站外智能定向业务分享

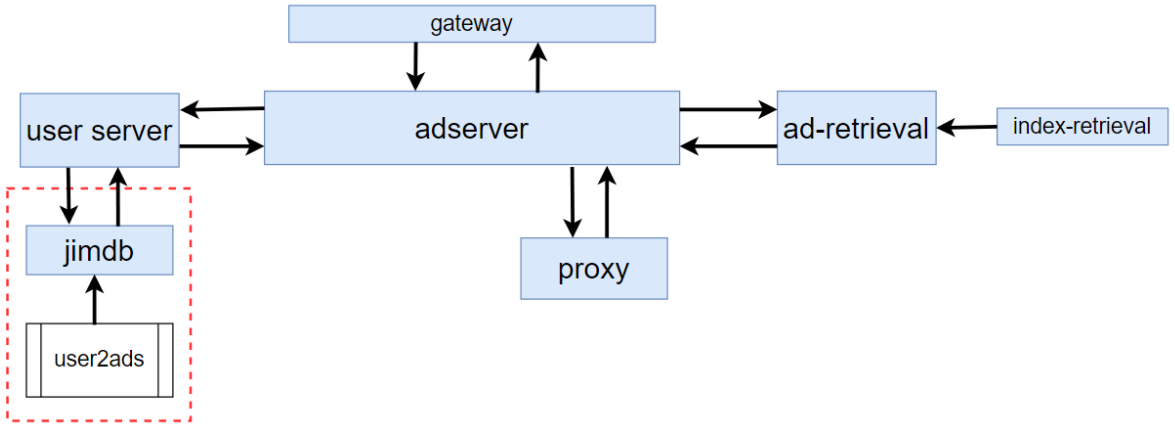
赵昆

JD.京东.COM

目录　CONTENTS

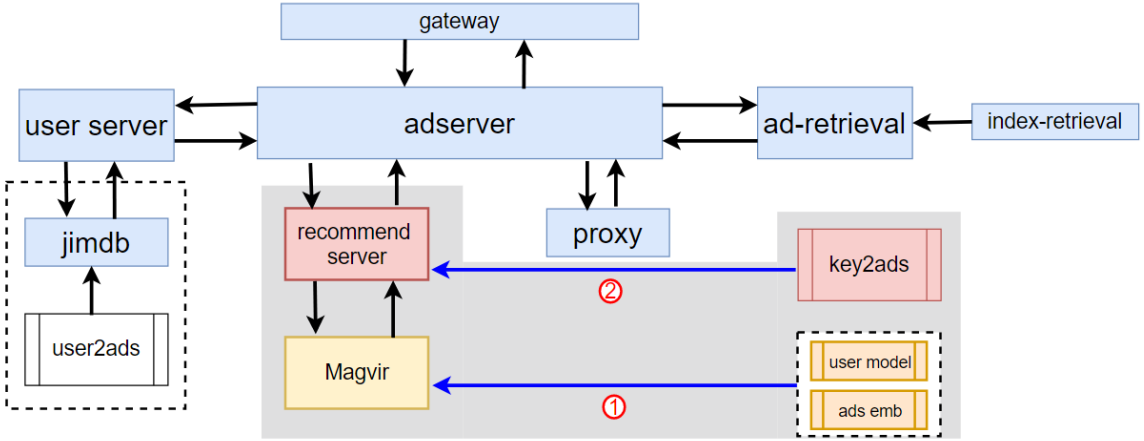
站外智能定向-线上逻辑

- 传统视觉定向召回方案



- 离线挖掘用户的站内外行为(浏览、加购等)，计算单行为和广告的相关性，得到user2ads词表，写入user server下的jimdb服务

- 基于re和向量化召回的方案



- 基于re的二阶段召回

离线算法挖掘key2ads词表，加载到re内存，线上请求根据用户实时行为数据产生key，在re进行检索完成召回过程

- 基于magvir的模型召回

