网站流量统计分析系统需求分析

网站流量统计是改进网站服务的重要手段之一，通过获取用户在网站的行为，可以分析出哪些内容受到欢迎，哪些页面存在问题，从而使网站改进活动更具有针对性。

常用的网站流量统计指标一般包括以下情况分析：

* 按在线情况分析

　　在线情况分析分别记录在线用户的活动信息，包括：来访时间、访客地域、来路页面、当前停留页面等，这些功能对企业实时掌握自身网站流量有很大的帮助。

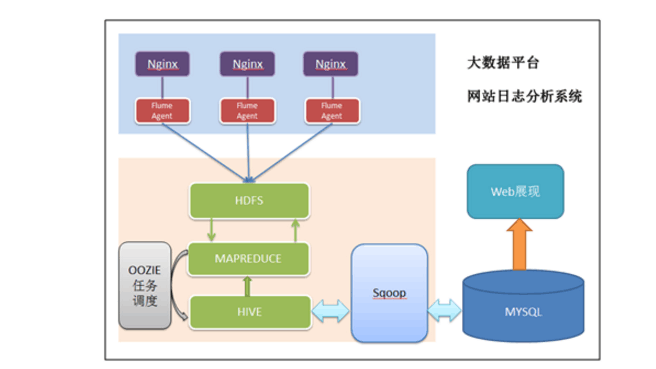
* 按时段分析

　　时段分析提供网站任意时间内的流量变化情况.或者某一段时间到某一段时间的流量变化，比如小时段分布，日访问量分布，对于企业了解用户浏览网页的的时间段有一个很好的分析。

* 按来源分析

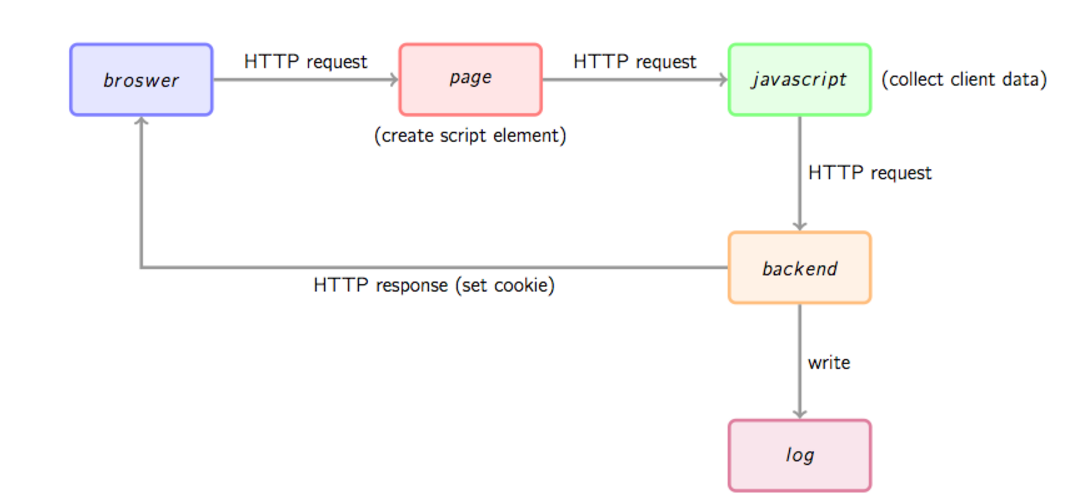
      来源分析提供来路域名带来的来访次数、IP、独立访客、新访客、新访客浏览次数、站内总浏览次数等数据。这个数据可以直接让企业了解推广成效的来路，从而分析出那些网站投放的广告效果更明显。

1. **系统架构设计**



1. **数据源分析**
   1. js埋点

编写特定的js脚本，需要做日志分析的web项目的相关页面上需要通过<script>标签引入该js文件。该js文件会去访问一个nginx或apache搭建的日志服务器上的一个web页面，这样nginx或apache会在access.log文件中记下相应的请求记录。



* 1. access.log日志格式分析

Apache Server安装后,通过在安装目录下的conf目录下查看httpd.conf文件，能够确定日志文件的所在位置。 Apache的日志文件分为两类，分别是访问日志和错误日志。默认情况日志文件的目录在apache安装目录下的logs目录下,分别是access\_log和error\_log 。

access\_log为访问日志,记录所有对apache服务器进行请求的访问,它的位置由CustomLog指令设定,其内容和格式由LogFormat指令来设定 。例如：

CustomLog logs/access\_log combined(日志别名)

LogFormat "%h %l %u %t \"%r\" %>s %b \"%{Referer}i\" \"%{User-Agent}i\"" combined（日志别名）

error\_log错误日志,记录下任何错误的处理请求,它的位置和内容由ErrorLog指令控制,通常服务器出现什么错误,首先对它进行查阅,是一个最重要的日志文件 ,例如：

ErrorLog logs/error\_log

下面抓取一条访问日志记录，对其格式进行分析:

192.168.42.1 - - [10/Jan/2016:22:11:48 -0800] "GET / HTTP/1.1" 403 5039 "-" "Mozilla/5.0 (Windows NT 10.0; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/46.0.2490.80 Safari/537.36"

* 192.168.42.1 (%h)

请求服务的客户端的ip地址

* - (%l)

这项一般没用，可忽略

* - (%u)

这项是用户记录用户HTTP的身份验证,一般的网站这一项为空白

* [10/Jan/2016:22:11:48 -0800] (%t)

访问日期和时间，最后的+0800表示服务器所处的时区为东八区

* "GET / HTTP/1.1" (\"%r\")

这一项整个记录中最有用的信息,首先,它告诉我们的服务器收到的是一个GET请求,其次,是客户端请求的资源路径,第三,客户端使用的协议时HTTP/1.1

* 403 (%>s)

这是一个状态码,由服务器端发送回客户端,它告诉我们客户端的请求是否成功,或者是重定向,或者是碰到了什么样的错误

* 5039 (%b)

这项表示服务器向客户端发送了多少的字节,在日志分析统计的时侯,把这些字节加起来就可以得知服务器在某点时间内总的发送数据量是多少

* "-"(\"%{Referer}i\")

告诉服务器我是从哪个页面链接过来的,没有值时可能是直接打开网页的原因

* "Mozilla/5.0 (Windows NT 10.0; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/46.0.2490.80 Safari/537.36"(\"%{User-Agent}i\")

这项主要记录客户端的浏览器信息

1. **网站指标分析**
2. **PV**

PV(page view)，即页面浏览量，或点击量。通常是衡量一个网站的主要指标。

用户对一个页面A进行一次刷新(按F5)或者一次通过网址访问，该页面A的pv就会增加1。

1. **UV**

UV(unique visitor)，即独立访客数。指访问某个站点或点击某个链接的不同访客人数。一日内访问某个网站或者网页的不同用户数量。同一个用户对一个网站或者一个网页的多次访问记为贡献一个uv。从访客来到您网站到最终关闭网站的所有页面离开，计为1次访问。若访客连续30分钟没有新开和刷新页面，或者访客关闭了浏览器，则被计算为本次访问结束。

