R&S[24] | 浅谈Query理解和分析

原创 机智的叉烧 CS的陋室 2020-04-27

熟悉却不知道歌名系列，听下去。

Roundabout [In the Style of Yes]

The Karaoke Channel - The Karaoke Channel - The Best Of Rock Vol. - 77

往期回顾：

- 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
- R&S[17] | 手把手搞推荐[6]：回顾整体建模过程

搜索是一个系统，大小不好说，但肯定是五脏俱全，我做的比较多的就是query理解和分析，这次给大家重点讨论一下这块内容。

久违的懒人目录：

- query理解的目的。
- 例子。
- query理解的内容。
- query理解的操作。
- 背后的知识。
- 软实力支撑。
- 总结。

query理解的目的

query理解是整个搜索系统中最上游的一环，负责的是从query中提取信息，从而了解用户希望通过这个query搜索出什么，泛泛的说完了，剩下来看详情。

query理解，决定了下游的搜索召回策略。底层数据从技术上，有各种类型的数据库需要检索；从算法策略上，也有多种召回的方案，例如高准确的、高召回的等等，要用什么策略，这要取决于query理解的结论。

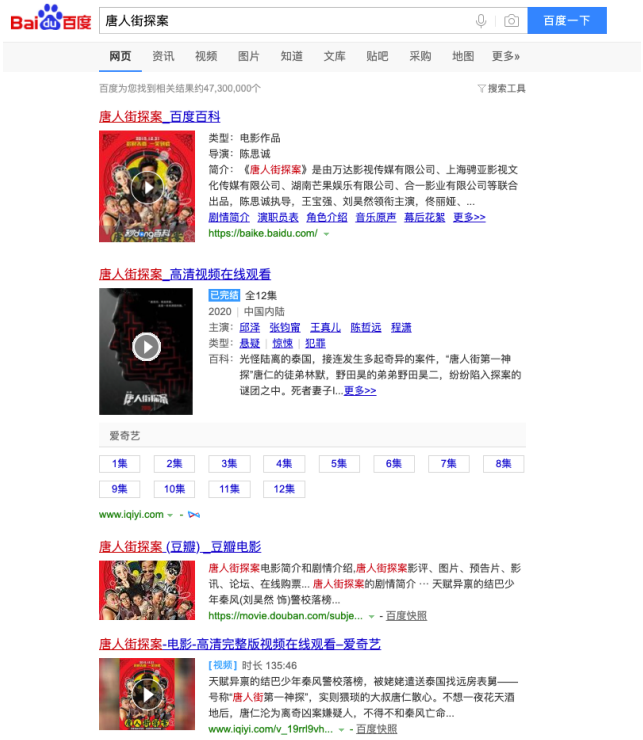
例子

要把一个事情说清楚，举例是一个很好的方法。来一个query：唐人街探案（此条广告5毛，记得删掉括号）

直观的看，这里有一个核心词——唐人街探案，其核心意图就是想看看唐人街探案的相关内容吧，来看看系统内干了些啥：

> 纠错：初步看来，没有错误，过。
>
> 意图识别和实体识别：有唐人街探案这个实体，常见的首先是一部电影，最近还上了网剧，从热度上看，由于网剧比较新，所以用户在近期更可能看的是网剧，当然信息不足，不代表用户真的就只想看网剧，所以电影的东西也要给一些，最大限度保证满足需求。

