网易数帆社区 Lv3
2018年11月09日 阅读 754

搜索意图识别浅析

对于搜索引擎来讲，很多情况下只需要用户在搜索输入框内输入所需要查询的内容就可以了，其余的事情就全部交给搜索引擎去处理。理想的情况下，搜索引擎会优先返回用户想要的结果。理想很丰满，但总会存在一些骨感的现实，用户通过搜索无法找到最想要的结果。如果应用中压根不存在用户搜索的内容，倒还可以理解。反之的话，就是一个大写的尴尬。本文主要谈论和解决的是令人尴尬的问题。

为什么会搜索不到

1、不同的用户对同一种诉求的表达往往是有差别的，往往会存在一种比较常见的现象，用户输入的query并不能清晰准确的表达需求。

2、搜索系统对用户query的理解能力较弱，无法挖掘出用户的真实需求。

3、召回结果集的排序不合理，可能用户需求的内容被排在后面而未曝光。

用户作为一个使用主体，其知识水平和表达能力会有差异，当不同用户想搜索同一个商品时所输入的query会存在差别，具体如下所示：

澳佳宝blackmores蔓越;痛经;圣洁;圣洁莓精华;澳洲佳宝;泌尿系统感染;
blackmores澳佳宝;澳大利亚blackmores;蔓越莓澳佳宝;女性保养;蔓越莓;绝经;
蔓月莓;blackmores圣洁莓;月蔓;裸版胶囊;;女人;澳洲;月经;蔓越莓胶囊澳洲;
胶囊;女士痛经;蔓越梅胶囊;澳佳美;妇科炎症;blackmore;卵巢囊;澳洲blackmores
女性;宫颈;女性美容;澳佳宝圣洁莓;妇科保养;内分泌;manyuemei;调理月经;美少女;
越曼莓;澳洲blackmores澳佳宝;女士宝;思瑞;澳洲的蔓越莓;blackmores蔓越莓;子宫;
女性妇科;子宫保养;圣莓洁;羊胎素胶囊;澳洲杜虫;蔓越莓精华胶囊;blackmore蔓越莓;
美国自然之宝蔓越莓;蔓越莓精华;女宫;保护子宫;奥佳宝;blacksmore;痛经少女;泌尿;
女人妇科;佳宝;blackmores圣洁;曼越梅;女生内分泌;圣洁酶;澳佳;蔓越莓胶囊;炎症;
蔓越酶;蔓越;妇科;blackmores圣洁莓精华;莓;月洁莓;蔓越梅;澳洲蔓越莓;澳洲奥佳宝;

可见，对于同一个商品往往会对应不同的query，相对精确的有“蔓越莓胶囊欧洲”、“blackmore蔓越莓”；品牌优先的有“blackMores”；功效优先的有“女士痛经”，"泌尿系统感染"；输入错误的有"蔓越梅"，输入别名的有"圣洁莓"；输入较模糊的有“妇科”，“炎症”。所以说用户的输入一般会存在表达差异，词汇差异，需求明确性差异等。

要想解决这些问题就需要通过用户输入的query来获取用户的真实需求，本文把对用户输入的理解称为QueryParser，包含：query切分（分词），query意图识别，query改写(query扩展/query纠错/query删除等)，接下来本文主要针对query意图识别和query改写结合在考拉海淘搜索中的具体应用来和大家聊聊。

1.query意图识别

本文主要针对垂直搜索进行介绍，不同的垂直引擎中的query会有自己的特点。像去哪儿网的日志中肯定有很多“城市a到城市b的机票”这种pattern的query，而电商网站中肯定大部分是“产品/品牌/

1.1意图识别的难点

1、输入不规范，前文中已有介绍，不同的用户对同一诉求的表达是存在差异性的。

2、多意图，查询词为："水"，是矿泉水，还是女生用的化妆水。

3、数据冷启动。当用户行为数据较少时，很难获取准确的意图。

4、没有固定的评价标准。pv,ipv,ctr,cvr这种可以量化的指标是对搜索系统总体的评价，具体到用户意图的预测上并没有标准的量化指标。

1.2 意图识别的方法

1.2.1 词表穷举法

这种方法最简单暴力，通过词表直接匹配的方式来获取查询意图，同时，也可以加入比较简单并且查询模式较为集中的类别。

- 查询词：德国[addr] 爱他美[brand] 奶粉[product] 三段[attr]

当然查询模式是可以做成无序的。这种意图识别的方式实现较为简单，能够较准确的解决高频词。由于query一般是满足20/80定律，20%的query占据搜索80%的流量。但是，80%得长尾query是无法通过这种方式来解决的，也就是说这种方式在识别意图的召回可能只占20%。同时，需要人工参与较多，很难自动化实现。

