基于计费系统的校园用户行为分析与建模

北京交通大学 计算机与信息技术学院

贾卓生 周爱娟

2019-11-13

主要内容

一 论文研究背景与意义

二 URL混合分类算法

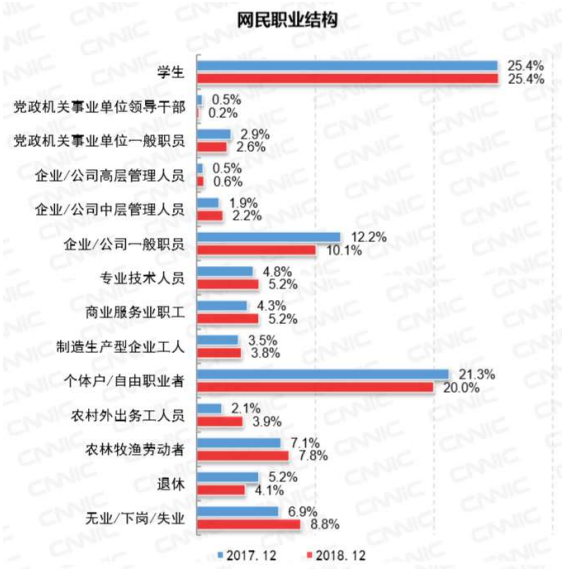
三 用户行为分析模型设计

四 校园用户行为分析系统的设计与实现

五 结论

一·论文研究背景与意义

研究背景



◆我国互联网普及率越来越高，而且学生群体占比最高；

◆以北京交通大学为例，在校师生三万余人，通过城市热点计费系统每天产生海量的日志文件；

◆通过对计费系统日志数据的深入挖掘分析，实现网页URL分类体系的建立，从而获取校园用户行为特征分类，建立校园用户行为模型。这不仅为**学生个人**全面了解自己的网络使用情况提供依据，同时有利于**校园管理者**及时了解在校学生的思想动态和行为模式。

研究意义

校园网络管理者：

◆从宏观的角度掌握整个校园网络的使用情况，协助网络管理部门进行校园网络服务优化与效率提升；查看校园用户整体的兴趣趋向，更好地掌握学生的上网行为模式；

◆通过挖掘出的学生异常行为可以对其网络资源限制或者进行有效的心理指导。

校园网络用户：

◆ 校园网络用户本身可全面查看并评估自己的历史上网行为。

数据来源

数据来源于**北京交通大学Dr.com计费系统的**用户访问日志。本课题采集全校的用户访问日志，其中每天产生的日志在70GB左右，数据量非常的庞大。

```
hive> select * from user_action_log_tmp0314 limit 10;
OK
2015-12-10 (4) 11:06:24 api.pallas.tgp.qq.com/core/tcall?callback=jQuery17205517296250978811_1449717245728&p=%5B%5B17%2C%7B%22player_list%22%3A%5B%7B%22qquin%22%3A%
22U1444391671628965 3    AQEAAAFF        000000000000    219.242.252.90  80      2       140.207.69.31   64564
2015-12-10 (4) 11:06:25 123.125.110.22  ABAEFPMA        000000000000    172.29.68.38    80      1       123.125.110.22  9795
2015-12-10 (4) 11:06:25 s3.qhimg.com/!01fc2b8a/check.css        AJAEFJIE        000000000000    219.242.241.13  80      2       60.207.246.98   33172
2015-12-10 (4) 11:06:25 inews.gtimg.com/newsapp_ls/0/115490689/0?tp=webp        ABAEFPMA        000000000000    172.29.68.38    80      2       123.125.110.22  9790
2015-12-10 (4) 11:06:25 inews.gtimg.com/newsapp_ls/0/115479188/0?tp=webp        ABAEFPMA        000000000000    172.29.68.38    80      2       123.125.110.22  9791
2015-12-10 (4) 11:06:25 inews.gtimg.com/newsapp_ls/0/115479634/0?tp=webp        ABAEFPMA        000000000000    172.29.68.38    80      2       123.125.110.22  9792
2015-12-10 (4) 11:06:25 211.90.27.8     ABEEAFFE        000000000000    59.65.168.224   80      1       211.90.27.8     51939
2015-12-10 (4) 11:06:25 inews.gtimg.com/newsapp_ls/0/115611555/0?tp=webp        ABAEFPMA        000000000000    172.29.68.38    80      2       123.125.110.22  9793
2015-12-10 (4) 11:06:25 140.205.243.66  AQAEAIIB        000000000000    219.242.244.111 80      1       140.205.243.66  54245
2015-12-10 (4) 11:06:25 short.weixin.qq.comhttp://short.weixin.qq.com/cgi-bin/micromsg-bin/getemotionlist       ABEAEFMM        000000000000    219.242.117.212 80 1
82.254.114.108  55174
Time taken: 0.49 seconds, Fetched: 10 row(s)
```

字段名	日志内容
log_reqtime	用户请求时间
log_requrl	用户请求URL地址
log_username	用户登录账号
log_usermac	用户MAC地址
log_userip	用户IP地址
log_userport	用户端口号
log_destip	目的IP地址
log_destport	目的端口号

URL特征

URL 的语法格式为：
protocol://hostname[:port]/path[;parameters][?query][#fragment]

URL结构	含义
Protocol（协议）	用于指定使用的传输协议；
Hostname（主机名）	指存放资源的服务器的域名系统DNS主机名或者IP地址；
Port（端口号）	为整数，省略时使用方案的默认端口；
Path（路径）	由零或多个/符号隔开的字符串，表示主机上一个目录或文件地址；
Parameters（参数）	用于指定特殊参数的可选项；
Query（查询）	可选，用于给动态网页传递参数；
Fragment（信息片段）	用于指定网络资源中的片断。

