用户行为数据分析 项目计划书

2011/5/4

修改记录

版本	修改日期	修改人	修改内容	审核人
V1.01			创建，初稿	

V1.01			数据采集方式分析	
V1.01			数据分析模型 需求描述和示例	
V1.01			网站用户身份识别；web 日志缺陷;漏斗模型	
V1.01			数据分析模型与数据库表的对应关系	
V1.01			WEKA 开源数据挖掘工具	
V1.01			统计分析系统项目周期和项目开发进度 Analysis.mpp	

目录

一、 项目背景

数据挖掘技术是近年来计算机技术发展的热点之一。通过对历史积累的大量数据的有效挖掘，可以发现隐藏的规律或模式，为决策提供支持，而这些规律或模式是不能够依靠简单的数据查询得到，或者是不能在可接受的时间内得到。这些规律或模式可以进一步在专业人员的识别下成为知识。数据挖掘面对的任务是复杂的，通常包括**分类**、**预测**、**关联规则发现**和**聚类分析**等。

企业网站的绩效考评就是指企业网站访问情况的绩效考评，在网络营销评价方法中，网站访问统计分析是重要的方法之一，通过网站访问统计报告，不仅可以了解网络营销所取得的效果，而且可以从统计数字中发现许多有说服力的问题。网站访问量统计分析无论对于某项具体的网络营销活动还是总体效果都有参考价值，也是网络营销评价体系中最具有说服力的量化指标。

销售预测在提高企业的经济效益及决策支持水平方面占有重要的地位。随着企业信息化水平的提高，企业销售数据的日益丰富，管理者对其中隐藏的销售预测信息的渴望日益强烈。用传统的方法来分析这些海量数据中的销售信息非常困难，已不能适应时代的要求。如何找到更好的方法挖掘出销售数据中隐藏的销售预测信息。

二、 相关术语

1. Web 数据挖掘

Web 数据挖掘建立在对大量的网络数据进行分析的基础上，采用相应的数据挖掘算法，在具体的应用模型上进行数据的提取、数据筛选、数据转换、数据挖掘和模式分析，最后做出归纳性的推理、预测客户的个性化行为以及用户习惯，从而帮助进行决策和管理，减少决策的风险。

Web 数据挖掘涉及多个领域，除数据挖掘外，还涉及计算机网络、数据库与数据仓储、人工智能、信息检索、可视化、自然语言理解等技术。

1）Web 数据挖掘分类

Web 数据挖掘可分为四类：Web 内容挖掘、Web 结构挖掘、Web 使用记录挖掘和 Web 用户性质挖掘。

其中，Web 内容挖掘、Web 结构挖掘和 Web 使用记录挖掘是 Web1.0 时代就已经有了的，而 Web 用户性质挖掘则是伴随着 Web2.0 的出现而出现的。

2.1 Web 内容挖掘(WCM，Web Content Mining)

2.2 Web 结构挖掘(WSM，Web Structure Mining)的基本思想是将 Web 看作一个有向图，他的顶点是 Web 页面，页面间的超链就是图的边。然后利用图论对 Web 的拓扑结构进行分析。

2.3 Web 使用记录挖掘(WUM，Web Usage Mining)

Web 使用记录挖掘也叫 Web 日志挖掘或 Web 访问信息挖掘。它是通过挖掘相关的 Web 日志记录，来发现用户访问 Web 页面的模式，通过分析日志记录中的规律，可以识别用户的喜好、满意度，可以发现潜在用户，增强站点的服务竞争力。

Web 使用记录数据除了服务器的日志记录外， 还包括代理服务器日志、浏览器端日志、注册信息、用户会话信息、交易信息、Cookie 中的信息、用户查询、等一切用户与站点之间可能的交互记录。

Web 使用记录挖掘方法主要有以下两种：

(1) 将网络服务器的日志文件作为原始数据，应用特定的预处理方法进行处理后再进行挖掘；

(2) 将网络服务器的日志文件转换为图表，然后再进行进一步的数据挖掘。通常，在对原始数据进行预处理后就可以使用传统的数据挖掘方法进行挖掘。

2.4 Web 用户性质挖掘

Web 用户性质挖掘是伴随着 Web2.0 的出现而出现的。基于 RSS、Blog、SNS、Tag 以及 WiKi 等互联网软件的广泛应用，Web2.0 帮助人们从 Web1. 0 时代各大门户网站“填鸭”式的信息轰炸，过渡到了“人人对话”，每个普通用户既是信息的获取者，也是信息的提供者。[4] 面对 Web2.0 的诞生，Web 数据挖掘技术又面临着新的挑战。

如果说 Web 使用记录挖掘是挖掘网站访问者在各大网站上留下的痕迹，那么 Web 用户性质挖掘则是要去 Web 用户的老巢探寻究竟。在 Web2.0 时代，网络彻底个人化了，它完全允许客户用自己的方式、喜好和个性化的定制服务创造自己的互联网，它一方面给予互联网用户最大的自由度，另一方面给予有心商家有待发掘的高含金量信息数据。通过对 Web 用户自建的 RSS、Blog 等 Web2.0 功能模块下客户信息的统计分析，能够帮助运营商以较低成本获得准确度较高的客户兴趣倾向、个性化需求以及新业务发展趋势等信息。有关 Web2.0 下的数据挖掘正在进一步的研究中。

2) Web 数据的特点

1)**异构数据库环境**。Web 上的每一个站点就是一个数据源，每个数据源都是异构的，因而每一站点的信息和组织都不一样，这就构成了一个巨大的异构数据库。

2)**分布式数据源**。Web 页面散布在世界各地的 Web 服务器上，形成了分布式数据源。

3)**半结构化**。半结构化是 Web 上数据的最大特点。Web 上的数据非常复杂，没有特定的模型描述，是一种非完全结构化的数据，称之为半结构化数据。

4)**动态性强**。Web 是一个动态性极强的信息源，信息不断地快速更新，各站点的链接信息和访问记录的更新非常频繁。

5)**多样复杂性**。Web 包含了各种信息和资源，有文本数据、超文本数据、图表、图像、音频数据和视频数据等多种多媒体数据。

3) 典型 Web 挖掘的处理流程

包括如下四个过程：

1)**查找资源**：根据挖掘目的，从 Web 资源中提取相关数据，构成目标数据集，Web 数据挖掘主要从这些数据通信中进行数据提取。其任务是从日标 Web 数据(包括 Web 文档、电子邮件、电子文档、新闻组、网站日志、网络数据库中的数据等)中得到数据。

2)**数据预处理**：在进行 Web 挖掘之前对“杂质”数据进行过滤。例如消除数据的不一致性；将多个数据源中的数据统一为一个数据存储等。预处理数据的效果直接影响到挖掘算法产生的规则和模式。数据预处理主要包括站点识别、数据选择、数据净化、用户识别和会话识别等。

3)**模式发现**：利用挖掘算法挖掘出有效的、新颖的、潜在的、有用的及最终可以理解的信息和知识。常用的模式发现技术包括：路径分析、关联规则挖掘、时序模式发现、聚类和分类等技术。

4)**模式分析**：利用合适的工具和技术对挖掘出来的模式进行分析、解释、可视化，把发现的规则模式转换为知识。

4) 常用的数据挖掘技术

6.1 路径分析技术

我们通常采用图的方法来分析 Web 页面之间的路径关系。G＝(V，E)，其中：V 是页面的集合，E 是页面之间的超链接集合，页面定义为图中的顶点，而页面闻的超链接定义为图

中的有向边。顶点 v 的人边表示对 v 的引用，出边表示 v 引用了其他的页面，这样形成网站的结构图，从图中可以确定最频繁的访问路径。路径分析技术常用于进行改进站点的结构。如 70%的用户访问 / company / product 时，是从 / company 开始，经过 / company / new / company / products / company / product。此时可以将路径放在比较显著的地方，方便了用户访问，也提高了该产品的点击率。

6.2 关联规则技术

关联规则挖掘技术主要用于从用户访问序列数据库的序列项中挖掘出相关的规则，就是要挖掘出用户在一个访问期限(Session)，从服务器上访问的页面文件之间的联系，这些页面之间并不存在直接的参引(Reference)关系。使用关联规则可以发展很多相关信息或产品服务。例如：某信息 A 和 B，同时被很多用户浏览，则说明 A 和 B 有可能相关。同时点击的用户越多，其相关度就可能越高。系统可以利用这种思想为用户推荐相关信息或产品服务。如当当电子书店就采用了这一模式用以推荐相关书目。当你选择某本图书时，系统会自动给你推荐信息，告知“很多读者在购买此书时还购买的其他书目”。ACM 数字图书馆也采用了这一思想，推出信息推荐服务“Peer to Peer”。

6.3 序列模式挖掘技术

序列模式数据挖掘技术就是要挖掘出交易集之间的有时间序列关系的模式。它与关联挖掘技术都是从用户访问下的日志中寻找用户普遍访问的规律，关联挖掘技术注重事务内的关系，而序列模式技术则注重事务之间的关系。发现序列模式，便于预测用户的访问模式，有助于开展基于这种模式的有针对性的广告服务。依赖于发现的关联规则和序列模式，能够在服务器方动态地创立特定的有针对性的页面。以满足访问者的特定需求。

6.4 聚类分类技术

分类规则可挖掘出某些共同的特性，而这一特性可对新添加到数据库中的数据项进行分类。在 Web 数据挖掘中，分类技术可根据访问用户而得到个人信息、共同的访问模式以及访问某一服务器文件的用户特征。而聚类技术则是对符合某一访问规律特征的用户进行用户待征挖掘。发现分类规则可以识别一个特殊群体的公有属性的描述，这种描述可以用于分类新的检索。如政府机关的用户一般感兴趣的页面是 / company / product。聚类可以从 Web 访问信息数据库巾聚集出具有相似特性的用户群。在 Web 事务日记中聚类用户信息或数据项能够便于开发和执行未来的市场战略。这些事务信息可以用在：在找出用户共同兴趣后，进行合作式信息推荐，共同体的成员町以互相推荐新的滚动信息；自动给一个特定的用户聚类发送销售邮件，为用户聚类动态地改变一个特殊的站点等。

