# APP首页推荐

## 背景

使用百姓网APP的行为大致分两种：一种是目的性比较强的，如想要找二手iPhone 4S的直接关键词搜索，想要租房的则会直接浏览房屋出租或日租/短租类目下的信息；另一种是没有明确需求的，如老用户习惯性地打开看看，或者新用户尝鲜时随便翻翻。本项目主要是**针对后一种行为**的。

新版的百姓网APP将增加“首页”板块（iOS版的之前已经内部偷跑了），用户打开APP后首先看到的就是“首页”的信息列表。如此一来，不管将要进行何种行为，用户的目光都会首先短暂停留在“首页”。若能通过**有效的推荐**，让用户更快地发现自己感兴趣的内容，提高用户在“首页”的浏览-点击转化率，不仅能提高老用户对APP的使用频率，还能提高新用户保留和继续使用APP的概率。

## 调研

受屏幕和操作所限，移动端应用更重视内容的有效性，所以大部分项目都或多或少会用到或将要用到推荐系统或推荐算法：

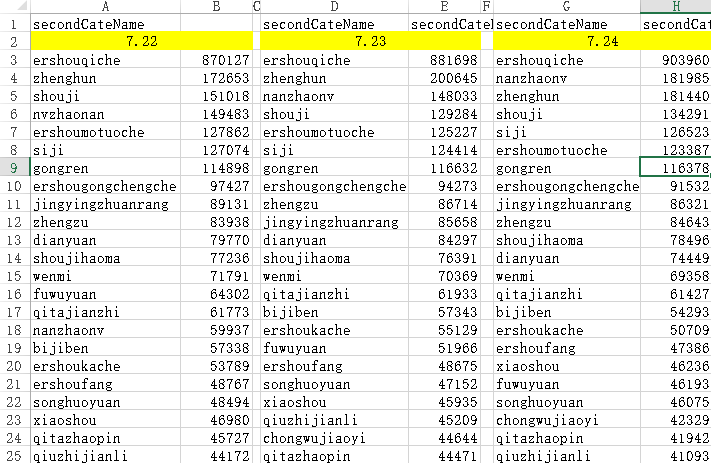
1. **Foursquare**：**根据用户当前的位置来推荐相关的信息。**比如用户在一个饭店里面，就会推荐一些菜品。如果在一个景点，就会推荐附近其他好玩的景点，好吃的饭馆等；
2. **Facebook/Twitter**：**根据用户当前位置来推送广告。**比如用户所在位置附近有商家投放了广告，而通过用户的历史信息知道该用户可能会对该广告有兴趣，则给用户推送该广告；
3. **豆瓣电台**：**根据用户的收听记录来推荐音乐。**根据用户的“红心”记录，（可能会综合考虑“跳过”记录），通过**协同过滤**可以推测用户对音乐的喜好，从而推荐用户可能感兴趣的音乐；
4. **知乎发现**：**推荐热门问答或热门用户。**筛选并展示过去 24 小时内被赞同最多的答案和用户，以及由编辑推荐不那么热门，但有价值的回答和用户；
5. **Amazon：推荐用户可能喜欢的商品。**根据用户的浏览记录、购买记录、给购买物品的打分推荐以及根据与用户兴趣相似的用户的记录推荐。

## 方案设计

结合调研结果和百姓网的无社交、信息即广告、分类完备、时效性较强等特点，考虑从两方面进行推荐：

1. **地理位置**：现在的“首页”显示的已经是附近3公里的信息了，但这些信息的类目分布是和**发布信息**的类目分布相关的，而对于浏览“首页”的用户来说，这样的分布并不是最合理的。相对合理的做法是从**浏览信息**的用户角度出发，**给用户推荐附近浏览得最多的类目的信息**，这样的推荐既适用于冷启动，也有其考虑地域差别的合理性。
2. **浏览记录**：由于百姓Ad信息时效性较强、特色化较弱的天然属性，不太可能出现像电商网站上面那种热门浏览Ad推荐，但是分析每日的数据可以发现存在热门浏览类目。

**7． 22/23/24单日类目浏览数量排行榜**



因此可以从两方面入手，一方面分析用户最近的若干条浏览记录（比如200条），得到浏览最多的前三个类目（数量可调整）及其在这三个类目中的相对百分比，根据这个百分比从数据库中获得相同百分比的三个类目中的Ad作为推荐结果，对于冷启动的情况，则可以返回前K个热门类目中的相同百分比的Ad。这种方法结合了个人化推荐(personalized recommendation)和项推荐，相对于社会化推荐(social recommendation)推荐的准确度更高，但是扩展性不强，有些用户不怎么使用的且比较冷门的类目中的Ad可能永远不会出现在推荐结果中。

