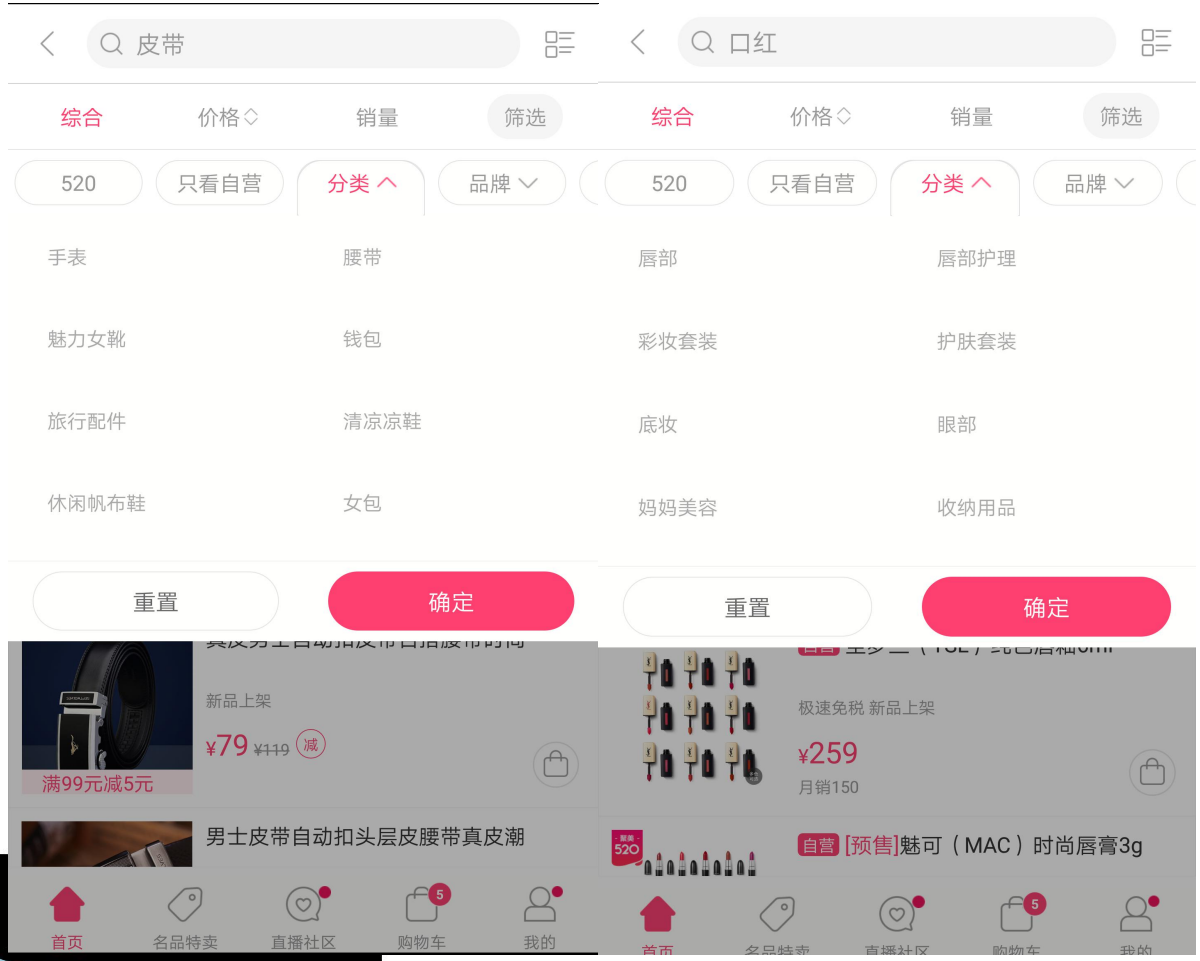
基于支持度与置信度的搜索词商品类别匹配

赵海臣

背景

- 在搜索词过程中，我们会对搜索词的商品类别进行分类匹配，以提升用户对目标商品的搜索体验，并且对搜索结果进行校正。
- 对搜索词的商品类别匹配目前主要有两种方式：
  - 基于经验的规则判断
    - 类似于用户画像，对商品/类别建立相应的画像信息，利用搜索技术进行NLP匹配；
  - 基于用户反馈数据判断
    - 类似于协同过滤，通过利用用户的反馈数据，间接形成人脑的经验分类器。