1. **VV**

名词：VV = Visit View（访问次数）  
说明：访客的访问次数，用以记录所有访客1天内访问了多少次您的网站。当访客完成所有浏览并最终关掉该网站的所有页面时便完成了一次访问，同一访客1天内可能有多次访问行为，访问次数累计；

1. **Bounce Rate跳出率**

Bounce Rate说明了用户进入网站后是否对网站的内容产生了兴趣,如果用户进入了一个网站的页面，而没有点击该网站的其他页面就离开了，即称跳出。这个指标一般针对某个网站。

1. **独立ip**

这里指的IP地址就是给每个连接在互联网上的计算机分配的一个32位ipv4地址。(现在已经有64位地址)申请公网IP是需要钱的，因此出现了**NAT(网络地址转换协议)**。

　　局域网内的本地IP地址可以通过NAT转换为集体的公网的IP地址。这样集体只需要申请少量IP地址，局域网内就可以拥有上百台机器。最典型的应用在公司、网吧，这些单位只有少量的公网IP地址。

1. **平均在线时长**

指人们访问网站的平均停留的时间。例如一个网站有3个访问，一个停留了2分钟，一个停留了10分钟，一个停留了0分钟，那么time on site则是4分钟。不过，值得注意的是，网站分析工具上统计的时间与实际用户在网站上停留的时间肯定是不同的。人们访问网站最后一个页面的时间长短不会被网站分析工具统计到。原因很简单，因为一般的网站分析工具不统计人们离开一个网站的精确时间，而只能记录他访问这个网站倒数第二页的精确时间，这样最后一个被他访问的页面的停留时间实际上被完全忽略掉了。这种安排意味着两点，第一，网站分析工具统计到的网站浏览时间总是小于网站被打开在浏览器中的时间（尽管浏览器打开页面未必意味着你真的每分每秒都在看它）；第二，只访问了一个页面的visit，在网站上的停留时间计为0。

