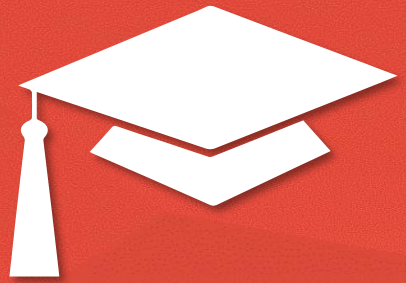
桂林电子科技大学
GUILIN UNIVERSITY OF ELECTRONIC TECHNOLOGY

“达观杯”推荐算法比赛汇报

团队 你看见过我的小熊吗

队长 徐绍凯

队员 侯嘉婷、王海洋



赛题解释

- 赛题要求：

  提供了候选用户在某三天对于资讯内容的行为数据、候选资讯数据、全量资讯数据；据此预测每个候选用户在第4天会产生行为的资讯列表。

- 数据情况：

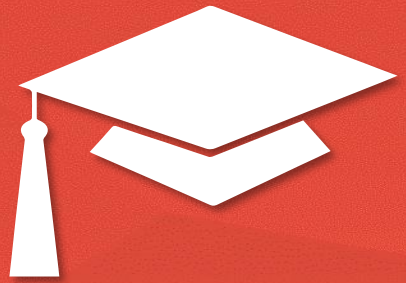
  train.csv：训练集，某三天内用户行为数据
  test.csv：测试集，部分用户第四天的行为数据，不包含待推荐用户
  candidate.txt：待推荐用户ID
  news_info.csv：候选资讯内容，包含资讯ID，所属类别，产生时间
  all_news_info.csv：全量资讯内容

●训练数据含义说明：

列名	描述	数据类型
user_id	用户唯一ID	string
item_id	资讯唯一ID	string
cate_id	资讯类别ID	string
action_type	用户行为类型	string
action_time	行为发生时间，秒级时间戳	int

●资讯数据含义说明：

列名	描述	数据类型
item_id	资讯唯一ID	string
cate_id	资讯所属类别ID	string
timestamp	资讯创建时间，秒级时间戳	int



