个性化视频推荐系统@土豆

明洪涛
htming@tudou.com

内容提纲

- 推荐系统的目标
- 土豆网的数据与算法
- 推荐系统的流程
- 遇到的问题及其解决方法
- 实际转化率数据分享

推荐系统的目标

视频推荐系统定义：

根据用户的兴趣，为用户推送与他兴趣相关的视频。

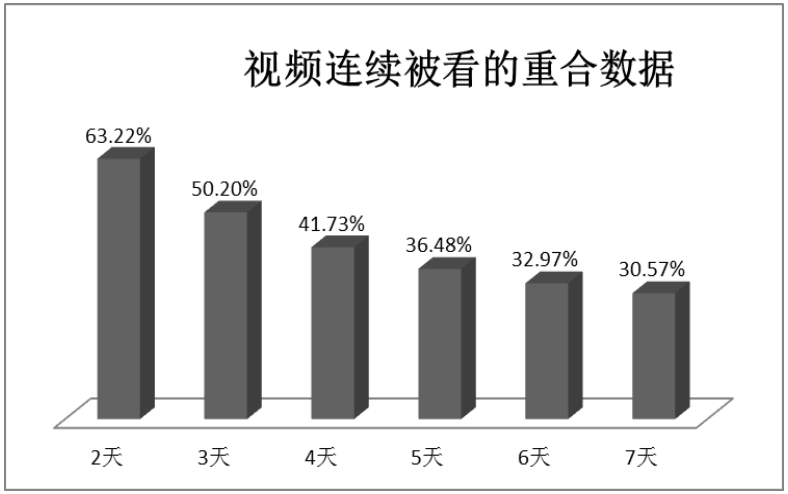
服务用户群：

进入土豆网看视频娱乐，但是不明确知道自己想要看什么视频，无法用词描述出来。与搜索、分类导航有差异。

考核指标：转化率=点击数/页面PV

数据与算法

被观看的总视频数千万/天左右，7天重合30%



隐性数据（挖、埋、收藏、评论、转帖）几十万/天，平均每个视频得到0.025/天，可推荐的视频平均0.043/天，用户信息可忽略，非登陆用户的历史记录随用户cookie的清空而消失。50%用户每天看两个视频。