5) Web 商业智能 BI（Business Intelligence）

深入分析访问数据，从访问数据中挖掘财富。

http://www.web-ia.cn/

1、文本挖掘技术和聚类模型分析的网站自动分类；

2、用户访问兴趣聚类；

3、用户等级自动分类；

4、用途分析；

5、新产品推广预测分析和精算分析；

等一系列基于数学模型的 True BI 决策分析工具，帮助企业进行产品 BI 分析、用户 BI 分析、服务质量测评、新产品市场预测与分析等一系列 True BI 服务。

一、 **异常访问分析** 一般情况下，正常的用户访问网站都是通过浏览器（IE、FireFox 等）向网站发送 URL 请求，操作是一个手动平缓的过程。 所谓“异常访问”，是指不是通过浏览器，而是通过程序进行的一个高速机械化的连续 URL 请求过程。这包括不良程序黑客攻击、搜索引擎蜘蛛程序对网站的访问等。 **“异常访问”主要包括 5 个功能：异常访问分析、搜索引擎访问分析、发生错误分析、异常 URL 分析、时段访问分析。** 通过“异常访问分析”，可以让用户发现异常访问行为和访问规律，通过对 URL 请求 频度、服务器处理时间、请求流量等时序图形趋势分析，确定黑客攻击点，排查软件错误、诊断服务器处理能力、网站 Internet 带宽限制“瓶颈”所在点。

二、 **频道关联分析** 频道关联分析应用对象是内容管理者。网站在内容服务层面被抽象为**“频道--子频道--内容”**，组成**“网站结构树”**。数据挖掘的经典故事是**“啤酒和尿布”关联**发现，说的是对某个商场的数据挖掘发现，购买啤酒的人有很多同时购买尿布。关联分析的目的，是发现在一个事物中，各个元素的关联关系，通过关联关系的发现，指导“关系设置”，进而引导事物向有利于管理者主观倾向的方向发展。 Web-DM 中的“频道关联分析”，针对 Web 的具体应用情况，对经典的“关联分析”算法进行了改进，使关联分析速度更快，分析结果也更加有效。 简单的结果可能不能给管理者更多的指导。Web-DM 不仅仅简单地给出关联分析中的“支持度”和“置信度”指标，在此基础上，提出了“置信差”指标，进一步提高关联分析结果的可用性。在给出关联分析技术指标的同时，给出包含关联项的访问 Session，使用户可以更加详细观察和研究关联分析的结果。

三、 **特定关联分析** “频道关联分析”是在内逻辑层面的关联分析，对于“广告”和用户特别关心的 Page 关联分析是网站管理者希望掌握的数据。 哪些 Page 对于广告的贡献有多大？看广告的人更多的看了哪些 Page？ 特别推出的内容与网站的其他 URL 有哪些关联？关联程度如何？ Web-DM 的“特定关联分析”给出深入分析结果，同时以简单直观的形式展示给用户。

以提供新闻或本地新闻为主的门户网站，
管理人员关心网站总体访问情况，整体访问趋势，内容编辑人员关心热门新闻和冷门新闻以及 TOP 排名，
经营人员关心访问者从哪个频道登录网站、从哪个频道的哪个页面离开网站，其访问行为呈现什么规律，
设计人员关心网站频道的如何设置以及页面版面的如何布局，
维护人员关心错误是怎么产生的、如何跳转的、网站是否收到恶意攻击等。

商务网站 主要针对在网站上已经注册的客户群，作为网站的经营者不仅要掌握用户在网上

关心哪些商品，更重要的是要掌握匿名用户怎么变成注册用户，转化率是多少，匿名用户是直接访问的还是通过搜索引擎链接来的，购买行为如何，营业额是多少等。对于电子邮件市场推广，通过沉默用户分析其沉默时间，根据发出量、返回量、成交量来判断市场推广效果。对于广告市场推广，通过曝光量、点击量、成交量来反映市场推广的效果。

2. 网站流量统计

流量统计是什么
是指通过各种科学的方式，准确的纪录来访某一页面的访问者的流量信息，目前而言，必须具备可以统计：统计独立的访问者数量（独立用户、独立访客）；可以统计独立的 IP 地址数量；可以统计页面被刷新的数量。其他附加信息。

3. 统计指标/术语

页面浏览数（page views）
PV(page view)，即页面浏览量，或点击量;通常是衡量一个网络新闻频道或网站甚至一条网络新闻的主要指标。
高手对 PV 的解释是，一个访问者在 24 小时(0 点到 24 点)内到底看了你网站几个页面。这里需要强调:同一个人浏览你网站同一个页面，不重复计算 PV 量，点 100 次也算 1 次。说白了，PV 就是一个访问者打开了你的几个页面。
PV 之于网站，就像收视率之于电视，从某种程度上已成为投资者衡量商业网站表现的最重要尺度。
PV 的计算:当一个访问者访问的时候，记录他所访问的页面和对应的 IP，然后确定这个 IP 今天访问了这个页面没有。如果你的网站到了 23 点，单纯 IP 有 60 万条的话，每个访问者平均访问了 3 个页面，那么 PV 表的记录就要有 180 万条。
影响 PV 的因素：
新闻发布的时间
访问的周期
突发事件

独立访客数 （unique visitor）
UV(unique visitor)：指访问某个站点或点击某条新闻的不同 IP 地址的人数。
在同一天内，UV 只记录第一次进入网站的具有独立 IP 的访问者，在同一天内再次访问该网站则不计数。独立 IP 访问者提供了一定时间内不同观众数量的统计指标，而没有反应出网站的全面活动。

每个访问者的页面浏览数（Page Views per user）

Page Views per user: 这是一个平均数，即在一定时间内全部页面浏览数与所有访问者相除的结果，即一个用户浏览的网页数量。这一指标表明了访问者对网站内容或者产品信息感兴趣的程度，也就是常说的网站“粘性”。

·重复访客者数（repeat visitors）

repeat visitors：重复访问者。是指在一定时期内不止一次访问一个网站的独立用户。

浏览数 Page Views： 网页(含文件及动态网页)被访客浏览的次数。Page View 的计算范围包括了所有格式的网页，例如：.htm、.html、.asp、.cfm、 asa、cdx、htmls、shtm、shtml、txt 等等，可以由用户根据实际情况自己设定。

访问数 Visits： 也称为登陆数，一个登陆是指客户开始访问网站到离开网站的过程。其中：相邻两次点击页面时间间隔在 30 分钟以内（系统默认 30 分钟，用户可以修改默认值）为一次登陆，大于 30 分钟为两次登陆。

用户数 Unique Visitors： 也称为唯一客户数，是指一天内访问本网站的唯一 IP 个数。

点击数 Hits： 是指日志文件中的总记录条数。

停留时间 Visiting Times： 也称为访问时长，是用同一个访问过程中最后一个页面的访问时间减去第一个页面的访问时间，得到此访问在网站上的停留时间。

首页浏览数： 网站首页被访客浏览的次数。

过滤浏览数 Filter Page Views： 网站中的某些页面并不是独立的页面，而是附属于某个页面，如滚动条页面就是附属于首页的页面，用户可以将这些附属页面设置为过滤页面，过滤页面被访客浏览的次数即为过滤浏览数。

有效浏览数 Effective Page Views： 去除过滤页面后的其他所有页面被访客浏览的次数，即有效浏览数=浏览数-过滤浏览数。

平均访问浏览数： 一次访问平均产生的浏览数，即平均访问浏览数=浏览数÷访问数。

重复访问数 Returning Visits during a day： 一天内访问两次以上的用户数。

曝光数： 广告弹出次数。

广告点击数： 用户点击弹出广告的次数，即 Click 数。

返回数： 通过电子邮件进行市场推广时，用户通过点击邮件中的链接地址访问网站的次数。

注册数： 用户通过电子邮件和广告访问本网站，并最终转换为注册用户的数量。

返回率： 广告弹出后，被用户点击的程度，即返回率=点击数÷曝光数×100%。

客户转化率： 客户转化率包含两方面含义：用户通过广告访问本网站，并最终转化成注册用户的程度，即客户转化率=注册数÷点击数×100%；用户通过邮件上的链接地址访问本网站，并最终转化成注册用户的程度，即客户转化率=注册数÷返回数×100%。

发送字节数： 从服务器端向客户端发出的字节数。

接收字节数： 服务器端从客户端接收的字节数。

总字节数： 是发送字节数和接收字节数的总和，即总字节数＝发送字节数＋接收字节数。

行为/路径： 在一个访问过程中，客户访问过的所有页面的轨迹称为路径，或称为行为。

特定行为： 由用户自行定义的行为，包含若干行为步骤，其中行为步骤不受限制，即可以任意设定行为步骤。进而分析出满足设定行为的发生次数及各个步骤之间的转化率。

特定行为转化率： 在特定行为中，两个步骤之间的转化率。

行为入口： 客户开始访问网站的第一个页面。在 Web-IA 中，根据入口给出典型行为分析。

行为出口： 客户访问网站的最后一个页面。在 Web-IA 中，根据出口给出典型行为分析。

沉默时间： 注册用户最后一次访问网站到分析日的天数。

沉默用户： 在沉默时间内未访问网站的注册用户。

重复访问用户比例： 一天内访问两次以上用户占总用户数的比例，该值越大表明用户品质越高，理想值为 100%。

用户粘着度指数： 一天内的总访问数与总用户数之比，该值越大表明用户品质越高。

重度访问用户： 按每次访问的停留时间划分，把停留时间超过 20 分钟的用户归为重度访问用户；也可以按照每次访问产生的浏览数划分，把一次访问浏览超过 10 个页面的用户归为重度访问用户。对于重度访问用户，包括以下四个指标，每个指标值越大，表明用户品质越高。

重度用户比例（次数）=（浏览数≥11 页面的访问数）÷总访问数

重度用户比例（时长）=（>20 分钟的访问数）÷总访问数

重度用户指数=（>20 分钟的浏览数）÷（>20 分钟的访问数）

重度访问量比列=（>20 分钟的浏览数）÷总浏览数

轻度访问用户： 按每次访问的停留时间划分，把停留时间不超过 1 分钟的用户归为轻度访问用户。对于轻度访问用户，包括以下三个指标，每个指标值越小，表明用户品质越高。