数据分析

●赛题分析

✓用户冷启动问题

由于待推荐用户在训练集中均有记录，所以题目中不存在用户冷启动问题。

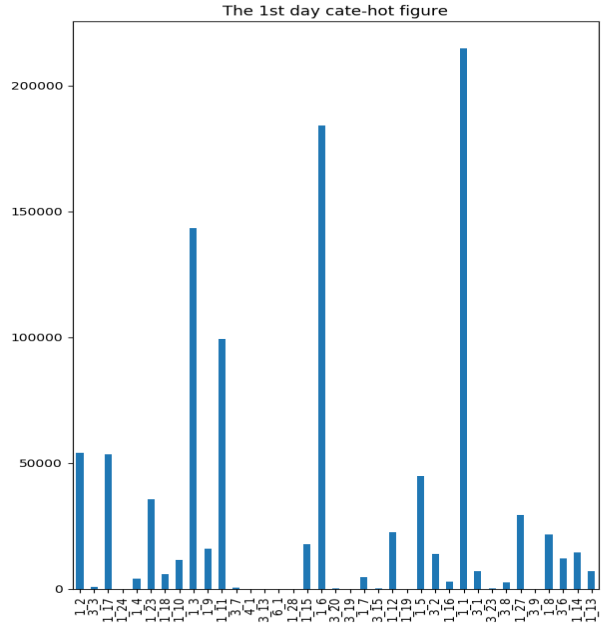
✓Item冷启动问题

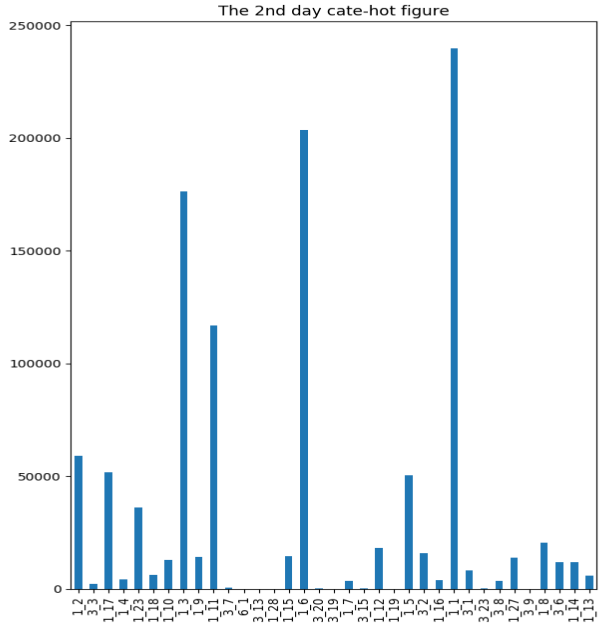
由于新闻资讯有很强时效性，对资讯的推荐不同于其他商品的推荐。

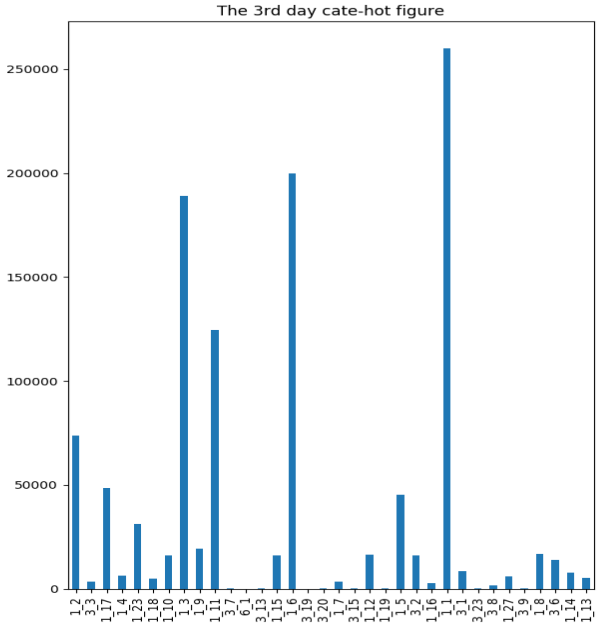
新产生的资讯在前三天中不会有任何浏览记录，赛题中仅提供了其类别信息和产生时间，并未提供item的具体信息。对于具体的某一类资讯中，最终将哪些item推荐给用户很难抉择。