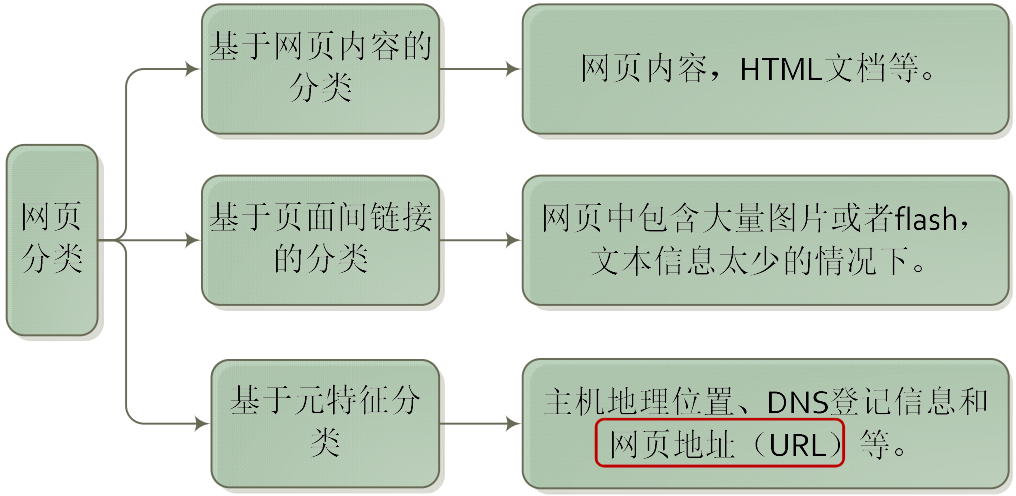
例如：

http://finance.sina.com.cn/roll/index.d.html?cid=56941&page=1

通过查看网页URL格式，发现真正有区分能力的是Host（主机名）和 Path（路径）这两个字段，而其他字段对于网页分类的贡献程度几乎可以忽略不计。

二·URL混合分类算法

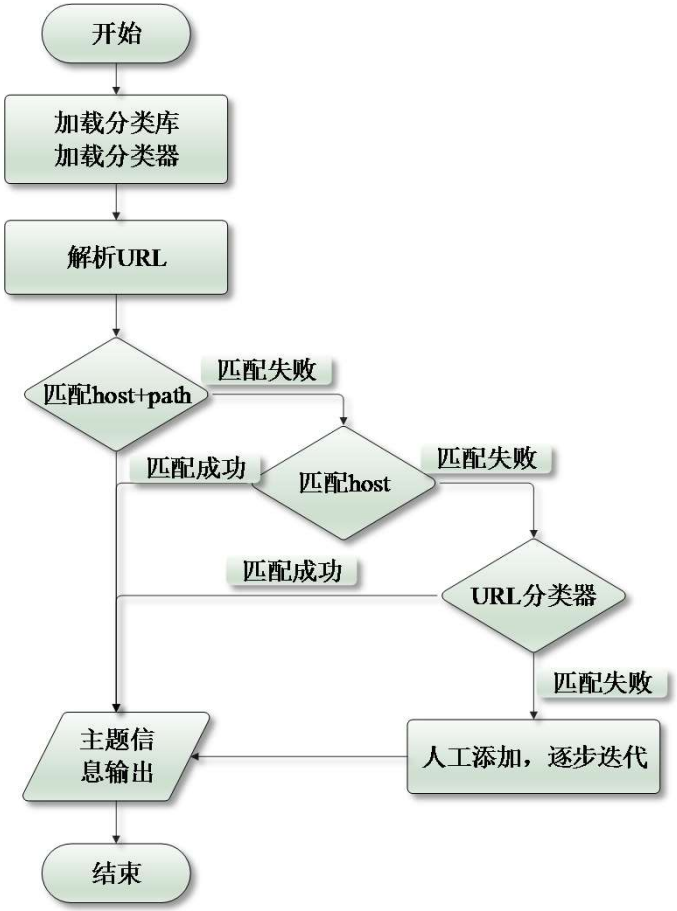
1. URL混合分类模型设计



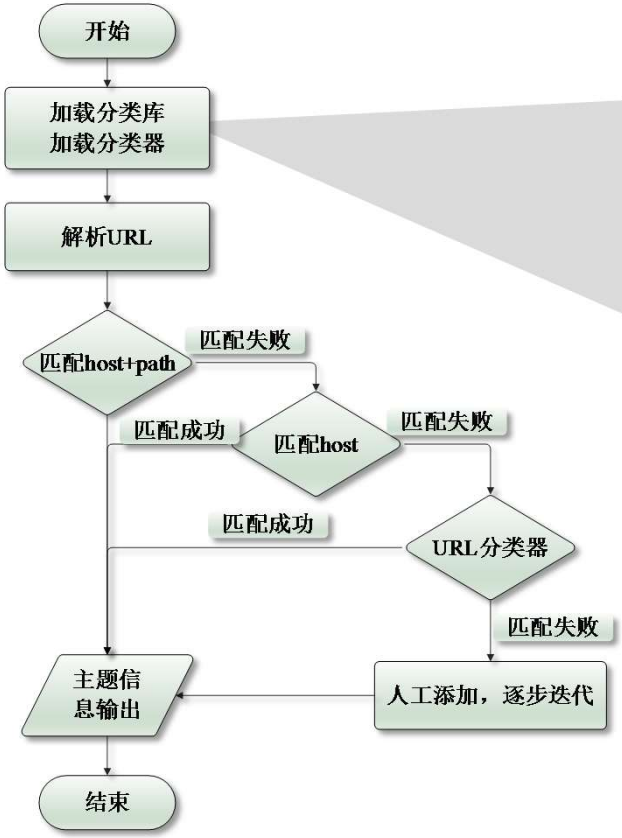
分类方法	算法	优缺点
传统网页分类方法	对于每一条URL都去爬取页面内容，然后分词、特征选取，再进行算法计算。	准确率较高；但效率太低，复杂度太高。
本文采用的网页分类方法	URL混合分类算法（URL分类库+基于N-Gram语言模型的主题分类器）	效率高，适合海量日志的分析处理

二、URL混合分类算法

URL混合分类算法



URL混合分类算法--分类库



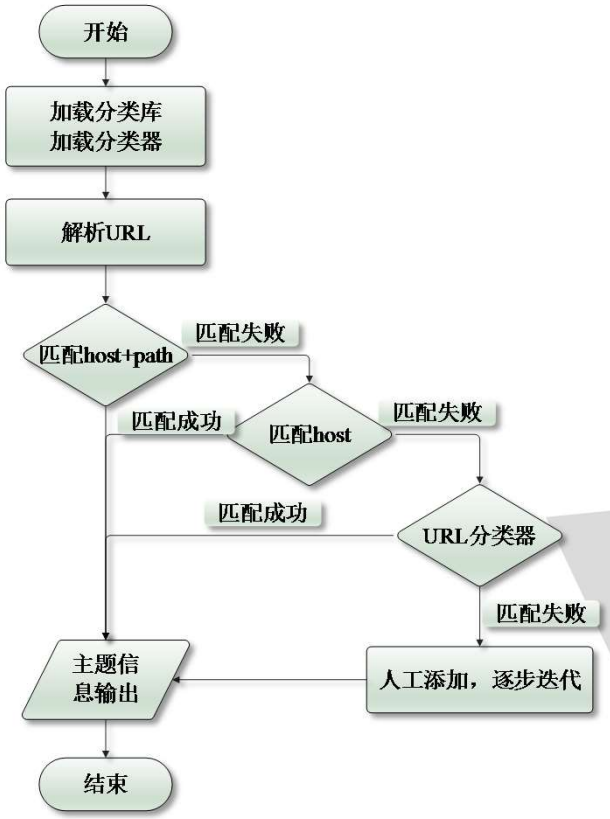
校园用户行为分析模型

◆URL分类库的主题集分为娱乐、体育、游戏、购物、新闻、经济、计算机、求职、搜索引擎等共15类，基本可以反应校园用户的兴趣趋向。

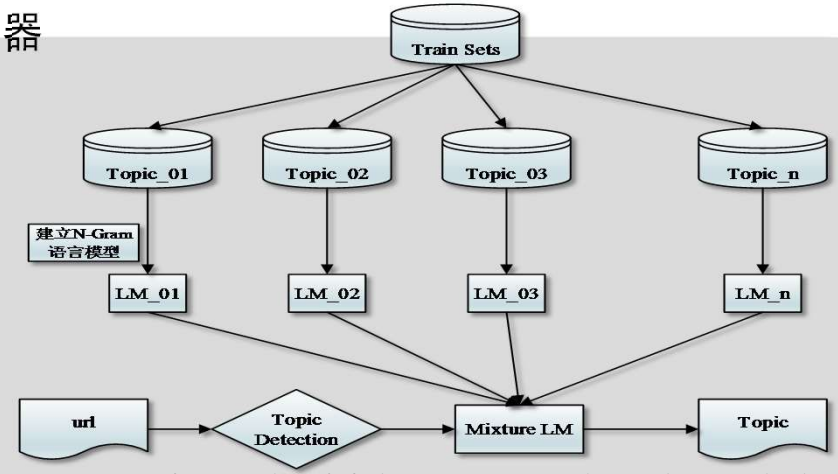
◆采用Python脚本分布式爬取开放式分类目录网站（ODP），保存格式为<URL，主题类别>。

id	url	topic
1	music.163.com/#/discover/toplist	娱乐
2	china.nba.com/?gr=www	体育
3	jx3.xoyo.com/gn/index.html?to=sj	游戏
4	qiang.suning.com/?safp=d488778a.homepage1.99345513343.1	购物
5	channel.chinanews.com/cns/cl/cj-msrd.shtml	新闻
6	finance.ce.cn/jjpd/index.shtml	经济
7	bj.lianjia.com/zufang/	生活
8	food.qm120.com/zi9d/dsyp/	健康
9	www.45it.com	计算机
10	www.jiaoyou.com/login.php	交友
11	www.yjbys.com/zhaopinhui/beijing_xiaoyuan/	求职
12	www.dsti.net/Information/ViewPointList	军事
13	www.baidu.com	搜索引擎
14	www.chinavalue.net/MiniBlog	博客论坛
15	www.wanfangdata.com.cn/index.html	教育文化

URL混合分类算法--基于N-Gram LM的URL分类器



校园用户行为分析模型



- N-gram模型：基于马尔科夫假设,即下一个词的出现仅依赖于它前面n-1个词。

$$P(\omega_i \mid \omega_1\omega_2 \ldots \omega_{i-1}) = P(\omega_i \mid \omega_{i-n+1} \ldots \omega_{i-1})$$

