# 前沿重器[4] | 腾讯搜索的Quer理解如何直击心灵

原创 机智的叉烧 CS的陋室 2020-10-18

收录于话题
#搜索 10 #自然语言处理 14 #对话 9 #前沿重器 3

**李白**

李荣浩 - 模特

**【前沿重器】**

全新栏目，那么栏目主要给大家分享各种大厂、顶会的论文和分享，从中抽取关键精华的部分和大家分享，和大家一起把握前沿技术。具体介绍：仓颉专项：飞机大炮我都会，利器心法我还有。

往期回顾

- 前沿重器[1] | 微软小冰-多轮和情感机器人的先行者
- 前沿重器[2] | 美团搜索理解和召回
- 前沿重器[3] | 平安智能问答系统
- NLP.TM[38] | 对话系统经典：检索式对话
- SIGIR20最佳论文：通往公平、公正的Learning to Rank!

说起搜索，腾讯腾讯应该并不算领先，但是受益于这篇来自腾讯的分享，我们来看看腾讯的研究成果，看看如何完整地理解用户的搜索Query，并找到用户需要的东西。

先把分享放在这里：https://zhuanlan.zhihu.com/p/112719984，来自腾讯技术工程。

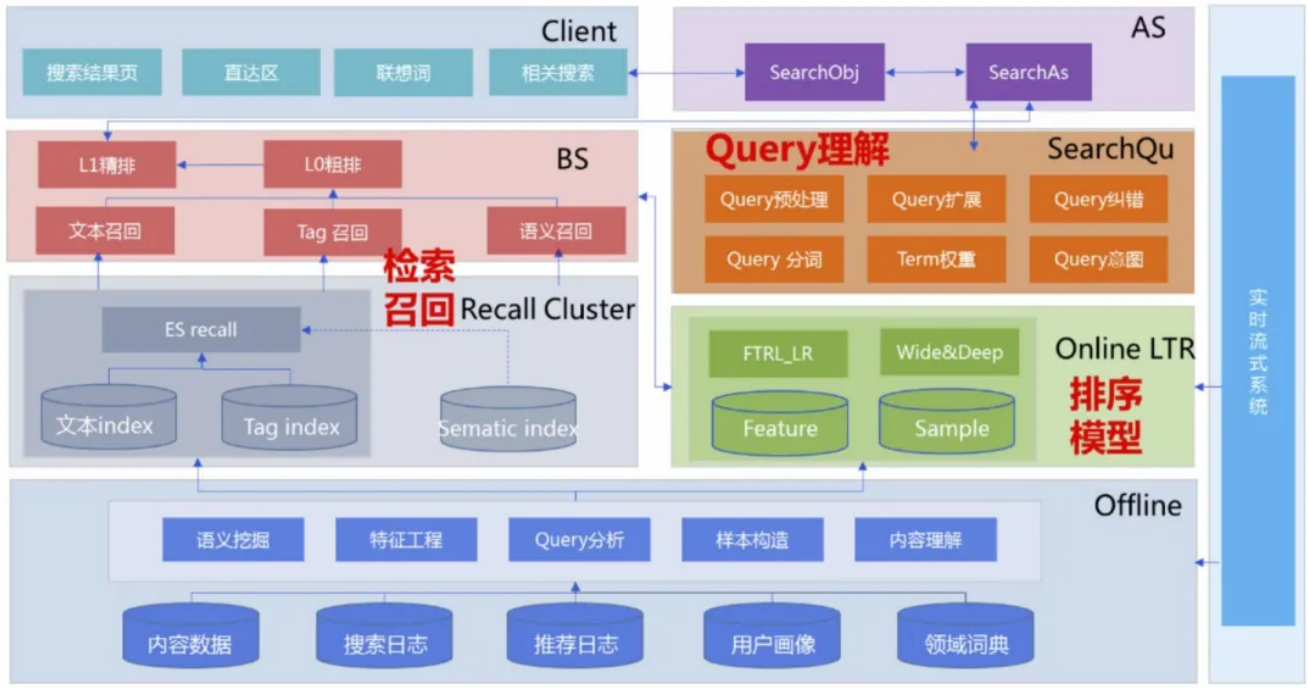
懒人目录：

- 搜索工程架构
- Query理解
  - 预处理
  - 分词
  - 改写
  - term分析
  - 意图识别
  - 其他操作
- Query理解结果的应用