test数据为待推荐用户之外的部分用户第四天的真实数据，所以依据test数据可以得知新item中用户感兴趣的item。

●描述性统计

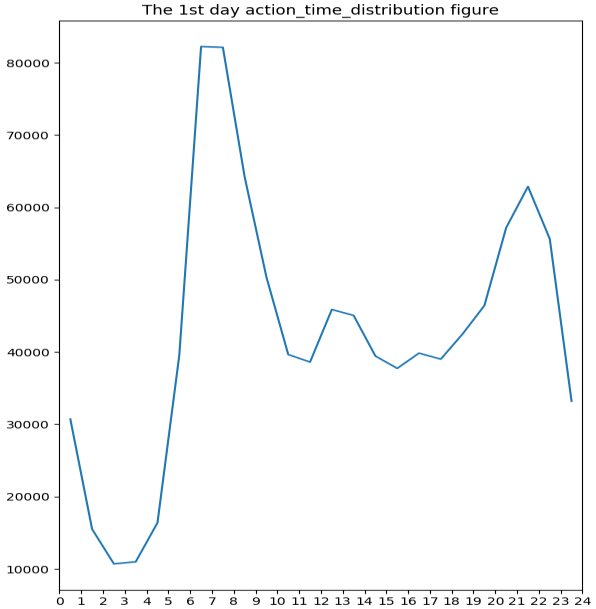


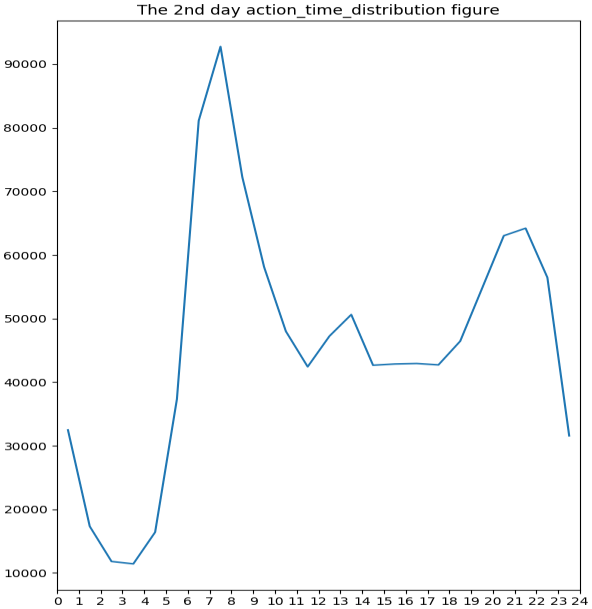


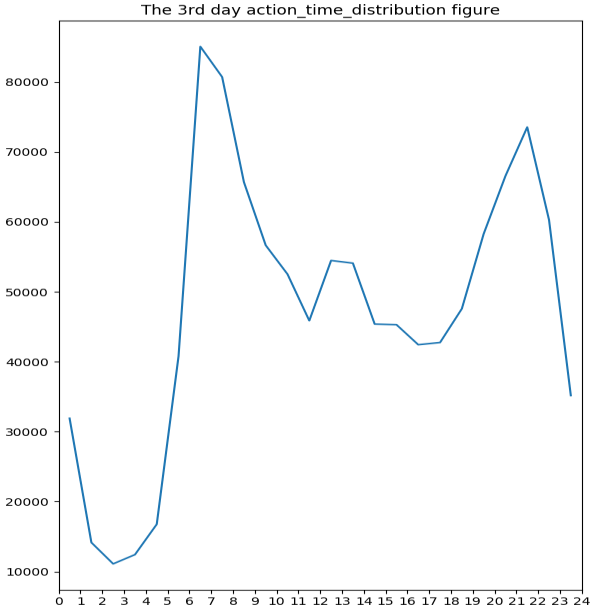


从资讯类别流行度分布图可以看出，资讯类别在每天的流行度分布是极其相似的，不同的资讯类别有其各自的用户群体。

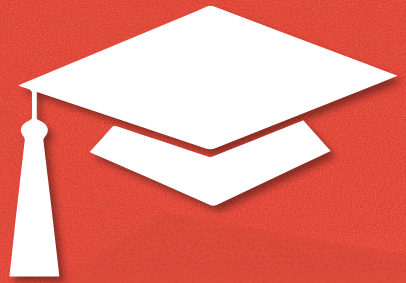
- 描述性统计







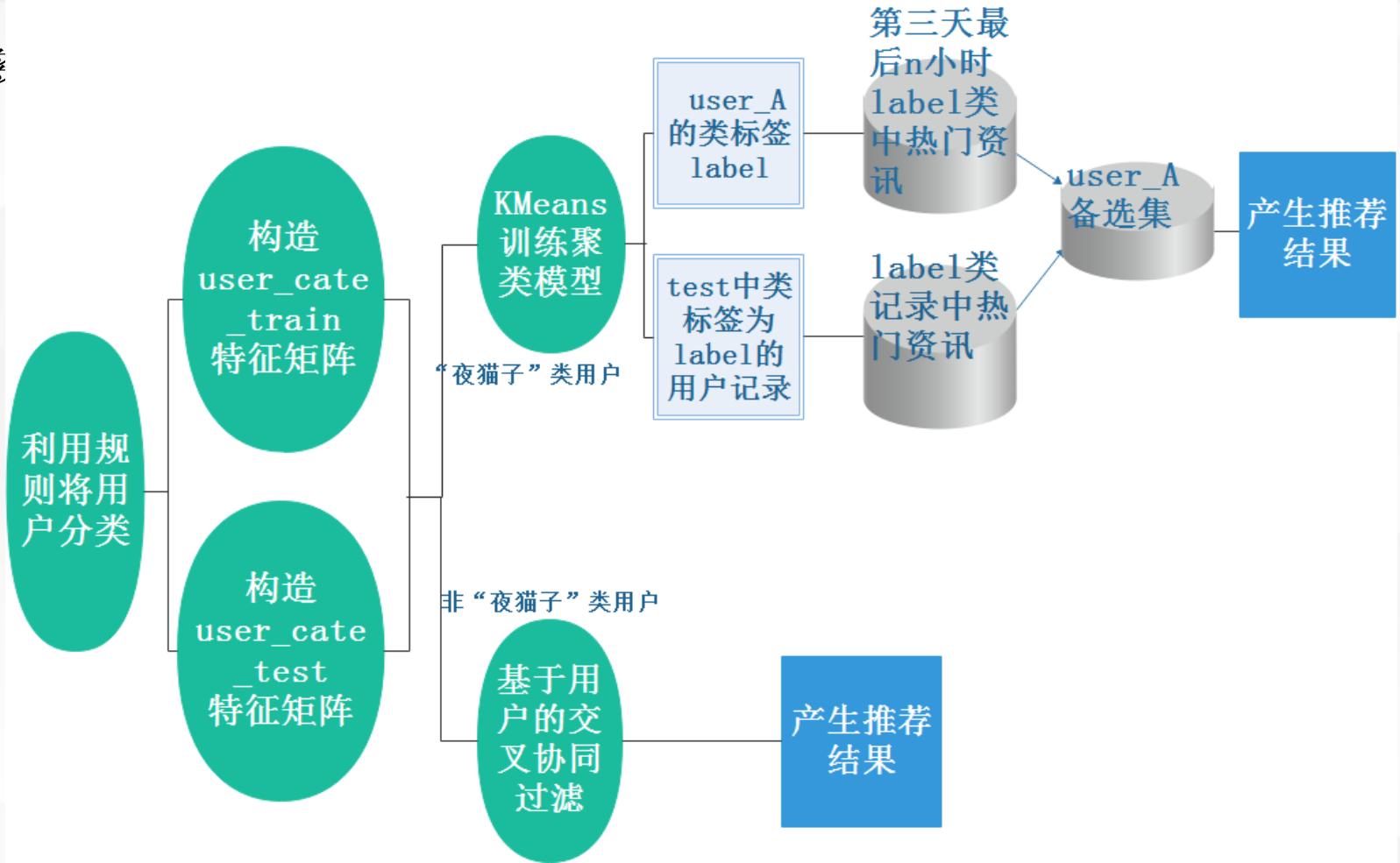
从用户行为时间的分布来看，凌晨三点是个转折点，在三点之前，用户的行为数量逐渐减少，在两点与三点之间达到最小，从三点以后用户行为数量又开始逐渐增多，在六点到七点之间达到顶峰。