轻度用户比例=（0-1 分钟的访问数）÷总访问数

轻度用户指数=（0-1 分钟的浏览数）÷（0-1 分钟的访问数）

轻度访问量比例=（0-1 分钟的浏览数）÷总浏览数

拒绝率： 一次访问只访问一个页面的访问次数占总访问数的比例，比例越小，表明用户品质越高。

拒绝率（一个页面）=只访问 1 个页面的访问数÷总访问数

拒绝率（首页）=只访问首页的访问数÷总访问数

地区： 访问客户的来源地区，是根据 IP 地区对照表，查询访问客户的 IP 地址落在哪个 IP 区段内，而得到其对应的地区。地区包括国内地区和国外地区，国内地区以省为单位，国外地区以国家为单位。

时段： 按照一天 24 个小时自然时间段进行划分。

趋势： 趋势分为两种，第一种是以时段为单位的一天 24 小时发展趋势。第二种是以日为单位的周、月、以及指定区间发展趋势。

IP 地址： IP 地址由 4 个数组成，每个数可取值 0～255， 各数之间用一个点号"."分开，例如： 202.103.8.46。

页面： 网站中的所有格式的网页(含文件及动态网页)，例如：.htm、.html、.asp、.cfm、 asa、cdx、htmls、shtm、shtml、txt 等等，可以由用户根据实际情况自己设定属于页面的文件格式。

特定页面： 对于需要特殊分析的页面，通过设置，从众多页面中独立出来，进行特定分析的页面。

过滤页面： 网站中的某些页面并不是独立的页面，而是附属于某个页面，如滚动条页面就是附属于首页的页面，用户可以将这些附属页面设置为过滤页面。过滤后的浏览数方能真正反映网站的访问情况。

离开页面： 客户访问网站的最后一个页面。

未定义页面： 页面功能没有定义的页面，即没有归类到任何频道的页面。

频道/栏目： 将网站中的各种内容根据功能归类，划分出若干逻辑上的频道或栏目。

网站： 网站是由 Web Server 组成，专业版一个网站只有一个 Web Server，企业版和商务版一个网站至少由一个 Web Server 组成。

热点： 将一个网页中包含的各个链接根据功能归类划分出若干板块，比如新闻板块、财经板块、体育板块、科技板块等，每个板块成为一个热点。进而分析出该页面上的各个热点板块被点击的情况。

汇总： 对多网站的分析进行汇总。
同期比较： 对任意两个日、周、月、以及指定区间的浏览数（或访问数、或用户数、停留时间）进行比较。比较对象可以是页面、频道、栏目、广告、地区等。
聚合： 对日期的聚合，比如周聚合就是将 7 天的数据合在一起为一个分析项，聚合目的就是以聚合项为单位分析网站发展的趋势。
环比： 在趋势分析中，当前日期数据与上一日期数据的比成为环比。
跳转： 状态代码为 302 的访问请求。
热门： 最受欢迎的页面或频道，即浏览数排名前若干位（可由用户自行定义）的页面或频道。
冷门： 最不受欢迎的页面或频道，即浏览数排名后若干位（可由用户自行定义）的页面或频道。
广告： 通过在别的网站上弹出窗口等方式介绍本网站的一种商业活动。
邮件： 通过发送电子邮件，邮件中包含链接地址，吸引用户通过点击邮件中包含的链接地址访问本网站，实际上也是广告的一种。
搜索引擎： 在互联网上为您提供信息"检索"服务的网站。
关键字： 通过搜索引擎"检索"的内容。
Excel 输出： 将分析结果以 Excel 表格形式输出。
网站拓扑结构： 网站的拓扑结构是由网站汇总、网站分析和频道分析三类节点构成。其中，网站汇总下可以有部门汇总，网站分析下可以有子网站，频道分析下可以有子频道。用户根据网站拓扑结构，来查询所需要的分析结果。
匿名用户： 登陆网站不用确认身份，便可访问网站内容的用户。
认证用户： 通过身份认证后，方可访问网站内容的用户。一般情况，用户通过注册成为认证用户。
日志文件： 在 Web-IA 中，日志文件是指被分析网站的工作日志。
浏览器： 客户端通过什么浏览器访问网站。
操作系统： 客户端通过什么操作系统访问网站。
运营商： 客户端接入互联网的服务提供商，比如中国电信、中国网通、教研网等。
接入方式： 客户端接入互联网的方式，比如拨号、专线、ISDN、ADSL 等。
状态代码： 也称作错误代码，是为服务器所接收每个请求（网页点击）分配的 3 位数代码。

4. 用户分析 -- 网站用户的识别

用户分析是网站分析中一个重要的组成部分，在分析用户之前我们必须首先能够识别每个用户，分辨哪些是”New Customer”，哪些是”Repeat Customer”。这样不但能够更加清晰地了解到底有多少用户访问了你的网站，分辨他们是谁（用户 ID、邮箱、性别年龄等）；同时也能够帮助你更好地跟踪你的用户，发现它们的行为特征、兴趣爱好及个性化的设置等，以便于更好地把握用户需求，提升用户体验。

通常当你的网站提供了注册服务，而用户注册并登陆过你的网站，那么用户可以更容易地被识别，因为网站一般都会保存注册用户的详细信息；但是你的网站并不需要注册，而用户的行为以浏览为主，这是用户识别就会显得较为困难，下面提供了几种常用的用户识别的方法：

识别用户的几种方法

当用户并未注册登录的情况下，识别用户的唯一途径就只剩下用户浏览行为的点击流数据，通常情况下它们会保存在 WEB 日志里面。而 WEB 日志本身存在的缺陷可能导致用户识别的不准确性，所以我们在选择用户识别方法的过程中，在条件允许的情况下尽量选择更为准确的方法：

1) 基于 IP 的用户识别

IP 地址是最容易获取的信息，任何的 WEB 日志中均会包含，但其局限性也较为明显：伪 IP、代理、动态 IP、局域网共享同一公网 IP 出口……这些情况都会影响基于 IP 来识别用户的准确性，所以 IP 识别用户的准确性比较低，目前一般不会直接采用 IP 来识别用户。

获取难度：★

准确度：★

2) 基于 IP+Agent 的用户识别

同样基于最简单形式的 WEB 日志，我们可以增加一项——Agent，来提高单一 IP 方式识别用户的准确性。Agent 也是 WEB 日志中一般都会包含的信息，通过 IP+Agent 的方式可以适当提高 IP 代理、公用 IP 这类情况下用户的分辨度，同时通过 Agent 还可以识别网络爬虫等特殊“用户”，但同样准确度也欠高。

获取难度：★

准确度：★★

3) 基于 cookie 的用户识别

当你通过自定义 Apache 日志格式或者 JavaScript 的方法获得用户 cookie 的时候，其实你已经找到了一个更有效的用户识别的手段。cookie 在未被清除的其前提下可以认为是跟某个访问客户端电脑绑定的（一个客户端有可能包含多个 cookie），所以用 cookie 来标识用户其实指的是用户使用的客户端电脑，而并非用户本身。

用 cookie 识别用户的方法当然也存在缺陷：最常见的就是 cookie 被清除而导致用户无法与原先记录实现对应；同时由于客户端电脑会被共用，或者用户会在不同的电脑上访问你的网站，这个时候 cookie 就无法直接对应到该用户了。

获取难度：★☆

准确度：★★☆

4) 基于用户 ID 的用户识别

基于用户 ID 的用户识别是最为准确，因为一般情况下用户不同共享他的用户 ID，所以我们可以认为数据中的 userid 唯一地指向该用户，几乎不存在偏差。当然要使用用户 ID 来识别用户是需要一定的前提条件的：网站必须是提供用户注册登录服务的，并且可以通过一些手段在点击流数据中记录 userid。

获取难度：★★

准确度：★★★

所以对于一个需要用户 ID 注册登录的网站来说，用户唯一标识符的选择可以遵从以下顺序：当用户注册登录时以 userid 为准，当用户在未登录状态浏览时以用户的 cookie 为准，当用户未登录且 cookie 无法获取的情况下以 IP+Agent 为准；这样就能从最大程度上识别唯一用户。

这里推荐一个网站日志中 cookie 项的自定义设置方法，以便更好地识别用户。cookie 是从用户端存放的 cookie 文件记录中获取的，这个文件里面一般在包含一个 cookieid 的同时也会记下用户在该网站的 userid（如果你的网站需要注册登陆并且该用户曾经登录过你的网站且 cookie 未被删除），所以在记录日志文件中 cookie 项的时候可以优先去查询 cookie 中是否含有用户 ID 类的信息，如果存在则将用户 ID 写到日志的 cookie 项，如果不存在则查找是否有 cookieid，如果有则记录，没有则记为”-”，这样日志中的 cookie 就可以直接作为最有效的用户唯一标识符被用作统计。当然这里需要注意该方法只有网站本身才能够实现，因为用户 ID 作为用户隐私信息只有该网站才知道其在 cookie 的设置及存放位置，第三方统计工具一般很难获取。

5) 获取用户信息的途径

通过以上的方法实现用户身份的唯一标识后，我们可以通过一些途径来采集用户的基础信息、特征信息及行为信息，然后为每位用户建立起详细的 Profile：

1) 用户注册时填写的用户注册信息及基本资料；
2) 从网站日志中得到的用户浏览行为数据；
3) 从数据库中获取的用户网站业务应用数据；
4) 基于用户历史数据的推导和预测；
5) 通过直接联系用户或者用户调研的途径获得的用户数据；
6) 有第三方服务机构提供的用户数据。

6) 识别并获取用户信息的价值

通过用户身份识别及用户基本信息的采集，我们可以通过网站分析的各种方法在网站是实现一些有价值的应用：

基于用户特征信息的用户细分；
基于用户的个性化页面设置；
基于用户行为数据的关联推荐；
基于用户兴趣的定向营销；

参考：http://webdataanalysis.net/data-collection-and-preprocessing/

5. WEB 日志的作用和缺陷

Avinash Kaushik 将点击流数据的获取方式分为 4 种：log files、web beacons、JavaScript tags 和 packet sniffers，其中包嗅探器（packet sniffers）比较不常见，最传统的获取方式是通过

WEB 日志文件（log files）；而 beacons 和 JavaScript 是目前较为流行的方式，**Google Analytics 目前就是采用 beacons+JavaScript 来获取数据的**，我们可以来简单看一下传统的网站日志和 beacons+JavaScript 方式各自的优缺点：