- 加一平滑法：又称拉普拉斯平滑法，保证每个N-grams在训练语料中至少出现1次。

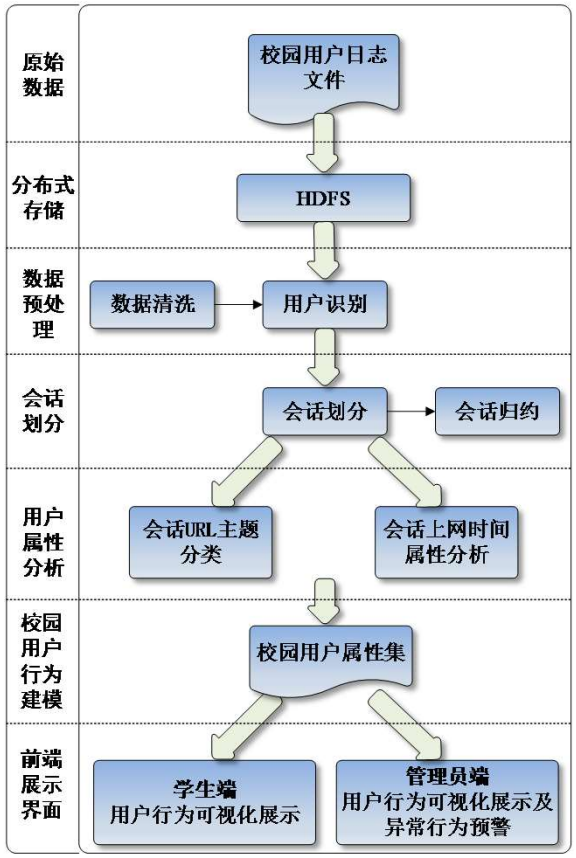
$$\rho_{c_j}(w_i \mid w_{i-n+1} \ldots w_i) = \frac{count(w_i \mid w_{i-n+1} \ldots w_i) + \gamma}{count(w_i \mid w_{i-n+1} \ldots w_i) + V \times \gamma}$$

V-训练集中所有N-Gram的个数

γ-用于控制分配的未知字符串序列的概率比重

三·用户行为分析模型设计

用户行为分析模型整体框架



校园用户行为分析模型

◆Step1：将采集到的原始数据批量上传存储到HDFS中；

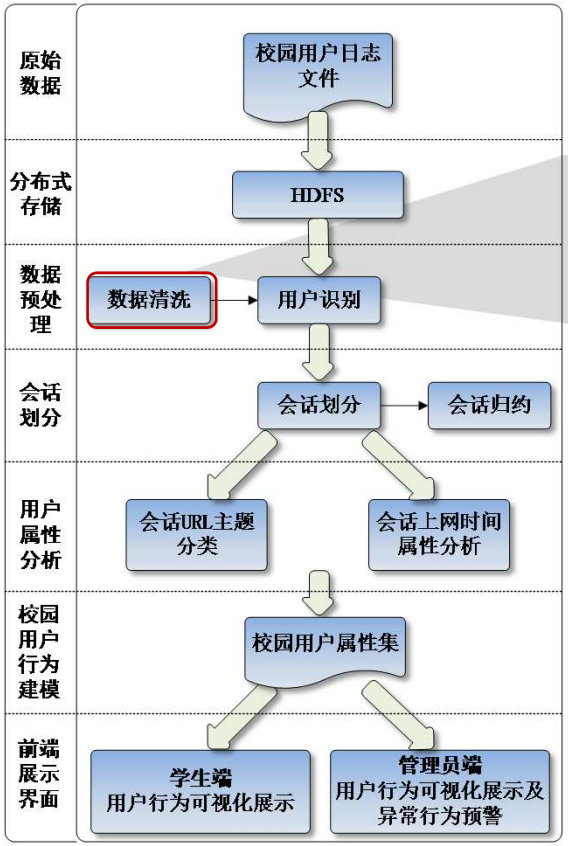
◆Step2：数据预处理。对输入的数据先进行数据清洗，然后根据IP地址的不同进行用户识别；

◆Step3：会话划分。根据设置好的时间阈值将每个用户访问的url地址由多个会话集表示；

◆Step4：用户属性分析。分为两部分，一部分是会话URL主题分类，并计算用户的主题兴趣度；另一部分是分析上网时间属性。最终得到校园用户上网行为属性集；

◆Step5：前端展示。将用户上网行为分析结果分别面向学生和管理员进行可视化展示。

数据清洗



校园用户行为分析模型

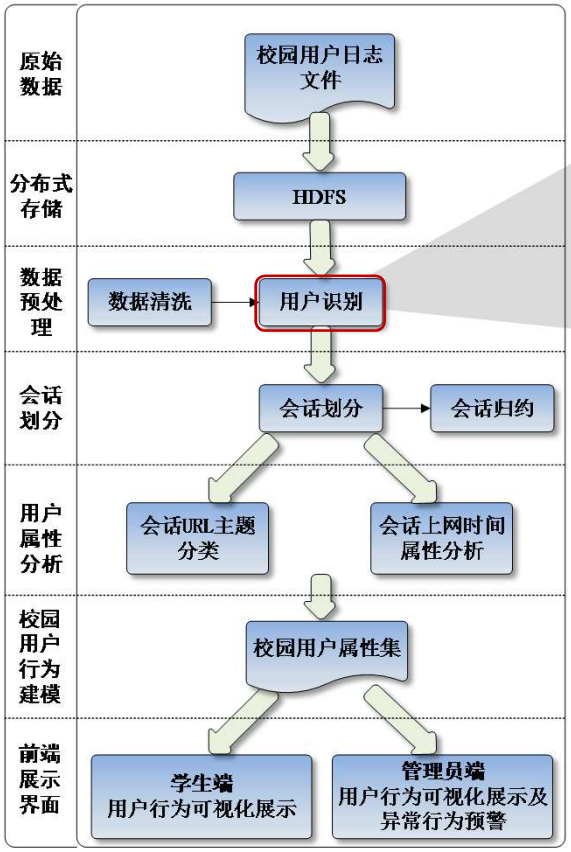
我们在以下三个方面对数据进行清洗：

◆ **（1）网页自动加载的记录。**清除掉自动加载的静态资源文件（包括jpg、gif、png、css、js等非HTTP文本）。

◆ **（2）冗余的URL记录。**清除在网络情况不稳定情况下多次刷新该页面产生的对同一URL的多次请求。

◆ **（3）日志无效字段。**清除掉原始日志中对于网页分类没有任何价值的无效字段，并提取URL的host和path部分，最后保留时间戳、URL的host和path、用户IP三个属性。