搜索词商品类别匹配


皮带
综合
价格
销量
筛选
520
只看自营
分类
品牌
手表
腰带
魅力女靴
钱包
旅行配件
清凉凉鞋
休闲帆布鞋
女包
重置
确定
新品上架
¥79 ¥119 减
满99元减5元
男士皮带自动扣头层皮腰带真皮潮
首页
名品特卖
直播社区
购物车
我的
口红
综合
价格
销量
筛选
520
只看自营
分类
品牌
唇部
唇部护理
彩妆套装
护肤套装
底妆
眼部
妈妈美容
收纳用品
重置
确定
极速免税 新品上架
¥259
月销150
自营 [预售]魅可（MAC）时尚唇膏3g
首页
名品特卖
直播社区
购物车
我的

用户对搜索商品的期待与反应

- 用户获取信息的方式：
  - 搜索引擎：目标明确，主动
  - 推荐系统：目标模糊，被动
- 用户的搜索行为本质上是人脑经验对目标商品的词分类：
  - 目标明确，心中有一个明确的目标商品或商品范围
  - => 通过人脑分类获得经验分类词，并进行搜索
  - => 期待能出现心中的目标商品或商品范围内商品
  - => 三种搜索结果的行为
    - a. 看到目标商品或商品范围内商品 => 点击进入 (成功)
    - b. 没有看到目标商品或范围内商品，但感兴趣 => 点击进入(失败)
    - c. 没有看到目标商品或范围内商品，且不感兴趣 =>不点击(失败)

三种搜索结果行为分析

- 用户看到了目标商品或商品范围内商品 => 点击进入 (成功):
  - 这种结果是搜索的目标，通过用户的经验分类词，成功预测到用户的心理需求
    - 假设1 ：用户所输入的搜索词是对商品的准确词分类；
    - 假设2 ：用户搜索完点击的商品是用户认为的符合搜索词类别的商品。

- 数据表现：
  - 点击/曝光ctr高 => 置信度高
  - 点击量大 => 支持度高

三种搜索结果行为分析

- 没有看到目标商品或范围内商品，但感兴趣 => 点击进入(失败):
  - 这种结果是属于介于成功与失败的搜索结果，能够给予用户一定的惊喜，但不符合大部分其它用户的兴趣。
- 没有看到目标商品或范围内商品，且不感兴趣 => 不点击(失败):
  - 明显失败的搜索结果，大部分用户都不感兴趣。

- 数据表现：
  - 点击/曝光ctr高 => 置信度低

数据的容错

- 在数据的采集与清理过程中，不可避免产生错误，会对计入错误来源信息。
- 假设数据来源的错误计数是小概率事件，那么通过支持度(点击次数)能够对错误信息进行有效过滤。

筛选策略

- 分层筛选策略：
  - 细颗粒度更容易区分，因此对商品层次(细颗粒)进行筛选；
    - 商品层次(细颗粒)策略
      - 支持度、置信度、综合信心值
  - 粗颗粒度根据量化指标进行筛选，对商品类别(粗颗粒)实行量化辨识。
    - 类别层次(粗颗粒)策略
      - 对商品层次(细颗粒)设定较松弛的阈值，再对过滤后的商品进行统计学总结，给出连续量化的“类别信心值”