1) WEB 日志文件

优势：简单方便，不需要修改网页代码，可以自定义日志格式；较多的现成的日志分析工具的支持（AWStats、Webalizer 等）；获取网络爬虫数据的唯一途径；可以收集底层数据供反复的分析。

缺陷：数据的质量较低，网站日志包含所有日志数据，包括 CSS、图片、脚本文件的请求信息，所以过滤和预处理来提升数据质量必不可少；页面缓存导致浏览无日志记录，这个是比较致命的。

2) beacons+JavaScript

优势：只需要在页面代码中操作，不需要配置服务器；数据的获取有较高的可控性，可以只在需要统计的页面植入代码；能够获取点击、响应等数据；不需要担心缓存等的影响，数据的准确度较高；可用第三方 cookie 实现多网站跟踪比较。

缺陷：当浏览器禁止接收图片或者禁用 JS 时，都可能导致数据获取的失败；只在应用服务层操作，无法获取后台的数据；对图片、文件等请求信息的获取难度相对较大；过多地 JS 可能导致页面性能的下降，虽然这方面的影响一般可以忽略。

无论通过何种方式，最终数据都是通过日志文件来记录的，只是通过 JS 可以更容易控制想要获取的数据，并通过在 URL 带参数的方式记录到日志文件中共解析和统计。所以底层的数据形式无非就是记录在日志文件中的那几项，在 WEB 日志格式一文中，已经对网站日志的类型和组成做了基本的介绍，这里就再来解析下 WEB 日志中各项对网站数据分析的作用，以及存在的不确定性和缺陷。

3) 日志的不准确性

WEB 日志在技术层面的获取方式及各类外部因素的影响使基于网站日志的数据分析会存在许多的不准确性，下面来介绍下 WEB 日志中那些项目可能造成数据的不准确，以及造成这些缺陷的原因。

a) 客户端的控制和限制

由于一些浏览网站的用户信息都是有客户端发送的，所以用户的 IP、Agent 都是可以人为设置的；另外 cookie 可以被清理，浏览器出于安全的设置，用户的可以在访问过程中限制 cookie、referrer 的发送。这些都会导致用户访问数据的丢失或者数据的不准确，而这类问题目前很难得到解决。

b) 缓存

浏览器缓存、服务器缓存、后退按钮操作等都会导致页面点击日志的丢失及 referrer 的丢失，目前主要的处理方法是保持页面信息的不断更新，可以在页面中添加随机数。当然如果你使用的 JavaScript 的方法，那么就不需要担心缓存的问题。

c) 跳转

一些跳转导致 referrer 信息的丢失，致使用户的访问足迹中断无法跟踪。解决方法是将 referer 通过 URL 重写，作为 URL 参数带入下一页面，不过这样会是页面的 URL 显得混乱。

d) 代理 IP、动态 IP、局域网（家庭）公用 IP

IP 其实准确性并不高，现在不止存在伪 IP，而且局域网共享同一公网 IP、代理的使用及动态 IP 分配方式，都可能使 IP 地址并不是与某个用户绑定的，所以如果有更好的方法，尽量不要使用 IP 来识别用户。

e) session 的定义与多 cookie

不同的网站对 session 的定义和获取方法可能差异，比如非活动状态 session 的失效时间、多进程同时浏览时 sessionid 的共享等，所以同一个网站中 session 的定义标准必须统一才能保证统计数据的准确。cookie 的不准确一方面是由于某些情况下 cookie 无法获取，另一方面是由于一个客户端可以有多个 cookie，诸如 chrome、Firefox 等浏览器的 cookie 存放路径都会与 IE 的 cookie 存放路径分开，所以如果你是用不同的浏览器浏览同一网站，很有可能你的 cookie 就是不同的。

f) 停留时间

停留时间并不是直接获取的，而是通过底层日志中的数据计算得到的，因为所有日志中的时间都是时刻的概念，即点击的时间点。这里不得不提的是一个 session 的最后一个页面的停留时间是无法计算得到的，可以来看一下停留时间的计算过程：

假设一个用户在一个 session 里面依次点击了 A->B->C 这 3 个页面，并在点完 C 之后关闭了浏览器，或者长时间的禁止导致了 session 的中断。那么我们可以从日志中获得的数据为 3 个页面的点击时间（HitTime），假设 A、B、C 点击时间分别为 HTA、HTB、HTC，那么 A 和 B 页面的停留时间（StayTime）就可以通过计算得到：STA= HTB-HTA，STB= HTC-HTB，而因为我们无法获取 session 结束的时间，所以 STC 是无法通过计算得到的，所以一般 session 最后页面的停留时间是 0，而 session 得停留时间，即一次访问的时间（Time on site）是 HTC- HTA，其实是从打开第一个页面到打开最后一个页面的时间间隔，也是不准确的。

另外，我们也无法获知用户在浏览一个页面的时候到底做了什么，是不是一直在阅读博客上的文章或者浏览网站上展示的商品，用户也有可能在期间上了个厕所、接了通电话或者放空的片刻，所以计算得到的停留时间并不能说明用户一直处于 Engagement 的状态。

参考：http://webdataanalysis.net/data-collection-and-preprocessing/effect-of-weblog/

6. 漏斗模型（Funnel Model）

漏斗模型不仅显示了用户在进入流程到实现目标的最终转化率，同时还可以展示整个关键路径中每一步的转化率。

单一的漏斗模型对于分析来说没有任何意义，我们不能单从一个漏斗模型中评价网站某个关键流程中各步骤的转化率的好坏，所以必须通过趋势、比较和细分的方法对流程中各步

骤的转化率进行分析：

趋势（Trend）：从时间轴的变化情况进行分析，适用于对某一流程或其中某个步骤进行改进或优化的效果监控；
比较（Compare）：通过比较类似产品或服务间购买或使用流程的转化率，发现某些产品或应用中存在的问题；
细分（Segment）：细分来源或不同的客户类型在转化率上的表现，发现一些高质量的来源或客户，通常用于分析网站的广告或推广的效果及 ROI。

所以，漏斗模型适用于网站中某些关键路径的转化率的分析，以确定整个流程的设计是否合理，各步骤的优劣，是否存在优化的空间等。试着去了解用户来你的网站的真正目的，为他们提供合理的访问路径或操作流程，而不是一味地去提高转化率。

7. 目前提供此服务产品/企业

北京蓝太平洋科技开发有限公司 http://www.webdss.com/
（目前公司就购买的此产品 IIS 日志分析）

般若网络科技有限公司 http://www.web-ia.cn/ Web 商业智能 Bi，深入分析访问数据，从访问数据中挖掘财富。

WEKA 怀卡托智能分析环境（Waikato Environment for Knowledge Analysis） 开源软件。
官方网址：http://www.cs.waikato.ac.nz/ml/weka/
WEKA 作为一个公开的数据挖掘工作平台，集合了大量能承担数据挖掘任务的机器学习算法，包括对数据进行预处理，分类，回归、聚类、关联规则以及在新的交互式界面上的可视化。
WEKA 提供的接口文档可以实现自定义的数据挖掘算法。

三、 项目目的

四、 项目需求

1. 页面统计

页面 PageView 访问数
页面 Ref 跳入
页面 UserView 用户数

页面 IP 独立 IP 数
页面 Out 跳出

2. 用户行为指标

用户行为指标主要反映用户是如何来到网站的、在网站上停留了多长时间、访问了那些页面等，主要的统计指标包括：

- 用户在网站的停留时间；
- 用户来源网站（也叫“引导网站”）；
- 用户所使用的搜索引擎及其关键词；
- 在不同时段的用户访问量情况等。
- 用户区域分布
- 用户对在线支付功能的使用情况
- 用户对在线教室功能的使用情况

• 用户浏览网站的方式 时间 设备、浏览器名称和版本、操作系统

用户浏览网站的方式相关统计指标主要包括：

- 用户浏览器的名称和版本；
- 用户浏览器的版本分布；
- 访问者电脑分辨率显示模式；
- 用户所使用的操作系统名称和版本；
- 用户所在地理区域分布状况等。

3. 潜在用户特征分析

（ / 网易邮箱广告 分区域/分学历/分职业投放 ）

1、最常浏览该网站的用户性别分布
2、最常浏览该网站的用户年龄分布
3、最常浏览该网站的用户学历分布
4、最常浏览该网站的用户婚姻状况分布
5、最常浏览该网站的用户职业分布
6、最常浏览该网站的用户收入分布
7、最常浏览该网站的用户区域分布

最后：那些人是我们潜在的用户？

首页
有多少 PV 数？UV 数？有多少独立 IP 访问？ 都从那些网址跳入？用户又跳出到哪里去了？

频道/栏目首页 - 网站栏目分类的优势
有多少 PV 数？UV 数？有多少独立 IP 访问？ 都从那些网址跳入？用户又跳出到哪里去了？

新/旧功能页面 - 听课行为 / 记笔记行为
有多少 PV 数？UV 数？有多少独立 IP 访问？ 都从那些网址跳入？用户又跳出到哪里去了？

用户都是用那个页面进入到听课功能页面的？是否和我们设定/假象的用户行为一致？

4. 指定 User Cookie 的分析

用户浏览轨迹特征；
用户页面时长特征；
用户操作轨迹特征；

新学员/老学员的操作行为差异？
功能选择差异？
浏览/关注点差异？

5. 用户趋势分析

用户浏览的趋势；
使用产品的趋势；

五、 项目系统设计

如何获取流量统计信息

获取网站访问统计资料通常有两种方法：一种是通过在自己的网站服务器端安装统计分析软件来进行网站流量监测；另一种是采用第三方提供的网站流量分析服务。两种方法各有利弊，采用第一种方法可以方便地获得详细的网站统计信息，并且除了访问统计软件的费用之外无需其他直接的费用，但由于这些资料在自己的服务器上，因此在向第三方提供有关数

据时缺乏说服力；第二种方法则正好具有这种优势，但通常要为这种服务付费，虽然也有一些免费网站流量统计服务，但由于在功能方面会有一定的限制，或者通常需要在网站上出现服务商的标识甚至广告，对于商业网站来说使用免费服务肯那个不太合适。此外，如果必要，也可以根据需要自行开发网站流量统计系统。具体采取哪种形式，或者哪些形式的组合，可根据企业网络营销的实际需要决定。

