用户行为去噪

赵海臣

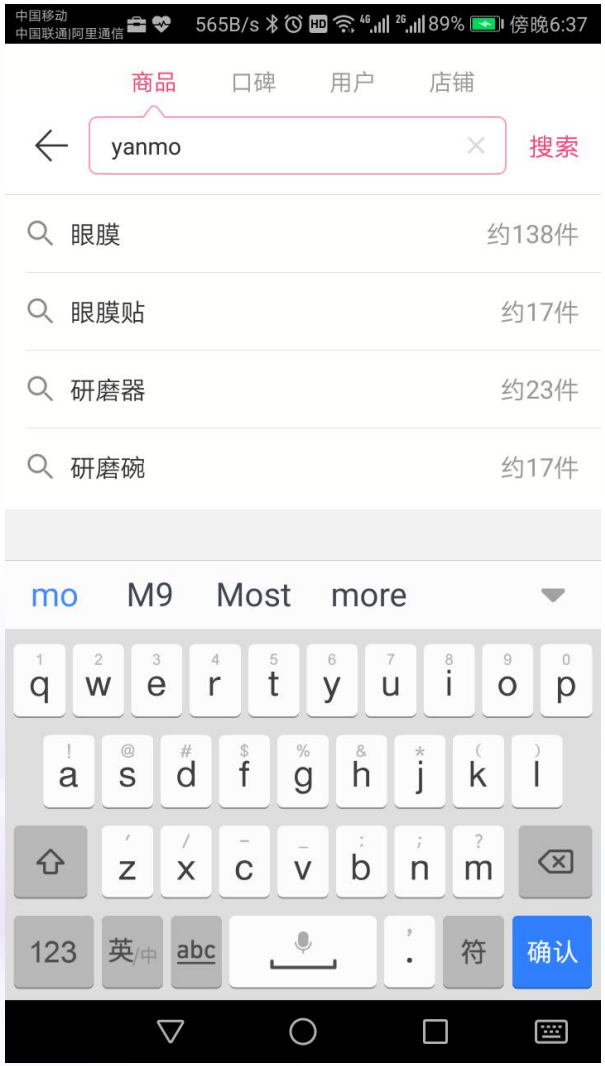
背景

对于搜索来说，假设数据采集没有问题，用户的搜索关键词的一些错误输入、错误拼写等缘故会干扰我们对用户搜索数据的统计。

- 我们做搜索关键词的推荐时，使用boolean cosine distance计算关键词相似度的时候，发现一个奇怪的现象：
  - 研磨器 - 眼膜贴 的相似度很高
  - 眼膜贴 - 研磨器 的相似度很高

  这是完全通过用户搜索反馈得出的相似度，即用户在同一天内同时搜索这两个词的概率很高，可是这两个词并没有任何的关系。
- 研磨器(Yan Mo Qi)和眼膜贴(Yan Mo Tie)的拼音关系很近，所以我们认为是用户输入拼音的联想导致了这两个词汇会经常被同时搜索
  - 用户想搜眼膜贴(YanMoTie)，可是输入了YanMo后直接点击联想词，结果联想词是”研磨器”(YanMoQi)

聚美APP上的拼音联想



聚美优品
JUMEI.COM

用户行为去噪

- 用户行为去噪的一个基本方针是：增加用户的成本，用户更慎重、成本更高的操作的噪声远比随意的、成本低的行为所带的噪声低。
  - 所以我们换用”搜索完并且有点击“行为的搜索词重新进行距离计算，料想用户如果是错误输入，则并不会点击对应的商品，而是会修改输入，对正确的搜索结果进行点击。

结果

- 将用户所有搜索关键词换成用户”搜索并点击“的搜索关键词之后，研磨器和眼膜贴的关联完全被消除。
  - 具体的结果请参考文件夹下的Excel表格文档。

THE END

THANK YOU!