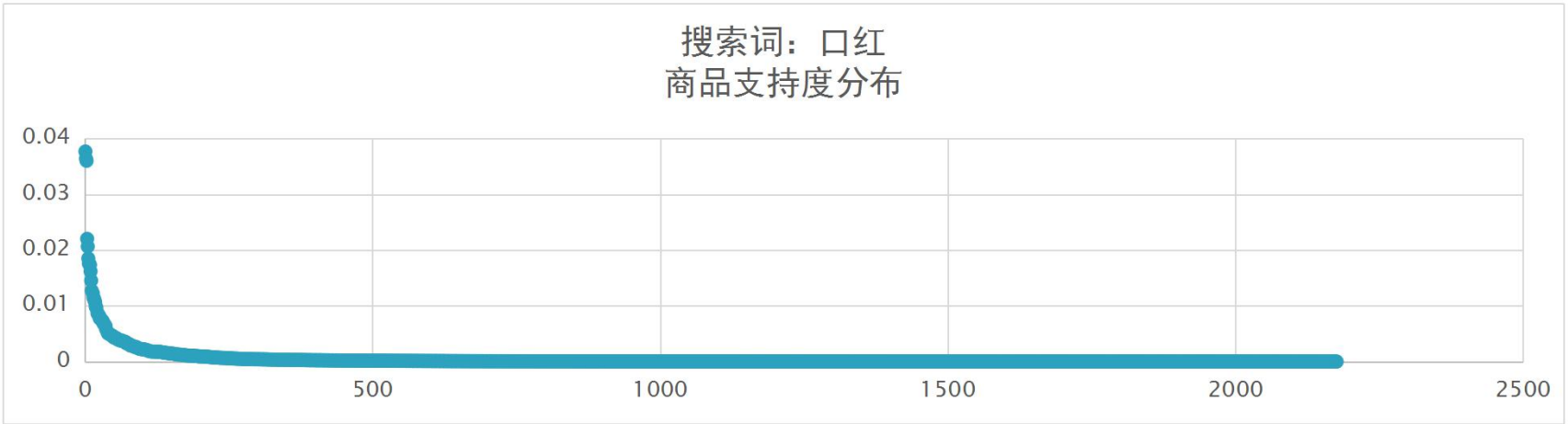
支持度定义

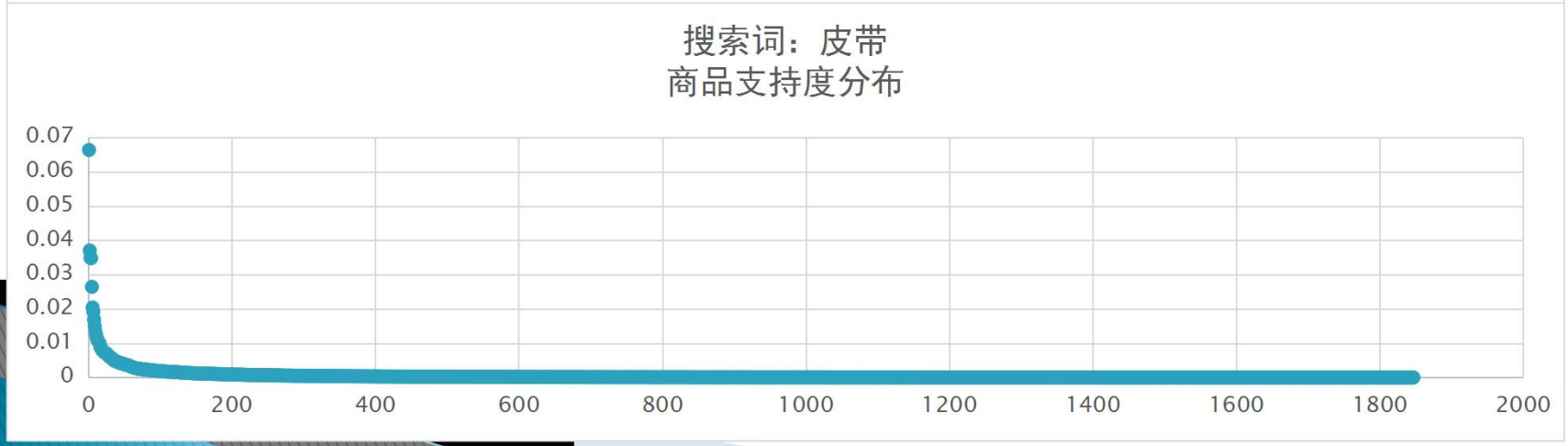
- 支持度是对用户兴趣可靠性的一个指标，使用“点击量”作为支持度的度量。
  - 同一搜索词下，某个商品的点击概率符合正态分布。

