Item-based CF的几种优化方式

机器学习与推荐算法 2020-06-19

作者 | 小雨姑娘

来源 | 知乎

链接 | https://zhuanlan.zhihu.com/p/148951213

编辑 | 机器学习与推荐算法

题主之前参加了KDDCUP2020，CIKM2019等几个经典的推荐系统比赛，发现大部分优胜方案都采用了传统的Item-CF召回方式，在通过一些trick进行微调后同样可以取得甚至超过Embedding+Faiss，Self-Attentive Sequnetial Model。这里总结了可以提高Item-CF召回效果的几种方案，供大家参考。

1. 基于时间维度的优化

1.1 时间权重优化

由于用户的兴趣变化较快，在召回时更改时间权重可以使推荐系统更加关注于用户的当前兴趣，其中在KDD2020比赛中前排大部分采用了以下策略[1]：

$$Sim(i,j) = |U_i \cap U_j| * \sum_{u \in U_i \cap U_j} (t_{u,i} - t_{u,j})^k$$

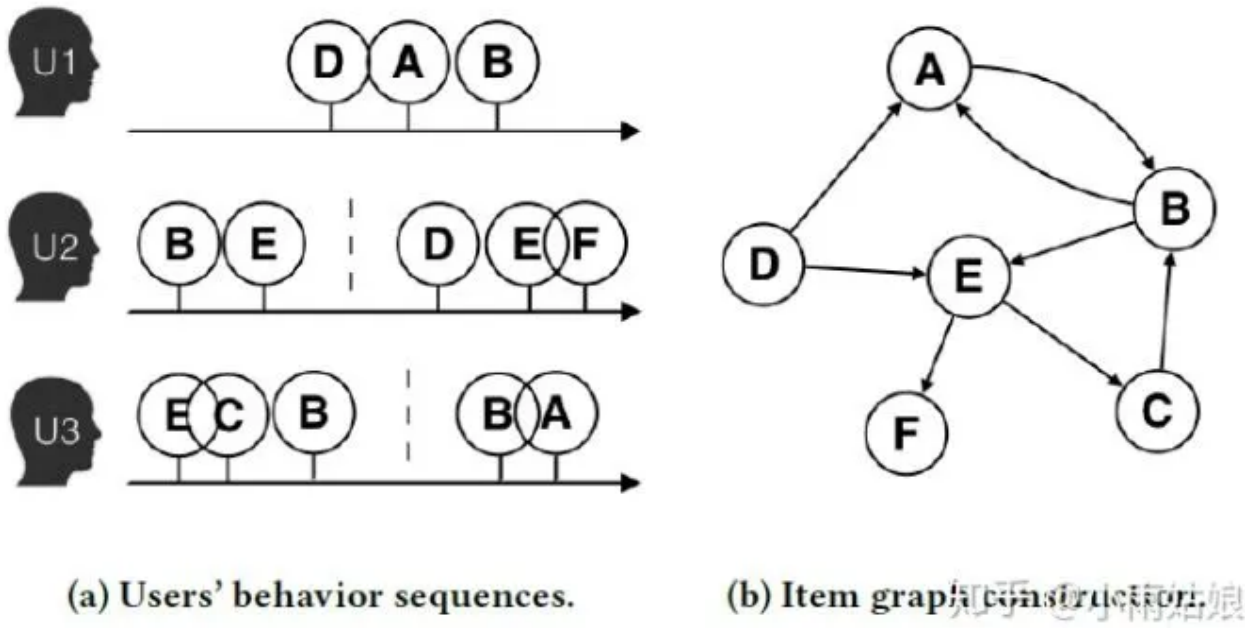
即在计算两个商品相似度时，增加一个权重表示用户选择商品 i 与选择商品 j 的时间差。这个策略的核心思想是如果用户选择两个商品时的时间间隔越近，则两个商品的相关性更强。例如用户下单手机后又马上下单了手机壳，则手机壳和手机之间的权重更大，手机跟昨晚用户下单的马桶刷之间权重更低。

1.2 用户session优化

另一种更加直观的方法是根据时间把用户的评分分成几个session，每个session中计算相似度：

$$Sim(i,j) = \sum_{t=0}^{T} |U_i^t \cap U_j^t|$$

即根据时间把用户的评分分为T个阶段（通常是30分钟或一两个小时），分阶段计算相似度最终求和。在阿里巴巴18年KDD论文[2]中有使用该方法，如下图(a)所示



(a) Users' behavior sequences. (b) Item graph construction.

其次在KDD2020比赛中也听说前排队伍使用该方法极大地提高了召回效果。该方法的核心思想是认为用户在时间段内的行为具有一定的关联，举个例子你的女朋友买口红一般会短时间内只看看很多口红，买粉底会短时间内只看粉底。

2. 基于序列次序的优化

2.1 序列次序权重优化

基于序列次序权重的方法与时间权重优化的思路类似，策略是为用户行为序列中临近的商品设置更高的权重：

$$Sim(i,j) = |U_i \cap U_j| * \sum_{u \in U_i \cap U_j} (loc_{u,i} - loc_{u,j})^k$$

其中 $loc_{u,i}$ 是用户 u 对商品 i 在行为序列中的位次。核心思想是用户序列中更近的商品的关联性更高。[1]

2.2 基于序列的单向相似性优化

这种方法假定用户的行为具有一定的序列意义，例如用户在购买手机后可能会追加购买相应型号的手机壳（正次序），但用户购买手机壳后却往往不会购买手机（反次序）。其中一种方法是为

反次序增加一个衰减因子：

$$Sim(i,j)=\begin{cases}|U_i\cap U_j| & \text{if } l_i<l_j\\ |U_i\cap U_j|*\lambda & \text{if } l_i>l_j\end{cases}$$