## 具体方案

1. **地理位置（仅上海。。）**
   1. 从APP\_TRACKD44ATA导出前1天，，（**用前3天或前一周的数据更好**）上海范围内（**纬度：31.14 - 31.345；经度：121.36 - 121.58**）的包含GPS或/viewAd动作的浏览记录：

|  |
| --- |
| **select**  UDID**,**  bx**.**dbo**.**fHiveKV**(**attr**,** 'secondCateName'**)** **as** secondCateName**,**  bx**.**dbo**.**fHiveKV**(**attr**,** 'GPS\_geo'**)** **as** GPS\_geo  **from** APP**..**TRACKDATA\_20130720  **where** **(**to\_action**=**'GPS' **or** to\_action**=**'/viewAd'**)** **and** UDID **in** **(**  **select** **distinct** UDID  **from** APP**..**TRACKDATA\_20130720  **where** to\_action**=**'GPS' **and** attr **>** 'GPS\_result1GPS\_geo(31.140000' **and** attr **<** 'GPS\_result1GPS\_geo(31.345000'  **)**  **order** **by** UDID |

* 1. 把导出的csv数据处理成可导入solr的格式。由于导出的csv仅按UDID排序，对于同一个UDID的记录，仅考虑含有GPS信息之后的类目，认为该类目是定位成功后在该GPS位置浏览的；另外，字段间分隔符换成**分号**是因为经纬度类型是以逗号分隔的，处理后的字段类型分别是id(string)/category(text\_general)/store(location)。处理前后的对比如下：

|  |  |
| --- | --- |
| origin.csv | solr.csv |
| UDID,secondCateName,GPS\_geo  Id1,,"(31.222395,121.370865)"  Id1,,"(31.222395,121.370865)"  Id1,qitazhaopin,  Id1,qitazhaopin,  Id1,qitazhaopin,  Id1,qitazhaopin,  Id2,,"(31.256516,121.564156)"  Id2,,  Id3,,"(31.151415,121.323746)"  Id4,shouji,  Id4,,"(31.325829,121.466400)"  Id4,shouji,  Id4,shouji, | id;category;store  Id1;qitazhaopin;31.222395,121.370865  Id1;qitazhaopin;31.222395,121.370865  Id1;qitazhaopin;31.222395,121.370865  Id1;qitazhaopin;31.222395,121.370865  Id4;shouji;31.325829,121.466400  Id4;shouji;31.325829,121.466400 |

* 1. 把处理后的csv导入solr中，就可以查询指定位置附近3公里浏览最多的10个类目了：

http://ip:port/solr/select?fl=category&q=\*:\*&fq={!geofilt}&sfield=store&pt=31.247443,121.489879&d=3&facet=true&facet.field=category&facet.limit=10

* 1. 在ad\_recommend方法里加入上述查询，把$allAds里不属于上述类目的结果去掉，作为最终结果返回。

1. **浏览记录**
   1. 每次首页发送推荐算法请求时，根据本地保存的用户浏览记录，计算出用户最近浏览最多的前三个（数量可调整）类目及所占百分比，比如200条历史记录中ershouqiche 有100条，nvzhaonan有40条，shouji有20条，其余40条为其他类目，则ershouqiche、nvzhaonan、shouji大的百分比为5/8，2/8，1/8。
   2. 推荐算法中根据这三个类目及百分比从三个类目中获取最新的Ad。比如最后要返回40条Ad，暂且不考虑地理定位部分的Ad，那么分别返回25条ershouqiche、10条nvzhaonan，5条shouji的Ad
   3. 如果发现用户历史记录太少（比如冷启动），则只根据地理位置推荐。

**3 推荐结果**

将根据地理位置推荐的Ad和根据浏览记录推荐的Ad合并排序去重后返回给用户。针对用户刷新频率过快的问题，第一可以在前端做判断是否要重新获取数据还是默认已存在的结果即为最新推荐，第二可以在算法中做些cache处理，加快响应速度。

## 后续工作

1. 跟踪评估推荐结果，调整参数，如纳入统计的公里数、推荐的类目数等。还可以考虑比较**统计所有trackdata**和**只统计发生在首页的trackdata**的推荐效果。
2. 调整查询类目数，比如3、4、5，或者调整热门类目查询的类目及其百分比（试试热门类目随机推荐和固定推荐的效果对比），对trackdata进行分析采用最优策略。