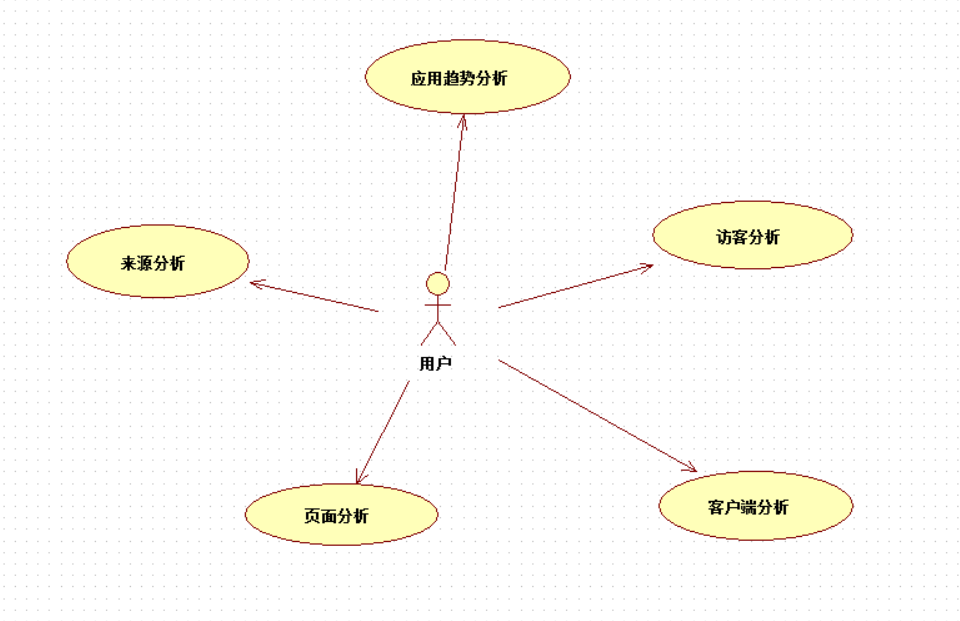
1. **新独立访客**

当日的独立访客中，历史上首次出现的访问网站的访客为新独立访客。

1. **访问深度**

访客的去重的访问页面数量。

1. **主要功能分析**



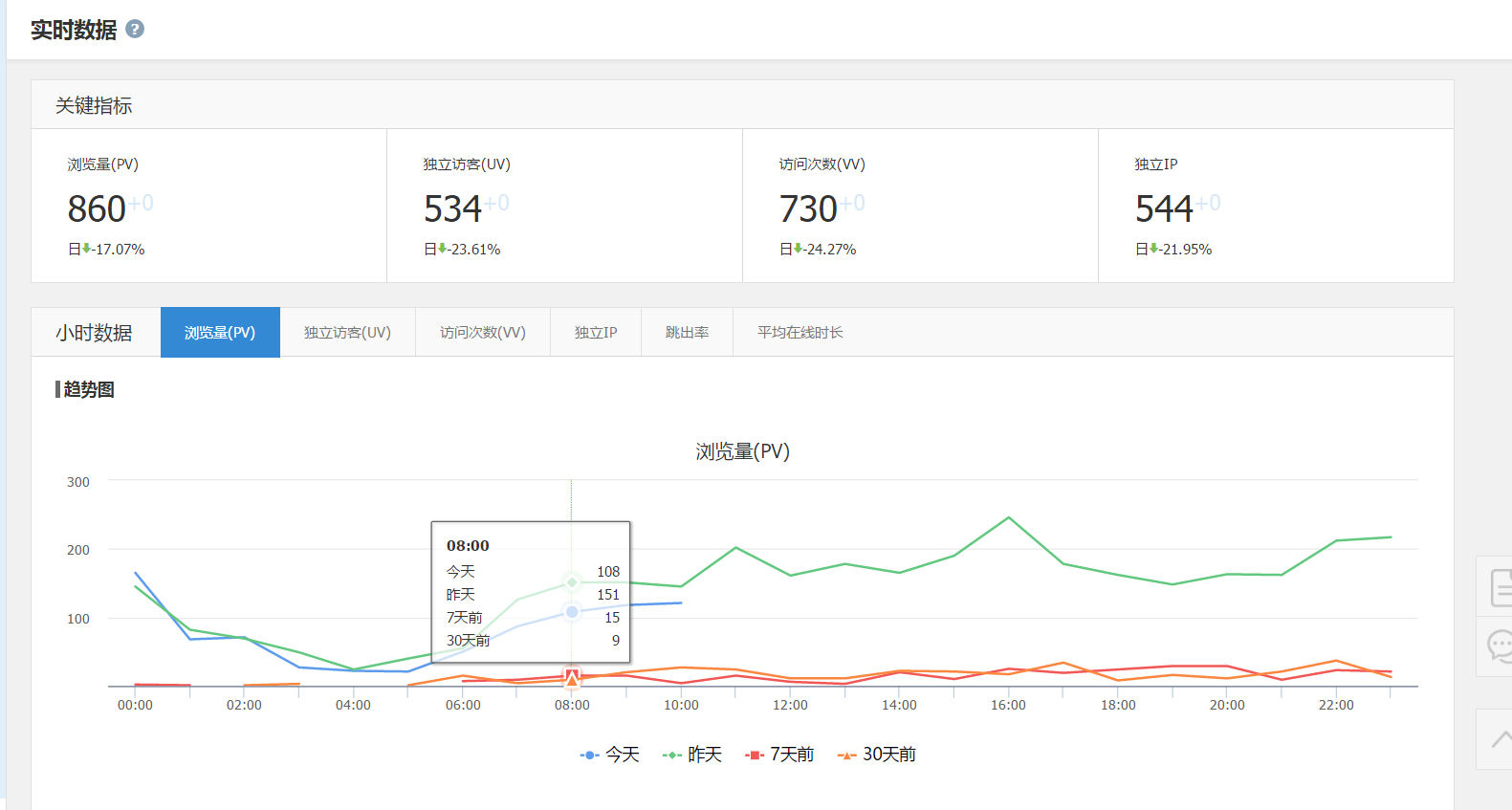
1. **应用趋势**

以时间粒度(按小时、天、月）来分析各个流量指标（PV、UV、VV、独立IP、跳出率、平均在线时长）

* 实时数据

页面功能点分析如下：

* 显示当天的数据
* 关键指标与前一天的对比
* 关键指标的趋势图
* 以小时为单位的流量指标的明细数据的分页显示





* 历史趋势

页面功能点分析如下：

* 针对浏览量、独立访客、访问次数、独立IP、跳出率、平均在线时长等关键指标以天为单位进行趋势图分析，其中分析的时间段的可以设置，以30天为限。
* 以小时为单位的流量指标的明细数据的分页显示





1. **访客分析**

以访客(UV)为粒度来分析各个流量指标，包括以下流量指标：

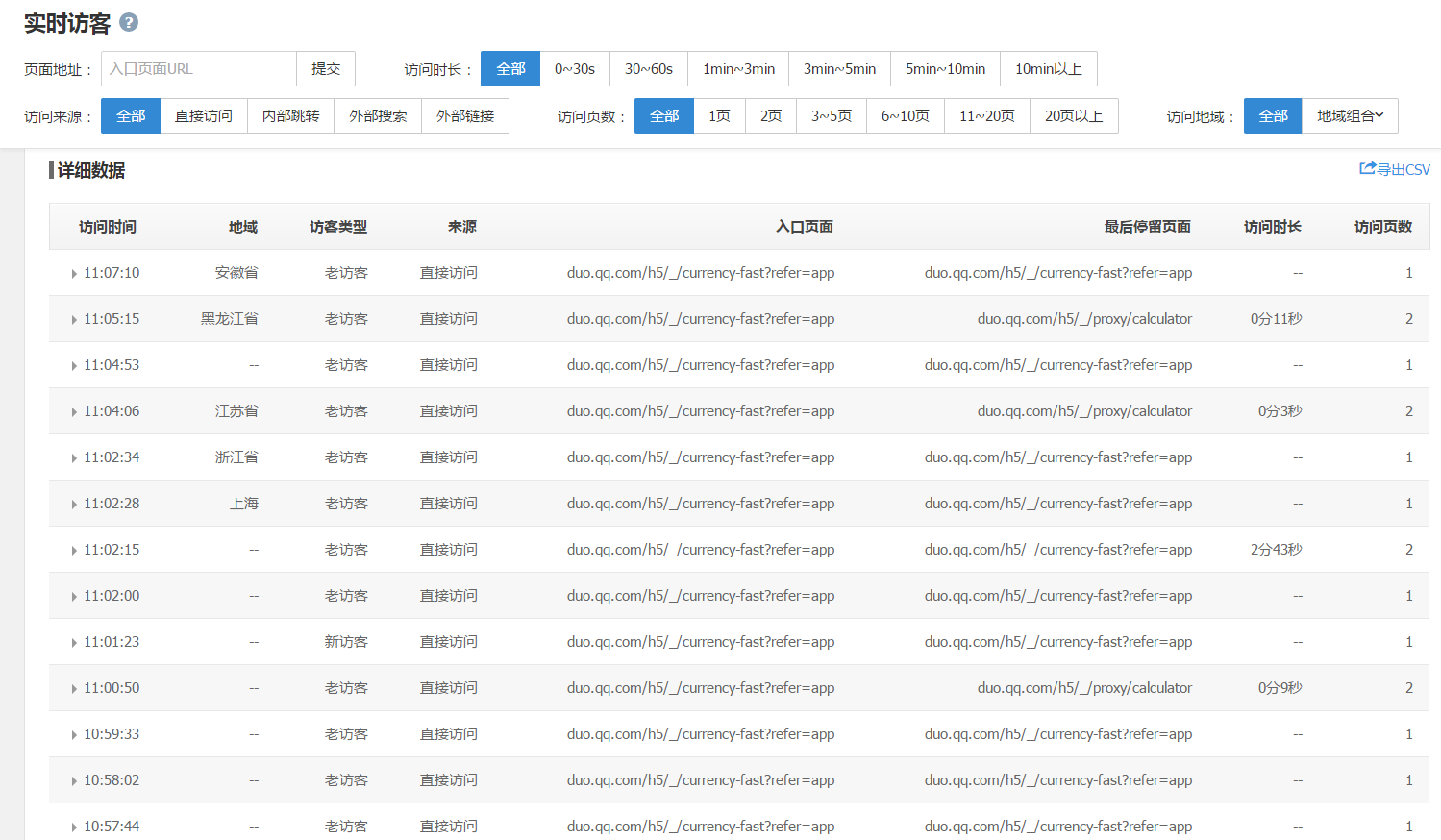
|  |  |
| --- | --- |
| **指标名称** | **说明** |
| 访问时间 | 指UV的开始访问时间 |
| 地域 | 可以包括省和市 |
| 访客类型 | 老访客或新访客 |
| 来源 | 取值包括：直接访问、内部跳转、外部搜索、外部链接 |
| 入口页面 | 该uv访问的第一个页面 |
| 最后停留页面 | 即离开页面, 当某个UV 在30分钟内没有再访问网站页面，它之前的访问页面就是最后停留页面 |
| 访问时长 |  |
| 访问页数 | 访问了几个页面，即访问深度 |