在线流量统计网站有哪些？

Google 网站访问统计（Google Analytics）：http://www.google.com/analytics/

ITSUN 网站流量统计：http://www.itsun.com

51yes 网站流量统计：http://count.51yes.com

六、 项目详细设计

1. 数据收集

目前日志服务/ 格式：

IIS： 主站 http://www.chinaacc.com （ ASP 语言 ）
Negios： 论坛 http://bbs.chinaacc.com （ PHP 语言 ）
Apache/Tomcat： 博客 http://blog.chinaacc.com （ PHP 语言/ Java / JSP 语言）
SysLog： (C 语言)
其他格式日志：自定义的日志格式

Apache 服务器:

mod_uid : http://www.lexa.ru/programs/mod-uid-eng.html
mod_usertrack :
http://httpd.apache.org/docs/2.0/mod/mod_usertrack.html
http://www.chedong.com/blog/archives/001077.html

Nginx 服务器:

http://wiki.nginx.org/NginxHttpUserIdModule

数据的收集方式：

1、 在各个应用上通过 JS 程序收集数据，统一访问日志服务器，记录日志，做日志分析；
（推荐：将数据服务和数据捕获分离、数据格式统一）

2、 由各个应用服务器配置日志，保持格式的简单一致性，再汇总日志做分析；

日志包含信息：

时间	访问 URL	来访 IP 地址	来访来源	用户唯一标识
Date	URL	IP	Ref	UserCookieID

统计时间范围为： 每小时的 00 分钟 - 59 分钟

注：要考虑用户操作的时间的不确定性。用户可能在 25 分-下一小时的 10 分在操作，以绝对的时间范围来分析，分析数据会有偏差。

2. 数据模型

1) 统计 PV 量(趋势)

统计要素：Date + URL

描述： 统计时间范围内 Date，访问 URL 的浏览量 PV（汇总数）；

例如：09:00 - 10:00 之间，访问博客首页 http://blog.chinaacc.com 的有 259 次；

数据库表：log_date_collect "汇总统计站点的每小时/30 分钟的 pv/ 独立 ip/ uv"

2) 消重 统计独立 IP 量 / IP 的平均访问页面量(趋势)

统计要素：Date + URL + IP

描述：统计时间范围内 Date，访问 URL 的这一 IP，访问页面次数（更详细记录）；所有次数的总和大约等于 PV；

例如： 09:00 - 10:00 之间，访问博客首页 http://blog.chinaacc.com 的 IP 为 192.168.1.102 有 5 次访问；

描述：多个 IP 的平均访问页面数 = IP 访问页面次数总和/IP 总数

描述：统计时间范围内 Date，访问 URL 的 IP 数（汇总数）；统计时间范围内 Date，一个 IP 的多次访问只记算为一次（消除重复 - 消重处理）；

例如：09:00 - 10:00 之间，访问博客首页 http://blog.chinaacc.com 的独立 IP 有 212 个；

数据库表：log_date_collect "汇总统计站点的每小时/30 分钟的 pv/ 独立 ip/ uv"

3) 消重 统计独立 UV 量 / UV 的平均访问页面量(趋势)

统计要素：Date + URL + UserCookieID

描述： 统计时间范围内 Date，访问 URL 的这一用户 UserCookieID，访问页面次数（更详细记录）；所有次数的总和大约等于 PV；

例如：09:00 - 10:00 之间，访问博客首页 http://blog.chinaacc.com 的用户 UserCookieID，访问了 5 次；

描述：统计时间范围内 Date，访问 URL 的这一用户 UserCookieID 数（汇总数）；统计时间范围内 Date，一个 UserCookieID 的多次访问只记算为一次（消重处理）；

例如：09:00 - 10:00 之间，访问博客首页 http://blog.chinaacc.com 的 UserCookieID 为 190 个。

数据库表：log_date_collect "汇总统计站点的每小时/30 分钟的 pv/ 独立 ip/ uv"

4) 统计 URL 的访问来源 Ref 的量 / Ref 排行(趋势)

统计要素：Date + URL + Ref

描述： 统计时间范围内 Date，访问 URL 页面的是从那些页面 Ref 跳入，跳入量统计（更详细记录）；

例如：09:00 - 10:00 之间，访问博客首页 http://blog.chinaacc.com，来源是从 bbs 点击过来的有 43 次；

描述：跳入排行

例如：最多的是用 bbs 点击过来来访问博客首页的，排行第一，43 次；

5) 统计 Ref=URL 的去访 URL*/跳出的量 / 去访/跳出排行(趋势)

统计要素：Date + URL* + Ref=URL

描述：统计时间范围内 Date，访问 URL 页面的人，都又去了哪些页面 Out 跳入，跳入量统计（更详细记录）；

例如：09:00 - 10:00 之间，访问博客首页 http://blog.chinaacc.com，点击去访问 bbs 的有 68 次点击；

描述：跳出排行

例如：从博客首页点击去访问 bbs 的最多，排行第一，有 68 次；

6) 统计分析/预测/规律 特定用户的行为(趋势)

统计要素：UserCookieID + URL + Date

描述：根据用户的访问历史记录(更详细记录)来总结规律

例如：09:00 － 10:00 之间，用户 UserCookieID 访问记录列表，统计/预测/聚类/分类，做“啤酒和尿布”的规则整理。

7) 统计新访客/老访客(趋势)

统计要素：UserCookieID + Date

描述： 统计新访问用户，统计昨天也访问的用户(每天的 UserCookieID 的消重记录，详细记录)

例如：对比昨天的 00:00 －24:00 和今天的 00:00 － 24:00 之间，UserCookieID 的重复出现次数（老用户），第一次出现为新用户。

8) 页面平均停留时间 / 页面平均时长 (趋势)

统计要素： UserCookieID + URL

描述： 用户 UserCookieID 在页面 URL 的停留时间（更详细记录），用户在那些页面上停留时间最长

例如：用户 UserCookieID-1 在我的网校我的家 听课页面上停留了 1:30 小时

描述：在页面 URL 上的停留时间最长的排行

例如：在全部页面中，在听课页面停留时间最长

描述：计算用户 UserCookieID 在页面 URL 的停留时间的平均值(汇总计算值)

例如：09:00 － 10:00 之间，访问博客首页 http://blog.chinaacc.com 的

9) 搜索引擎列表

统计要素：Ref

描述： 统计外部网站（搜索引擎-Google/baidu/soso/youdao/sogou + 其他站点 sina/163/sohu/QQ）跳入统计站点的 pv 量 / IP 量
例如： 从百度跳入统计站点 269 次/独立 ip 100 个

描述： 外部站点跳入排行
例如：从百度跳入统计站点的 pv 最多排名第一； 从 sina 跳入统计站点的独立 IP 最多排名第一；

10) 搜索引擎关键词

统计要素：Ref 的关键词参数
描述：从全部搜索引擎网站跳入统计站点的关键词参数 统计跳入次数
例如：关键词“会计考试” 排行第一 跳入 500 次

11) 搜索引擎关键词(各搜索引擎)

统计要素：Ref 的关键词参数(区分搜索引擎)
描述：从区分搜索引擎网站跳入统计站点的关键词参数 统计跳入次数
例如：百度关键词“会计考试” 排行第一 跳入 500 次
Google 关键词“CPA”排行第一 跳入 423 次

12) 老用户回头率（用户黏性）

统计要素：UserCookieID + Date
描述：最近一月的用户 UserCookieID 在最近是否有访问网站，占比多少
例如：最近一月的用户 UserCookieID 有 34%的回头率，就是说有大概 34%的用户会在最近一个月里有再次访问网站的动作

13) 新增用户增加/流失（用户黏性）

统计要素：UserCookieID + Date
描述：最近一月的用户 UserCookieID 在第一次访问网站的用户，占比多少，
例如：最近一月的用户 UserCookieID 有 66%的新增用户率，就是说有大概 66%的用户在最近一个月里是第一次访问网站

描述：最近一月的用户 UserCookieID 是否有在第二个月里有访问网站，占比多少，
例如：最近一月的用户 UserCookieID 有 53%的新增用户流失，就是说有大概 53%的用户在上个月访问过网站后，在这个月里没有访问网站

14) 不活跃用户激活（用户黏性）

统计要素：UserCookieID + Date
描述：用户访问网站时间间隔较长 / 用户访问网站页面较少
例如：用户进入网站后，待了一会就跳出或者关闭了网页；
用户进入网站后，访问了少量页面

15) 用户浏览深度（用户黏性）

统计要素：UserCookieID + URL
描述：统计用户网页访问路径 Path 的深度
例如：例如用户访问首页 - 访问频道页 - 访问二级栏目页 - 发表评论

16) 用户访问兴趣分析（用户黏性）

统计要素：Date + UserCookieID + URL
描述：统计一段时间内，全部用户访问网页的重合度
例如：例如一天内，全部的来访用户，大概有 56%用户访问了”我的网校我的家”页面

17) 性别结构(访客特征分析)

统计要素：性别
描述：男女比例
例如：

18) 年龄结构(访客特征分析)

统计要素：年龄分层
描述：年龄各个阶段占比
例如：

19) 学历结构(访客特征分析)

统计要素：学历分层
描述：学历各个层次占比
例如：

20) 收入结构(访客特征分析)

统计要素：收入分层
描述：收入各个层次占比
例如：

21) 操作系统类型(客户端信息)

统计要素：User-Agent
描述：操作系统占比
例如：

22) 操作系统语言(客户端信息)

统计要素：User-Agent
描述：操作系统语言占比
例如：

23) 操作系统时区(客户端信息)

统计要素：User-Agent
描述：操作系统时区占比
例如：

24) 浏览器(客户端信息)

统计要素：User-Agent
描述：用户使用的浏览器占比
例如：

25) 显示器颜色(客户端信息)

统计要素：User-Agent
描述：用户显示的颜色占比
例如：

26) 屏幕分辨率(客户端信息)

统计要素：User-Agent
描述：用户使用的屏幕分辨率占比
例如：

27) 国家/省份 - 地址位置(客户端信息)