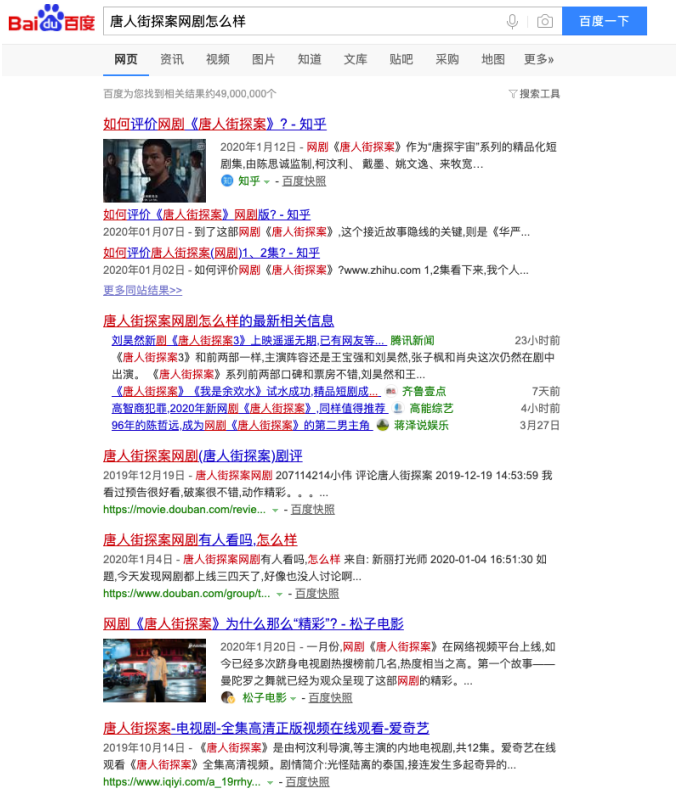
好了，以百度为例看看结果：



前4条，分别给的是百科、爱奇艺网剧、豆瓣电影评论、爱奇艺电影。基本上就覆盖了我上面的分析内容，用户只输入了一个简单的实体，就会给出精准的对应信息，百科了解概况，爱奇艺网剧满足近况，豆瓣电影有影评，最终补充了电影，满足更为全面的请求。

我们来复杂一些，升级为唐人街探案网剧怎么样。

唐人街探案还是一个实体，网剧和电影双意图，但是由于用户输入了网剧，有关电影的内容基本上就可以不出了，最后来了个怎么样，说明用户是更在乎影评，而非要看电视剧了，当然给电视剧了用户不会反感，属于弱意图了。好了，来看百度结果：



前5条都围绕着影评进行，可以说是分析的非常精准了，且前面几条也是比较出名的媒体给出的答案，知乎、新闻、豆瓣、松子电影，第六条很机智的给了爱奇艺的链接了，而且不是展开的，而是一个摘要形式的，大家可以对比一下上一条搜索的结果区别，从这里，大家就能理解，query理解具体做了些什么事情。

query理解的内容

那么，要做query，要做什么工作呢。仔细想想，其实主要就是下面几个：

- 纠错改写。针对用户输错的，没输入完全的，内容，进行修正。底层数据库只支持精准搜索，因此需要将query改写到正确的内容下。
- 意图识别。通过分析语义等方式，在一定的类目结构下，识别出具体意图。这个意图识别的目标，大家可以理解为告诉下游，需要在哪个库数据进行搜索。
- 实体识别。其实和意图识别一样，只不过，粒度更细，但是是词级别的分析，从query中抽取关键的实体，如果说意图识别是为了告诉下游该检索那个数据库，那实体识别就是为了告诉下游，在该数据库下，该检索哪些字段。
- 词权重问题。query里面有两个词，两个文档分别匹配到了其中一个词，那谁能靠前？这就要看匹配到什么内容更为重要。如家宾馆，匹配到一个如家酒店和五洲宾馆，如家酒店应该在前，这里就是为了解决这个问题。

query理解的具体操作

query理解下的所有内容，除了意图识别本身外，其实我都或多或少介绍过，简单的谈一下。

纠错改写

之前写过文章了，在这里：R&S[23] | 搜索系统中的纠错问题

简单地说，就这两个思路：

- 基于统计挖掘，分析最高频的正确答案，在用户错误的时候，分析他的真实意图，改写过去。
- 基于机器学习和深度学习，识别错误，改正错误。

意图识别

意图识别简单的理解，其实是一次文本分类，那么文本分类，我们把思路拓展开，其实也是两条路——传统方法和NLP。

- 传统方法想必很多人其实了解的并不多，但其实是搜索领域内非常常见，通过规则、词典、正则等方式进行识别，准确率高、速度快。
- NLP，通过语义分析的手段，文本分类，达到语义分析的目的。

