搜索query意图识别的演进

DataFunTalk 2020-11-15

以下文章来源于微信AI ， 作者jackhan

文章作者：jackhan
内容来源：微信AI

导 语

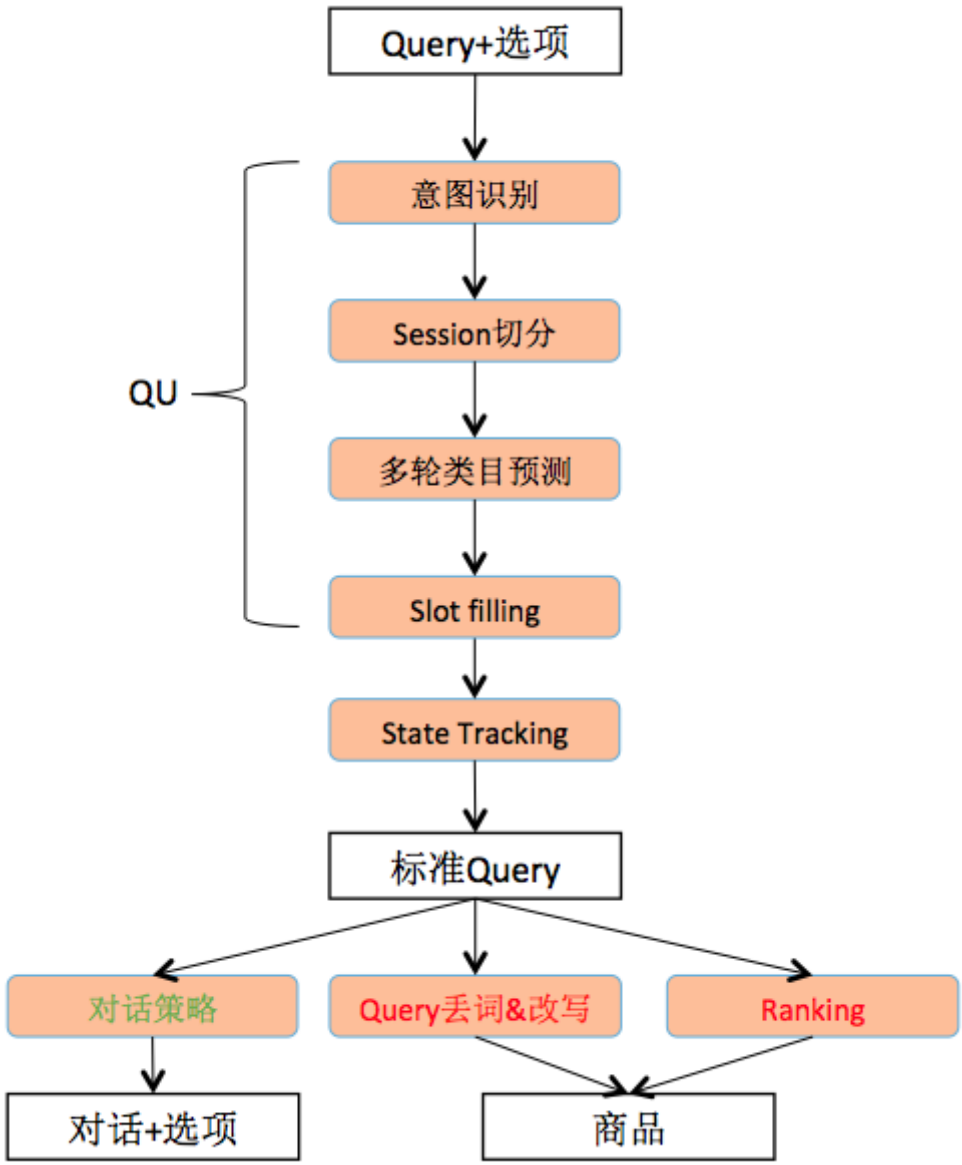
在如今搜索领域中，简单的关键词匹配已经无法胜任全量的query，如果能够识别出query的意图，对于返回类型多样性，提升相关资源占比以及关联相关结果更加有效。所以query的意图识别尤为重要，在一些垂直领域中，query意图识别也演变为类目识别。借着近期工作内容，梳理一下在垂直领域下query的类目识别。