数据处理后，推荐的视频库中有200万个视频，这些视频获得大约5千万/天的浏览量。

填充矩阵，数据的稀疏性在五万分之一，放弃了SVD、slopeone等算法。

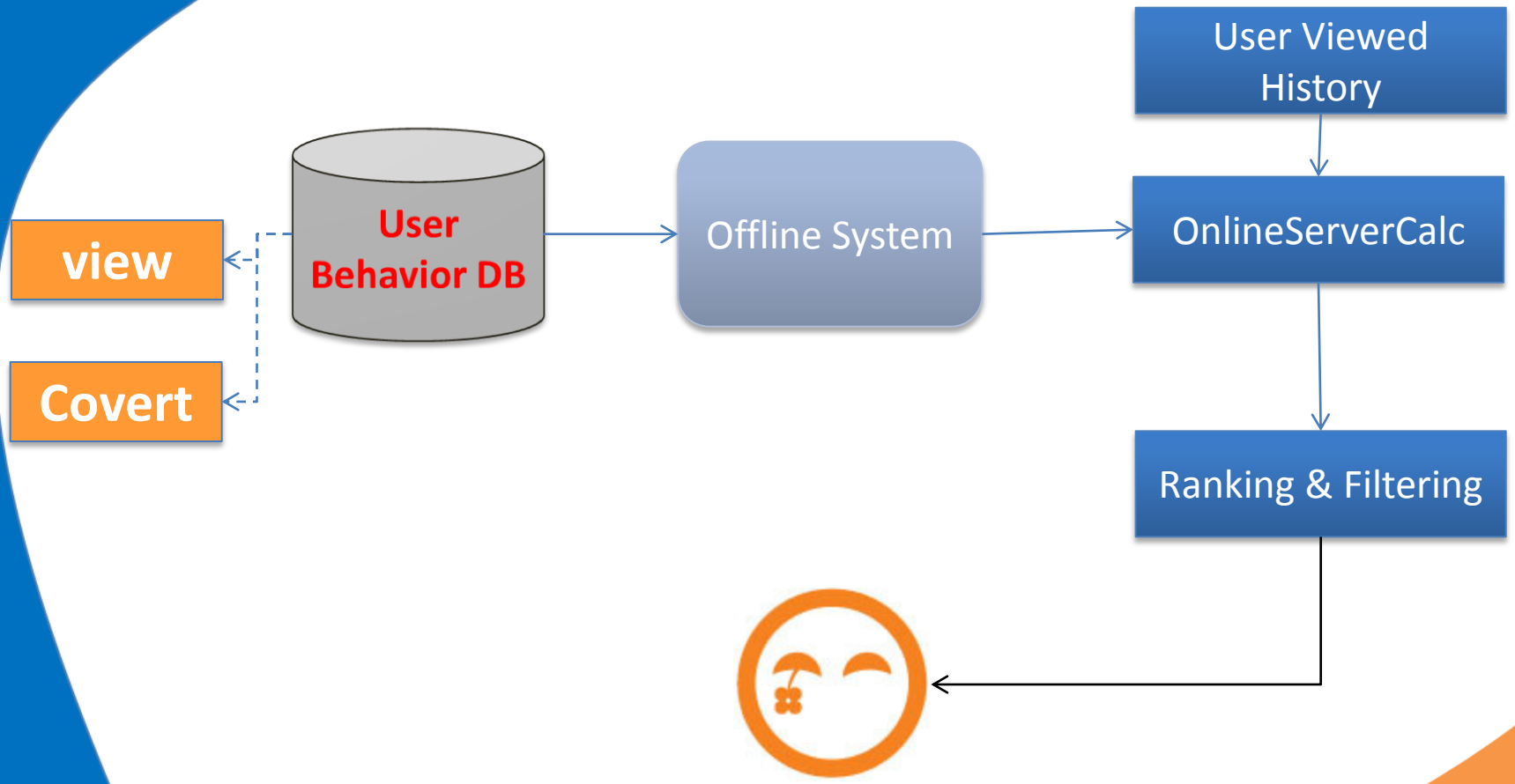
数据与算法

协同过滤：Item-based
假设：用户观看的视频是用户喜欢的。

	I1	I2	I3	…	…	In
U1						
U2		1		1		
:		1		1	1	
:			1		1	
Un		1	1		1	

$$Sim(Ix,Iy)=\frac{\sum_{(x,y,i)}^{n}(Ix \cup Iy) \in Ui}{\sum_{(x,y,i)}^{n}(Ix \in Ui) \cup (Iy \in Ui) - (Ix \cup Iy) \in Ui}$$

土豆推荐系统的流程


User Viewed History
User Behavior DB
Offline System
OnlineServerCalc
view
Covert
Ranking & Filtering

目前的应用到的产品

一、首页的“推荐给我”。
二、播放页下的“相关视频”。
三、频道首页：风尚、美容、女性、娱乐、原创频道。

即将上线的：搜索结果、排行榜下的推荐。

遇到的问题及其解决方法

冷启动问题
数据的稀疏问题
多样性问题
准确性问题

} 转化率的提高

一、系统服务性能问题

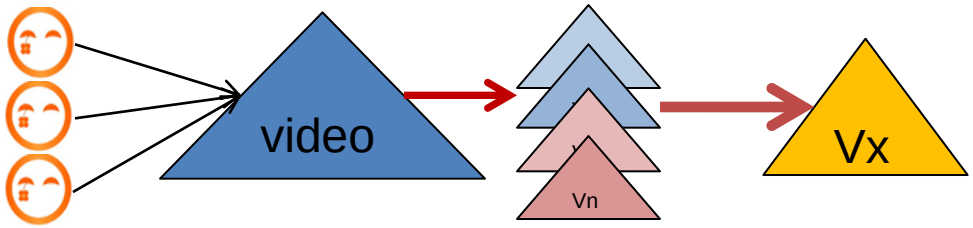
C/C++实现整套系统、通过Http对外服务.采用线下离线计算与线上实时计算相结合。