用户识别



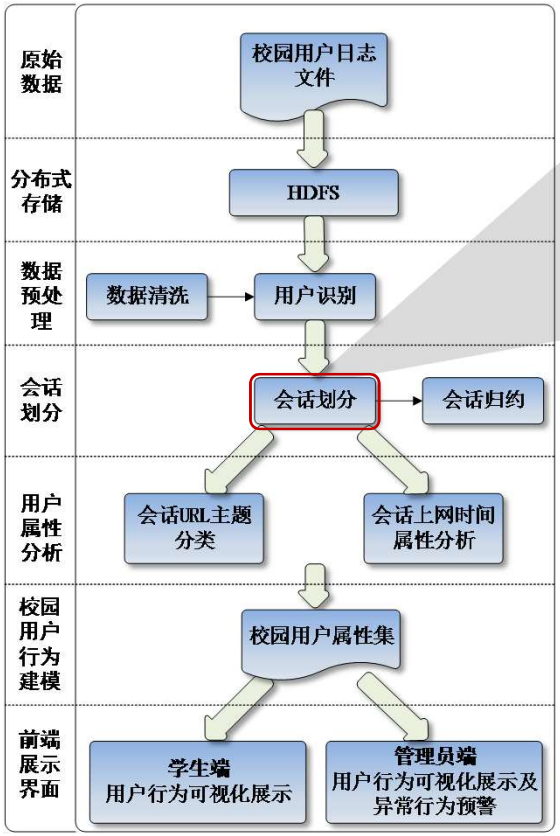
校园用户行为分析模型

◆因此本实验中采取一个IP地址视为一个用户的策略。

◆用户识别后每个用户的访问记录可表示如下形式：

$$\{(timestamp_1, url_1), (timestamp_2, url_2) \ldots, (timestamp_n, url_n)\}$$

会话划分



校园用户行为分析模型

- 根据时间窗口将用户所访问的URL列表进行会话划分。根据经验值，本实验设置会话窗口为30min。
- 会话划分之后，每个用户访问URL被划分成m个会话的集合。

$$S=\{s_1,s_2,\ldots,s_j\}\ (j=1,2,\ldots,m)$$

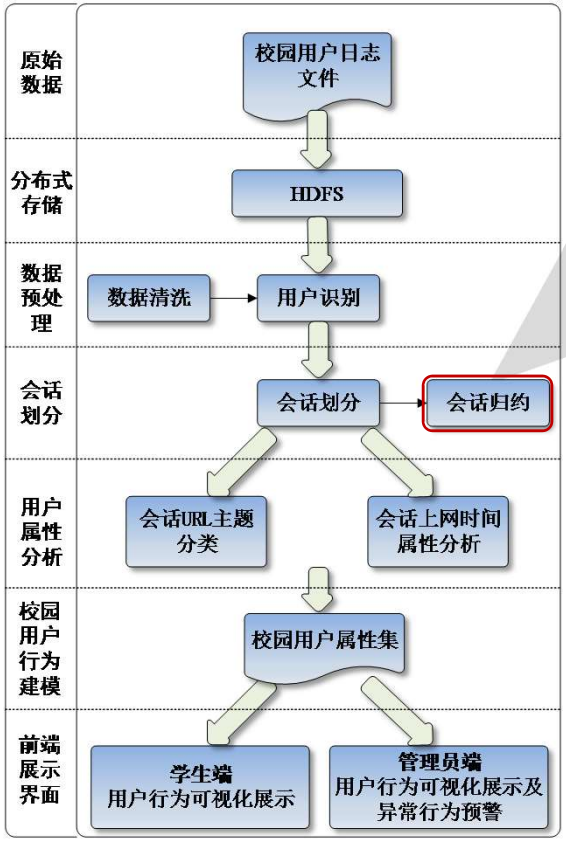
其中

$$s_j=\{ip_j,(url_1^j,stay\_time_1^j),(url_2^j,stay\_time_2^j),\ldots,(url_l^j,stay\_time_l^j)\}$$

