Embedding搜索能代替文本搜索吗?

原创 WePlayData WePlayData 2020-04-05

收录于话题
#视频搜索

2个

Embedding在人工智能中被广泛使用，万物皆可embedding，尤其在推荐系统中。比如airbnb将user和house表示成向量帮助用户获取更准确的house推荐；youtoube将user和video分别表示成向量，提供给用户最好的视频推荐。那在搜索系统中里是否也可以用embedding来表示query和doc，然后进行召回排序？如果此路行的通，那embedding搜索能否革了文本搜索的命？

Embedding有个好处是方便和统一，通过端对端的模型训练去学习item的表示。并且加大数据规模，加强计算能力，理论上可以学习到更好的表示。搜索的doc是多种模态的，有文章、视频、图片、音乐... 如果能用embedding统一表示这些多模态的doc，那后面则可以用统一的召回和排序算法。同时也可以避免不同模态doc的字段不可比的问题，比如视频、图片的文本信息相比于文章是比较少的。然而，笔者先验认为embedding目前还革不了文本搜索的命。下文分析embedding搜索面临的潜在问题。

Embedding的学习大概分为两种形式：

- 首先学习细粒度的embedding（词/像素...），再根据细粒度embedding去计算更长/复杂item的embedding（句子/文章/图片/视频...），实际中尤其是文本领域大都采用这种方式；
- 基于用户行为日志去学习item的整体embedding，比如用户的点击行为可表示成序列，那用户/item则是序列中的元素，那可采用普通的embedding训练方法去获取embedding表示。这种方法需要丰富的用户行为；

1. Embedding的稳定性：比如一篇文章多加上/减去几个无关的字，embedding是否和原embedding很接近，还是抖动很大？比如在电商搜索，item的修改是经常发生的（比如商品的营销标题），修改前后学习到的embedding是否稳定；比如在视频搜索，同一份视频会以加logo，在最前面插入一些广告等方式被复制变成新的视频，这时视频的内容仍然是相似的，embedding学习该怎么去保证学得到相似的表示？

2. 长尾的embedding：中长尾问题又是搜索中比较难解决的问题。在文本搜索里，长尾query难解决主要是没有资源或者query的语义难学习。但是当长尾query有少量资源时，文本搜索靠文本匹配一定能够将对应doc召回。那embedding是否做到？无论是上文提到的方法1还是方法2，对于中长尾都比较难去学习embedding。比如query="科

比”的embedding能学习很好，但query=“科比和麦迪单挑谁能更胜一筹”的embedding能否学好就不一定了。

3. 搜索强调精准：区别于推荐，搜索的特性是需要将最相关的item排在top位置。那embedding检索能够支持这种特性？文本搜索是可以做到，因为最相关的，通常文本匹配度都是匹配比较好的。

4. Embedding能否支持分层计算：文本搜索考虑耗时通常采用分层排序，对大量候选item的轻计算粗排序和少量item的重计算细排序，那么embedding能否支持这种特性还是embedding采用的单层排序就够了？

5. Embedding是单个还是多个：item可能不止一个字段，并且不同字段可能反映不同维度的信息。是学习单个统一的embedding还是为不同字段学习多个embedding？如果学习多个embedding，这些embedding该如何融合或者如何在召回和排序中使用。

6. 不同模态embedding可比性：如果通过统一模型同时学习不同模态item的embedding，得到的embedding是可比的。但大多场景很难同时学习或者学习复杂度高，并且多模态或多任务方法还不太成熟，单独去学习item的emebedding能否保证可比性？

7. Embedding搜索可控性：文本检索在一定程度上是可控的，因为召回的item和query是至少存在共同的词。Embedding通过向量相似度计算并不能保证一定存在共同的词，那会不会误召回一些很飘的item？碰到这种badcase该如何解决？

8. Embedding能否解决一词多义：embedding在多词一义上取得了不错的效果，但是对于消歧问题呢？目前来看，其还只能覆盖一些非常明显的歧义case，比如苹果、黎明...

9. Embedding扩展性：除了相关性，embedding是否能cover排序的其他因子呢？比如时新性、多样性...

Embedding或许有一天会在搜索中起着比较重要的作用，但目前仍处在尝试和发展的阶段，而且可能永远都不会代替文本在搜索中的作用。

相关阅读

1. Query理解 - 搜索引擎“更懂你”
2. 从搜一搜中检“相关性排序”的排序结果说起...
3. 搜索排序 = 相关性排序?