- bug: 但如果一个搜索词是热搜词，那么它对应的所有商品都会有一个较高的点击量。
  - 改进支持度定义：
    - 支持度 = 商品点击量/所有点击量。
  - 支持度的分布：该搜索词下，商品支持度符合长尾分布，值越大，该商品符合搜索词的概率越大。

支持度例子——长尾分布

搜索词对应的商品支持度符合长尾，说明用户在搜索某个词时呈现出很强的目标性。



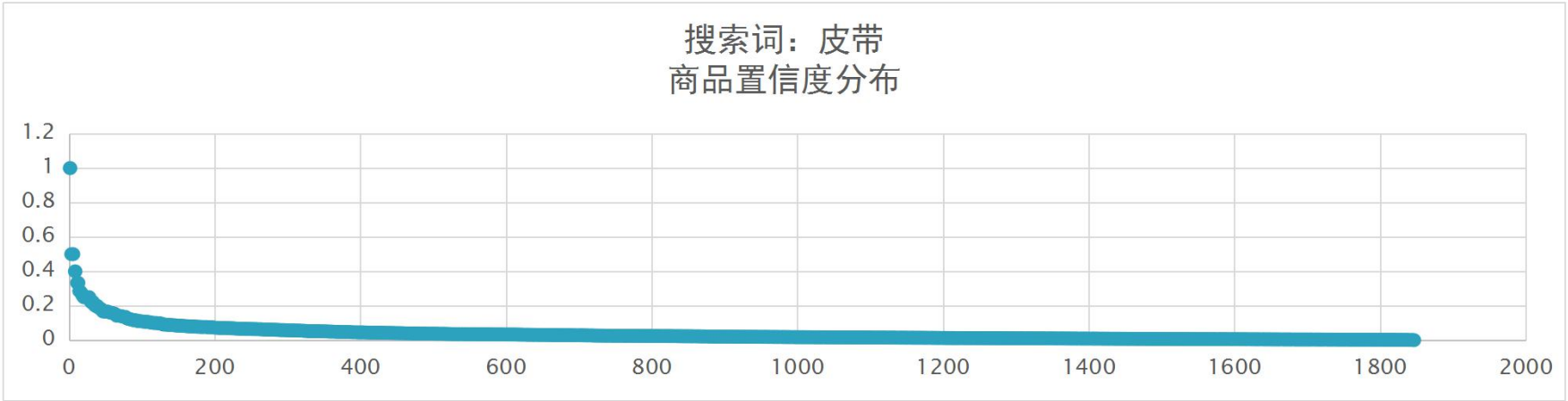


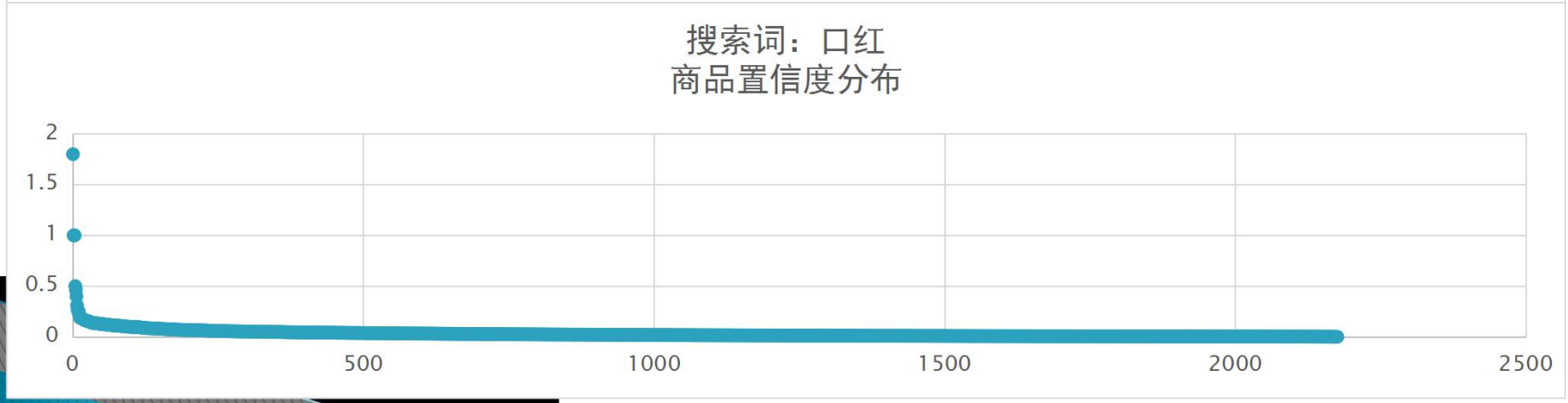
置信度定义

- 但支持度高不能完全判定用户的意图，也许只是搜索结果的异常导致某个错误的商品大量出现在用户眼前。
- 为了解决这个问题，引入置信度，对用户兴趣进行进一步的分析：
  - 置信度ctr = 点击次数/曝光次数
- 置信度表明用户在看到商品后，确认它与搜索词相关程度。

置信度例子——长尾分布

搜索词对应的商品置信度符合长尾，说明用户对某个搜索词下的商品兴趣分化很大。





综合值定义

- 一个商品是否符合搜索词，需要通过“支持度”+“置信度”联合界定，可以分别设置两个标准，但也可以尝试使用综合支持度与置信度：
  - 综合值 = 支持度 * 置信度

- 通过对搜索词的综合值进行排序，可以获得词的综合排序，通过对综合值设定某个规则，可以有效排除错误分类。

搜索词列表到商品类别的映射

- 对综合值设定阈值后，可以使用阈值以上对应商品的分类比例作为搜索词分类的信心值。
  - 信心值 = 阈值以上类别商品数/阈值以上商品总数