  - BS需求
  - 排序需求
- 小结

# 搜索工程架构

搜索的场景有很多，有QQ音乐、爱奇艺视频、应用宝等垂域的搜索，也有百度、头条、微信等搜索的开放域大搜，随着搜索场景的变化，架构或多或少也会有很大区别，但大都满足一个非常通用的结构：输入-理解-召回-排序-返回的结果，本篇分享为大家整理了一个更为详尽的版本，大家可以参考一下：




这张图给出了比较简洁而又完整的架构，不仅说清楚了各个模块如何交互，还展开了解析了各个模块需要做的事情，我想好好研读这张图，大家一定会有深入的了解。

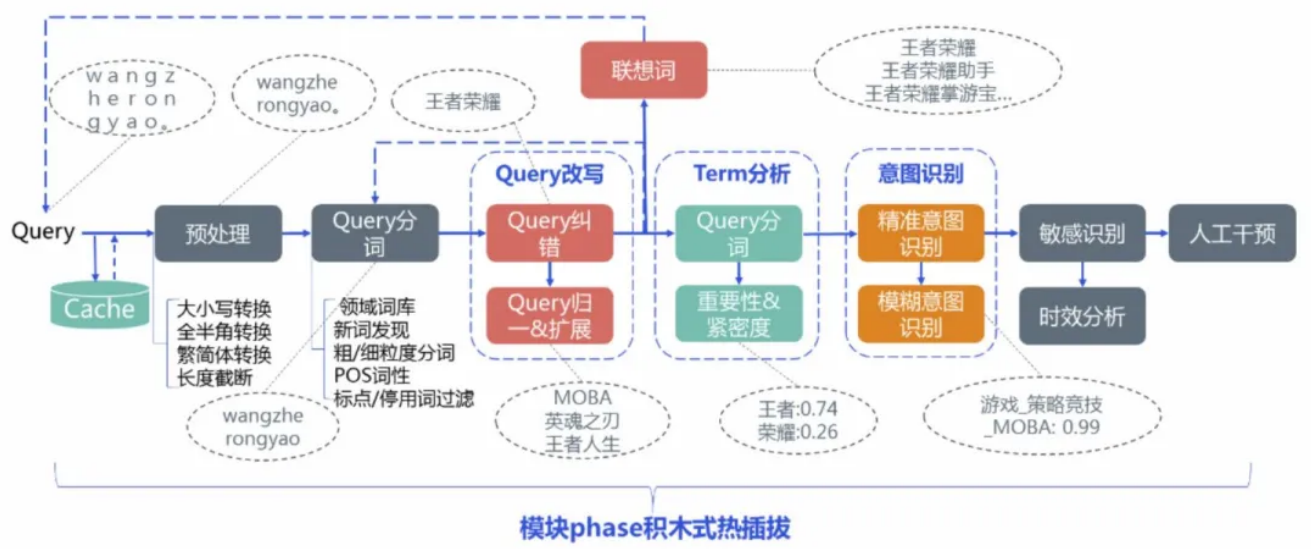
这张图足够清晰，我不需要做太多的重述，我认为更需要了解的是架构设计背后考量的因素和原理。我认为主要有如下几个点需要展开讨论一下，当然有新的理解欢迎评论区留言交流：

- 特地单独划分client层主要是因为有很多query可以采用更加基金的方式处理，如直达等，这些query的意图非常明确，不需要重走大量的流程，通过词典等方式快速配置快速完成，能给到用户极致的体验。
- AS(advanced search)和BS(base Search)虽都为搜索但是两者有所区别，前者追求更丰富的语义把控，而后者则直面检索，通过文本、意图、语义等方式构造多种检索逻辑完成搜索过程，直接请求的就是底层的搜索引擎（如大家熟悉的Elastic Search，当然还有比较新潮的向量召回引擎Faiss），两者关注点不同。从流程上看，BS应该是AS的一部分。
- 搜索是一个在线互联网产品，但是其离线的工作并不比在线要轻松，无论是用户侧的行为、query记录和处理还是物料的存取处理，都有非常多的任务，这些工作的直接体现就是BS检索的对象以及算法计算的特征。

- 和推荐一样，使用的同样是“召回-排序”的二段结构，召回阶段尽可能保证需要的内容能够被找到，排序阶段则负责把更加“相关”的内容展示给用户，而在这里，统一都是BS负责管理。

# Query理解

Query理解已经不是我第一次聊了，这篇文章也能看到Query理解的操作有哪些。虽然如此，这篇文章还是刷新了我的理解，无论是方法还是操作上都有不少的新东西。首先看一个pipline图：




