大数据技术创新大赛——个性化新闻推荐解决方案

本次大赛我们采用一种结合信息老化及核密度估计的协同过滤模型，最终取得了一定的成绩。该模型融合了时效性特征、兴趣特征，在这些特征的基础上结合基于用户以及基于项目的协同过滤方法，能够有效的在稀疏数据的情况下提高推荐效果，同时解决了协同过滤的冷启动问题以及高复杂度问题。

该模型的关键步骤如下：首先是用户兴趣建模，即用一定的数学模型来表示用户在整个项目空间的兴趣偏好；然后，利用该兴趣模型计算用户间的兴趣相似度，产生基于该相似性度量的邻居集；最后将目标用户的邻居所感兴趣的项目通过一定的推荐策略返回给用户.

但是这种模型的复杂度非常高，初期我们尝试过之后，就对其进行改进。**首先是根据基于用户的相似度来计算用户 u 的 TopN 最近邻的用户，然后基于 TopN 用户计算总的新闻的并集 U,在新闻并集的基础上，采用核密度估计的方法，对每一篇新闻进行兴趣度计算，最后在兴趣度的基础上结合新闻时效性（流行度）的值，综合加权，最终得到对于每一个项目的评分值，根据评分值以及该用户 u 的新闻阅读能力来推荐 TopK 用户。**

当然上面的模型只是一个大致的过程，下面将就每一部分进行详细的说明：

第一部分：选择用户 u 的 N 近邻用户，

在选择用户u的N(N我们取的是120)近邻用户时，用到了一些相似度计算的模型，先求用户u与某一用户v的新闻浏览的交集，Inter=news(u)∩news(v),根据交集，计算每一个交集里面的新闻的相似度加权值：$\text{Simi} = \sum_{item \in Inter} \frac{1}{alpha*(|t1-t2|)}$，这样就选择出用户u的TopN最近邻用户。

第二部分：根据用户u的N近邻用户，计算所有用户的新闻并集U,对U中所有的新闻进行兴趣分布(用户u的兴趣偏好)计算：设$X_1 \ldots X_1$为总体分布X的独立同分布，X的密度函数定义如下：

$$f(\mathrm{X}) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{\| X - X_i \|}{h})$$

其中 K 的定义如下：

$$K_g(\mathrm{z}) = \frac{1}{\sqrt{2\pi}h}\sum_{i \in I_u} r_{u,i} \times \exp\{\frac{z^2}{2h^2}\}$$

代入 f(X)中得如下表达式：

$$f_{P_u}(\mathrm{j}) = \frac{1}{|\mathrm{I}_u|\sqrt{2\pi}h}\sum_{i \in I_u} r_{u,i} \times \exp\{\frac{d_{i,j}^2}{2h^2}\}$$

其中

$$d_{i,j} = 1 - sim_c(\mathrm{i,j})$$

这里的 sim 代表新闻与新闻之间的相似度结果，这里在计算的时候采用的是用户 u 的已点击新闻 i 与待评估的新闻 j 同时被某一用户 v 阅读时，采用加权时间差相似度方法 $\sum_{i,j\in I_v} sim = \frac{1}{a(\mathrm{t_{v,i}} - \mathrm{t_{v,j}})}$，计算这两条新闻的相似度以及得出最后的距离，h 代表核窗宽，I_u 代表的是用户 u 的新闻点击的集合，最后求出 v 的兴趣分布值。

第三部分：求出每一条待评估新闻的时效性。

在求出用户对新闻的兴趣关注度之后，还需要考虑的是新闻的时效性问题，这里在在求解时效性时，采用了信息老化模型——负指数模型，$C(\mathrm{t_{t,}}\,\mathrm{t}_f) = \mathrm{e}^{-a(\mathrm{t}_t - \mathrm{t}_f)}$，其中 t_f 表示信息发布的时间，t_t 表示当前时间，$C(\mathrm{t}_t, \mathrm{t}_f)$ 表示信息在 t_t 时刻的影响力大小，a 代表的是信息的老化率系数，其中老化系数 $a = -In(0.5)/\mathrm{T}_h$,所以求得半衰期，就可以求得老化系数 a,根据半衰期的定义，半衰期等于信息最中间点击时刻减去信息的发布时刻，而本文有单个信息点击量的集合 s，这很容易在其中找出中间点击的时刻 t_0,信息的发布的时刻 t_f,所以信息的半衰期为 $T_h = t_0 - t_f$.

第四部分：加权融合

这部分主要是将第二部分得到的某一用户对候选新闻的兴趣分布以及这些新闻的时效性进行加权融合，最后得到的对候选新闻的评分，然后结合用户的阅读能力（一共阅读了 V 条新闻），进行 TopK 推荐，这里采用的融合权值是 $p(\mathrm{x}) = 0.8\mathrm{f(x)} + 0.2\mathrm{C(t,t}_j)$,推荐条目和阅读能力成正比，V/10 的上取整。

展望：如果可以对所有的新闻进行兴趣偏好计算的话，最终根据偏好，进行相似用户选择与评分，结果估计能更上一层，但是复杂度过高，所以下一步可以考虑如何在保证精度的同时保证运行时间复杂度相对较小。

主要参考文献：

王玉斌、孟祥武、胡勋 ,一种信息老化的协同过滤推荐算法;电子与信息学报，2013（10）。

王鹏、王晶晶、俞能海,基于核方法的 user-based 协同过滤推荐算法;计算机研究与发展，2013，50（7）。