问题背景

什么是意图识别

在搜索场景中，用户输入query，搜索引擎根据用户的搜索返回相关信息。例如用户搜索“苹果6”，这个query意图对应的类目为“手机通讯-手机”，返回的结果应该为“Iphone6、Iphone6s、Iphone7...”。用户搜索“红苹果”，对应的类目为“水果-苹果”，返回的结果大多数为“红富士苹果、红苹果”等。如果意图识别类目做的较为准确，最后返回的结果也不会相差太多。

常见的NLU任务为下图，意图识别是Query Understanding部分，作为NLU的第一步，结果也会影响后续部分。商品在关键词索引召回之后，在第一轮海选粗排阶段通过类目相关性，可以优先选择更相关类目的商品进入第二轮精排中。一方面保证排序的效率，使得排序在类目相关的商品集合上进行；另一方面从最上层保证类目的相关性，保证用户的体验效果。



NLU任务模块

意图识别的难点

由于query具有简短、缺失、时效等多个特点，所以在识别意图时也会存在很多问题，问题总结有：

- **输入不规范**

每个人对问题的描述方法不同，输入具有多样化。有的使用交流的口吻，例如“适合女朋友穿的红色的裙子”；有的使用关键词方法，例如“裙子 女朋友 红”。

- **多意图问题**

同样的query对应了多个意图，例如：搜索“水”是指矿泉水还是爽肤水；“苹果”（手机/水果）、“变形金刚”（玩具/电影）等也有多个意图问题。

- **时效性**

随着时间的推移，相同的query会改变其意图。例如用户搜索“Iphone X”现在很大可能对应了手机类目，再过三年，可能对应了手机配件或者维修，再过五年，可能对应的是Iphone X回收相关类目。这个例子可能不是很典型，再换一个季节性的例子。例如夏天搜索“衣服 女”，可能对应的是“裙子”、“T恤”等类目，冬天搜索“衣服 女”，可能对应的是“外套”、“衬衫”等类目意图，所以同一个query会因为不同的时间对应着不同的内容。

- **用户信息冷启动利用数据较少**

对于一个刚刚新建立的业务，是很难拿到大量的点击数据以及用户日志的，所以冷启动比较困难。如果对应了物品本来类目就不是很明确，那意图识别就会更加困难。

大家都是怎么做的

之前介绍了什么是意图识别以及意图识别会遇到的一些难点。现在垂直领域中的一些搜索服务，普遍是根据用户关键词来理解字面上的意图，另一方面会使用语义信息和用户历史行为做意图预测。在本文末尾会介绍垂直领域的两个意图识别案例作为借鉴。

常见意图识别处理办法

词表穷举法

词表穷举法顾名思义是使用既有词表对query做pattern映射，将query转化为固有模式的组合，便于匹配意图。这个过程首先是维护已有词表，词表中的数据对应到不同的实体类型，例如地名，品牌名，属性名，人名等。然后将Query中分词后的结果映射到对应的词表类型中。再匹配固定的查询模式，最后得到用户搜索意图。例如：

- 查询词语：澳洲[addr]cemony[brand]水乳[product]面霜[sub_product]

- 查询pattern: [brand]+[product];[addr]+[product]+[sub_product]

这种直接词表穷举的方式较为直接简单，在初期冷启动时可以运营同学来负责pattern的构建，以及查询匹配策略。在搜索Query中，也满足中长尾分布，头部query一般占有60%以上QV，可以使用硬匹配规则解决。中尾部Query虽然占有QV量低，但是Query量大，需要的运营成本高，在无法覆盖全面query时，词表穷举法的局限性就显露出来了。

在这个基础上也衍生出规则解析的方法，该方法可以把现有的query中的词汇映射为一些固定搜索类型，类型分类更加容易，当然在对类型分类时也需要人工参与。例如：

- 一英镑等于多少人民币。这个query可以转化为 [数字][币种]等于[数字][币种]