图就不细说了，直接展开来讲吧。

## 预处理

预处理的操作还是比较统一的，来来去去其实也就是这些事：

- 大小写。
- 半全角。
- 繁简体。
- 标点符号。
- 长度截断。

## 分词

搜索分词要是想做复杂，其实可以很复杂，绝对不是jieba那么简单，在特定场景下需要定制各种各样的分析，主要考量的是这几个的角度：

- 切词粒度。往大的切和往小的切。“滑雪大冒险”可以是“滑雪 大 冒险”也可以是直接一个专名“滑雪大冒险”。

- 领域问题。有些可能是领域内的词汇，我们需要注意识别，切开了反而没有意义了，上面滑雪大冒险的例子也是这样，如果在APP领域就不能切开。
- 新词发现。对于时效性比较高的搜索引擎，如微博等，需要快速识别新词以免由于切词错误导致下游无法搜索，如近期有的“水门弄斧”。

有关具体里面的方法，其实都还比较基础，复杂的模型也有但是在这块花费宝贵的耗时意义并不大，所以基于词典和简单的机器学习（HMM/CRF）之类的就能能够达到目标了，新词发现的话左右熵之类的就能快速挖掘出来。

## 改写

改写应该是一个比较复杂的工作了，在文章中主要分为3种情况：

- 纠错。针对句子中出现的错误进行修改。这块的方法我曾经写了不少文章讨论了。（NLP.TM[37] | 深入讨论纠错系统）
- 归一。某些内容有很多说法，把他们统一化有利于下游更加快速的搜到。实质上是一个同义词挖掘和同义句征程的过程，这个来源于挖掘积累，甚至是一些知识库的建立可以完成，按照文中的说法就是一个“积累”的过程。
- 拓展。这个主要通过用户行为进行挖掘即可完成基线，后续通过这些平行样本做一些语义相似度的训练，再用annoy等方式做query的向量召回也很方便。

## term分析

term分析是我自己比较熟悉的领域了，这一步的目标是深入到句子中的词汇中进行分析，包括实体抽取、重要性分析、紧密度分析等。

紧密度分析，我的理解，往难听的说，来源于对切词的一种不信任，往好听的说就是对切词具有更高的灵活性要求，切词很可能对某些词汇切得太散，所以需要紧密度分析把他们都合并起来，这种合并在后续构造检索语法树的时候（可以理解为检索语法构建）会发挥很大作用，例如上面的“滑雪大冒险”，合并以后，简单的“滑雪”的东西其实就不会出来了，这样对提升精准度很有好处。

有关重要性分析，我之前也有文章提到过，简单的基于TF-IDF、textrank之类的方式就能轻松处理，而复杂的用序列标注的方式处理也是有的，在这篇分享里，作者列举了很多可以用到的特征可供大家选择，至于模型则可以考虑BiLSTM+Attention（个人感觉attention本身已经是一种词权重计算）、Seq2Seq/Transformer：

> term词性、长度信息、term数目、位置信息、句法依存tag、是否数字、是否英文、是否停用词、是否专名实体、是否重要行业词、embedding模长、删词差异度、前后词互信息、左右邻熵、独立检索占比（term单独作为query的qv/所有包含term的query的qv和）、iqf、文档idf、统计概率

> $$P(t|d) = \frac{P(d|t) * P(t)}{P(d)}$$
>
> 以及短语生成树得到term权重等

实体识别（或者叫做槽位）在美团内把它归为term分析的一种，而在这篇腾讯的文章中则把它归到意图分类下的，当然这一方面取决于个人理解以及架构的设计。这个我也已经写过，此处也不赘述。

- 前沿重器[2] | 美团搜索理解和召回
- R&S[24] | 浅谈Query理解和分析

## 意图识别

意图识别在本文被成为query理解中**最重要却也最具挑战**的模块，并列举了原因：

- 用户输入不规范。
- 歧义与多样性。
- 用户的个性化差异。

好了来谈谈意图识别的操作。这里作者分为了2种，一种是精准意图，另一种是模糊的。前者我们只需要把他需要的放在他面前就好了，后者则需要更复杂的理解操作来保证返回的内容尽可能满足用户的需求。