统计要素：User-Agent
描述：用户 IP 所在的地区占比
例如：

28) 城市 - 地址位置(客户端信息)

统计要素： User-Agent
描述：用户 IP 所在的城市占比
例如：

29) 接入商(客户端信息)

统计要素：User-Agent
描述：用户访问网站使用的接入商占比
例如：

30) 场所(客户端信息)

统计要素：User-Agent
描述：用户访问网站的场所占比
例如：

3. 数据处理

4. 数据展示

1）参考网站

可访问一下链接，查看统计实例。
Google 网站分析： https://www.google.com/analytics/
百度统计：http://tongji.baidu.com
百度统计 功能说明：
http://support.baidu.com/tongji/?module=default&controller=index&action=detail&nodeid=2602
51 啦 统计： http://www.51.la/about.asp
51yes 统计： http://demo.51yes.com/all.aspx
CNZZ 数据专家统计： http://new.cnzz.com/v1/main.php?siteid=2799&s=main_stat

点击量： 记录每一小时的 IP 数和 PV 数，提供多种形式供用户对任意时间段进行查询。IP 数完全基于 24 小时 IP 防刷。

客户端： 记录来访者所处的地区、访问者的浏览器、操作系统、语言、时区、屏幕尺寸、屏幕色彩、IP 地址及 Alexa 安装情况，并可对这些数据按任意时间段查询。

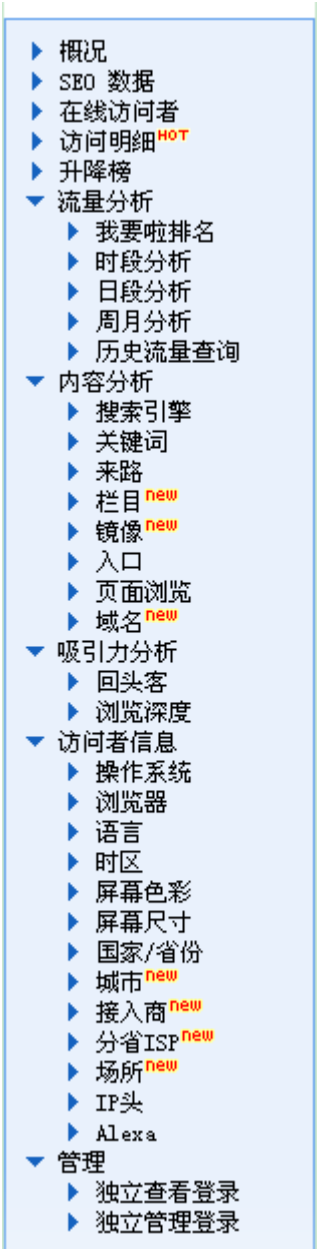
流量源： 记录点击来源，并根据来源对关键词和搜索引擎进行分析。可对来路信息按时间段和特征字查询，提供多种排序方式。

关键词： 精确的辨别并记录各大搜索引擎搜索进入时用户所搜索的关键词，兼容各种编码格式，无乱码，可按时间段和特征字查询分析，提供多种排序方式。

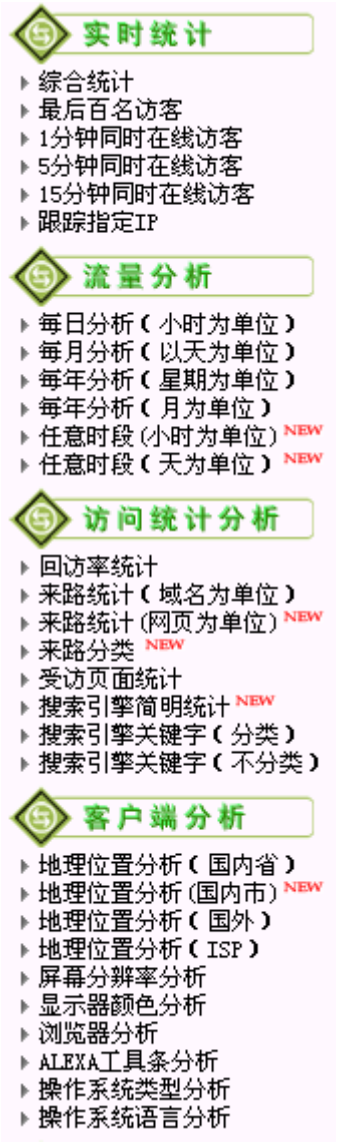
被访页： 记录用户进入网站时的网页被进入的次数（入口网址）和每个网页被浏览的次数。可按时间段和特征字查询，提供多种排序方式。

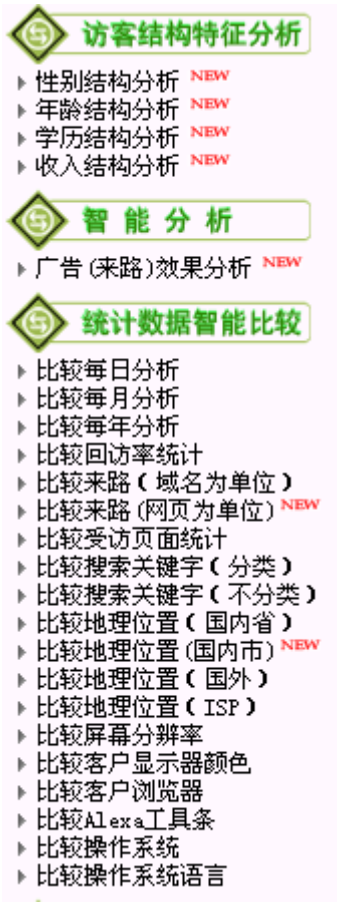
明 细： 访问明细和在线用户栏目细致到用户的全部信息，并可追踪任一用户的浏览记录。

1. 51 啦统计 功能菜单：


概况
SEO 数据
在线访问者
访问明细HOT
升降榜
流量分析
我要啦排名
时段分析
日段分析
周月分析
历史流量查询
内容分析
搜索引擎
关键词
来路
栏目new
镜像new
入口
页面浏览
域名new
吸引力分析
回头客
浏览深度
访问者信息
操作系统
浏览器
语言
时区
屏幕色彩
屏幕尺寸
国家/省份
城市new
接入商new
分省ISPnew
场所new
IP头
Alexa
管理
独立查看登录
独立管理登录

2. 51yes 统计 功能菜单：




访客结构特征分析
性别结构分析 NEW
年龄结构分析 NEW
学历结构分析 NEW
收入结构分析 NEW
智 能 分 析
广告(来路)效果分析 NEW
统计数据智能比较
比较每日分析
比较每月分析
比较每年分析
比较回访率统计
比较来路(域名为单位)
比较来路(网页为单位) NEW
比较受访页面统计
比较搜索关键字(分类)
比较搜索关键字(不分类)
比较地理位置(国内省)
比较地理位置(国内市) NEW
比较地理位置(国外)
比较地理位置(ISP)
比较屏幕分辨率
比较客户显示器颜色
比较客户浏览器
比较Alexa工具条
比较操作系统
比较操作系统语言


数据下载
统计数据下载
管理选项
修改注册资料
修改登陆密码
获得统计代码
清除所有统计数据
我要留言
退出

综合统计

蓝色为：页面总访问量（PV），同一访客的每次访问均被记录
紫红色为：网站独立IP访问量（IP），24小时内相同IP地址只被计算1次 详细说明 >>

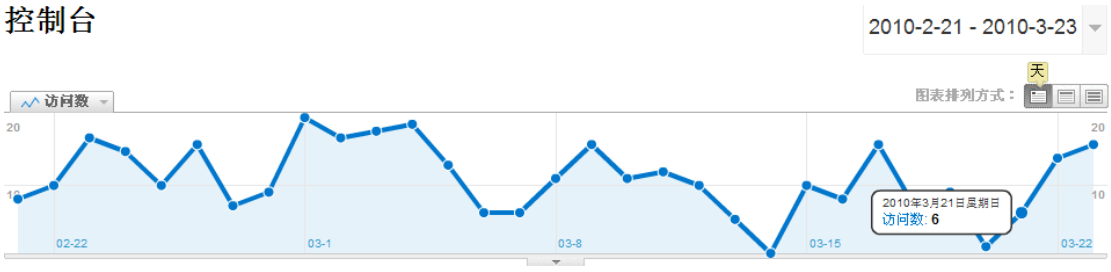
网站名称	51YES功能演示页面		
网站地址	http://demo.51yes.com		
注册日期	2005年03月26日		
今日新客户量	575	今日老客户量	89
同时在线人数	1（注：本数据1分钟刷新一次）		
5分钟同时在线	7（注：本数据5分钟刷新一次）		
15分钟同时在线	15（注：本数据15分钟刷新一次）		
预测今日访问量	4,032（1,212）		
今日访问量	2,181（664）	昨日访问量	3,086（1,114）
本周访问量	16,619（5,840）	上周访问量	26,566（7,513）
本月访问量	8,767（3,163）	上月访问量	159,152（37,217）
平均日访问量	15,278.07（1,805.00）		
平均周访问量	106,207.24（12,547.63）		
平均月访问量	438,989.93（51,863.53）		
本年访问量	971,778（208,694）		
总访问量	13,169,698（1,555,906）		
最大日访问量	54,200（3,077）2005年06月06日（注：本数据以页面总访问量（PV）为准）		
最大周访问量	257,647（21,461）2005年08月22日 ～ 2005年08月28日（注：本数据以页面总访问量（PV）为准）		
最大月访问量	978,234（86,138）2005年08月（注：本数据以页面总访问量（PV）为准）		