在分析时提供以下两种方式：

* 实时访客 : 当天的访客数据

页面功能点分析如下：

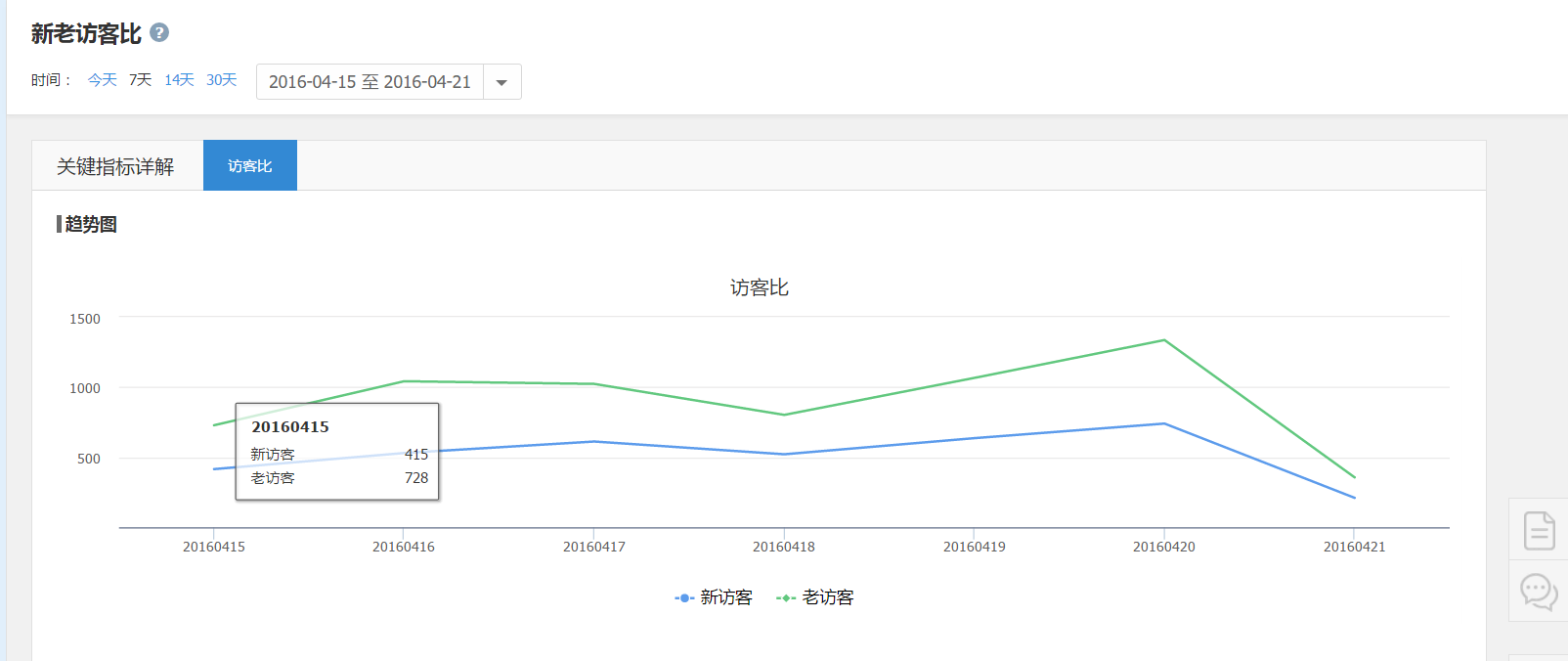
* 当天访客数据的分页显示
* 能根据“入口页面”、“访问时长”、“访问来源”、VV、访问地域等条件组合过滤当天的访客数据

