在 Airbnb 中实现实时个性化搜索 Embedding 技术的应用

Liam Huang*

2018 年 11 月 28 日

*liamhuang0205@gmail.com

万物皆可萌 Embedding

NLP 中的困难：用何种编码方式描述词之间的语义关系.

- one-hot;
- n-gram.

Word to Vector $\rightarrow$ Word Embedding.

推荐场景：将商品、新闻、视频等当做需要 Embedding 的实体.

特点：

- 高维实体映射到低维 embedding;
- embedding 表意空间内，相似实体有较近距离.

当我们在讨论 Embedding 的时候，我们在讨论什么?

- 希望 Embedding 表达什么?
- 如何让 Embedding 学到东西?
- 如何评估 Embedding 向量效果?
- 线上如何使用?

希望 Embedding 表达什么?

NLP 里的 Word Embedding：表意空间 $\rightarrow$ 不言自明 $\rightarrow$ 语义空间.

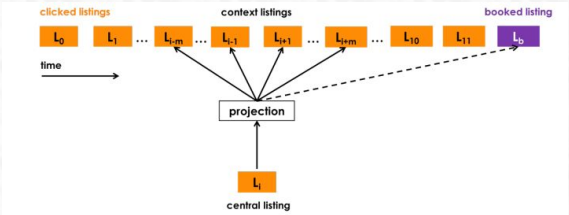
对于同样的实体，训练语料不同则表意空间不同．商品推荐场景：

- 用户的浏览兴趣空间 $\leftarrow$ 商品点击数据日志
- 用户的购买兴趣空间 $\leftarrow$ 商品购买数据日志

魔改 w2v，不如精细化构建语料：

- item embedding 没有 NLP-like 的明确上下文
- session 切分
  - Airbnb：30 分钟 time gap $\Leftrightarrow$ 连续两次点击
  - 点击列表：`dress 1`, `dress 2`, `phone 1`, `phone 2`
    - 针对 Query 下的兴趣：切分
    - 针对用户兴趣转移：不切分
- 短期 v.s. 长期 $\rightarrow$ 类比到 recsys 的用户画像
  - 短期兴趣 embedding $\leftarrow$ 点击日志
  - 长期兴趣 embedding $\leftarrow$ 预定日志

如何让 Embedding 学到东西?



- $s = (L_1, L_2, \ldots, L_n) \in S$ 表示一个 session $\rightarrow v(L_i) \in \mathbb{R}^{32}$
- Skip-gram model
  中心 item $\rightarrow$ 滑动窗口 $\rightarrow$ 预测窗口内其他 item 的点击概率（正例）
- 随机负采样
  中心 item $\rightarrow$ 全局随机负采样 $\rightarrow$ 预测负例的点击概率
- 全局上下文（booked-session 中）
  booked-listing 作为全局正例 $\rightarrow$ 预测 booked-listing 的点击概率
- 聚合搜索（业务强相关）
  中心 item $\rightarrow$ 本地城市随机负采样 $\rightarrow$ 预测负例的点击概率

最终优化目标

$$\arg\max_{\theta} \sum_{(l,c)\in\mathcal{D}_p} \log \frac{1}{1+\mathrm{e}^{-v^\top(c)v(l)}} + \sum_{(l,c)\in\mathcal{D}_n} \log \frac{1}{1+\mathrm{e}^{v^\top(c)v(l)}} + \log \frac{1}{1+\mathrm{e}^{-v^\top(l_b)v(l)}} + \sum_{(l,m_n)\in\mathcal{D}_{m_n}} \log \frac{1}{1+\mathrm{e}^{v^\top(m_n)v(l)}}$$

- l：在更新的中心 listing；
- $v(l)$：listing-l 的 Embedding 向量
- $\mathcal{D}_p$：正例对 (l, c)，$v(l)$ 和 $v(c)$ 在训练中会被推近
- $\mathcal{D}_n$：全局负例对 (l, c)，$v(l)$ 和 $v(c)$ 在训练中会被推离
- l_b：最终被预定的 listing
- D_{m_n}：本地负例对 (l, m_n)，$v(l)$ 和 $v(m_n)$ 在训练中会被推离

新 listing 冷启动：「求 3 取平均」

- 地理位置；
- 房源类别；
- 价格区间.

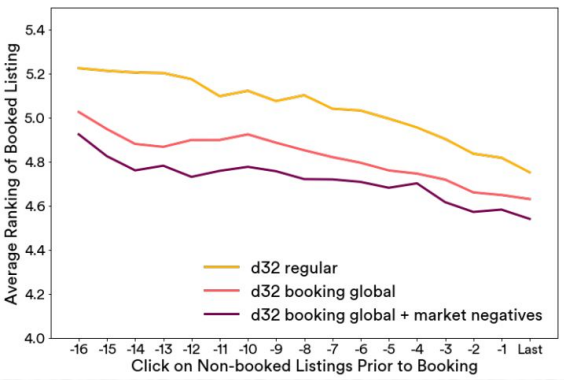
如何评估 Embedding 向量效果?

传统方式:

- 聚类;
- t-SNE;
- ……

Airbnb 创新的与业务结合的方式:

- 用户最近点击的 listing 列表;
- 包括被预定 listing 在内的候选列表;
- 余弦相似性排序;
- 观察被预定 listing 在排序中的位置.



线上如何使用?

相似房源推荐[1]

ABTest:

- Control: 调用 Airbnb Search 进行 rank;
- Treatment: 搜索 Embedding 相似度最高的.

结论:

- 点击率提升 21%;
- 下单率提升 4.9%.

[1]类似联播推荐.

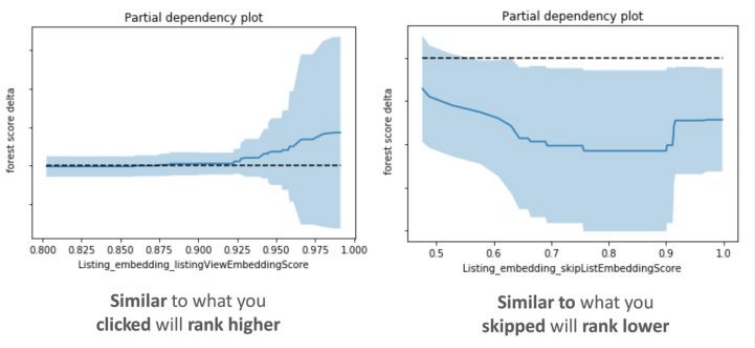
基于 Embedding 的实时个性化搜索

Based on Kafka, user level:

- H_c：近两周用户点击过的房源;
- H_s：近两周用户跳过的 highly ranked 房源.

针对待排序的房源 l_i，计算相似性:

- $\text{EmbClickSim}(l_i, H_c) \stackrel{\text{def}}{=} \max_{m \in M} \cos\left(v(l_i), \sum_{l_h \in m, l_h \in H_c} v(l_h)\right)$;
- $\text{EmbSkipSim}(l_i, H_s) \stackrel{\text{def}}{=} \max_{m \in M} \cos\left(v(l_i), \sum_{l_h \in m, l_h \in H_s} v(l_h)\right)$;



UGA

The opal codon.