搜索引擎的两大问题（1） - 召回

原创 XG数据 WePlayData 2019-03-25

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。本文首先介绍一下召回问题。

召回是对于输入query，能够高效的获取query相关的候选doc集合。召回对于搜索引擎起着致命性的作用。因为一旦相关的doc不能够被召回，即使后面的相关性排序做好的再好，也是徒劳。尤其在doc资源不是很丰富的搜索场景下，召回更是一个比较明显的问题。

首先面临的问题是索引粒度问题。我们知道召回是通过倒排索引求交得到的，当以词为粒度，粒度较细，召回的文章的数目较多，但也可能由于倒排过长把一些相关的结果误截断；当以更大的phrase粒度，粒度较粗，召回的文章相对更相关，但也容易造成召回的结果过少。

其次召回要能够保证有一定的召回文章数。query大部分模块都是为了解决召回问题，比如非必留，同义词，纠错。这是因为query和doc往往会存在描述不一致的问题。比如query是“如何考取广大的研究生？”，但大部分doc都是讲广州大学的研究生。因此需要将广大同义成广州大学才能正确的召回一些相关文章。用户query也会存在一些错误query，比如刘德花，这时系统需要将query纠错成“刘德华”，才能正确的召回一些相关文章。用户query也会存在和doc不是完全匹配的情况，尤其是对于长query，比如“无问西东电影的主演是谁？”，如果要求原搜索串完全命中，可能导致召回结果数过少或零结果。这里分析“电影”是一个冗余的信息，“是谁”是一个不重要的词，其参不参与倒排的求交并不影响召回doc的相关性，这时召回时可以直接把这2个词直接丢掉。

最后召回要保证结果的多样性。尤其是对于短query。因为相比于长query，短query往往是一些实体，召回doc数往往不是关键问题，用户也希望有一些惊喜的结果，避免搜索结构都是一些类似或重复结果。query事件扩展，query改写都是为了解决召回的多样性问题。比如当用户输入“武汉大学”，如果只是返回一些武汉大学的百科、高考录取信息，可能对用户并没有什么吸引力。这是如果能将武汉大学能和最近比较热的“武汉大学 樱花”、“武汉大学 和服”关联起来，可能会有侧重召回扩展内容相关的doc，增加结果多样性。

前面讲的召回还主要都是基于字面召回，深度学习的发展使得语义召回是现在研究的热点和流行的方法。其思路是分为将query和doc表示成embedding，然后基于embedding计算得到一些相似的doc。这种召回方式虽然能够召回一些相关doc，但其不能保证一些最相关的文

章被一定会被召回回来。一方面语义会漂移，另一方面embedding模型往往是黑盒模型，很难debug。

召回问题不仅在搜索里至关重要，在推荐，广告中同样是一个关键问题。并且不同应用的侧重点不太一样，比如搜索中更侧重召回doc的数量，推荐中更侧重召回结果的多样性。因此做好召回是保证后续模块的第一步。

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？