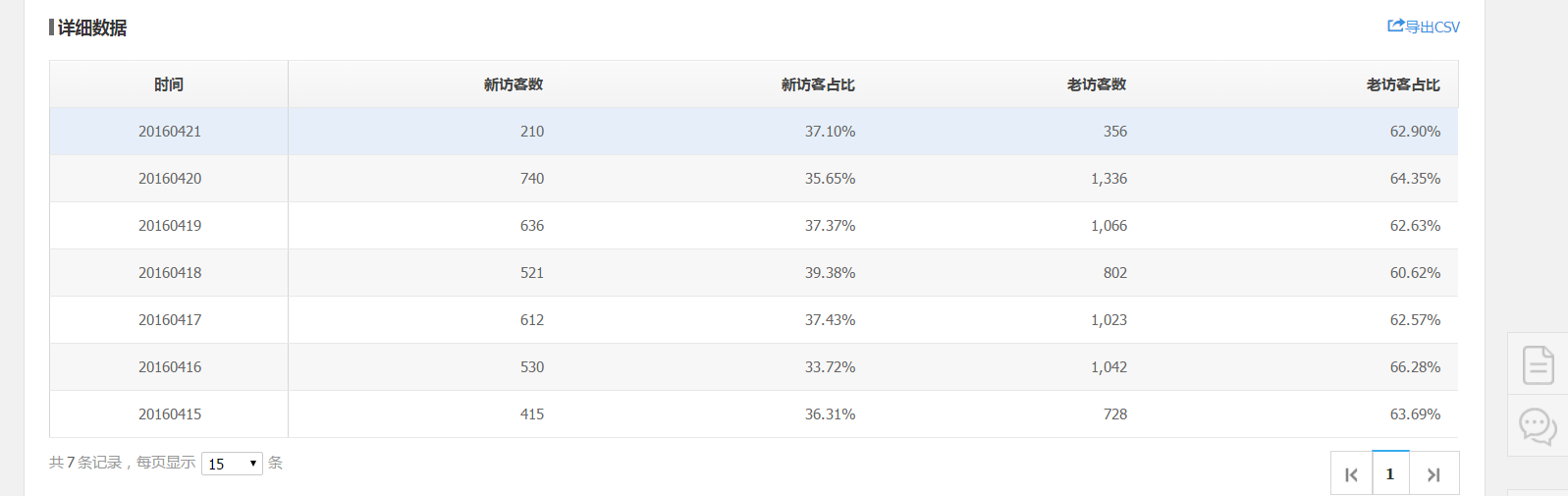


* 新老访客对比：按天显示新老访客的对比情况

页面功能点分析如下：

* 以天为单位显示新老访客的访客比趋势图,天的起始时间段可以设置，以30天为限
* 以表格形式，分页显示新老访客的明细记录

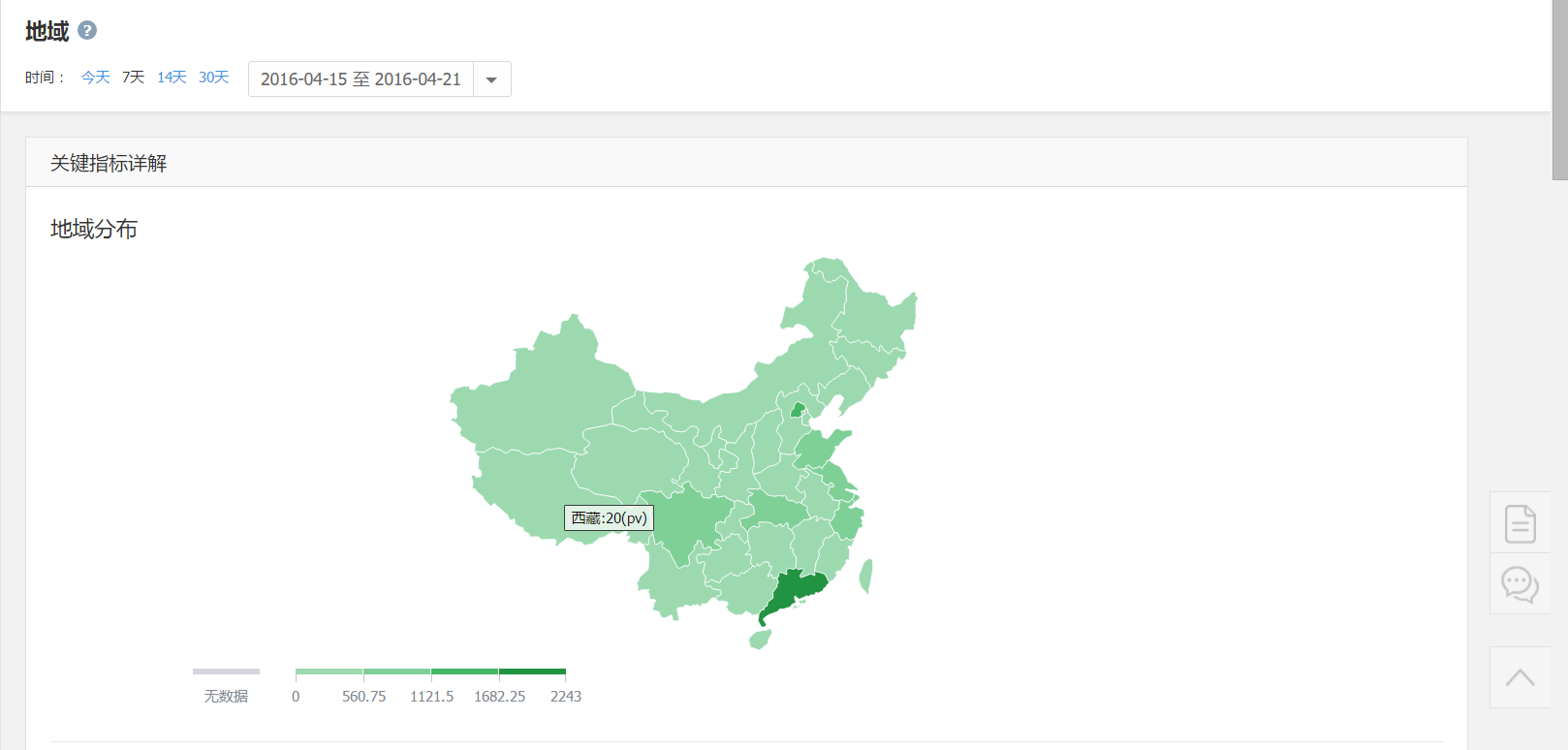




1. **客户端分析**

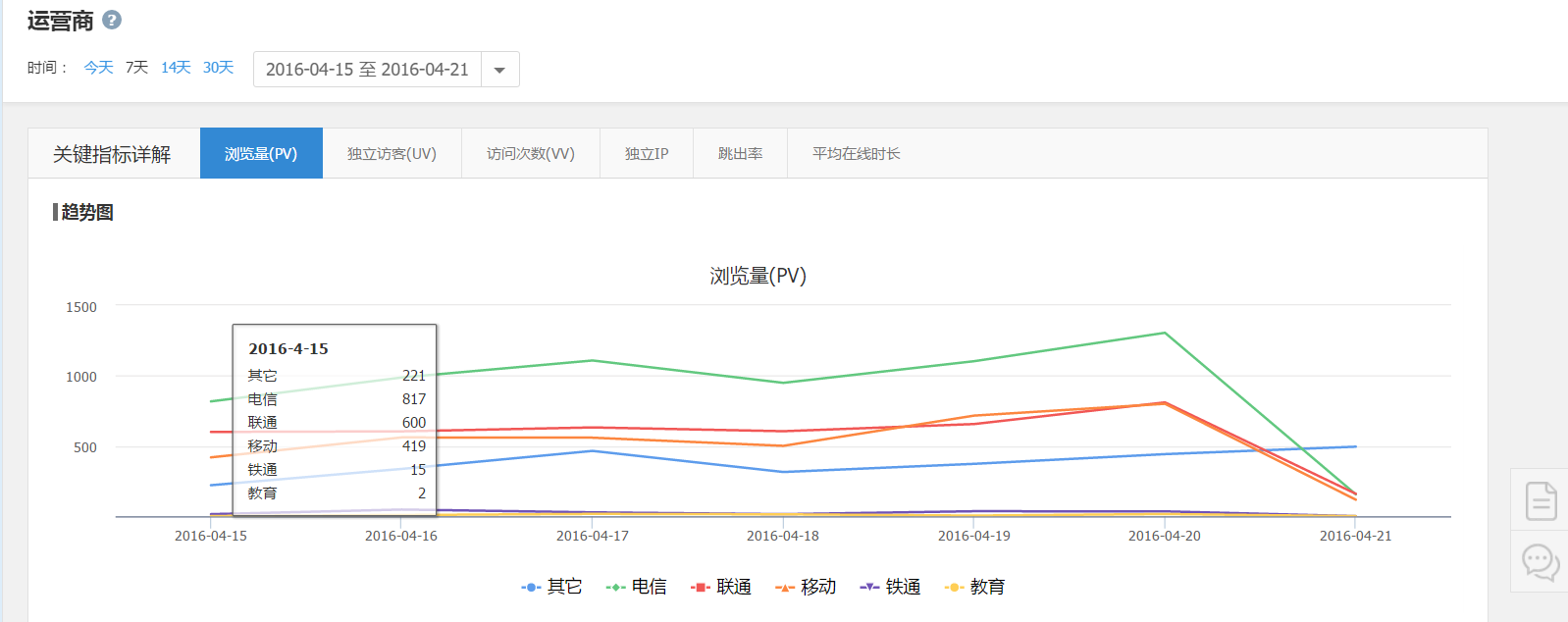