1.2.2 规则解析法

这种方法比较适用于查询非常符合规则的类别，通过规则解析的方式来获取查询的意图。比如：

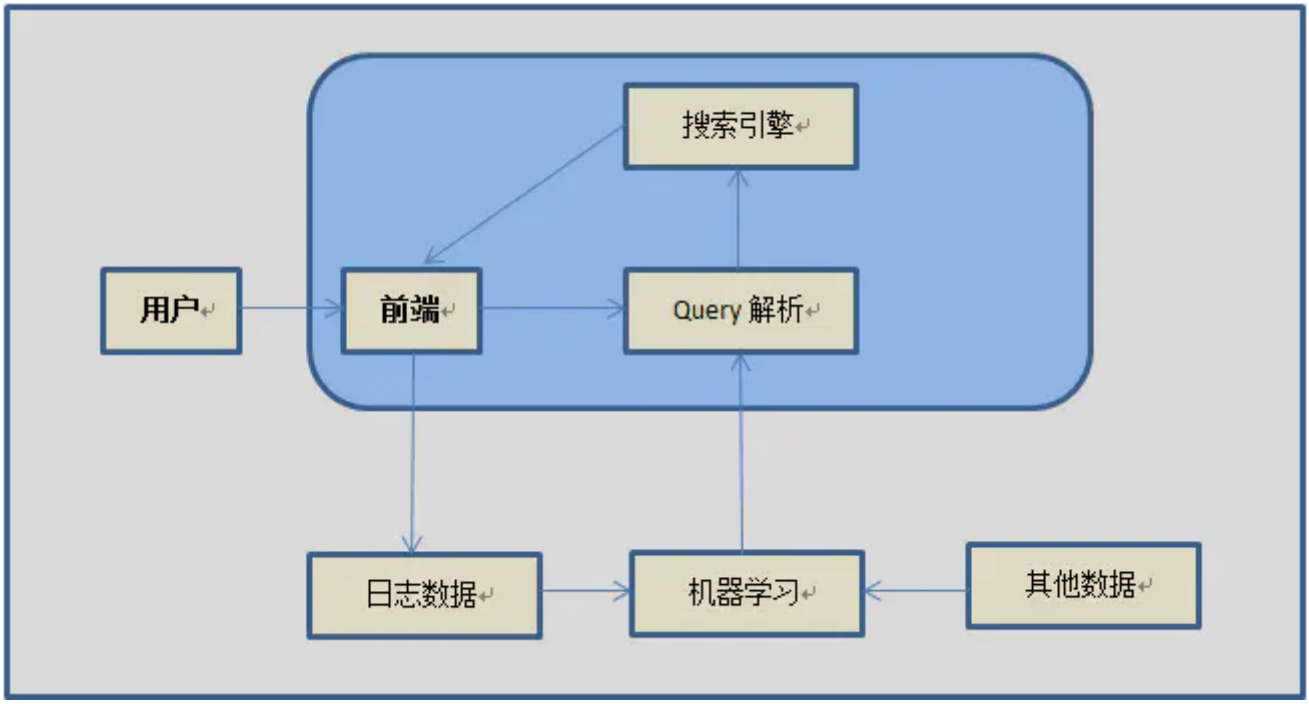
- 北京到上海今天的机票价格，可以转换为[地点]到[地点][日期][汽车票/机票/火车票]价格。
- 1吨等于多少公斤，可以转换为[数字][计量单位]等于[数字][计量单位]。

这种靠规则进行意图识别的方式对规则性较强的query有较好的识别精度，能够较好的提取准确信息。但是，在发现和制定规则的过程也需要较多的人工参与。

1.2.3 机器学习方法

意图识别其实可以看做是一个分类问题，针对于垂直产品的特点，定义不同的查询意图类别。可以统计出每种意图类别下面的常用词，对于考拉海淘而言，可以统计出类目词，产品词，品牌词，型号词，季节时间词，促销词等等。对于用户输入的query，根据统计分类模型计算出每一个意图的概率，最终给出查询的意图。 但是，机器学习的方法的实现较为复杂，主要是数据获取和更新较困数据的标注也需要较准确才能训练出较好地模型。

考拉海淘是一个电商类的产品，目前其搜索意图相对单一为产品购买。本文主要讨论考拉海淘中用到的query改写，类目相关，命名实体识别和Term Weight等内容。 考拉的搜索系统有大量的用户访问，我们希望通过对用户query的意图分析来提高搜索体验，目前，考拉系统的架构包含下图所示的几个部分：



2.1 实体词识别

通过对日志分析，将用户常用的搜索词分为以下四类：地址(澳洲)，品牌词(爱他美)，产品词(奶粉)，属性词(三段)。当用户输入query时，如果能准确的识别每个实体词，就能去索引里面精确匹配对应的字段，从而提高召回的准确率，在排序中也可以用到实体词进行优化。 举一个栗子：有一个商品的标题是"AYAM BRAND 雄鸡标 辣椒金枪鱼"，它的类目是"冷面/熟食/方便菜 其他熟食"。当用户搜"辣鸡面"的时候，通过单字逻辑召回这款商品。通过实体识别会得到这个商品的产品词是"金枪鱼"，而query要搜的产品词是"面"。这样就可以判断出其实这是一个误召回，进而可以将这个商品进行过滤或者是排序的时候放到较后的位置。

