R&S[26] | 搜索领域算法需要掌握的知识

原创 机智的叉烧 CS的陋室 2020-05-16

Eminem

Sunny Choi - Medleys, Vol. 1: Best Artists Best Songs

往期回顾：

- R&S[25] | 搜索中的意图识别
- R&S[24] | 浅谈Query理解和分析
- R&S[23] | 搜索系统中的纠错问题
- R&S[22] | 搜索系统中的召回
- R&S[18] | SIGIR2018：深度学习匹配在搜索与推荐中的应用

搜索领域在我的一些文章下，似乎让一些人产生了兴趣（窃喜），那么搜索领域需要掌握什么内容呢，这次我花时间总结了一下。

懒人目录：

前言

搜索早就不是新东西了，它的出现比机器学习深度学习可要早得多，也能蓬勃发展起来，截至目前，无论是百度谷歌必应这类的大搜，还是网易云音乐、应用宝等应用下的垂直搜索，都有一套比较成熟的方法。在我的视角看，搜索作为用户主动获取信息的重要入口，会长期存在，另外还有大量的人其实没有培养起搜索的习惯，随着慢慢培养需求会被激活，还是非常有发展前景的，而且成了推荐系统发展场景下的“灯下黑”，竞争其实并不如所想的激烈。

自然语言处理

自然语言处理在整个检索流程中都起到了重要作用，上游的query预处理、纠错拓展、意图识别，到下游的召回、精排，都要用到，简单句几个例子吧：

- 预处理和纠错，涉及到文本生成、文本共现挖掘等方面的技术。
- 意图识别，最直接的使用就是文本分类。

- 召回阶段，需要根据term weighting进行重要性算分，可供粗排使用。
- 精排部分，需要计算query和doc的匹配度，那就是文本匹配、语义相似度方面的内容了。

开始我的理解搜索中使用的是只有nlp中的NLU，即自然语言理解，如文本分类、实体识别、文本匹配之类，但是在查询拓展、纠错等方面，NLG，即文本生成，仍有一席之地。目前自己在文本分类和实体识别都点了一些技能点了，文本匹配尚且经验不足，没想到知道的越来越多，就发现自己不知道的也越多。

学习排序

在搜索引擎的初级阶段，一般就会从query理解整，在query理解模块逐步成熟，流量提升，doc增长后，就会开始建设排序模块，在已经具有优质内容的前提下，良好的排序能够为整体搜索体验带来新的提升，因此排序从这时候开始，将也会成为搜索系统的重要部分。

learning to rank其实是一个比较重要的课题，已经成为一个单独的科研课题，但是在现实应用中更多的工作关注于特征，对base模型是逐步提升迭代的，lr、xgboost之类的还是主流和首选，后面才有了基于问题的优化，很多工业界的模型其实都是基于业务场景进行的优化。

数据结构与算法

搜索系统涉及大量的计算，尤其是搜索方面的计算，因此需要构建大量方便检索的数据结构，其中大部分都和树有关，如trie树等，当然还有各种哈希等等，为的是能快速从海量数据中找到所需的数据，确实需要进行大量的操作，这些工作大部分可能是在工程团队手里，但是仍有部分内容会在算法的手里，例如词典的构建来协助提升准招。