精准意图来源于用户非常明确且高频的需求，如“下载王者荣耀”，这种唯一且精准的意图，如果同样采用模糊的方式去做，很可能因为个性化、物料的相似性而导致错误，但这个又是高频精准需求，所以我们需要更好地保护，这种问题可以把它归为QA，即问答来处理，一般做法其实来源于搜索（套娃？？），两套思路，分别是query理解+检索和语义召回，可以去召回与他相近的query这些都是一些常见的高频精准query，可以直接去整。这些操作有的地方也被叫做搜索直达，是一种非常激进的策略。

对于模糊意图，我们就需要步步为营，扎实理解好用户的意思，这就需要借助意图分类（可抽象为文本分类）和槽位提取（命名实体识别）来做了。

## 其他操作

文章还提到了两个工作：敏感识别和时效性分析。

敏感识别来源于对用户输入内容的不信任和对开放域物料的不信任，说白了用户可能输入黄反的东西，物料里也可能存在踩雷的东西，保证物料里面没有坏东西和保证用户输入没有坏东西都很重要，简单的可以用词典做匹配，当然复杂的还要结合语义分析，采用文本分类的方式来处理（个人经验来看文本分类能解的其实词典也可以额，类似弓虽女干的其实模型也很难覆盖）有很多很边缘的可能至今也不太好处理，所以现在的主流还是以可控性、准确性见长的词典。

至于时效性分析则也是一个复杂的问题，作者举的例子是同样是“最近上映的好看电影”、“疫情进展”之类的，在不同时间则会有不同的结果，作者将此类问题拆解成3种：持续时效性（“美食推荐”），周期时效性（“世界杯”），准/实时时效性（微博热搜之类的）来进行针对性的处理。

# Query理解结果的应用

本节并未在文章单独提出，但是个人认为应该结合搜索工程架构和Query理解及其在其中的位置来详细讨论下。

## BS需求

BS就是基础检索（base search），说着玄乎了，说简单点，计算机顶层再怎么智能底层也是非常头脑简单的，就是我们要在数据库里面找东西，需要构造一个类似SQL的检索逻辑才能够找到最优解，当然我们有特定的一些方式能让你搜得更快（构建索引），但是终究要构造那特定的形式才能完成搜索，所以我们需要充分理解Query，转化为计算机可以理解的格式才能够完成BS。

最传统的方式就是提槽、判断意图然后构造检索语法树来进行搜索，举个例子，query“唐人街探案的主角是谁”，我们要判断是一个百科意图或者是电影意图，然后按照类似“select actor from baike where type='movie' and item_name='唐人街探案'”的逻辑去进行搜索才行，所以我们要做query理解，进一步的我们通过词权重找到句子的关键词来协助进行相关性排序。另一方面，由于是文本匹配，如果数据库里面是“肯德基”而用户输入“KFC”，那也是无法匹配到的，所以我们也需要大量的改写逻辑。

复杂的，现在近邻索引也逐步发展，于是有了ball tree、annoy、hsnw这些以树或者以图为基础的索引，于是向量检索也成为了可能，通过语义向量召回的方式也能够召回一些内容，但首先我们也是需要通过向量生成模型来对他们进行向量化对吧，所以也是需要query理解模块。

## 排序需求

有了BS的召回结果，结果很多，我们如何排序呢，光靠语义相似度就够？肯定不可，我们要看召回的内容里特定的槽位有没有满足，意图有没匹配，还要涉及query里面关键词的重要度、流行度等等，语义相似度需要从一个排序模型降级为一个特征，这么多的特征结合机器学习就能有远优于语义相似度的效果（当然也可以以语义相似度为base添加其他特征concat起来也可，推荐不就是这么玩的么，狗头），所以，query理解是真的很重要。

## 小结

Query理解可以说是搜索引擎体现人工智能的最关键的一个模块了，也是算法所需要去关注的一个核心点，足够深入的理解和解析query，我们还能够挖掘出更加深层次的东西，从字面的简单需求到延伸隐含的需

求，这样就，能成为肚子里的蛔虫了~

## 我是叉烧，欢迎关注我！

叉烧，OPPO搜索算法工程师，主做Query理解，NLP方向。19届北科技统计学硕士（保研），17届北京科技大学信息与计算科学、金融工程双学位毕业，论文7篇，学生一作3篇，参与国家级及以上学术会议4次，优秀论文一次，国奖金。曾任去哪儿网大住宿事业部产品数据，美团点评出行事业部算法工程师。

微信个人公众号
CS的陋室

微信 zgr950123
邮箱 chashaozgr@163.com
知乎 机智的叉烧