3. CNZZ 数据专家 功能菜单

2) 趋势 – 曲线图趋势

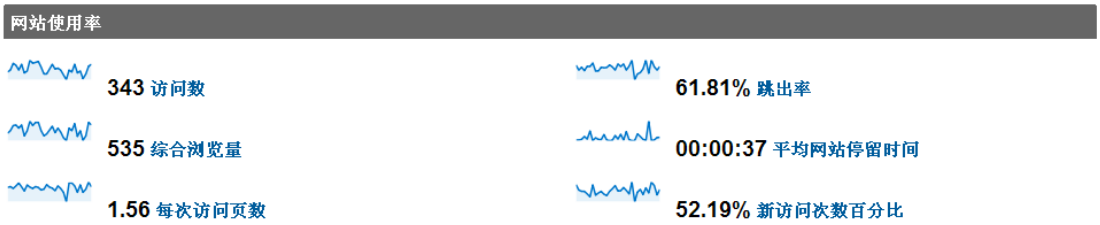
统计周期 天 / 周 / 月 / 年

访问数/PV 浏览量：

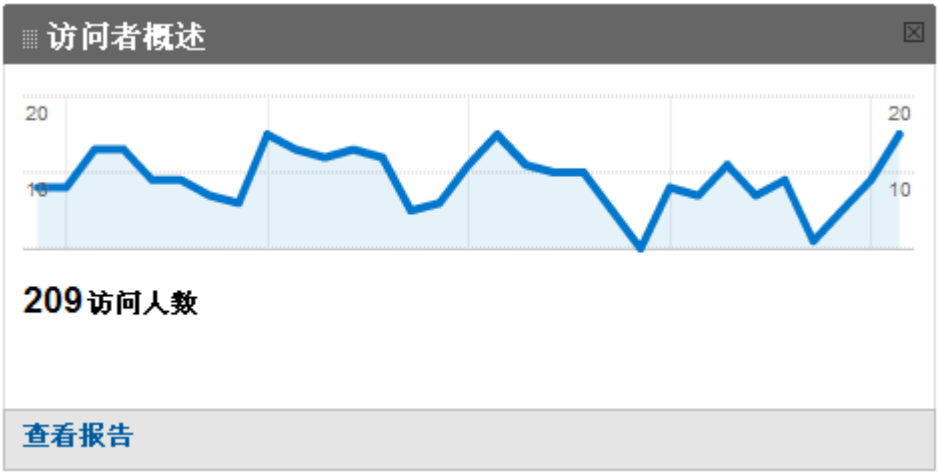


网站使用率：

访问数、跳出率、浏览量、平均停留时间、用户平均访问页数、新用户来访数/占比

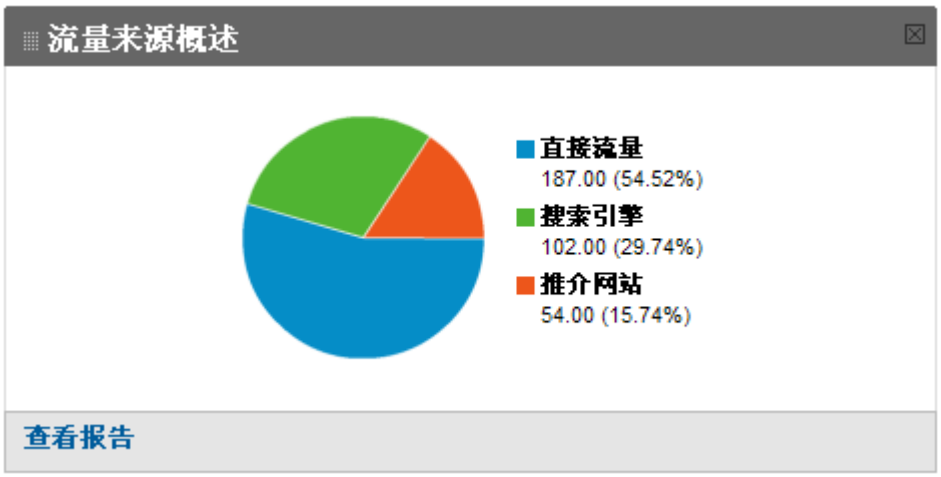


访问人数 UV 曲线图：



用户区域分布情况：

流量来源：



页面访问量排行：

内容概述

网页	综合浏览量	综合浏览量百分比
/fund/	121	22.62%
/fund/fund.php?url=http://biz.finance....	49	9.16%
/blog/html/1tbiDxE9SkiNUgNj5AAAs...	33	6.17%
/fund/fund.php?url=http://biz.finance....	30	5.61%
/fund/fund.php?url=http://biz.finance....	26	4.86%

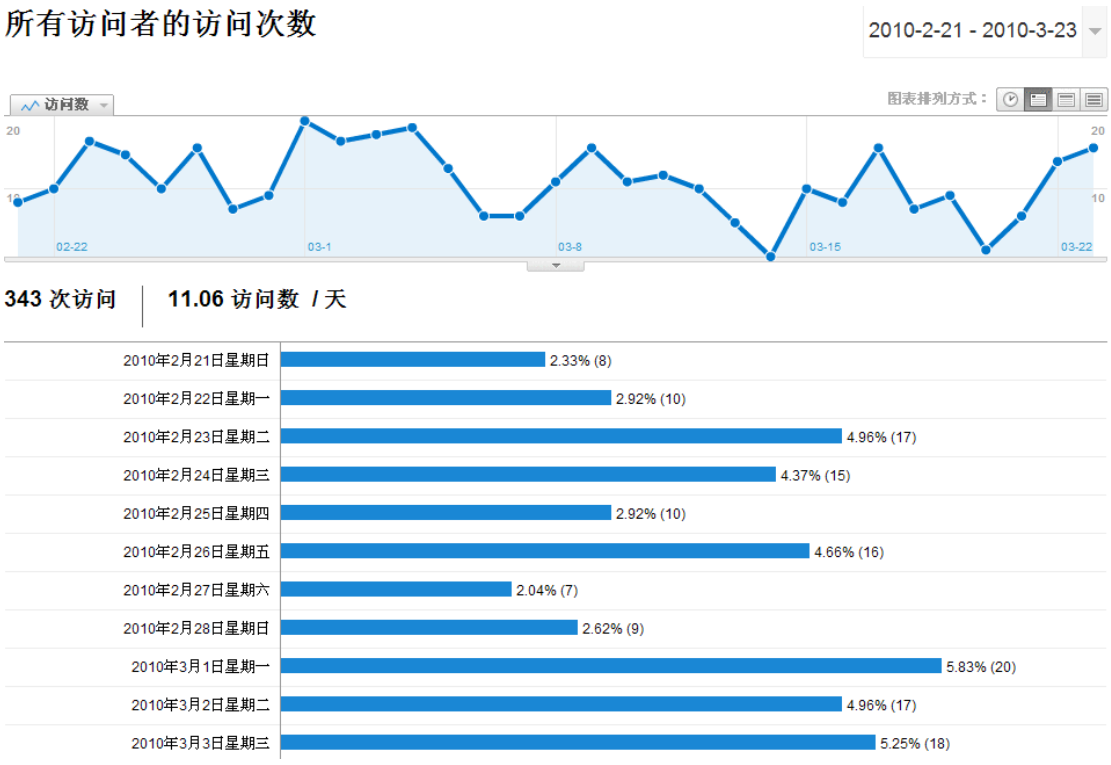
查看报告

详细展示：

访问者概述

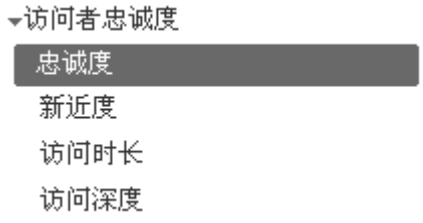
访问者访问次数统计：

所有访问者的访问次数



新访者与回访者：

3）忠诚度 / 用户黏性



访问者忠诚度

2010-2-21 - 2010-3-23

大多数访问的重复次数：1 次

来自此访问者的访问次数（包括当次访问）	访问者访问了 N 次的访问的次数	占所有访问的百分比
1 次	179.00	52.19%
2 次	19.00	5.54%
3 次	7.00	2.04%
4 次	7.00	2.04%
5 次	5.00	1.46%
6 次	4.00	1.17%
7 次	3.00	0.87%
8 次	3.00	0.87%
9-14 次	14.00	4.08%
15-25 次	21.00	6.12%
26-50 次	20.00	5.83%
51-100 次	12.00	3.50%
201+ 次	49.00	14.29%

访问时长

2010-2-21 - 2010-3-2

大多数访问持续的时间：0-10 秒

访问持续时间	这一时段的访问次数	占所有访问的百
0-10 秒	255.00	74.34%
11-30 秒	29.00	8.45%
31-60 秒	18.00	5.25%
61-180 秒	25.00	7.29%
181-600 秒	10.00	2.92%
601-1,800 秒	6.00	1.75%

访问深度

2010-2-21 - 2010-3-23

大多数访问的跟踪页数：1 次网页浏览

访问综合浏览量	达到此浏览量的访问的次数	占所有访问的百分比
1 次网页浏览	212.00	61.81%
2 次网页浏览	98.00	28.57%
3 次网页浏览	19.00	5.54%
4 次网页浏览	6.00	1.75%
5 次网页浏览	5.00	1.46%
6 次网页浏览	1.00	0.29%
7 次网页浏览	1.00	0.29%
8 次网页浏览	1.00	0.29%

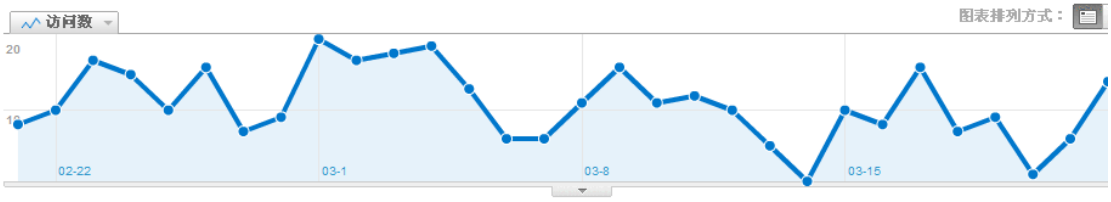
4）用户客户端 浏览器

浏览器功能

- 浏览器
- 操作系统
- 浏览器和操作系统
- 屏幕颜色
- 屏幕分辨率
- Flash 版本
- Java 支持

浏览器

2010-2-21 - 2010-3-



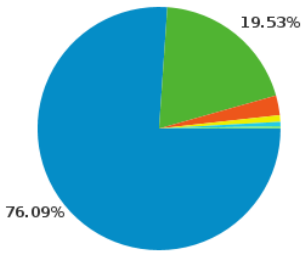
343 次访问使用了 6 浏览器

网站使用 | 第 1 个目标集 | 视图：

访问数	每次访问页数	平均网站停留时间	新访问次数百分比	跳出率
343	1.56	00:00:37	52.19%	61.81%
网站总额百分比： 100.00%	网站平均数： 1.56 (0.00%)	网站平均数： 00:00:37 (0.00%)	网站平均数： 41.69% (25.17%)	网站平均数： 61.81% (0.0