- 《哈利·波特》怎么样。这个query可以转化为 [书/电影/音乐]怎么样

统计分析

统计分析是针对现有用户行为日志所做的数据分析。用户在产品中进行浏览，停留，搜索，点击，收藏，购买时，会产生很多的用户行为信息，这些行为信息都能够辅助挖掘分析Query。比如前面的“《哈利·波特》怎么样”的例子，可能会遇到用户到底是想要搜索书还是想要搜索电影。经过分析最近用户数据后发现，在用户搜索完这个query后常常点击的是哈利波特这本书怎么样，而不是电影，如果没有书名号时“哈利波特好看吗”，可能更多的点击是跟电影相关，所

以就挖掘出对应的意图。这里的挖掘主要是离线的数据分析，在产品中可以体现在两个场景。Query Suggestion(搜索下拉框建议)和相关结果。

Query Suggestion是一个常见的场景，现有搜索引擎大多都支持下拉框推荐。首先将用户日志中的搜索与点击信息建立连接，针对不同的搜索query会有对应的最大可能意图，然后将其中非低质的query筛选出来作为备选query，按照具体的qv量，点击收益等各项指标对query打分，得到query的分值，该分值就是离线挖掘的结果。具体怎么上线呢？一般使用前缀的方式来存储现有的query，一般工程团队会使用ES的方式来构建前缀索引。在用户输入前缀时，可以得到有可能的整体query，按照之前计算好的分数来排序得到最终下拉框展示。推荐出合适的query出来。

相关搜索在如今的垂类搜索中更为常见。用户在一个session中搜索了一个query，没有点击，然后再次将该query修改了一部分结果，接着点击进入了。这里能够从session中挖掘出前后query其实具有同样的指向性，这样就能够挖掘出前query的意图。例如：用户首先搜索了“python分析数据”，用户没有点击，然后继续搜索“python数据分析”，点击了python数据分析这本书的购买链接，而且最终购买或收藏了该书。这个过程经过挖掘就可以得到“python分析数据”也具有搜索书的意图。这个策略在query改写上面应用比较广泛。

离线日志信息中包含的信息量非常多，可以找到多种挖掘方案来获取到用户的关联数据，例如用户协同过滤，物品协同过滤等方案能够准确返回给用户想要的内容。但是离线分析中包含的噪声也很多，在提炼时要摒除干扰信息，同时也要注意均衡各个类目，防止有的类目被多次搜索推荐，最终无法覆盖多样性。

机器学习方法

对query的意图识别实际可以看成是一个多分类问题。这个问题的label是搜索的类目，特征可以用到不同类目对应到的常见词语以及常见query，在用户输入新的query时，判断其属于不同类目的可能性大小。这里会遇到一个问题，类目很多怎么构建？因为在多分类中类目越多越难以预测成功。这里可以以多级类目的分类方式，先把query分为多个大类目，然后在大类目中预测小类目，大类目预测成功后整体偏离也不会太大，在小类目中要求也会小很多。与此同时，机器学习中的标注数据和模型更新也较为麻烦，需要大量的人工标注数据来训练，挖掘出来的正负样本以及特征需要清洗才能够得到较为干净的训练数据。一般使用适合多分类的算法分类：最大熵、FastText、TextCNN、BI-LSTM + attention等算法。其中FastText在面对文本多分类时具有容易实现，方便快速上线等优点。