离线训练双塔模型，magvir加载user model和广告物料emb，在线实时产生user emb，在物料emb中进行topk检索

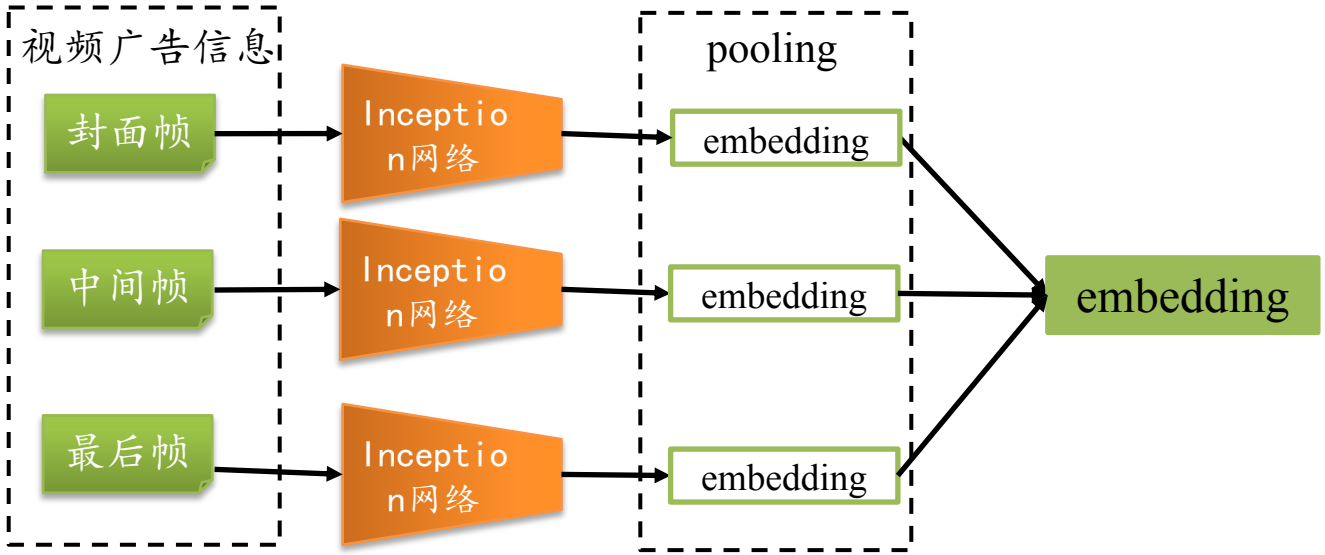
站外智能定向-分支汇总

- 基于图片/视频emb的召回分支（直投图片/视频渠道）
- 基于cidbrand emb的召回分支 （直投图片/视频渠道、AN渠道）
- 基于query的召回分支 （直投图片/视频渠道、AN渠道）
- 基于频繁二项集的召回分支（直投图片/视频渠道、AN渠道）
- 基于双塔模型的召回分支（直投视频渠道）
- 基于高商业人群的召回分支 （直投视频智能出价广告）
- 基于联邦学习的向量化召回分支 （直投视频渠道）

召回-基于视觉召回

- 图片emb处理

- 视频emb处理 (第一版)