实施方案

- 思路
  - ✓利用train数据可提取用户对资讯资讯类别的偏好；
  - ✓利用train数据可提取用户的行为时间习惯；
  - ✓利用test数据可提取新产生的热门资讯。
- 算法
  - ✓规则
  - ✓K-Means聚类
  - ✓协同过滤

●多


利用规则将用户分类
构造user_cate_train特征矩阵
构造user_cate_test特征矩阵
KMeans训练聚类模型
“夜猫子”类用户
非“夜猫子”类用户
user_A的类标签label
test中类标签为label的用户记录
第三天最后n小时label类中热门资讯
label类记录中热门资讯
user_A备选集
产生推荐结果
基于用户的交叉协同过滤
产生推荐结果

●实施方案

✓特征矩阵构建

•资讯类别作为特征，待推荐用户作为样本，values取训练集中对应用户在对应资讯类别中的前三天行为值总和。

•考虑到test数据中没有用户的行为类型，所以对于训练集特征矩阵，当用户对某一item有行为时，无论其是否有多个行为类型，均将其values取值为1。

•Test特征矩阵取与train特征矩阵相同的资讯类别作为特征，样本为test中所有用户。

✓基于规则将用户分为两类

- 凌晨3点是用户行为的一个明显分界线。

- 若用户在0点-3点之间的行为占全天总行为的a（0<a<1）以上，则将该用户归为“夜猫子”类用户。

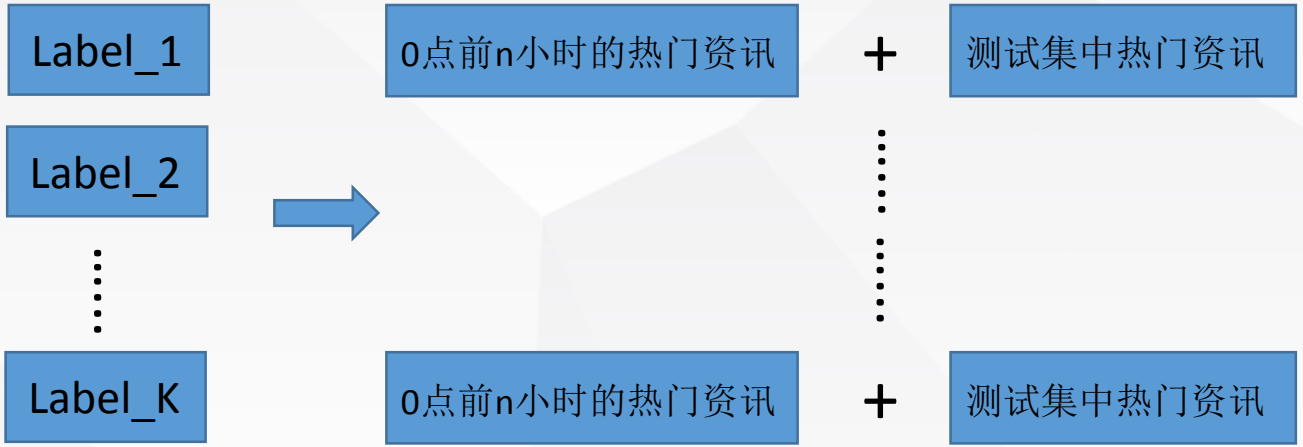
- 若用户在0点-3点之间的行为占全天总行为的a（0<a<1）以下，则将该用户归为非“夜猫子”用户。