分别以地域（省）、网络运营商、终端访问设备为粒度来分析各个流量指标

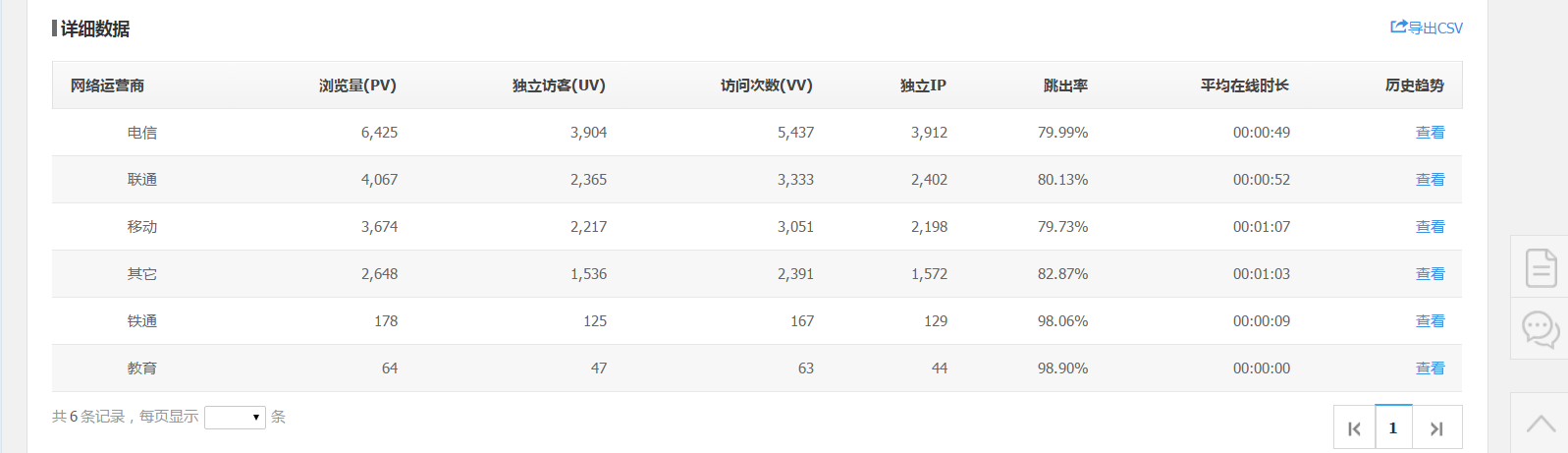
* 地域：以省为粒度按照天来统计PV值



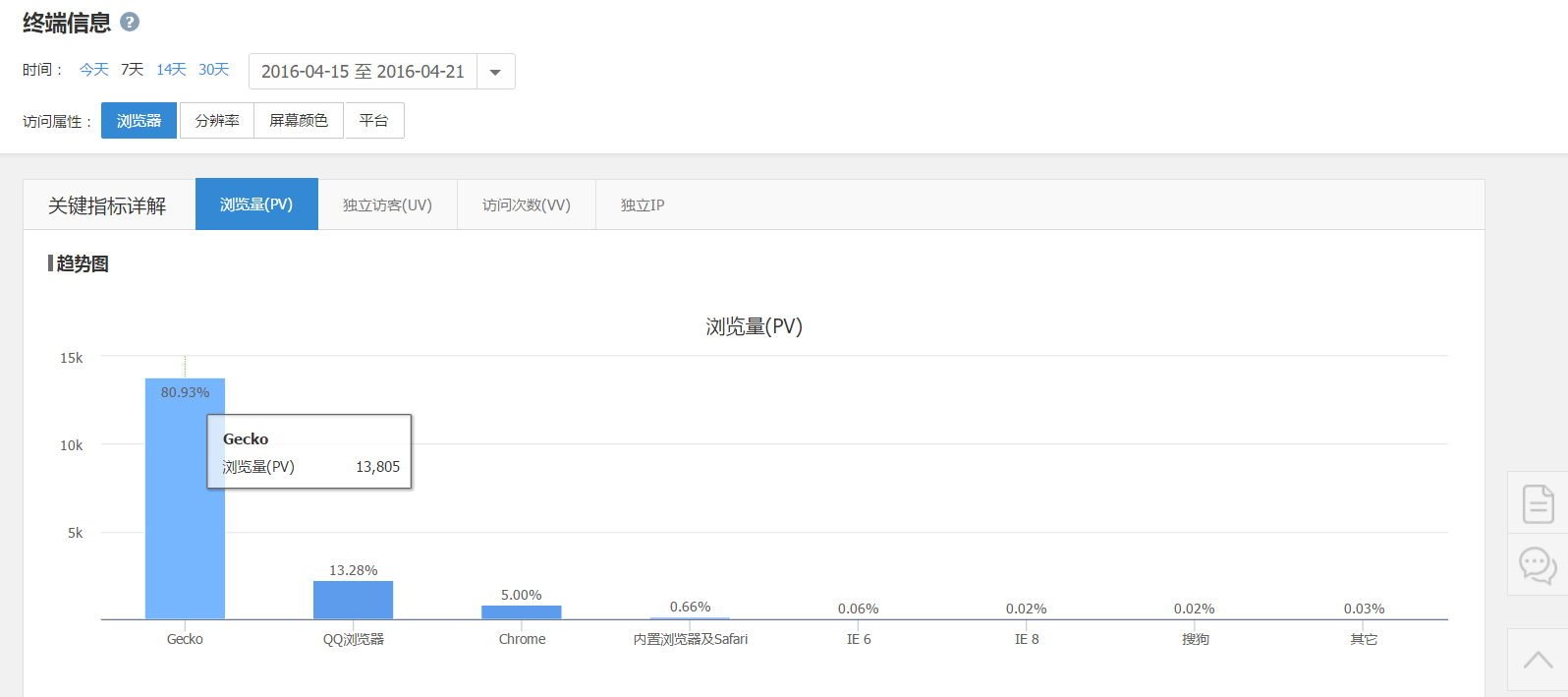


* 运营商：以运营商为粒度按照天来统计流量指标（PV、UV、VV、独立IP、跳出率、平均在线时长)，并能按天进行历史趋势分析





