

A wealth of data 数据财富

2020年3月5日 22:25

目标

确定可能的设计功能 → 评论筛选关键词

维度?

好: 宣传点

坏: 改进点

垃圾用户

时间趋势 → 关键词 freq 趋势 → 加权

问题要求

a) 评价与星级关联:

评价: score
星级: star / 加权各 50%? → point

同一商品根据 point 加权得到 mark

权重增加: 高于平均长度、绿标、高认可度
(但数值也要高)

权重降低: 太短、未购买、认可度低、

权重惩罚: (ban 用户 ID) 星级与评价相反,
如果评论很长, keyword 很少, 关键词众多但与众不同, 认可过低
那就 ban

b) 数据可靠性: 可能要加大低权重惩罚? 可能对于严重的要 ban

两方面: 恶意/善意比 和平均 mark?

拐点分析: 是否应加入低权重(占比)的曲线?

数量

c) 找前者拐点特征

拐点处: 绿标用户?

低权重用户?

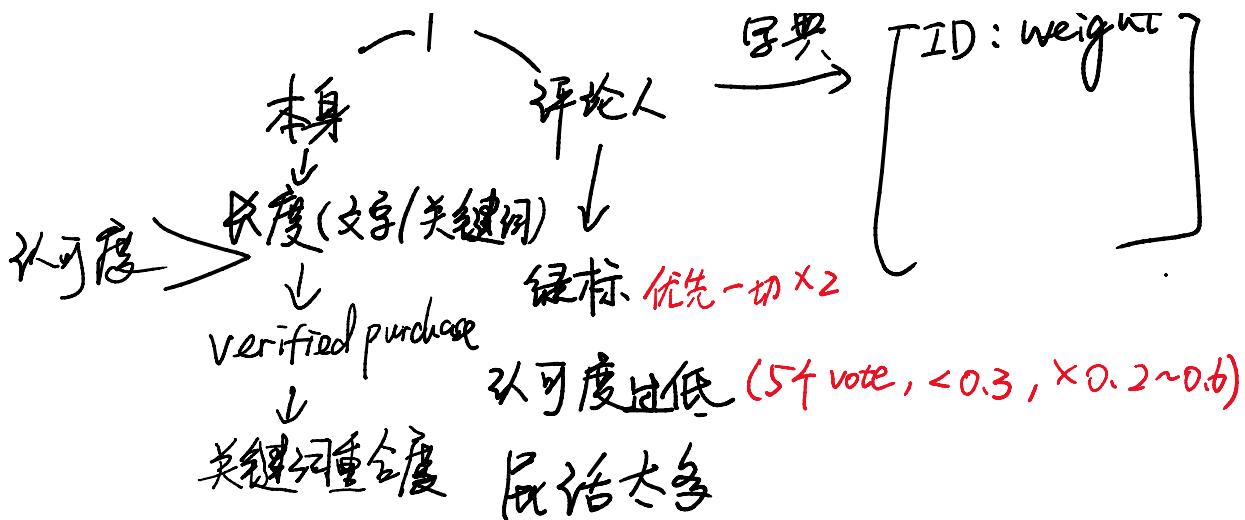
滚雪球?

d) 滚雪球

e) 情感分析(可提前写)

评价体系





认可度表 (5vote 以上)

< 0.3	$0.3 \sim 0.7$	$0.7 \sim 0.9$	> 0.9
$\times 0.2 \sim 0.7$, 用户级	$0.7 \sim 1$, 评论级	$1 \sim 1.2$	$1.2 \sim 1.5$

评论级 用户级