模型相关特征的准备

■ 词性 & 主体识别 & 属性 / 标签识别

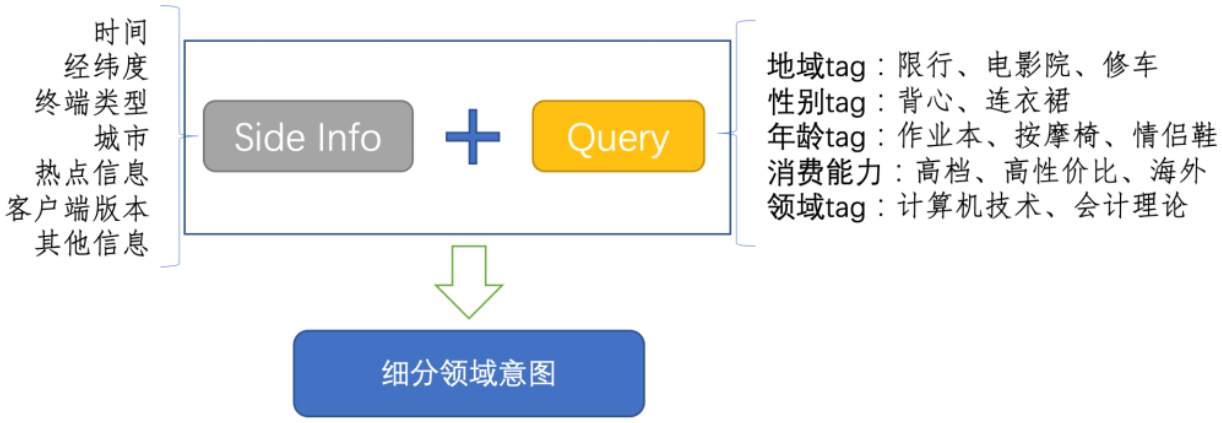
在搜索过程中，不同的term对于搜索的意义也不同，例如“桃子味的牙膏”，这里的桃子是修饰牙膏的，核心词为“牙膏”，核心词应该就有更高的查询分值。在意图类目识别时也应该根据核心词来确认。所以提前对query做词性分析，NER，以及对待搜索物品或类目做tag将会帮助检索。常见的序列词性标注有基于规则和基于统计以及CRF，Bi-LSTM+CRF或者带上预训练BERT的方法。标注的目的在于找到合适的query重心。目标在于识别出核心产品词语，常见的标注方式如下：

词语1	词语2	标注1	标注2	中心词
牙膏		名词		牙膏
桃子味的	牙膏	形容词修饰	名词	牙膏
薄荷	口香糖	名词修饰	名词	口香糖
跑步	鞋	动词修饰	名词	鞋
苹果	手机	名词	名词	苹果、手机
连衣裙	绣花	名词	短语修饰	连衣裙

核心词语权重

■ 词语画像&词语用户画像

有些词语在query中分数较低，可以根据query的词语打分来选出高分值词语。词语画像是指词语具有一定指向意义，例如“大嘴猴”、“古奇”等词语是特定品牌词，提前确定好品牌词的映射关系；词语的用户画像是指在query中能够对用户的身份做联想理解，例如“高档酒店”能够一定程度映射出用户身份，“洛丽塔”/“少女连衣裙”能够一定程度反映query指向的年龄，图片说明了能够根据query的个性化标注来完成浏览型query与转化型query的处理。所以对词语的画像能够帮助query更好的理解用户。



tag意图

- **其他特征**

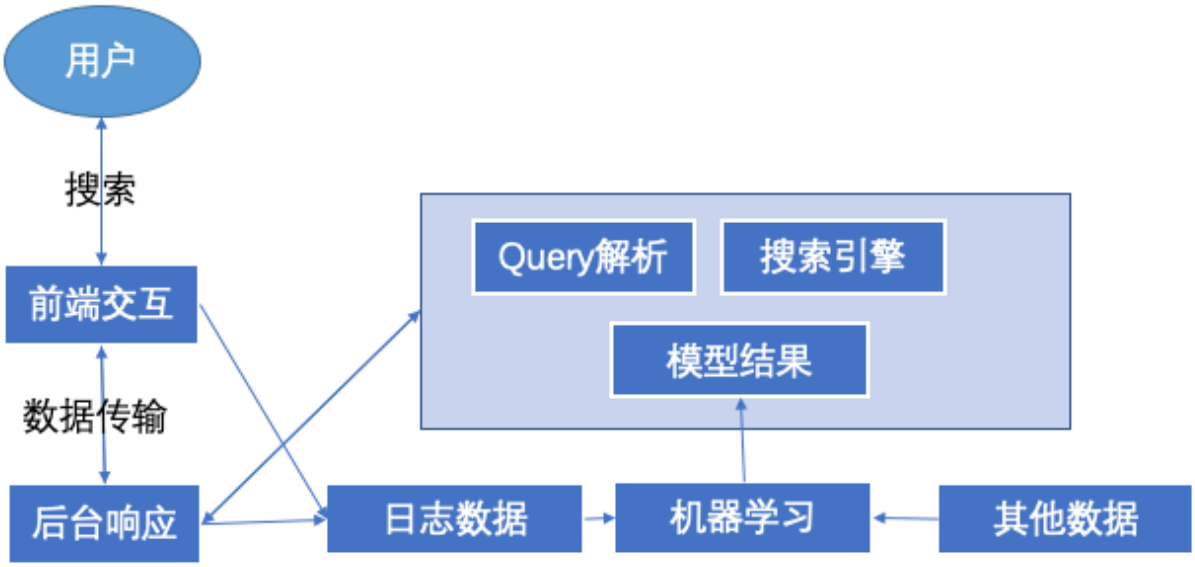
• Session内特征
– Query个数
– 翻页数目
– 点击情况
– 购买行为前浏览时长
• Query类型特征
– Query空格情况
– 命中关键词数目
– 无效字符清洗后剩余信息
– query长度
• 其他信息
– 搜索tab垂类类型：不同类型数目不同。例如服装鞋包类数目显著大于其他类型
– 搜索时间等side information

