协同过滤中的时间因素

赵海臣

传统时间衰减

- 传统时间衰减方式：
  - 时间窗口 time-window
  - 时间衰减 instance-decay
- 缺点：
  - 在一个有很多用户与商品的数据生态系统中，有很多时间相关的特征在同时产生作用，如果使用窗口或者衰减，将丢失很多信息。
- 目标：
  - 时间模型应该捕捉concept drift：
    - 舍弃临时的行为(噪声)
    - 捕捉用户长期的行为趋势

Concept Drift

- 推荐系统时间因素理论依据：Concept Drift
  - Concept drift 是个有趣的概念，User Modeling是个非常动态的过程，因为user的attributes会随着时间改变，model形状或比重，也需随之更新。例如： 当MLB季后赛时，user对棒球的新闻较有兴趣，当世界杯足球赛开打时，user又开始对足球的新闻有兴趣。这是个不难理解的过程，但对于会输入过去大 量的training data，来预测未来的machine learning方法来说，「时间」是个不好解决的敌人。
  - 以 数据量的角度来说，最新进的数据量一定小于过往累积的数据量，因此最新喜好的比重小，所以无法呈现最当下的user偏好；以质的角度来说， user的attributes未必乖乖地依照原先的设计分类或计分，有可能突然多出或修改了某个attribute的值，例如：user看多了运动类的 新闻，user有兴趣的新闻可能变成球员签约金的消息，它变成对篮球、撞球等球员的身价有兴趣，而原本设计的attribute却已无法描述。以上这些问 题，是输入再多的training data或error rate多低，都没法解决的。
  - 但透过比重的调整，如Webb & Kuzmycz将旧数据重要性调低；或用adjustable time window的方式来对新旧数据作动态区分，如根据影片的热门时间，调整间点，好片可能是一个月，烂片可能是一周

用户行为时间变化

- 用户行为时间变化可以分解成多种不同的concept drift，这些concept drift又分别有着不同的时间周期以及不同的变化方向。

- 我们需要捕捉到用户的这些不同的concept drift，并且modeling它们的各自变化趋势，以求获得更加精确的用户行为时间变化模型。

捕捉用户DRIFTING

- Time-window: 直接截取最近的某一段时间
  - 缺点：过于生硬，将失去用户的长期趋势信息。
- Time-decay: 使用渐变时间衰减函数对久远行为进行降权
  - 缺点：*根据论文介绍，当去除时间衰减函数的作用后，他们取得了最佳效果。虽然用户的行为确实在改变，但是很多久远兴趣仍然存在，并且影响着用户间、物品间的交互。
- Ensemble-learning: 使用多个预测器predictors一起进行预测。
  - 效果最佳

时间因素捕捉指导方针

- 尝试分析捕捉全时段的趋势，而非仅仅目前的短时行为(捕捉更稳定的行为趋势，规避短期噪声行为)。
- 捕捉多种concept drift：
  - User-dependent
  - Item-dependent
  - gradual drift
  - Sudden drift
- 对不同concept drifts进行建模后，需要将所有这些模型统一到一个框架中，从而获得更高层次的模型
- 不尝试对未来进行插值预测，虽然看起来很有用，但会很困难。

以ALS为例进行阐述

- 用户对某部电影的打分预测可以表示为：

$$\hat{r}_{ui} = \mu + b_i + b_u + q_i^T \left(p_u + |\mathrm{R}(u)|^{-\frac{1}{2}} \sum_{j \in \mathrm{R}(u)} y_j \right)$$

- Mu是全局平均值，bu是用户偏离度，bi是物品偏离度
  - 例如所有电影的平均打分是Mu=3.7
  - 泰坦尼克平均分是4.2，则bi=+(4.2-3.7)=+5
  - 用户A打出的平均分是3.3，则bu=+(3.3-3.7)=-4
- qi是物品特征矩阵，pu是用户特征矩阵，R(u)是用户打分矩阵

时间影响因素

- 物品的流行度会随时间变化： bi -> bi(t)
- 用户的行为倾向会随时间变化： bu -> bu(t)
- 用户的兴趣特征会随时间变化：pu -> pu(t)
- *物品的特征基本上不会随时间变化：qi

- 加上时间因素后，公式变为：

$$\hat{r}_{ui}(t) = \mu + b_i(t) + b_u(t) + q_i^T \left(p_u(t) + |\mathrm{R}(u)|^{-\frac{1}{2}} \sum_{j \in \mathrm{R}(u)} y_j \right)$$

The end