Estimating Position Bias without Intrusive Interventions 阅读笔记

2019 年 11 月 1 日

摘要

将点击的概率拆分为在位置 k 上的曝光概率和 <q,d> 相关性的乘积。目标是从点击日志中统计出在位置 k 的曝光概率作为 position bias。本文的亮点在于使用了不同 ranker 的日志，构造 Interventional Sets，将相同 <q,d> 在不同 ranker 下不同位置的曝光记录作为有效的统计数据。

1 统计方式

1.1 Position-Based Propensity Model (PBM)

点击事件可以拆分为两部分：

$$\begin{aligned}\Pr(\mathrm{C}=1|q,d,k) &= \Pr(E=1|k)\,\mathrm{rel}(q,d)\\ &= p_k\,\mathrm{rel}(q,d)\end{aligned}$$

这里的 $\Pr(E=1|k)$，即位置 k 上曝光的概率 (只和位置相关)，就是统计的目标。一种常用的统计方式是，将一部分排序在第 1 位的文档随机放置到第 k 个位置，统计第 1 位和第 k 位的点击率 [注 (个人理解)：这种交换的方式有一个亮点在于交换到第 k 位的 <q,d> 和在第一位的 <q,d> 的相关性的期望是相等的，所以位置点击率的比例能等同于位置曝光的比例，具体推导见原论文]：

$$\hat{c}_1^{1,k} = \frac{1}{n_1}\sum C_1^i \tag{1}$$

$$\hat{c}_k^{1,k} = \frac{1}{n_2}\sum C_k^j \tag{2}$$

用这两者的比值作为位置 k 的 position bias 的一种无偏估计:

$$\frac{\hat{p}_k}{\hat{p}_1} = \frac{\hat{c}_k^{1,k}}{\hat{c}_1^{1,k}} \tag{3}$$

然而，通过对排序结果随机交换并根据点击日志的统计点击率比例的方式会影响用户体验，因此只是一种理想的处理方式，本文通过不同 ranker 的历史点击日志来避开这种情况。

1.2 Interventional Sets from Multiple Rankers

在排序 ranker 经常更新的情况下，可以累积到不同 ranker 的历史点击日志，假定不同的 ranker 并不影响用户在搜索时时候的 query，在这种情况下，来构建 Interventional Sets:

$$\begin{aligned} S_{k,k'} := \{(q,d) : & q \in Q, d \in \Omega(q) \\ & \exists f, f' \, \mathrm{rk}(d|f(q)) = k \wedge \mathrm{rk}\,(d|f'(q)) = k'\} \end{aligned} \tag{4}$$

统计方式为在同一 <q,d> 在不同 ranker 下被排在不同的位置，此外论文中只选取了每个 query 下的前 10 的 doc。这样构造的集合中，可以得到同一个 query 在不同位置的点击率 [注 (个人理解): 这里得到点击率之后不能用位置点击率的比例代替位置曝光的比例]。此后将点击的期望拆分成曝光和相关性的乘积。

1.3 最大似然估计求解

统计出位置点击率之后，将位置曝光率和关联性作为位置变量，用最大似然估计求解:

$$(\hat{p}, \hat{r}) = \underset{p,r}{\mathrm{argmax}} \sum_{k \neq k' \in [M]} \hat{c}_k^{k,k'} \log\left(\hat{p}_k \hat{r}_{k,k'}\right) + \hat{\neg} c_k^{k,k'} \log\left(1 - \hat{p}_k \hat{r}_{k,k'}\right) \tag{5}$$

计算出的 $\hat{p}_k$ 就是目标位置曝光率

2 补充内容

可能的用法: 利用 examination probability 对排序完的排序做一次 rerank