✓K-Means聚类

•聚类目标：将资讯类别偏好不同的用户划分到不同的类中，对每一类中的待推荐用户生成备选资讯集合

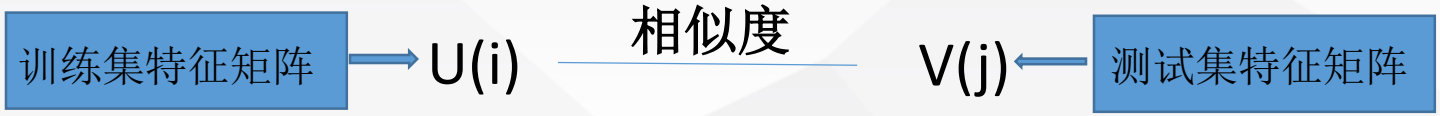
•聚类结果：将待推荐用户聚为K类，并将test中所有用户分别划分到对应的类当中。

✓K-Means聚类

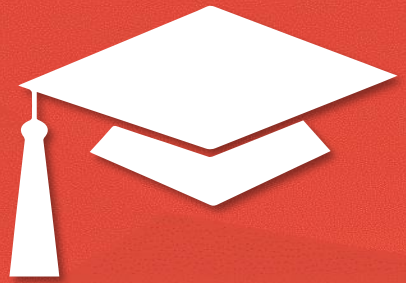


•对于“夜猫子”用户，他们的主要行为时间段集中在凌晨0点到3点之间，借助于热门资讯实时推送的思想：将0点前n小时内的热门资讯与新产生的资讯（即test中记录）作为备选集。从而对于每一类用户产生一个带筛选的备选集合，集合内item按流行度排序。

✓基于用户的交叉协同过滤



●对于非“夜猫子”用户，根据其日平均特征矩阵与test特征矩阵，求得非“夜猫子”用户与test中用户的相似度，取相似度最大的m个test用户记录中的item并集作为备选推荐资讯，将相似度换算为权重，求得每个备选item的得分。



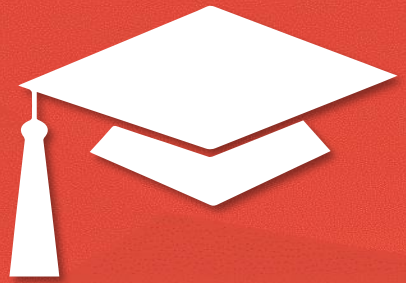
结果与评价

- 可调节参数
  - ✓ a：“夜猫子”用户分类规则所选取的比例阈值
  - ✓K：K均值聚类的类别数， 通过Calinski-Harabasz Index分数确定K值选取
  - ✓n：第三天最后n小时的热门资讯作为“夜猫子”用户的部分备选集
  - ✓m：协同过滤中选取的相似用户个数
- 评价方法

$$ap@5 = \sum_{k=1}^{5} P(k)/\min(m, 5)$$

$$MAP@5 = \sum_{i=1}^{N} ap@5_i/N$$

- ✓提交程序运行结果后，以官方评价系统做为结果评价得分



总结

●算法缺陷

手动调参

参数的选择没有建立合适的评价体系，仅仅依靠对数据的初步分析加手动调参来实现结果的优化，影响整体效率。

改进方法

进一步量化参数选择与算法结果的关系，对各步骤建立评价体系，并将其嵌入到完整的算法框架中。

Test数据

在真实场景中不会有类似于赛题中的test数据出现，故真实场景下无法通过test数据来获取新产生资讯的热度信息。

改进方法

在真实场景中可对资讯的内容进行提取，例如标题、摘要等，通过NLP等技术为用户和资讯添加标签，结合实时推荐技术，融合多种算法，实现新闻资讯的精准推荐。

●心得分享

✓要实现精准推荐的前提不仅仅要求对算法熟悉，还要求开发者要对用户数据有透彻的理解，只有理解数据，明确用户需求，才能将算法更好的应用于数据。

✓优秀开源框架的运用会使得算法效率得到很大的提升，在理解算法的基础上还需要掌握大数据框架的使用。

✓不要把思维仅仅局限在单一的方法上。

欢迎各位专家批评指正