在最极端的情况下使得lambda取0，则代表仅考虑正次序相似度。

3. 基于用户/商品热度的优化

比较经典的一种基于用户商品热度的优化方法来源于TF-IDF算法[3]：

$$Sim^{w}(i,j)=\frac{\sum_{u\in U}w_u\delta(i,j)}{\sum_{u\in U_i}w_u}$$

其中，是全体用户集合，Ui是对商品i感兴趣的用户集合；Wu代表用户u对相似性的贡献度，用户u的行为数量越多，则它的贡献度越低。

$$w_u=\frac{1}{\log(I_u)+1},\quad \delta(i,j)=\begin{cases}1, i\in I_u \text{ and } j\in I_u\\ 0, \quad else\end{cases},\quad I_u$$

策略的核心思想是兴趣广泛的用户的行为更难体现偏好，他的行为纪录多种多样，随机选出两个商品具有较高关联性的概率较低。

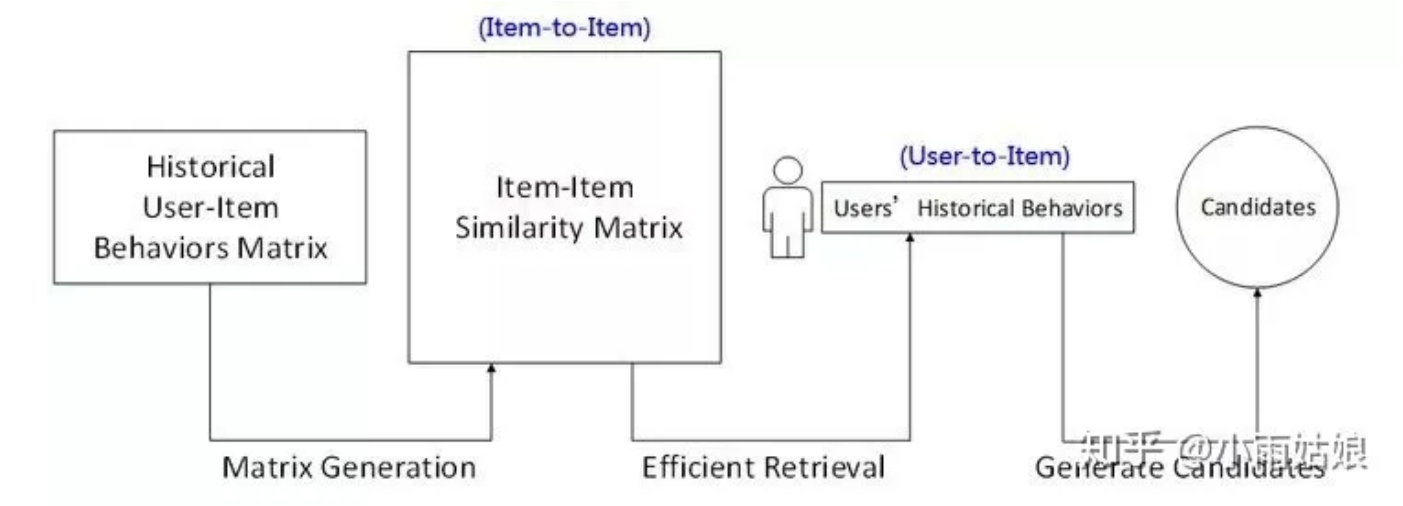
另一种基于用户商品热度优化的方法是，降低那些流行商品的相似性[4]：

$$Sim(i,j)=Confidence(i,j)=P(j|i)=\frac{|U_i\cap U_j|}{|U_i|}$$

此方法来源于关联规则分析中的置信度公式。核心思想是流行的商品被买的多，也就更有机会与其他商品计算相似度，因此要降低他们的相似度。

4. 对于Match的优化

以上方法都是作用在相似度计算环节，也就是召回的Item-to-Item环节，而对于Match的优化聚焦于召回的User-to-Item环节。



与上面的次序权重和时间权重类似，当通过用户历史行为进行matching时，应该让用户临近的行为具有更高的权重：

$$Score(\hat{i}) = \frac{max(t_u) - t_{u,i}}{max(t_u) - min(t_u)} * Sim(i, \hat{i})$$

其中 $t_{u,i}$ 代表用户u历史行为i出现的时间。

参考

1. KDD2020-baseline https://tianchi.aliyun.com/forum/postDetail?spm=5176.12586969.1002.12.
2. SDM https://dl.acm.org/doi/abs/10.1145/3219819.3219869
3. Using TF-IDF to Determine Word Relevance in Document Queries https://0bc297c6-a-62cb3a sites.googlegroups.com/site/caonmsu/ir/UsingTFIDFtoDetermineWordRelevanceinDocument(attachauth=ANoY7cojb_apOWGZxGpXGDoFo94TrILMf13c6yhvUO-SXAbzLg8t5-zdKgC3_gbnc39Fantv9sJDGGxu2ifKgmCaUu9phgd1EWYqeGipMH9HL6j49Oh_vEPUFypeMkEyH CkSa6sLcOm_JtYmfmC2Cx5MFCEWxv1uMFTXyUkwAprhXM6feYKqf-V-zjYFVchMpImcTG8sGdenKgYGQm5NBUiMX5YzCunLtTILQAZKjNMJz3OGmGJHeZu_4CWTq_LrO
4. CIKM2019冠军方案 http://zhuanlan.zhihu.com/p/91506866

推荐阅读

[0].推荐系统之FM与MF傻傻分不清楚
[1].WSDM2020推荐系统论文打包下载
[2].利用对抗技术来权衡推荐精度与用户隐私
[3].Context/Sequential/Session RS的区别