url_l^j 指的是用户在第j个会话中访问的第l个URL页面，

$stay\_time_l^j$ 指的是用户在第j个会话中访问第l个URL页面所停留的时间。

会话归约



校园用户行为分析模型

URL1: http://www.xinhuanet.com/fortune/201904/16/c_1210109898.htm

URL2: http://www.xinhuanet.com/politics/201904/14/c_1124364815.htm

两个URL的相似性计算公式：

$$Sim_{url} = \frac{\sum_{i=1}^{l} w_i}{\sum_{j=1}^{L} w_j}$$

◆定义$S_{url} > 0.7$表示两个两个网页同属于一个主题

$$Sim(url_1, url_2) = \frac{5+4}{5+4+3+2+1} = 0.6 < 0.7$$

◆两个网页的相似度存在如下特点：

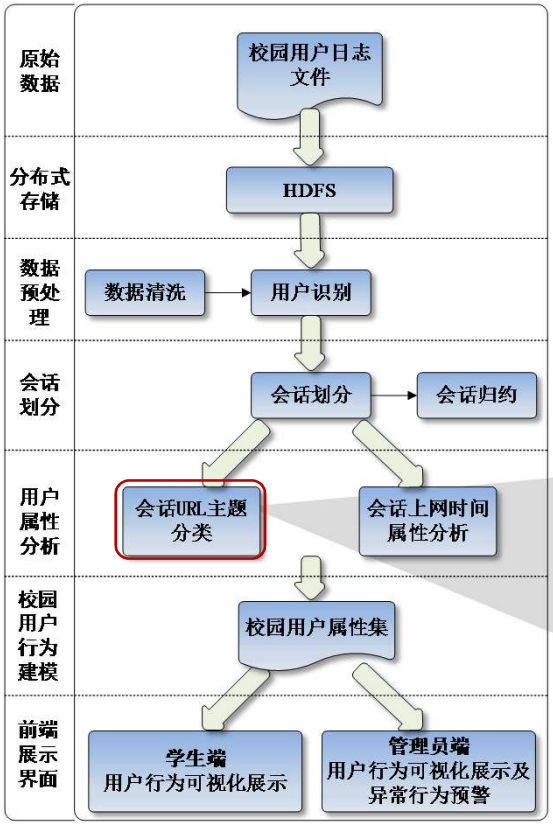
（1） $Sim_{url}=0$ 表示两个网页是两个完全不同访问页面；

（2） $0<Sim_{url}<1$ 表示两个网页的相似度介于 1 和 0 之间；

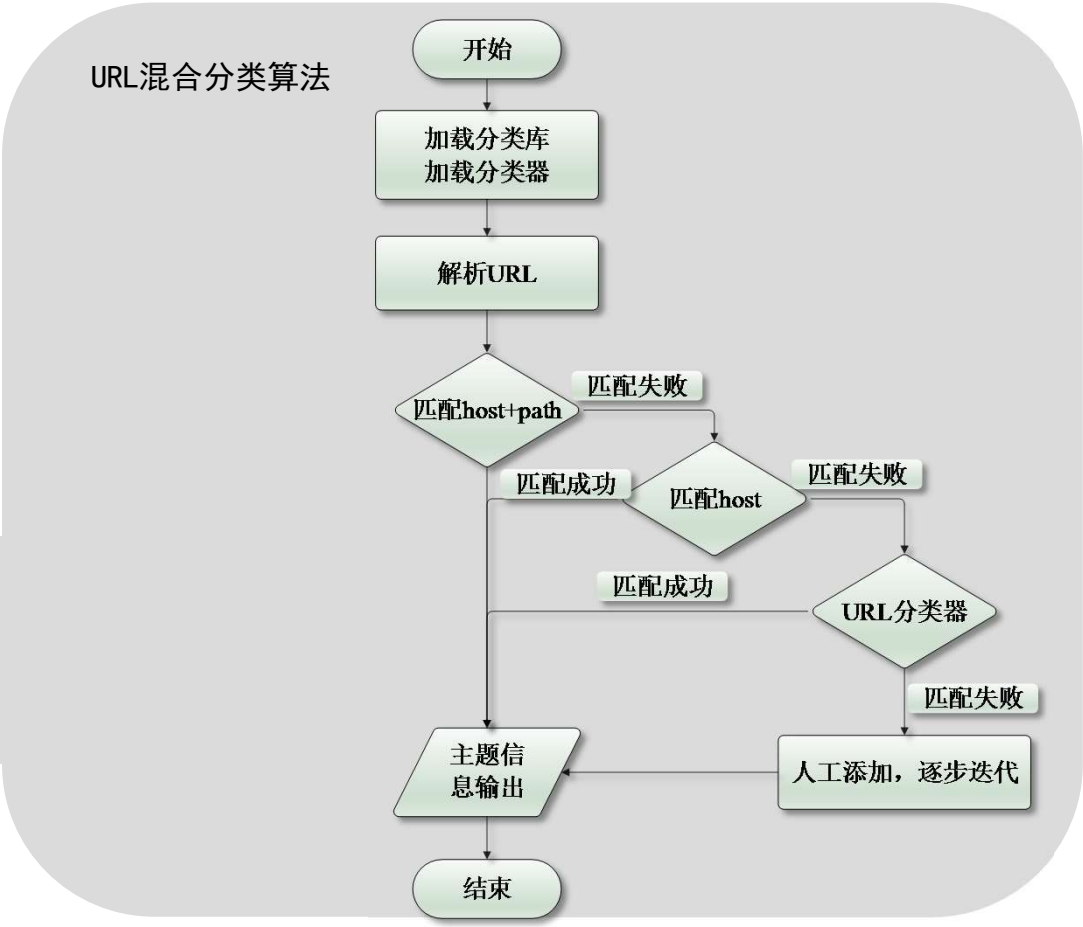
（3） $Sim_{url}=1$ 表示两个网页是连个完全相同的访问页面。

三、用户行为分析模型设计

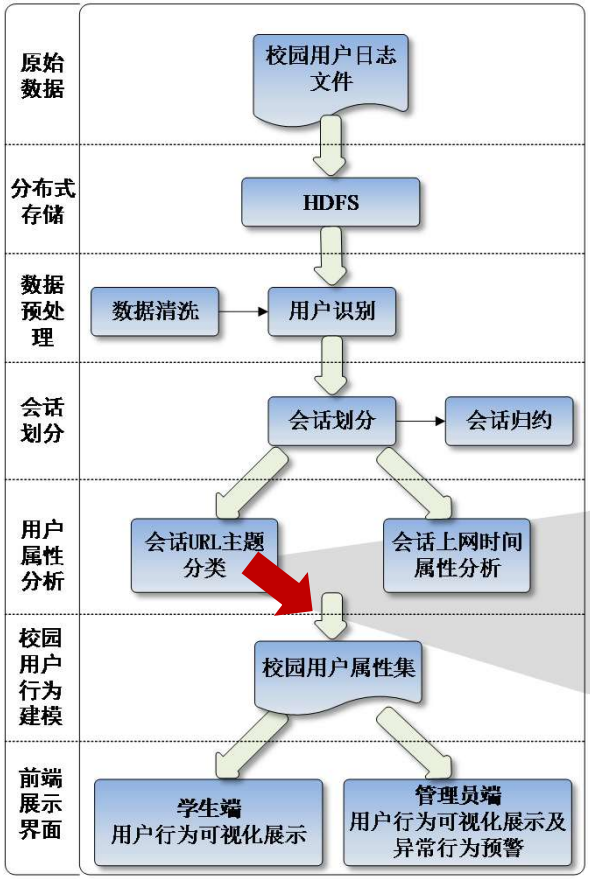
会话URL主题分类



校园用户行为分析模型



基于加权兴趣度的用户兴趣主题表示



校园用户行为分析模型

◆本实验从**访问时长**和**访问次数**两个方面对用户的兴趣主题进行表示。

定义兴趣度 1（w_1）：反映主题访问时长对兴趣度的影响因子。

$$w_1^i = \frac{t_i}{\sum_{i=1}^{k} t_i}$$

定义兴趣度 2（w_2）：反映主题访问次数对兴趣度的影响因子。

$$w_2^i = \frac{f_i}{\sum_{i=1}^{k} f_i}$$

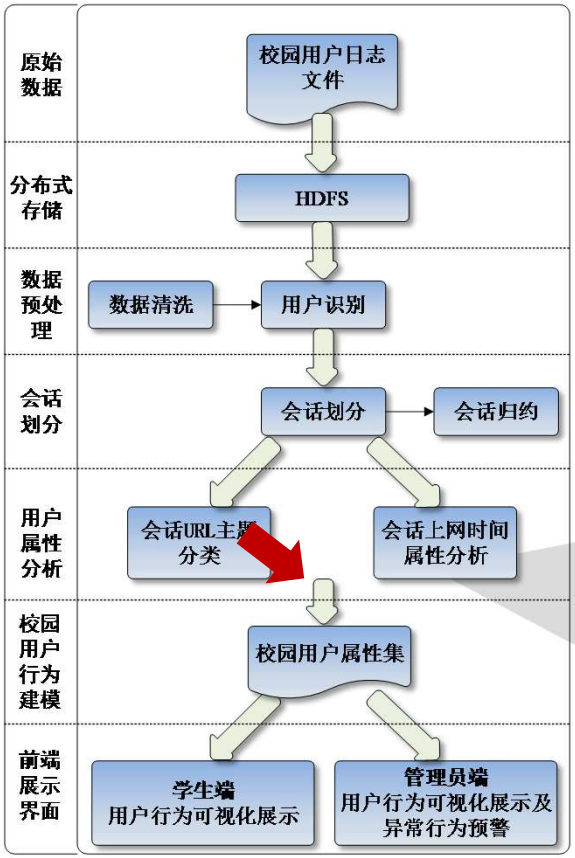
基于加权的思想，某个用户对某个主题的兴趣度表示为：

$$w_i = \rho \times w_1^i + (1-\rho) \times w_2^i$$

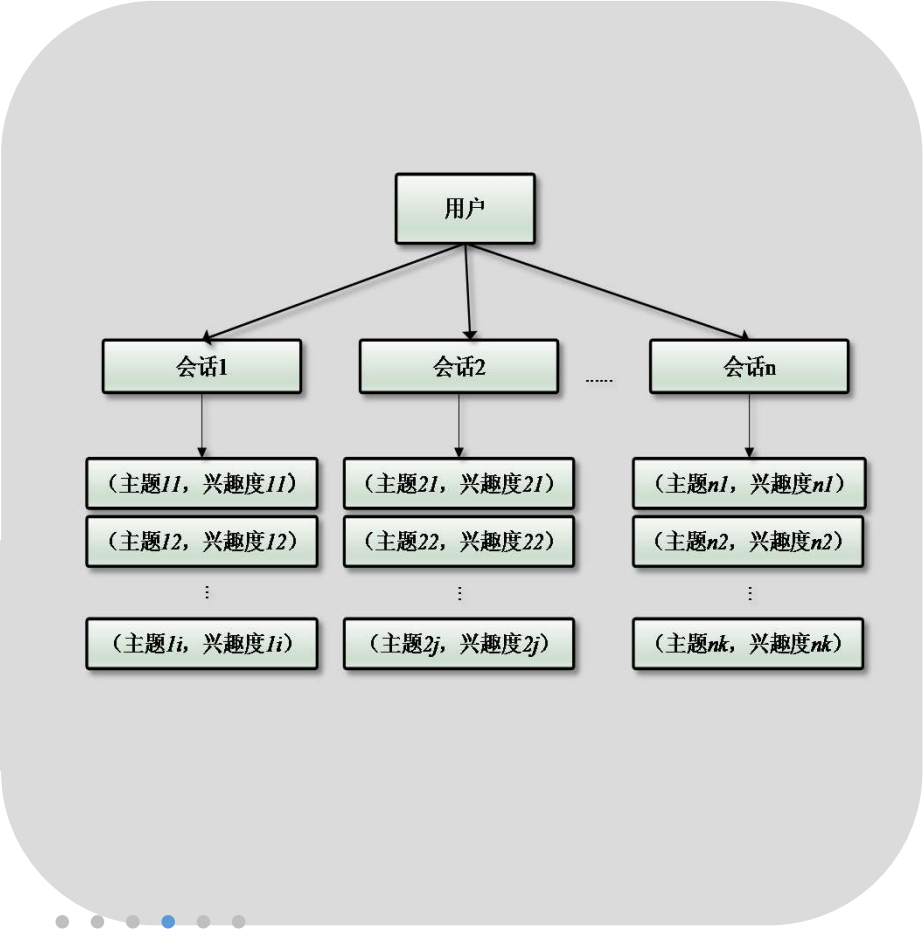
◆用户的一个会话的兴趣主题模型可以表示为：

$$s_i = \{(p_1, w_1), (p_2, w_2), \ldots (p_k, w_k)\}$$

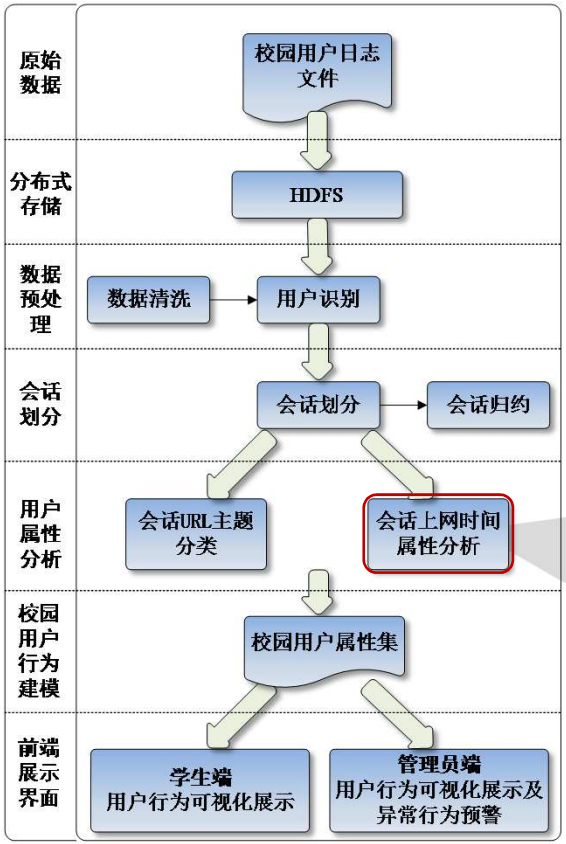
基于加权兴趣度的用户兴趣主题表示



校园用户行为分析模型



会话上网时间属性分析



校园用户行为分析模型

用户上网时长

每个会话可以表示为：

$$s=\{(p_1,t_1,f_1),(p_2,t_2,f_2),\ldots(p_k,t_k,f_k)\},0\le k\le 10$$

那么一个会话内用户的访问时长可表示为：