* 终端信息：以浏览器、分辨率、屏幕颜色、平台为粒度按照天来统计流量指标（PV、UV、VV、独立IP、跳出率、平均在线时长），并能按天进行历史趋势分析

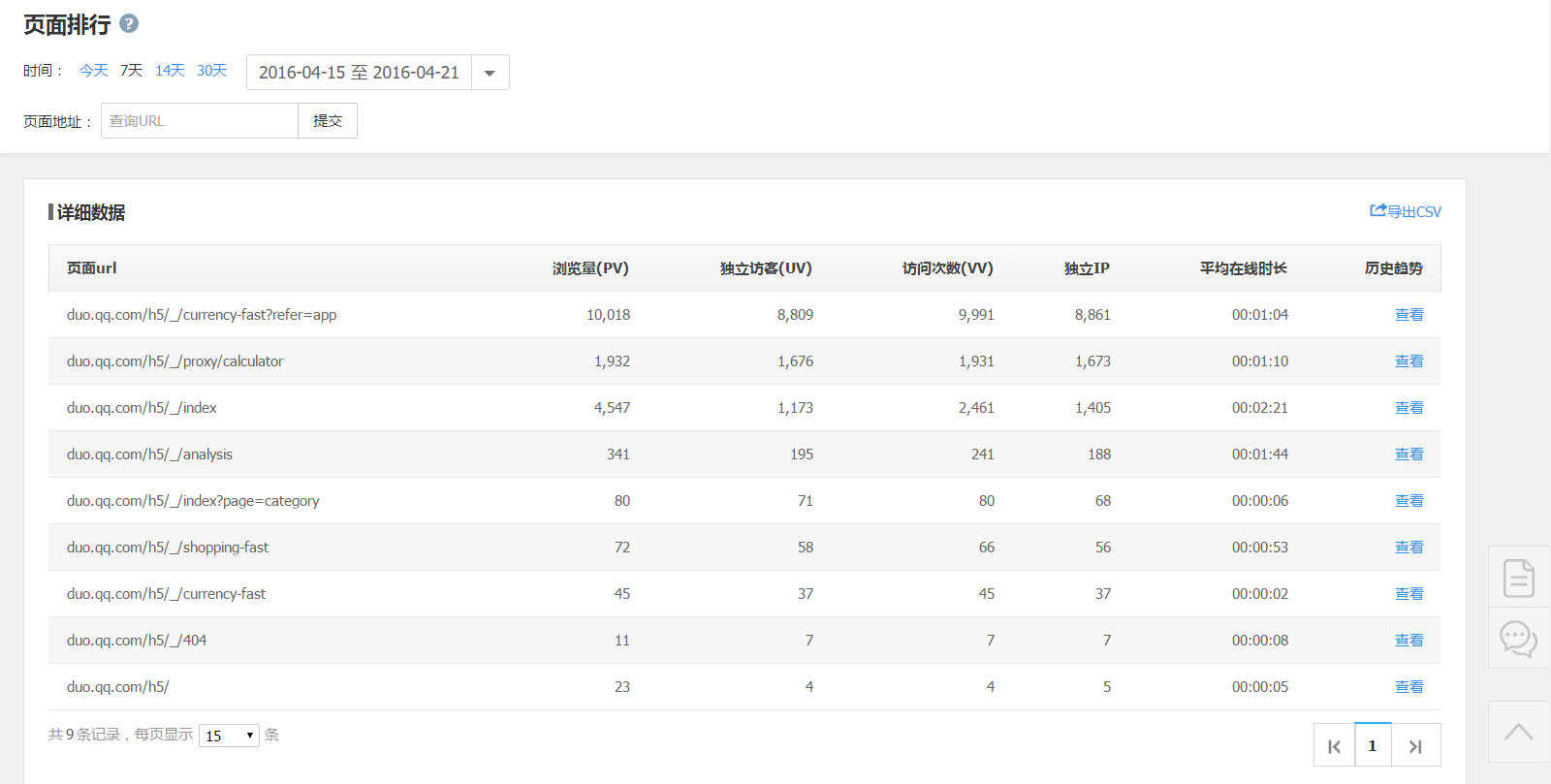




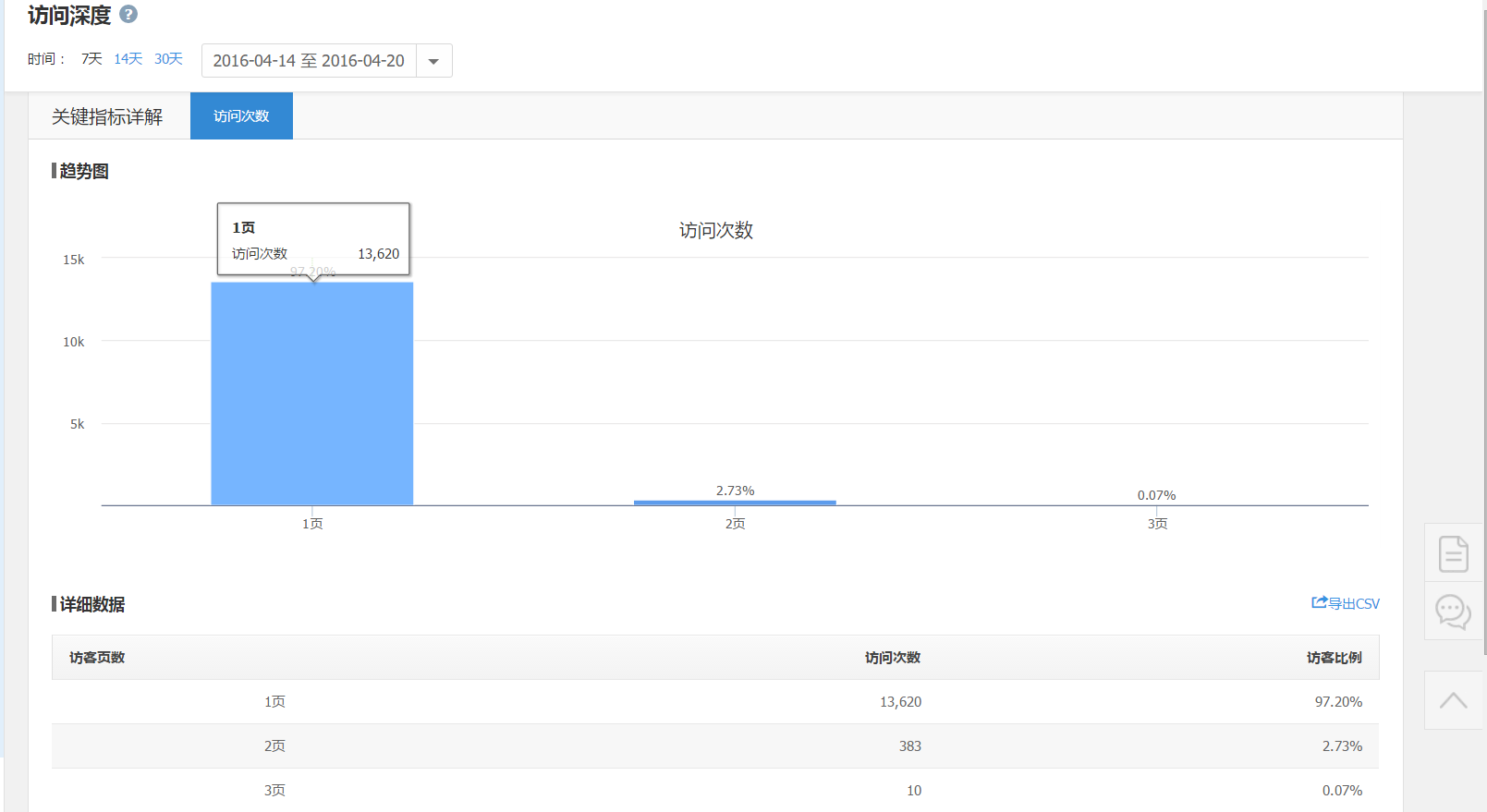
1. **页面分析**

以访问页面和VV为粒度按照天来统计分析各个流量指标

* 页面排行