- 爱 B-brand 他 I-brand 美 I-brand 奶 B-product 粉 I-product 三 B-attr 段 I-attr

训练出的模型对于地址，品牌词，产品词的识别准确率平均95%左右，英文属性词的识别准确率还有待提高，crf模型还有一个比较好的地方是具有一定的泛化能力。另外，模型的训练是使用考拉平台上的商品数据，所以对非考拉平台的产品和品牌识别的准确率也不理想。但是，最重要的是识别本平台已有的实体，尽可能准确的向用户展示最准确的商品搜索结果。

2.2 query改写

query改写包括：query纠错，query扩展，query删除，query转换。本文主要讨论在考拉中常用的query扩展，query删除和query转换。

2.2.1 query扩展

搜索召回依赖索引数据，商品数据依赖于编辑运营的录入，数据的完整性很难得到保障，也就是说很难从各个角度来描述这个商品。

还是用例子说明，一个商品的标题是“Fisher-Price 费雪 碎花儿童学步鞋”，由于用户输入的差异性存在，会有用户搜索"婴儿鞋"，"宝宝鞋"。很明显这个学步鞋恰恰用户所需的商品，但是因为数据的不完整性而无法被召回。这就是前文提到的有商品却无法展示给用户，这是最不希望遇到的情况。这时候就需要用到query扩展，我们会维护一个同义词扩展表，当用户输入一个query的时候，会进行同义词扩展，从而尽可能召回所有与用户相关的商品。

query删除一般的应用场景是在当用户输入query过多时导致无法正常召回，可以通过丢词的方式来筛选用户的query，从而召回与query最相关的商品。

依旧用例子说明，当用户的query为"卡乐比水果麦片"时，由于这款商品可能被下架，或者商品种类较少，通过query删除，可以把原query改写为“水果麦片”，进而可以召回其他品牌的水果麦片。query删除是需要用到实体识别的，因为要决定query中的哪些数据被删除才能对用户原意图造成的影响最小。像"卡乐比水果麦片"，通过意图识别得到”卡乐比“是品牌，”水果麦片“是产品，显然用户更需要的是水果麦片，而不是“卡乐比”其他类型的麦片。

2.2.3 query转换

会存在这样一种情况，确实没有商品是满足用户的明确需求。 比如，用户搜索"祖马龙"，考拉海淘并没有这款商品。也无法通过query同义词扩展和query删除来对原query进行处理。通过session数据可以发现，用户搜索“祖马龙”后会伴随着“香水”这个query出现，利用用户行为数据是可以挖掘出“祖马龙”和"香水"这两个query是相关的。当用户搜索"祖马龙"而无法召回时，是可以把query转换为"香水"来尽可能满足用户的需求。

2.3 类目相关

当用户搜索“Adidas”的时候，是想要搜索“运动鞋”，还是“衣服”，又或者是“沐浴露”。当然，你可能说不同的用户有不同的需求，这就涉及到个性化搜索的内容了，暂时不在本文的讨论范围内。如果用户行为数据足够多，直接使用统计分析就可以找到query对应的类目相关程度。当然，统计算法也是机器学习的一种。但是，仍有一部分问题是需要机器学习算法来完成的。

时会通过类目打散将衣服和沐浴露适当的掺杂在运动鞋中。

query的类目相关性是通过用户行为数据进行挖掘的，一些长尾的类目虽然与query相关，由于马太效应却无法被挖掘。比如query“面膜”所挖掘出的相关性类目为“男士面膜”/“女士面膜”/“面膜粉”等，而“孕妇面膜”这个类目却一直处于不相关的状态。其实，“男士面膜”/"女士面膜"/"面膜粉"/“孕妇面膜”在"面膜"这个维度都是相关的，我们通过虚拟类目的做法来解决这种长尾问题。离线将这四个类目归一为一个虚拟类目，当用户的query落在虚拟类目中的大部分类目时，认为这个query与虚拟类目包含的其他类目也具有相关性。

2.4 Term Weight

中文自然语言处理的第一步就是分词，分词的结果中，每个词的重要性显然应该时候区别的。Term Weight就是为了给这些词不同的打分，根据分值就可以判断出核心词，进而可以应用到不同的场景。比如，有一个商品的标题为“碗装保温饭盒套装”，通过Term Weight可以得到核心词为“饭盒”。当用户搜"碗"召回这个商品的时候，是可以根据term weight来进行排序降权的。

通过以上几点可以看出，query意图识别在一个搜索系统中是必不可少的，可以说query意图识别的精确程度高低决定着一次搜索质量的优劣。