$$t=t_1+t_2+\ldots+t_k$$

用户上网时间段

根据学校用户的学习生活作息将每天24个小时分为5个时间段，即：

00:00-06:00为凌晨，

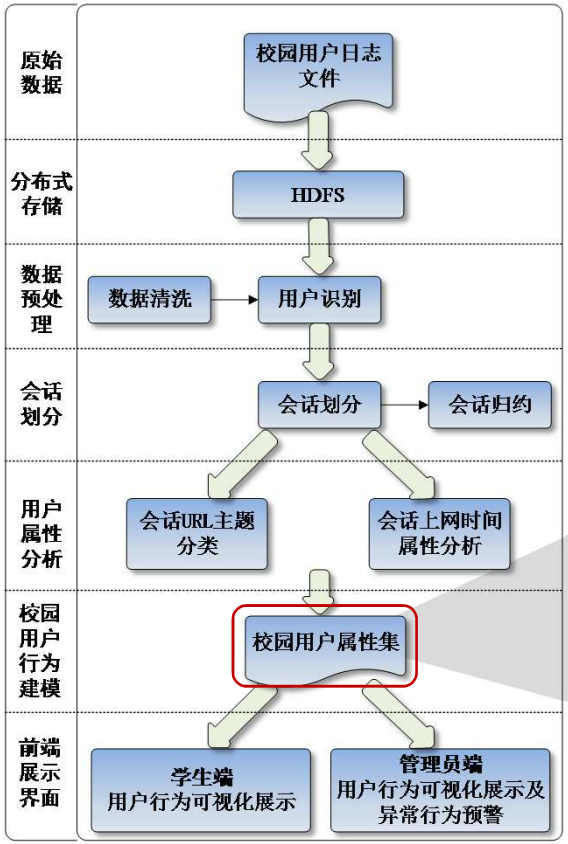
06:00-12:00为上午，

12:00-14:00为中午，

14:00-18:00为下午，

18:00-24:00为晚上。

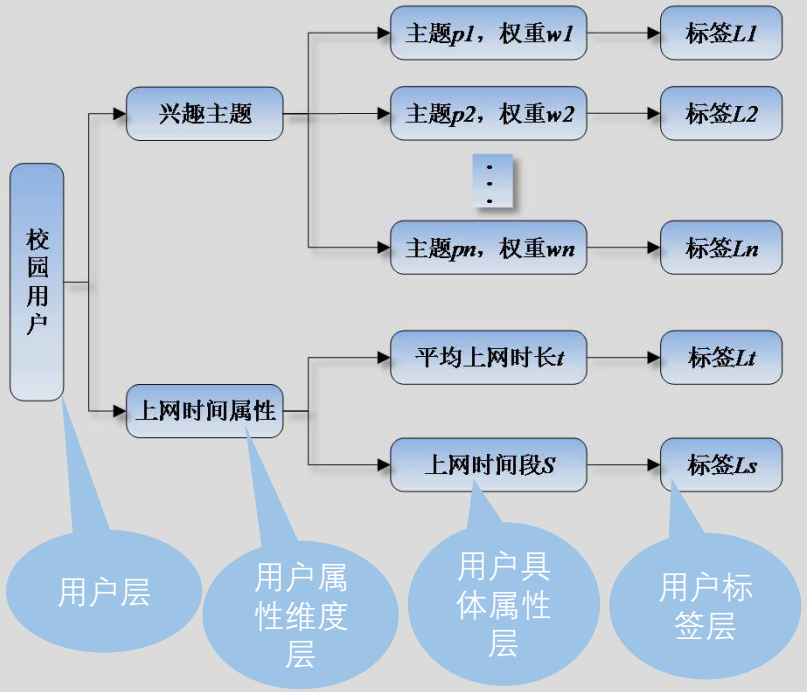
会话上网时间属性分析



校园用户行为分析模型

校园用户行为属性集表示为向量的形式：

$$Model=\{\{(p_1,w_1,L_1),(p_2,w_2,L_2),\ldots(p_k,w_k,L_k)\},t,L_t,S,L_S\}$$



基于用户属性的校园用户行为模型示意图

用户行为模型评估

Pearson相关系数

用于评定人工标注结果和模型生成结果的相关性。

$$r=\frac{\Sigma XY-\frac{\Sigma X\Sigma Y}{N}}{\sqrt{\left(\Sigma X^2-\frac{\Sigma X^2}{N}\right)\left(\Sigma Y^2-\frac{Y^2}{N}\right)}}$$

其中：
N指的是分类器中主题的个数，
X表示人工标注结果集中某用户对各主题的兴趣度，
Y表示模型生成结果中某用户对各主题的兴趣度。

排序准确率

指的是模型生成结果中按照兴趣度降序排列的主题顺序与人工标注的主题顺序的准确程度。

$$Accuracy=\left(\sum_{i=1}^{N}\frac{1}{1+\partial_i\left|rank_{c_i}-ideal\_rank_{c_i}\right|}\right)\Big/N$$

∂_i表示的是主题类i的调节因子，
$rank_{c_i}$表示的是模型生成结果中主题类的兴趣度排序，
$ideal_rank_{c_i}$指的是人工标注结果中主题类的兴趣度排序。

四·校园用户行为分析系统的设计与实现

系统需求分析

前端需求

校园用户

查看自己的历史上网记录。例如一天内对各主题兴趣度的变化趋势，一天内各主题占比等，可以有助于用户对自己上网行为有个全面的认识。

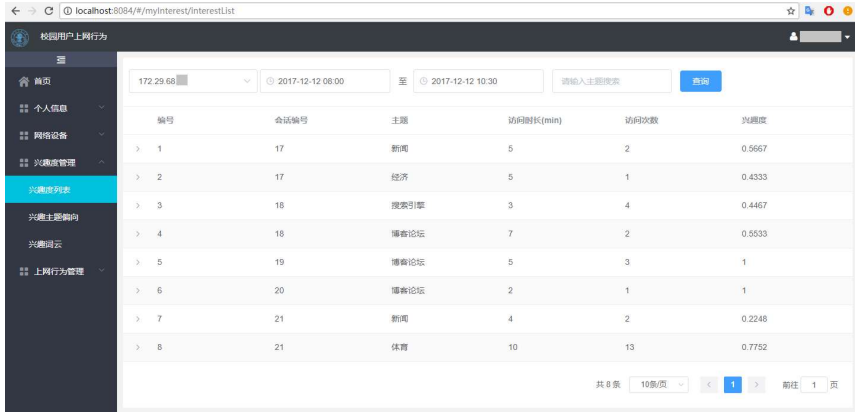
校园管理者

查看学校用户整体的上网行为，方便掌握学生的上网行为模式；
挖掘出学生的异常行为，比如沉迷于游戏，长时间不在校等。

后端需求

后端模块	介绍	所需技术
数据采集与存储	数据量较大，需要利用大数据组件进行分布式存储	Flume、HDFS
数据预处理	包括数据清洗、用户识别、会话识别、会话归约四部分	Hive、Spark SQL
用户会话主题识别	将会话内的URL输入URL混合分类器中，输出主题	Spark MLlib
用户兴趣度量表示	根据用户会话内的兴趣、访问时长、访问频次输出用户兴趣度列表	Spark SQL
用户上网属性分析	统计用户的上网时长、上网时间段等信息	Spark SQL
系统后端架构	MVC架构	SpringBoot、SpringCloud、Maven、Mybatis