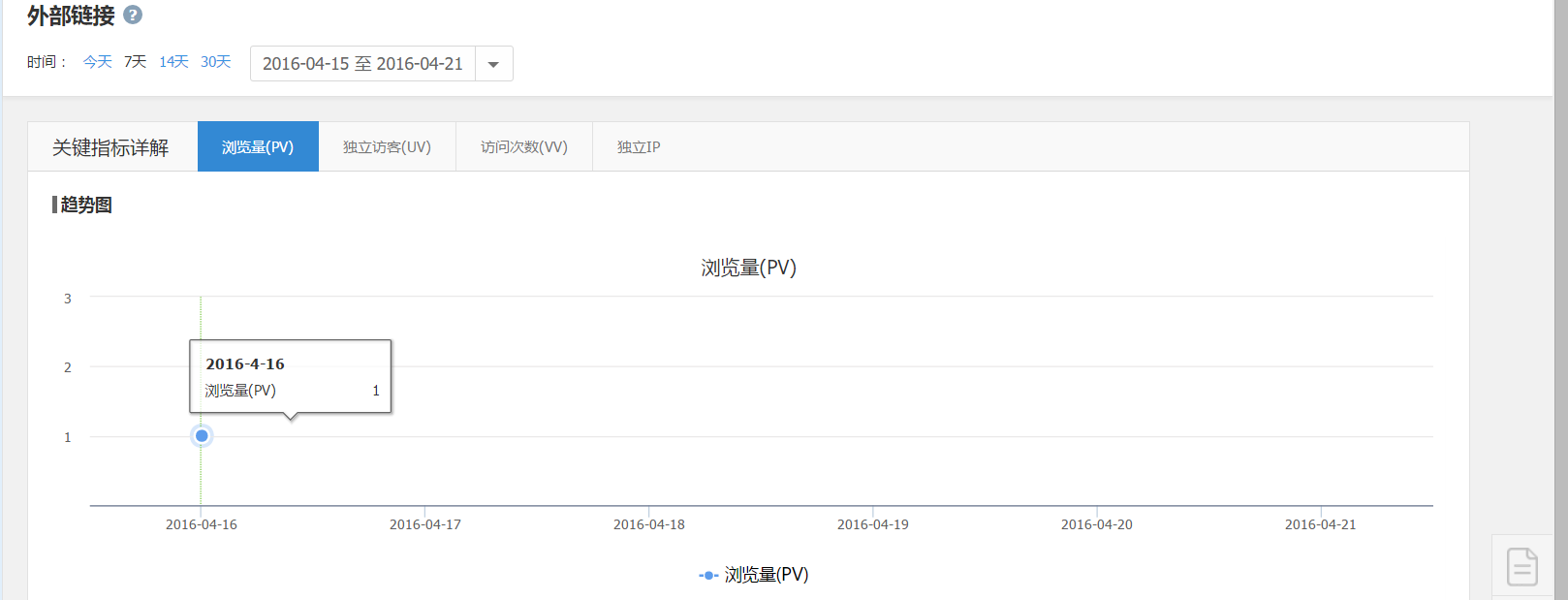
* 访问深度



1. **来源分析**

* 外部链接

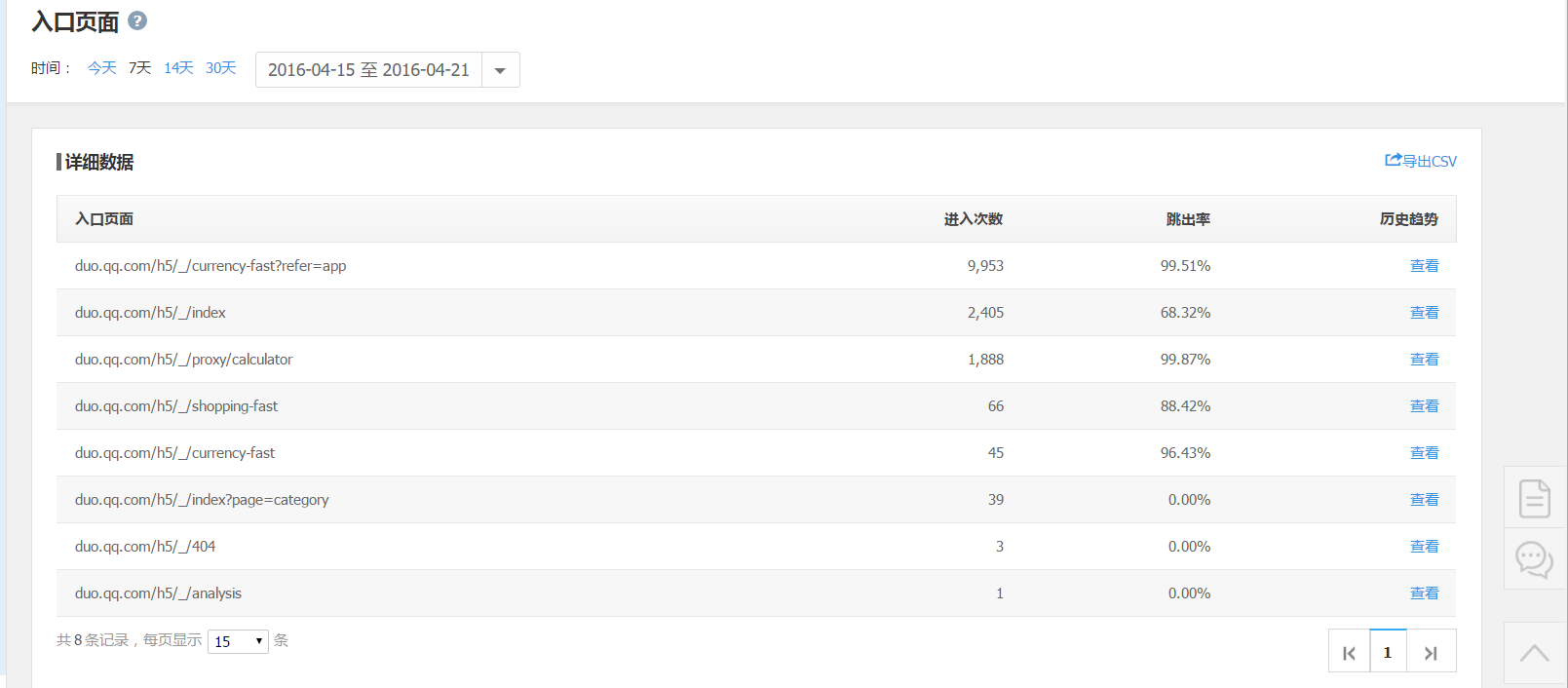
对访问来源中的外部链接的方式为粒度按照天来统计分析各个流量指标





* 入口页面

按照天来分析入口页面的进入次数和跳出率以及历史趋势



* 离开页面

按照天来分析离开页面的离开次数以及历史趋势



1. **数据模型**