实体识别

其实问题抽象出来，就是个难度高于文本分类的序列标注问题，搜索中的命名实体识别，我聊过的，在这里：NLP.TM[18] | 搜索中的命名实体识别。具体思路仍然分为两派，传统方法和NLP。

词权重问题

我还是聊过哈哈：NLP.TM[20] | 词权重问题。这个问题，在我多年（咳咳恩，别装）的经验下，感觉这是一个非常考验基本功的任务，对数据的理解、操纵，对程序的把握，都有很重的考验。

- 统计的方法，其中tfidf最为常见，而由于query的长度都不长，所以其实就是idf的计算了。
- NLP方法，其实就是序列标注问题的升级版了。

背后的知识

最近是在知乎上回答了一个问题：

一个合格的搜索算法工程师应该具备哪些能力?

https://www.zhihu.com/question/381003357/answer/1173551448

里面的内容不复述了，这里主要谈谈可能涉及的技术，无论是算法模型，还是工程技术。

算法上：

- nlp是跑不掉的，尤其是文本分类和序列标注，另外建议大家懂一些文本相似度的计算，会有很大机会用到。

- 紧跟着第一点来强调，对短文本的处理，非常重要。
- 简单复杂的统计和实验方法，在设置阈值、调优等等，非常重要。
- 低复杂度的，低消耗的算法，建议大家要会用，要懂原理，例如fasttext、word2vector。

技术上：

- python必会，java和c++至少会一个，尽可能会两个，这是门槛问题。
- 搜索领域常用的数据结构要懂，trie树之类的，当然基本的数据结构，链表、堆栈、树之类的，是基操。
- 会自己动手写一个可通信的服务，语言不定，要求对这个通信服务有概念，常见的是http请求，另外还有底层通信的grpc，这个其实对算法而言是更常用的。
- 大数据的操作，hadoop、hdfs、spark，当然还包括mysql。

软实力支撑

和选电脑一样，不能只看面板指标，和选男女朋友一样，不能只看外表，那么在搜索领域，最好有哪些软实力呢？

- 需求和问题的分析能力。这个分析，除了针对这个问题算法角度的分析，还要尽可能考虑多方，前后端数据端等等，算法计算一个东西需要什么，上游能不能给你，下游需要一个什么结果，你需要怎样才能输出，这都是很大的问题，这里面出错了，你大概率就白干了，时间有了kpi还没有，加班通宵就有了。
- 方案的选择能力，脑子里多几个方案，在面对问题时能提出针对性方案。这要求脑子里有方案，且知道这几个方案优缺点，且还能分析出问题的特性，所以其实没那么简单。
- 性能的敏感度。工程项目，对性能要求尤其高，这是一个可用性的问题，你总不能让用户点了搜索，10秒钟不出结果吧（网络问题除外），自己可以试着去百度看看，自己能忍受多长时间的加载。速度只是一方面，还有内存。
- 数据敏感度。模型策略出了问题，能快速找到，那怎么这里就有几个问题了，首先怎么判断有没有问题，然后是有没有什么方案可以快速处理，处理后能提升多少，有没有什么代价，这都是需要分析、预判的，这都依赖了数据敏感度。

总结

简单的就说这些吧，有关这块的理解，我推荐一个资料吧，有助于大家理解这个问题：

- 美团机器学习实践。第八章1、2节。

由于这个课题很大，我也不好详细提出一个方案，几段代码，而是带大家去理解这个问题下的思路是什么样的，怎么去解决，这是一个能独当一面的算法工程师需要知道的，而非每天盯着几篇论文，重现了一篇论文就高兴不已的样子，说实话我以前就是这样的，但现在发现，这样做远远不够。细品这句话，不是说这么做不对，而是，不够。

倘能生存，我当然仍要学习。——鲁迅。