一些产品的意图识别方案

注：本文中涉及到的其他企业信息均来自互联网公开资料。

案例一：某海淘平台

某海淘平台作为一个电商类产品，搜索较为单一，最终目标都指向浏览产品和产品购买。该平台在对Query处理时主要做了命名实体识别，Query改写，计算词语权重；在离线数据中，也分析了日志信息，挖掘出用户行为包含的query意图信息。其平台搜索部分的架构图整体简单清晰，详细如下：



Query处理

■ 实体识别

用户的搜索常常能够归结于对某些特定的实体词的搜索。一般有四类搜索：地址类（如香港），品牌名（如苹果），产品名称（如手机），属性词语（如12 红）。当用户输入一个query后，首先对query做实体识别，明确每一个词语代表的真实实体含义，就可以提前对query做分类处理，得到目标类目，然后从目标类目中召回合适的结果。举个例子：用户搜索词语为“辣鸡面”，有一个商品为“AYAM BRAND 雄鸡标 辣椒金枪鱼”，商品类目为“熟食-其他熟食”，在使用单字召回时，是可以直接召回这个商品的，但是如果经过实体识别，知道用户的搜索词语中前面两个字是修饰最后一个字“面”，说明用户主要目的是为了搜索面相关的类目，那就可以在面相关的类目中召回结果，而不用召回这个商品或者将这个商品的位置排的相对靠后。

在前面的部分已经介绍过实体识别的一些方法，实体识别主要采用序列标注的方法来完成，现有的序列标注都会采用不同的特征抽取，在模型的后面接一个CRF用于条件限制并将标注结果输出。在该平台的识别中，实体识别在中文的表现优于英文属性词，同事训练数据使用该平台自身的商品数据，所以识别效果具有一定的针对性。

query改写

Query改写作为搜索引擎中较为重要的模块，一般包含有纠错，扩展，删除无效词，转换。

纠错可以根据Session前后的点击情况，将前输入的query纠正为后输入并点击的query，这个方法可以在离线完成。例如用户首先输入“好用插线版”，发现召回的效果不佳，用户更改为“好用插线板”后，曝光了满意的结果且用户点进去后停留时间较长，说明用户对后者的结果满意，经过数据离线挖掘出这样的序列对，将前面的query与最后一个query匹配度较高的作为纠错pair对。

扩展是指将现有的商品扩展成更多的一些描述。一般商品会有一个title和描述，在搜索引擎召回时会根据文本匹配关键字，或者使用query对应的类目去尽量召回合适的内容，如果有的商品具有多义性，就容易召回不到合适的商品。例如商品标题为“瑜伽瘦身衣”，用户搜索“健美衣”，这两个通常是互相关联的，在用户搜索时，将query扩展成“瑜伽健美衣”增加一路召回，就会召回更多样性的结果。

删除无效词是指query中有的词语是无效的。例如“的、空格、连续重复的无意义词”，这些词语一般都不具有明确含义。在常见的搜索引擎中，会把这类的词语成为非必留词，也就是这些词语可以扔掉后再召回。