	浏览器 无	访问数	访问数
1.	Internet Explorer	261	76.09%
2.	Firefox	67	19.53%
3.	Chrome	9	2.62%
4.	Opera	3	0.87%
5.	Safari	2	0.58%
6.	Mozilla Compatible Agent	1	0.29%

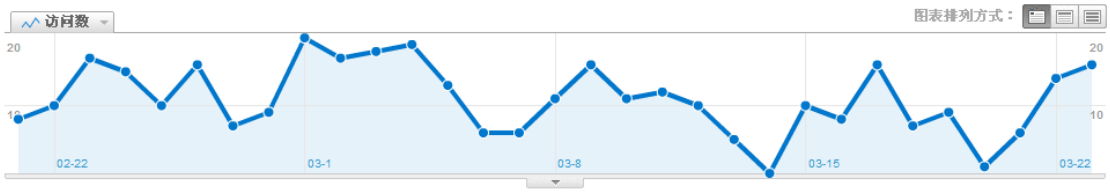
浏览器 占总数的比例： 访问数



5）来源分析：Ref 分析、 站内/站外、站外统计

流量来源概述

2010-2-21 - 2010-3-23

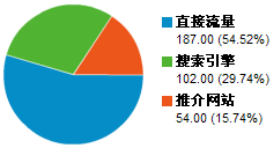


所有流量来源总共带来了 **343** 次访问

54.52% 直接流量

15.74% 推介网站

29.74% 搜索引擎



排名靠前的流量来源

来源	访问次数	访问次数百分比
(direct) ((none))	187	54.52%
google (organic)	69	20.12%
baidu (organic)	33	9.62%
sogou.com (referral)	23	6.71%
soso.com (referral)	16	4.66%

查看完整报告

关键字	访问次数	访问次数百分比
error 1040 (08004): too many conne...	18	17.65%
php xml2array	4	3.92%
error 1040 (08004) too many conne...	2	1.96%
亚太优势基金净值估算	2	1.96%
嘉实300走势图	2	1.96%

查看完整报告

站外来源 关键词分析：

关键字

2010-2-21 - 2010-3-23

搜索通过 78 关键字 发出了 102 次 合计 访问

显示： 合计 | 付费 | 非付费

网站使用 | 第 1 个目标集 | 视图：

访问数	每次访问页数	平均网站停留时间	新访问次数百分比	跳出率
102	1.16	00:00:28	97.06%	87.25%
网站总额百分比： 29.74%	网站平均数： 1.56 (-25.83%)	网站平均数： 00:00:37 (-24.58%)	网站平均数： 41.69% (132.81%)	网站平均数： 61.81% (41.17%)

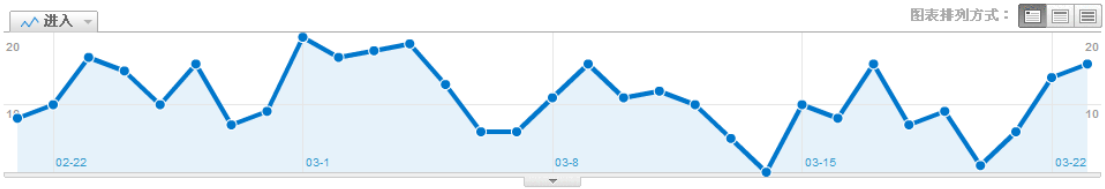
	关键字 无	访问数 ↓	每次访问页数	平均网站停留时间	新访问次数百分比	跳出率
1.	error 1040 (08004): too many connections	18	1.22	00:00:30	94.44%	88.89%
2.	php xml2array	4	1.00	00:00:00	100.00%	100.00%
3.	error 1040 (08004) too many connections	2	1.00	00:00:00	50.00%	100.00%
4.	亚太优势基金净值估算	2	1.00	00:00:00	100.00%	100.00%
5.	嘉实300走势图	2	1.50	00:00:24	100.00%	50.00%
6.	易方达基金走势图	2	1.50	00:00:04	100.00%	50.00%
7.	融通新蓝筹净值表	2	1.00	00:00:00	100.00%	100.00%
8.	融通新蓝筹基金净值表	2	1.50	00:00:22	100.00%	50.00%
9.	"error 1040 (08004): too many connections"	1	1.00	00:00:00	100.00%	100.00%
10.	"error 1040 (08004): too many connections" + mysql	1	1.00	00:00:00	100.00%	100.00%

过滤“关键字”： 包含 开始 高级过滤器 转到: 1 显示行: 10 1 - 10/78

访问量/浏览量最大的页面 ：

最常见目标网页

2010-2-21 - 2010-3-23



343 次访问通过 67 网页 进入网站

目标网页

视图：

进入	跳出	跳出率
343	212	61.81%
网站总额百分比： 100.00%	网站总额百分比： 100.00%	网站平均数： 61.81% (0.00%)

	网页 无	进入 ↓	跳出	跳出率
1.	/fund/	91	12	13.19%
2.	/blog/html/1tbiDxE9SkiNUgNj5AAAs6.html	31	29	93.55%
3.	/fund/fund.php?url=http://biz.finance.sina.com.cn/fund/real_jz.php?fund...	26	26	100.00%
4.	/fund/fund.php?url=http://biz.finance.sina.com.cn/fund/real_jz.php?fund...	24	20	83.33%
5.	/fund/fund_datevalues.php?fundid=200008	20	14	70.00%
6.	/blog/html/1tbiMhs4SklwOiW07AABsk.html	11	11	100.00%
7.	/fund/funddetail.php?fundid=161601	9	7	77.78%
8.	/fund/fund_datevalues.php?fundid=180012	8	4	50.00%
9.	/fund/fund.php?url=http://biz.finance.sina.com.cn/fund/real_jz.php?fund...	7	4	57.14%
10.	/fund/fund_datevalues.php	7	0	0.00%

过滤“网页”： 包含 开始 高级过滤器 转到: 1 显示行: 10 1 - 10/67

6）用户行为

七、 项目约束

- 时间约束：
- 资源约束：

八、 项目资源

- 人力资源

角色	职责	人员列表

项目负责人	负责项目进度控制、协调团队内外事务，保证项目正常可控的开展	林杨
系统架构师	负责项目技术架构	林杨
系统开发工程师	负责数据库程序开发、监控系统、日志采集和分析系统	待定
WEB 开发工程师	负责 WEB 应用的开发（注册、登录、Blog 列表、Blog 操作、发表日志、个人设置）管理系统	待定
UI/UE 工程师	负责项目的 UI 设计、页面构建，UE（用户体验）可参考目前运行中的 Blog 系统	待定
测试支持工程师	负责系统测试	王芳…
运营支撑工程师	负责系统设备准备、系统部署、系统运维	

- 设备资源（投入服务的最小集合，随着用户数的增加，2～3 设备数量将逐步增加）

编号	设备	配置需求	用途	数量
1	开发服务器	至少 4G 内存，1～2 颗 CPU（4 核+2.33GH） 300G 以上硬盘	用于开发、单元测试、性能测试	1
2	WEB 服务器+DB 服务器	至少 8G 内存，1～2 颗 CPU（4 核+2.33GH），600G 以上硬盘	用于注册、登录、用户登录后的各种操作，用于提供数据库服务	1
3	数据库备份服务器	至少 8G 内存，1～2 颗 CPU（4 核+2.33GH），300G 以上硬盘	用于提供数据库备份服务	1

注：关于配置需求，可参照配置标准化文档和咱们公司的实际情况调整。

九、 项目周期

开发周期：本项目分为两部分开发，分别为数据统计分析和用户行为分析。

以下为数据统计分析系统的开发周期。

本次开发周期的大概需要 52 人日(需求分析，系统分析，详细设计，系统开发，系统测试，系统部署)

阶段	任务描述	标准工时（1 人/日）
需求开发（0 个工作日）	**1.整理系统需求**	
	2.本版本功能以及用户体验、服务指标	
	3.管理需求	

	4.系统需求（技术需求）	
	需求审核	
系统设计（0 个工作日）	**1.总体系统设计**	
	2. 数据库设计	
	3. 管理中心设计	
	4.WEB 展示设计、以及页面构建	
	设计审核	
数据准备（0 个工作日）	**数据收集**	
数据分析（0 个工作日）	**日志数据结构分析,简单消除重复/汇总/导入数据库中**	
数据验证（X 个工作日）与系统开发工作并行	**分析程序和日志文件比对，验证数据分析程序的准确性**	
系统开发（33 个工作日/1 人，包含功能单元测试） 考虑不能完全并行调整为 33 个工作日 也要考虑开发人力的不完全并行	**1.统计参数设置**	0
	2.统计站点管理	0.5
	3.统计页面管理	0.5
	4.趋势分析	0
	5.访问量 PV 统计	2
	6.独立 IP 访问统计	2
	7.独立 UV 访问统计	2
	8.跳入页面排行	2
	9.跳出页面排行	2
	10.搜索引擎排行	2
	11.搜索引擎关键词排行	2
	12.新用户增量	2
	13.老用户回访	2
	14.用户流失率	2
	15.用户分析	0
	16.用户地址位置分布统计	2
	17.用户上网方式统计	2
	18.用户操作系统统计	2
	19.用户浏览器语言统计	2
	20.用户分辨率统计	2
	21.用户 Flash 版本统计	2
	22.行为分析	0
	23.用户浏览记录	0
	24.系统管理	0
	25.系统用户管理	0

	26.角色管理	0
	27.权限节点管理	0
页面工程师（0 工作日 0 人/并行开发）		
系统测试（12 个工作日）	集成测试	6
	性能测试	6
	第一次系统测试	2
	第二次系统测试	2
	第三次系统测试	2
系统部署（7 个工作日）	冻结发布	2
	各个服务单元部署	5

十、 项目交付

类别		
文档	项目计划书	
	系统架构设计	
	系统概要设计	
	部署运营使用说明	
	性能测试报告	
代码	程序代码	
	发布包制作代码	

十一、 其他信息