- 例如设置综合值阈值为0.00001，搜索词“皮带”的分类值为：

分类	信心值	分类	信心值
腰带	0.863	清凉凉鞋	0.0026
手表	0.0336	其它	0.0026
旅行配件	0.0258	魅力女靴	0.0026
皮具礼盒	0.0103	收纳用品	0.0026
其它饰品装饰	0.0076	旅行包套装	0.0026
时尚单鞋	0.0052	手链/手镯	0.0026
戒指	0.0052	家居饰品	0.0026

分值的意义

- 搜索词的分类结果，不能简单界定为“是/否”，但是它的正确性是可以通过连续变量来衡量的：
  - 皮带对应的类别，腰带0.863的信心度显然要比手表0.0336高，所以皮带是属于腰带类别，也符合人的经验；
  - 而皮带手表的存在显然让手表获得了比旅行配件0.0258(有一些皮带错误分到旅行配件)，皮具礼盒(皮带盒子)相对要高一些的信心值。

搜索词列表到商品类别的映射

- 例如设置综合值阈值为0.00001，搜索词“口红”的分类值为：

分类	信心值	分类	信心值
唇部	0.903	彩妆工具	0.0028
唇部护理	0.0417	宝宝护肤	0.0028
彩妆套装	0.0278	妈妈美容	0.0028
收纳用品	0.0083	其它	0.0028
护肤套装	0.0056		

容错性——对新增分类的处理

- 黑名单方式：为了防止新加入的商品以及分类被错误过滤，因此对正向商品分类进行取反操作，只记录反向过滤黑名单。
- 白名单方式：设置default值，查不到的分类自动取default作为白名单。

The end