✓图片召回

基于用户浏览的商品图片emb和广告图片的emb的余弦值进行召回

✓视频召回

视频广告处理成成图片emb，召回过程和上面一致

召回-基于视觉召回

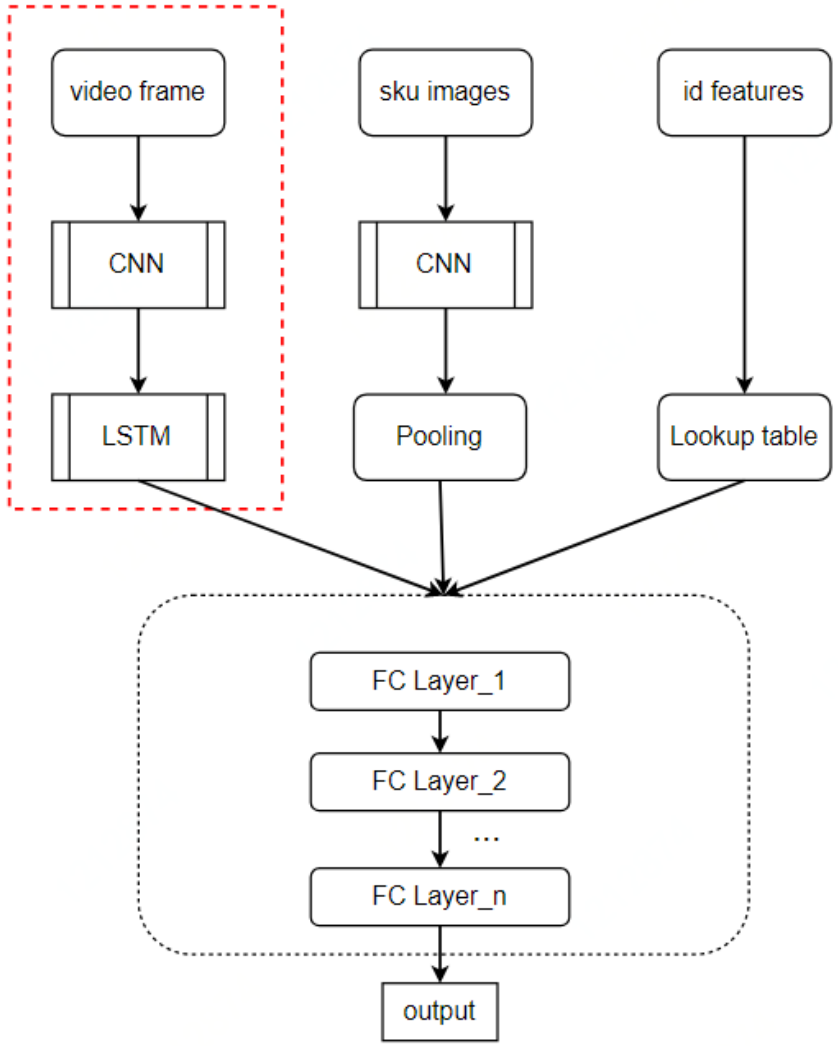
- 视频emb处理 (第二版)

特征处理

- ✓ **视频处理**：去掉黑帧之后，每隔一秒取一帧，取前15帧
  （目前已升级成关键帧提取，基于ffmpeg工具）
- ✓ **图片特征**：用户最近十个浏览行为
- ✓ **离散特征**：一些离散特征

模型CNN+LSTM建模

- ✓ 结合业务训练的有监督模型
- ✓ 视频的emb序列经过LSTM，考虑帧与帧之间的序列关系
- ✓ 基于用户视频行为来召回视频



召回-基于cidbrand emb的召回分支

- 方案
  - ✓ 基于全站用户的浏览数据，构建session数据，基于fasttext训练得到cidbrand emb，生成召回词表
- 具体实现
  - ✓ 基于用户行为数据，根据session切分规则，得到session序列数据，序列每个元素是sku对应的cid3和brand
  - ✓ 基于fasttext，训练session数据得到cidbrand emb (fasttext的字符粒度优化)
  - ✓ 广告emb：基于广告(跟单sku、落地页sku)的cidbrand，根据emb计算得到cidbrand2ads 词表，进行服务
- 字符粒度优化
  - ✓ 数据：

```
1 9755_8701 9755_8701 9755_8701 9755_8701 9755_8701 9755_8701
2 1355_200300 9719_246913 9719_52719 9719_213580 9719_93324 9719_233905 9719_197611 9719_235786 655_12669 655_12669 655_12669 655_12669 655_12669 655_12669 655_12669 655
  _12669
3 17399_162992 741_13680 741_13680
4 1300_14331 1300_11026 1300_11026 16965_9639 1300_14331 1300_11026 1300_11026 1300_11026 1300_11026 1300_11026 1300_11026 1300_11026 1300_11026 1300_14331
5 9763_3552 12417_8981 12347_12669 12347_231757 6190_23944
```