前端可视化模块（学生端）—兴趣度管理模块

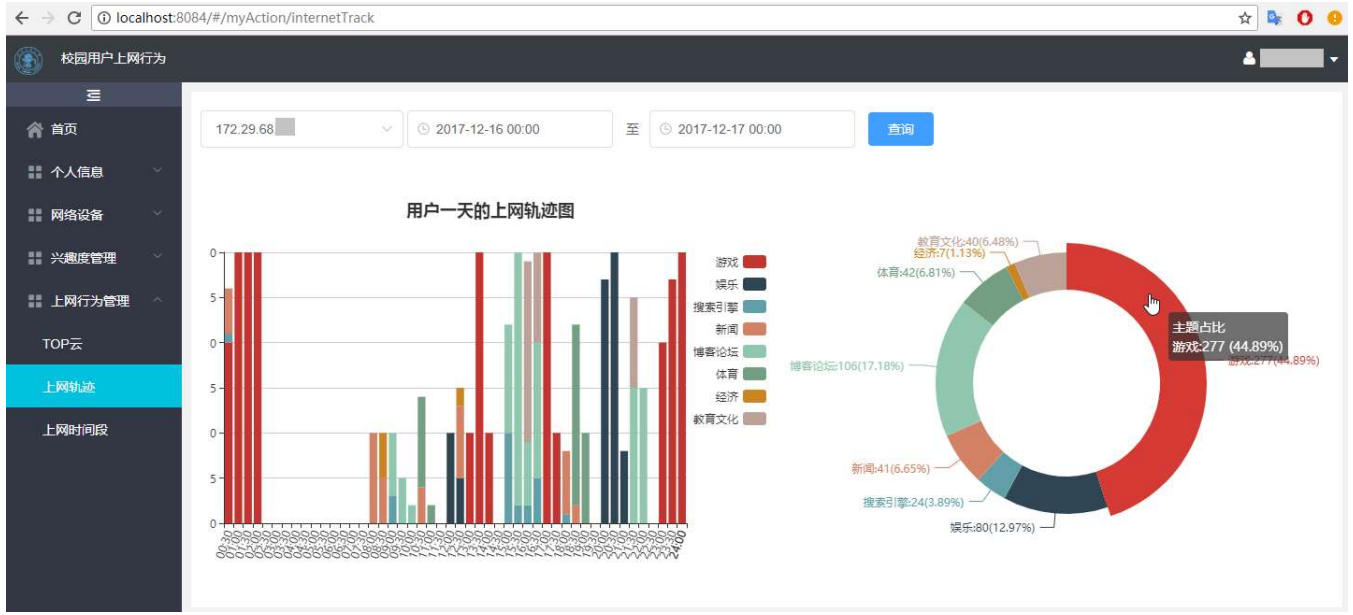


◆ 学生端：可通过选择登录设备的IP地址、开始时间、结束时间、主题来查看这段时间自己的访问主题的兴趣度信息

用户主题兴趣倾向图

◆ 兴趣度管理模块不仅可以供用户个人查看自己短时间内的兴趣变化，还可以查看一周、一个月或者几个月内用户的兴趣主题总趋势，发现该用户的兴趣偏向于经济、新闻、体育、论坛等主题。

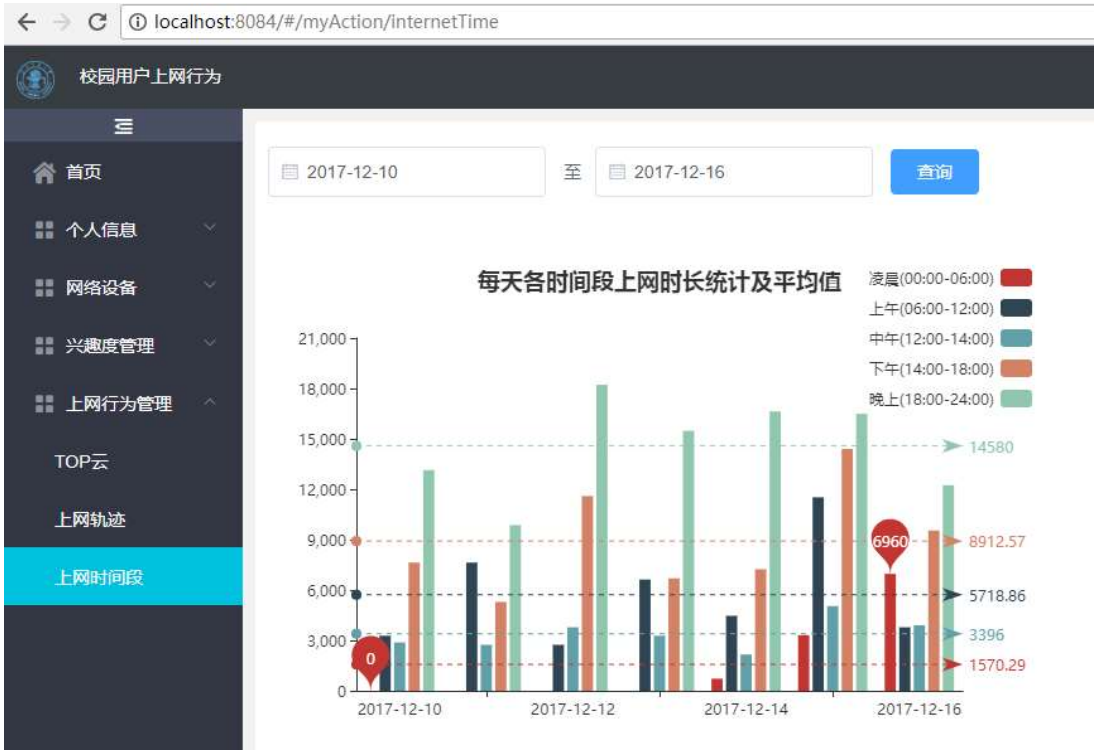
前端可视化模块（学生端）—上网行为管理模块



学生端的上网轨迹图

◆学生端可以通过选择IP地址、时间段查看对应用户的上网轨迹图以及各主题的访问时长在整个上网时长中的占比图。

前端可视化模块（学生端）—上网行为管理模块



用户各时间段的上网时长统计

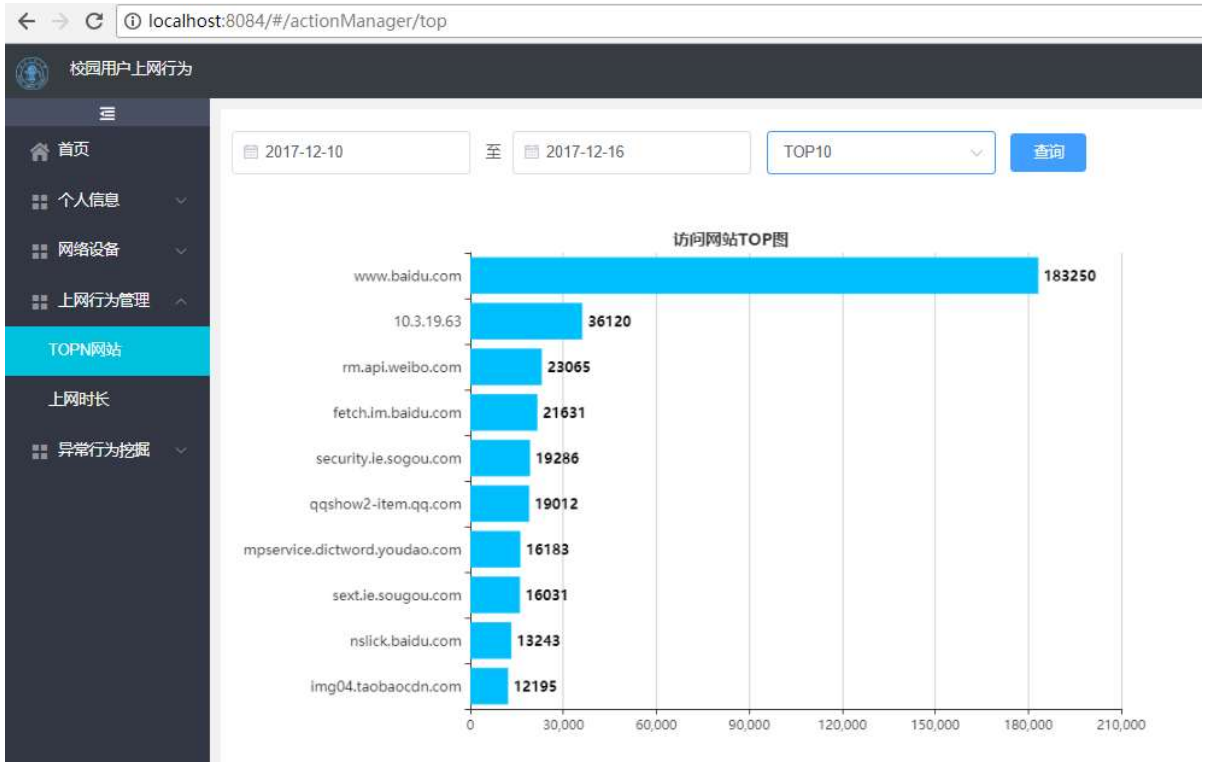
◆用户不仅可以查看自己一段时间内的上网轨迹，还可以查看自己在各个时间段的上网时长。在本实验中我们将一天二十四个小时分为五个时间段，即0时到6时为凌晨，6时到12时为上午，12时到14时为中午，14时到18时为下午，18时到24时为晚上。

