | 0.1% |
| 哥斯达黎加 | 4 | 0.1% |
| 科特迪瓦 | 3 | 0.1% |
| 白俄罗斯 | 3 | 0.1% |
| 格鲁吉亚 | 3 | 0.1% |
| 摩洛哥 | 3 | 0.1% |
| 尼日利亚 | 3 | 0.1% |
| 奥地利 | 3 | 0.1% |
| 埃塞俄比亚 | 3 | 0.1% |
| 卡塔尔 | 3 | 0.1% |
| 加纳 | 3 | 0.1% |
| 保加利亚 | 3 | 0.1% |
| 马其顿 | 2 | 0.0% |
| 立陶宛 | 2 | 0.0% |
| 肯尼亚 | 2 | 0.0% |
| 苏丹 | 2 | 0.0% |
| 波斯尼亚和黑塞哥维那 | 2 | 0.0% |
| 巴勒斯坦 | 2 | 0.0% |
| 希腊 | 2 | 0.0% |
| 孟加拉 | 2 | 0.0% |
| 塔吉克斯坦 | 2 | 0.0% |
| 塞尔维亚 | 2 | 0.0% |
| 塞尔维亚和黑山 | 2 | 0.0% |
| 塞浦路斯 | 2 | 0.0% |
| 多米尼加 | 2 | 0.0% |
| 卢森堡 | 2 | 0.0% |
| 亚美尼亚 | 2 | 0.0% |
| 克罗地亚 | 2 | 0.0% |
| 乌兹别克斯坦 | 2 | 0.0% |
| 黑山 | 1 | 0.0% |
| 阿塞拜疆 | 1 | 0.0% |
| 阿尔及利亚 | 1 | 0.0% |
| 阿尔巴尼亚 | 1 | 0.0% |
| 赞比亚 | 1 | 0.0% |
| 贝宁 | 1 | 0.0% |
| 纳米比亚 | 1 | 0.0% |
| 缅甸 | 1 | 0.0% |
| 老挝 | 1 | 0.0% |
| 莫桑比克 | 1 | 0.0% |
| 萨尔瓦多 | 1 | 0.0% |
| 蒙古 | 1 | 0.0% |
| 牙买加 | 1 | 0.0% |
| 特立尼达和多巴哥 | 1 | 0.0% |
| 玻利维亚 | 1 | 0.0% |
| 海地 | 1 | 0.0% |
| 洪都拉斯 | 1 | 0.0% |
| 毛里塔尼亚 | 1 | 0.0% |
| 文莱 | 1 | 0.0% |
| 斯洛文尼亚 | 1 | 0.0% |
| 斯里兰卡 | 1 | 0.0% |
| 巴拉圭 | 1 | 0.0% |
| 尼日尔 | 1 | 0.0% |
| 尼泊尔 | 1 | 0.0% |
| 坦桑尼亚 | 1 | 0.0% |
| 塞内加尔 | 1 | 0.0% |
| 喀麦隆 | 1 | 0.0% |
| 博茨瓦纳 | 1 | 0.0% |
| 危地马拉 | 1 | 0.0% |
| 厄立特里亚 | 1 | 0.0% |
| 冰岛 | 1 | 0.0% |
| 加蓬 | 1 | 0.0% |
| 伊拉克 | 1 | 0.0% |
| 伯利兹 | 1 | 0.0% |
| 佛得角 | 1 | 0.0% |
| 乌拉圭 | 1 | 0.0% |

国内国外比例：