二、数据问题

爬虫数据、非典型性用户和视频数据。用群体用户的浏览来区分。

遇到的问题及其解决方法

脏视频：多个用户上传同一个视频、用不同的标签不同的视频ID号。使用视频MD5进行数据融合处理。



数据量的选择与清理数据问题？通过离线评测方式。通过用户观看量和视频被观看量来进行过滤。

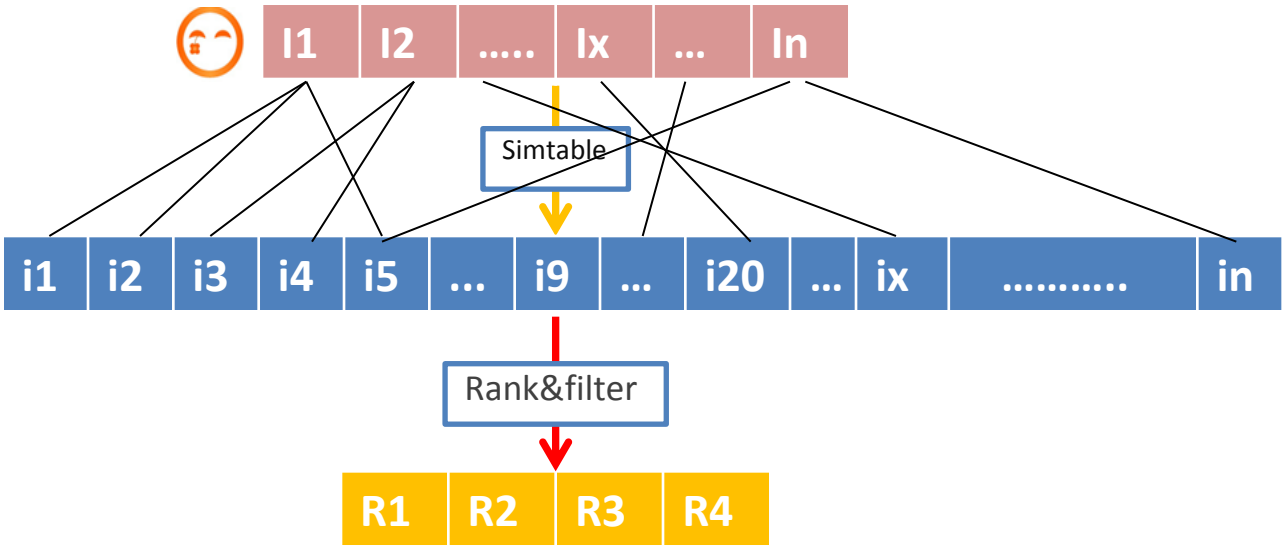
CleanDataByUser(Cmin,Cmax)

CleanDataByItem(Vmin,Vmax)

三、覆盖率、准确性和转化率三者的关系。覆盖率越小，准确性越高，冷启动问题就越严重，整体转化率越低。

遇到的问题及其解决方法

四、怎么推荐？



Rank:根据视频的质量和相似性来排序，视频的质量涉及到视频的浏览量，用户对视频的动作等数据。

Filter：过滤用户已经观看的，限定与用户兴趣不相关频道的视频个数。**过滤掉已经推荐。**

遇到的问题及其解决方法

相似度的问题： 设定阀值，在计算时只保留N列，在排序前安比例去掉相似性差。

itemid	item1:sim	item2:sim	item3:sim	item4:sim
2445424	25914738:0.01143	7799720:0.010	124437085:0.01053	124437946:0.001001
9252575	25410570:0.15122	15685303:0.18	22914131:0.117647	189613237:0.000422
5582692	3975273:0.101212	69879414:0.07	140725740:0.07401	127722333:0.000142