前端可视化模块（学生端）—首页

用户行为标签化展示

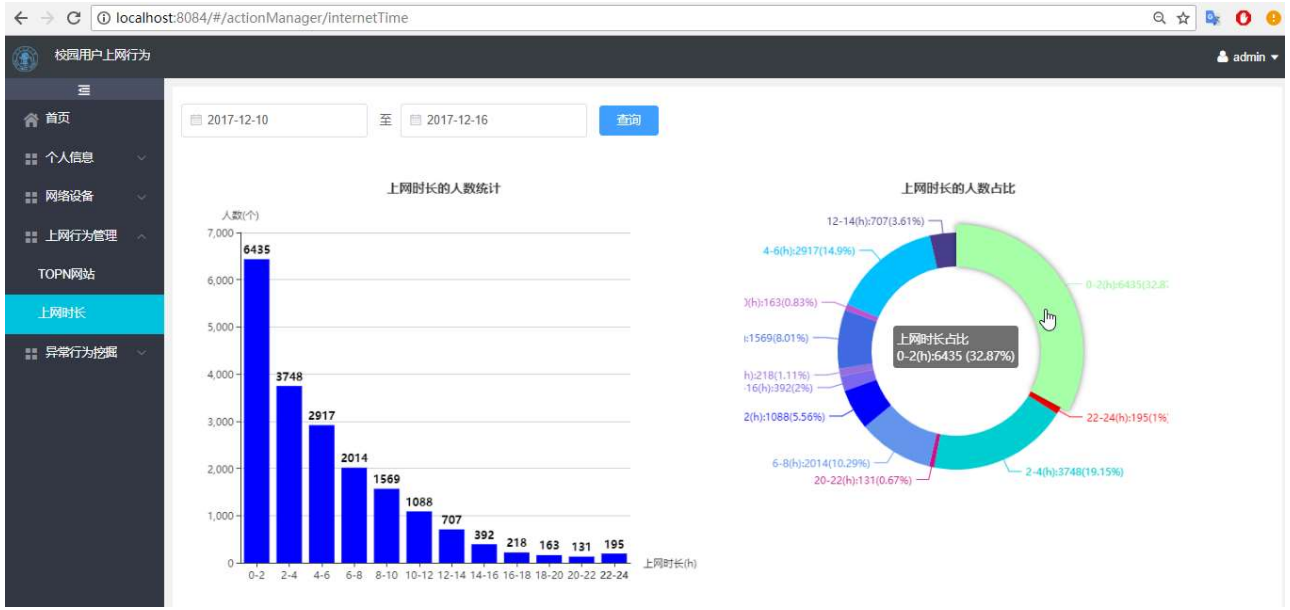
根据用户的兴趣倾向以及上网时长、时间段等其他特征，我们用标签来综合评估用户的行为特征。

前端可视化模块（校园管理者端）—上网行为管理



所有校园用户访问网站TopN图

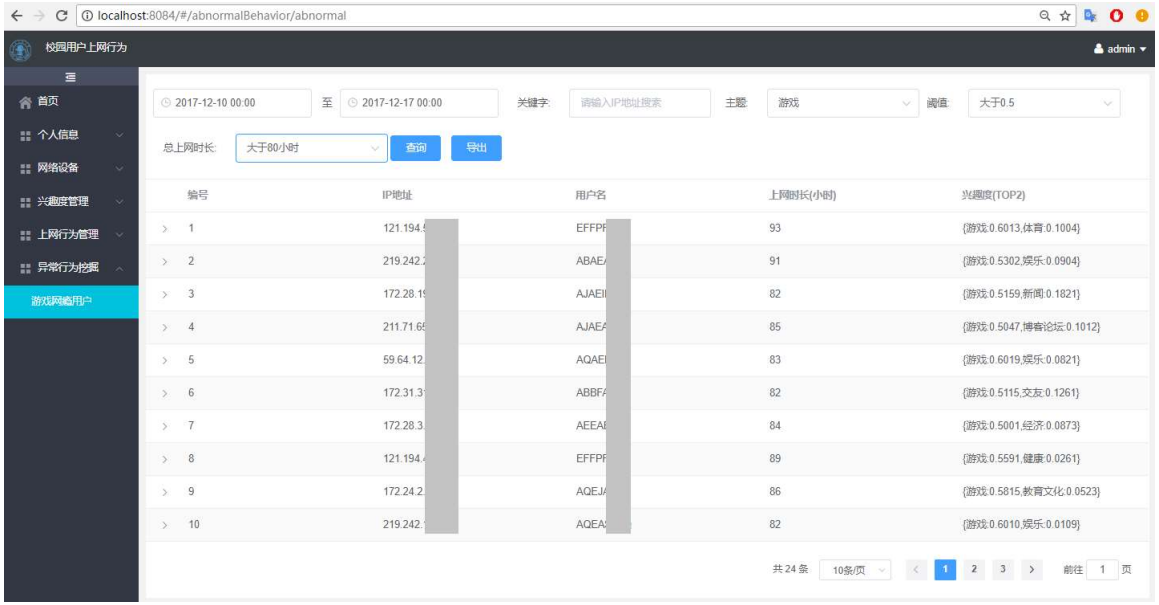
前端可视化模块（校园管理者端）—上网行为管理



所有校园用户访问网站TopN图

校园网络管理者可以查看一段时间内整个校园用户上网时长的人数，以及各上网时长的人数占比情况，以更好地了解校园用户的上网情况。

前端可视化模块（校园管理者端）—异常行为挖掘



异常行为挖掘模块

◆通过用户的上网时间段、上网时长与用户兴趣的结合，可以挖掘出潜在的用户异常行为。比如说挖掘出长时间沉溺游戏的同学，可以通过对其网络资源进行限制等措施，降低沉迷游戏的可能性。

◆通过设置查询的时间范围、异常行为检测的阈值、上网时长阈值等条件，查看异常行为列表。

五·结　论

五、结 论

◆校园网络作为互联网的一个重要组成部分，使用数据挖掘技术对校园用户网络行为进行分析一方面**使校园用户更加了解自己**，另一方面**可以为校园管理者提供十分科学有效的数据支撑**。

◆校园用户数量很大，用户访问的Web网页复杂多样，本文使用数据挖掘技术高效地从海量校园日志数据记录中挖掘校园用户的行为模式，找出了用户的兴趣点，追踪用户的上网轨迹及行为习惯，一方面**方便用户对自己的上网行为的认知**，另一方面也便于**网络管理员掌握学生的兴趣，进而个性化推荐，另外可以及时发现学生的异常行为**，从网络资源限制或心理疏导等方面着手解决，具有十分重要的意义和价值。

谢谢！

感谢各位专家的聆听！