  - ✓ 字符粒度优化：

按fasttext默认的字符粒度的n-gram没有意义，但是按照"_"分割，直接学习cid3，brandid, cid3_brandid三种粒度信息，可以减少无意义的ngram的噪音干扰，突出用户行为意图

召回-基于query

- 方案
  - ✓ 基于用户搜索行为和商品的交互数据，建模query和类目的映射关系，根据类目和广告挂靠，完成query2ads的挖掘
- 实现细节
  - ✓ 基于query场景的行为数据，基于fasttext训练query2cid3的分类模型（搜索广告组数据）
  - ✓ 解析站外广告物料，根据广告的跟单sku的cid3信息，构建cid2ads映射关系解析站外广告物料，根据广告的跟单sku的cid3信息，构建cid2ads映射关系
  - ✓ 根据query2cid和cid2ads映射关系得到query2ads召回词表

召回-频繁二项集

京东

- 方案
  - ✓ 基于用户自然浏览/购买行为数据，挖掘sku2skus的关联规则

支持度：支持度表示的是项集X和Y同时出现在购买记录中的频繁程度，用于剔除偶然项集

$$s(X \to Y) = \frac{N(X,Y)}{N}$$

置信度：置信度指的是Y出现在包含X的力矩中的频繁程度，用于衡量规则的可靠性（防止头部效应）

$$c(X \to Y) = \frac{N(Y|X)}{N(X)} = \frac{N(X,Y)}{N(X)}$$

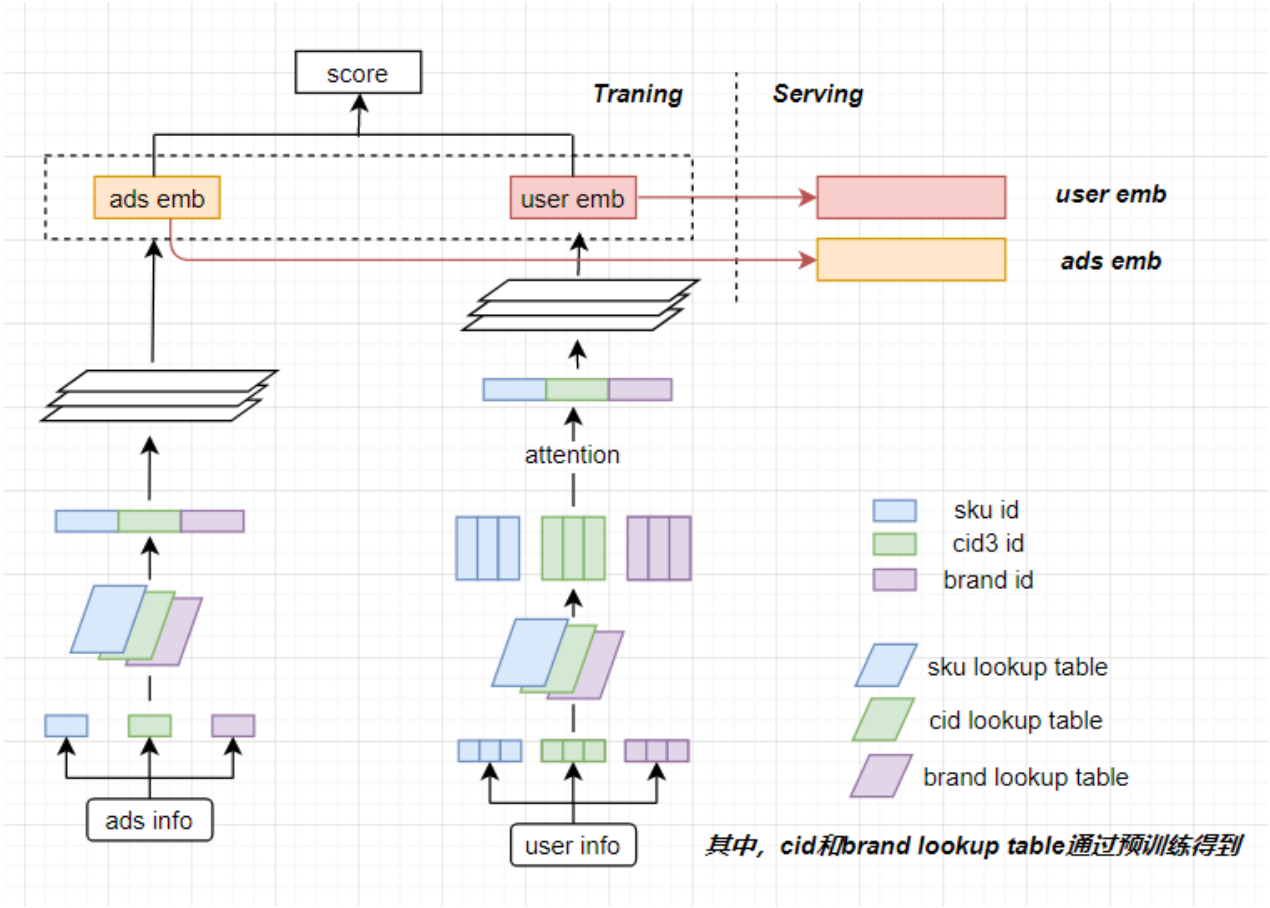
- 数据构建
  - ✓ 构造用户[browse session]->purchase数据流，挖掘browse->purchase频繁项集
  - ✓ 线上以用户浏览行为作key进行召回触发