公式的选择（意义不大）： 余弦、修正余弦等距离公式。

用户的兴趣： 分长短兴趣和频道的偏好，最近观看视频反映用户短兴趣，长期的一个观看历史记录反映的用户长兴趣，可以对用户长兴趣建模。

过热和过冷视频，直接CleanData的时候处理掉。

遇到的问题及其解决方法

五、评估方法：常用的MAE、CTR。
离线评测时使用过CTR，推荐与点击比30%左右，与实际点击比相差10倍。现在使用的是线上实时CTR，分流量做A/Btest。

六、产品、位置、转化率

遇到的问题及其解决方法

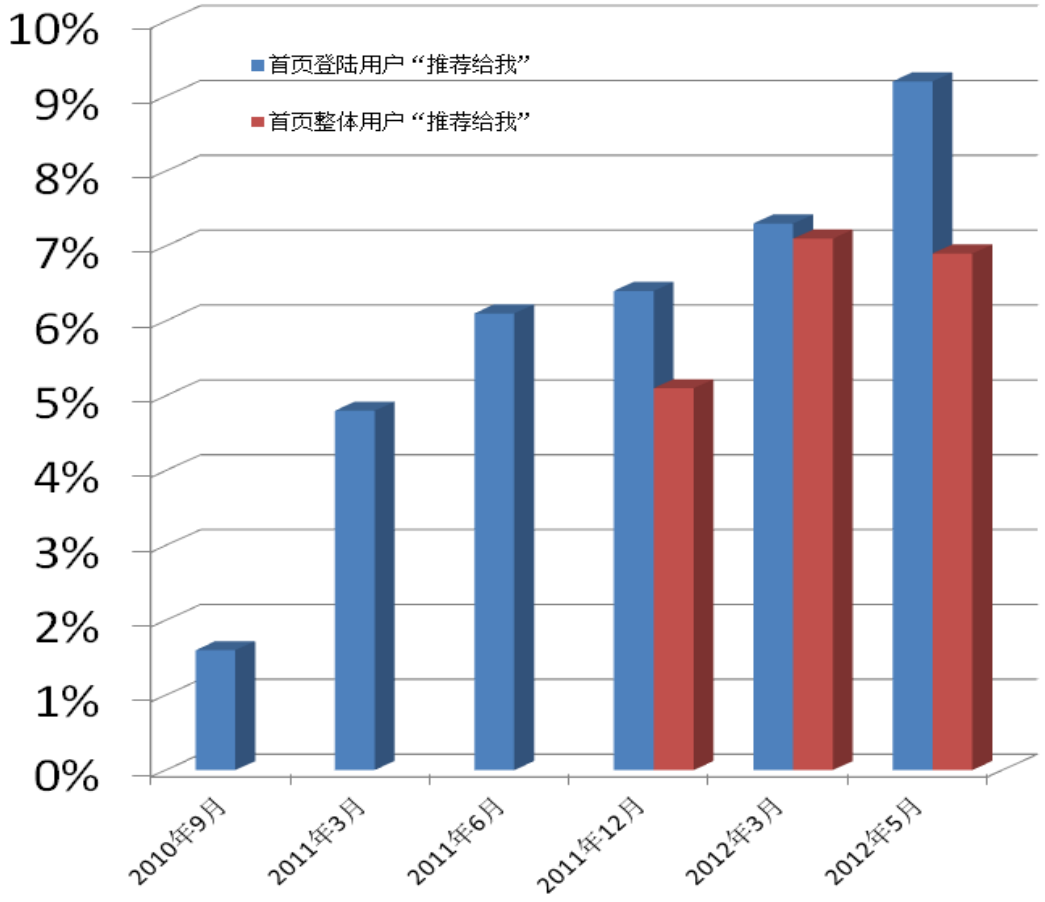
位置优势？ 同一位置不同算法的PK，转化率是其他算法的三倍。

推荐与美女PK

VS

失败的经验……

实际转化率数据分享



在当天使用推荐的用户中，
有40%左右的用户是二次以上用户。

优化无止境，细节决定成败

Q&A

THANKS!

推荐讨论QQ群：167485994

微薄：http://weibo.com/htming