Query转换是指对Query做常见的同义词切换，能够更好的召回合适的结果。因为每个人对于一个事物的形容是有偏差的，要使用同义词连接的方式把所有的同义词都映射为一个词语，再拿这个词语去召回，就能够在整个平台中统一标准。例如：中英文切换(Apple-苹果)、特殊叫法切换(古奇-Gucci)等。

物品类目

提前将Query映射到不同的类目中，并根据离线挖掘出的该query更加可能的类目分布，在召回排序的时候按照类目来做区分效果会更好。例如用户搜索“李宁”，有可能是搜索“鞋子”，也有可能是搜索“衣服”，或者是搜索“名人”，在离线挖掘中能够发现搜索这个query的用户更多的点击落在了鞋子上，少量落在了衣服上，由于没有名人相关的商品，所以这个query对应的类目是鞋子>衣服>名人。在召回与排序的时候，可以根据类目情况来对结果做出反馈。

同时由于类目体系是运营同学自建的，也可以弥补一些长尾分布的类目的不足，该平台采用了将相关的类目构建成虚拟类目的方式，这样召回结果会同时召回长尾类目内容。个人感觉这种召回方式也有局限性，因为长尾类目是搜索量小的类目，如果每次遇到虚拟类目时都会考虑长尾类目的加入，可能会影响大多数优质类目的结果，这个度上面需要仔细把握。

- **词权重**

在分词过程中，同时关注每一个词语的权重，词语的权重一般可以根据ngram考虑query的语法结构或者每一个term在分词中对query整体的贡献程度以及统计信息等来获得。在得到词语权重后，可以以重要程度比较高的内容来作为整个query的有效信息，同时也可以扔掉无意义信息。例如“桃子味的牙膏 女士”这里面比较重要的词语是“牙膏”，所以在用户搜索时，抓住关键信息并召回，同时根据额外信息用于排序展示，效果更佳。

微信搜索意图

微信的内部搜索也会有query的意图识别并分类，用户可以在微信首页中搜索聊天记录、朋友圈、公众号、小程序等，在搜一搜中还可以搜索公众号文章，新闻，问答，视频，微博等其他内容。那么这些内容都是怎样判断用户的确切搜索需求并且将合适的内容按照一定顺序展现给用户呢？这里面也有搜索意图识别的过程。

一般把搜索意图识别作为Query处理的一部分，Query处理中首先对Query做分词，词权重，改写，扩展等基础操作；然后根据query内容判断每一个意图的可能性程度，在微信中就会把不同类型的意图需求当做一个类目，在判断时也会引入其他信息，例如用户所在地，搜索历史和外部信息等内容；最后根据不同的类目意图去召回合适的内容，然后排序展现给用户。

一般处理流程

垂直搜索领域对于query类目意图识别较为重要，因为这决定了最终的返回结果与关联信息。常见的处理流程为使用规则来cover一定的query，剩下的query拿到query文本特征与用户行为特征作为模型输入，特征直接使用，具有时序性的点击信息可以用RNN结构来进一步提取特征，接着训练多个不同的离线模型，将多个离线模型合并使用ensemble的方式联合训练，得到最终意图识别分类模型。

参考文章

1:https://www.infoq.cn/article/user-search-intention-recognition “搜你所想”之用户搜索意图识别”

2:https://www.infoq.cn/article/V037TeLVfa-KhwL6WoIK “解读电商搜索——如何让你买得又快又好”

3:https://segmentfault.com/a/1190000014849907 “交互搜索中的自然语言理解技术”

4:https://yq.aliyun.com/articles/420506 “基于DNN+GBDT的Query类目预测融合模型”

（上下滑动查看）

今天的分享就到这里，谢谢大家。

在文末分享、点赞、在看，给个3连击呗~

文章推荐：

Angel：深度学习在腾讯广告推荐系统中的实践

社群推荐：

欢迎加入 **DataFunTalk 搜索算法** 交流群，跟同行零距离交流。**识别二维码**，添加小助手微信，**入群。**

关于我们：