召回-基于双塔模型的向量化召回

- 背景
  - ✓ 根据用户的行为兴趣序列与广告/商品的互动行为，建立用户和广告/商品的双塔模型
- 两个召回思路
  - ✓ user&ads双塔：

    基于线上用户与广告交互行为，训练模型（渠道人群以及广告特点可学习；数据稀疏，不易扩展）
  - ✓ user&sku双塔：

    基于全站用户自然行为，训练模型（基于自然数据训练，广告emb进行挂靠；数据充分，易扩展）



智能定向-待解决问题

- 召回框架问题

➢ user server侧的一路召回

问题：（1）用户行为不实时，占用离线资源

（2）线上召回架构不统一，且无分支标记，不方便数据分析，

方案：迁移re（但站外行为数据，线上无法获取）

➢ Re侧的一路召回

问题：（1）挖掘key2ads相关性词表，ads是material粒度，由于渠道限 制，会导致存在很多无效物料，挤占队列和资源 （内存）

方案：广告位合并之后会有缓解；临时方案，re做三级挂靠

- 基础能力建设

➢ 完善分支效果的离线评估流程

➢ 明确渠道划分，加强算法能力的迭代迁移

智能定向-待解决问题

- 智能定向算法方向
  - 不同单元效果差异较大（消费/转化）

    问题：（1）基于模型召回结果集中，badcase较多

    （2）智能定向人群量级不均匀

    方案：（1）多样性控制、相关性控制

    （2）基于单元粒度的动态扩量调节 召回策略支持调节扩量量级
  - 根据cpc动态调整阈值策略

    问题：目前高cpc广告，会限制扩量量级（大促期间不合适）

    方案：为减少线上影响，可考虑逐步放开
  - 广告信息挖掘不充分

    对活动、店铺广告落地页信息进行挖掘
- 智能定向产品方向
  - 产品形态单一，无法得到不同单元的不同需求（扩量、高点击、高转化、拉新）
  - 结合智能出价的具体方案（是单独产生分支词表，还是不区分智能出价和非智能出价）

智能定向总结的cf

1.1 智能定向整体概况介绍

站外召回现状梳理20210121

1.2 lookalike的相关介绍

抖音lookalike触发优化 基于user emb的

1.3 粗排相关的介绍

站外粗排 初版粗排（后验词表）

头条直投ctr粗排模型方案 正在进行，线上方案粗排模型

二 智能定向目前的召回分支和相关任务链接

站外智能定向在线全量任务汇总

三 智能定向线上相关

站外智能定向线上配置以及线上代码开发cr

四 智能定向的一些后期规划相关

召回策略支持调节扩量量级 初步方案

https://cf.jd.com/pages/viewpage.action?pageId=432058873 智能定向产品调研

谢谢！

Thank You！