搜索引擎的两大问题（2） - 相关性

原创 XG数据 WePlayData 2019-03-27

一个完整的搜索引擎往往包含了比较多的复杂模块，每个模块相互作用、兜底组成了我们使用的搜索引擎。抽象起来，召回和相关性是搜索系统里最重要的两个功能。本文主要介绍一下相关性问题。

相关性是为了计算query和返回doc的相关程度，也就是doc中的内容是不是满足用户query的需求。因此相关性计算就需要**充分的理解query和doc**。比如从quey角度，需要知道query中的哪些词比较重要，有没有实体成分，意图是什么？是要找药品的介绍还是要找药品的购买网站？query的时效性如何？是要找最新的事件新闻还是一般的事件介绍？从doc角度来看，需要理解doc的话题类型，doc的核心词，doc的文本质量，是否是标题党，是否是推销、广告、色情等页面？

充分的理解query和doc是非常有难度的，更进一步去计算query和doc相关性计算也存在很大挑战，比如以下几种场景：
1） query是歧义的，当用户搜索苹果时，是要找水果还是要找苹果手机？
2） query的意图不完全体现在term的匹配上，比如用户搜索“北京到上海的火车票”，doc“北京到上海的火车票的乘车体验”，虽然query完全紧邻命中doc，但用户要找的是火车票购买，并不是该doc；
3） query和doc的mismatched term对相关性也有很大的影响，而传统的相关性计算只考虑了matched term对相关性的贡献；
4） query和doc很多时候需要从语义维度来判断是否相关，比如query“苹果手机多少钱？”和“iphone xs max的官方标价？”；

从计算场景来看，query是变化的，需要在线动态计算，因此通常都是一些简单快速的方法，做轻一些；doc相对静态的，偏离线运算，因此可以使用很复杂的模型事先把doc的相关属性计算好，做重一些。

从计算方法来看，主要分为**字面相关性和语义相关性**两个维度。字面相关性主要是根据term的匹配度来计算相关性，一个不足是无法处理一词多义或者多词一义，并且会忽略词之间的顺序，常用的方法是BM25方法。语义相关性是近些年来研究的热点，像SVD，Topic Model，Embedding等等都是为了计算query的doc和语义相关性。其核心思想在于分别将query和doc标称一个低位稠密向量，然后用其cosine距离表示其相似性。Emedding是最近常用的方法，类似word2vec，doc2vec，sent2vec，lstm等等。不过embedding最早出现的还是word embedding，在word的embedding表示上效果比较好。如何学习长文本、有oov的文本的embedding一直没有得到很好的解决。最近比较流行的bert进行了相关实验，

效果也没有想象中的好。语义相关性的另外一个缺点是不太具有解释性，出现badcase只能大概猜个可能的原因。虽然embedding是个趋势，但字面相关性仍然是一个不可或缺的模块，起个断后的作用。

相关性模块扩展开来，可以认为是任意两个item的相关性计算。比如推荐是为了计算user和item的相关性，广告是为了计算user和ad的相关性。因此做好相关性计算是保证用户体验最核心的一步。

相关阅读

1. Query理解 - 搜索引擎"更懂你"
2. 搜索引擎新的战场 - 百度、头条、微信
3. 当我们关注舆情系统时，我们忽略了什么？
4. 搜索引擎的两大问题（1） - 召回