【调研】意图识别

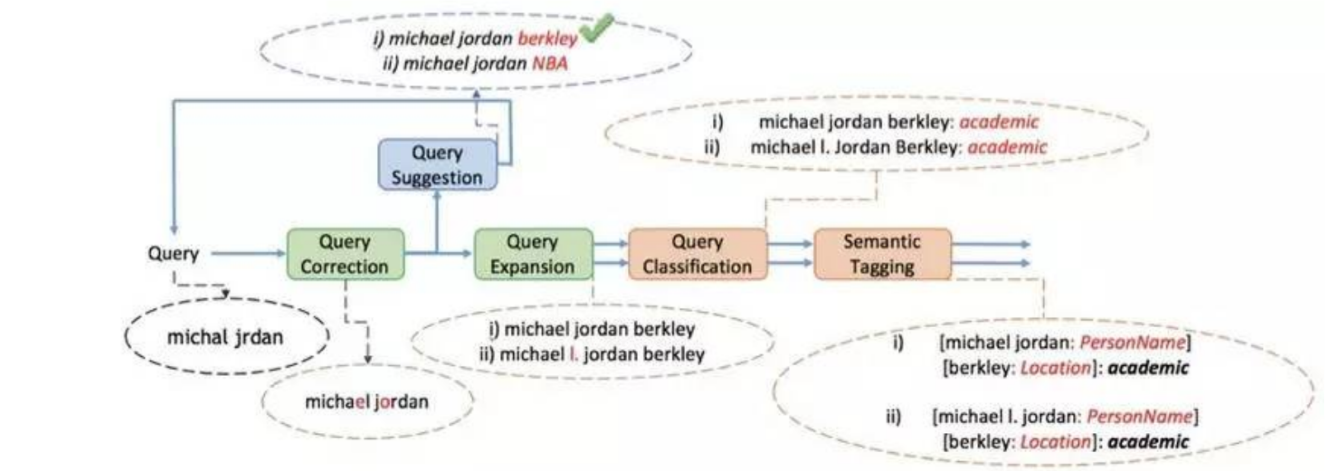
来自wiki迁移页面路径：深研BDG_DST / 智能检索 / 以文搜图 / 【调研】意图识别

现有应用

大多数用户在进行信息查询时，并不能十分准确地用查询语句表达自己的查询意图，识别用户意图才能最终满足用户需求。用户意图识别在通用/垂直搜索领域，chatbot 的NLU模块都有广泛的引用。

- 在chatbot方面，普遍的流程是1）对domain/intent 分类；2）对slot进行填充。

- 在搜索引擎方面，调研侧重query理解，进而改善搜索结果，涵盖 query 纠错、query拓展、query意图分类等，如下为一个输入“michal jrdan”的具体实例



难点分析

- 输入不规范：“Michael Jordan” 输成了“michal jrdan”
- 多意图：相同的查询意图，不同的用户输入可能完全不一样，如“仙剑奇侠传”，可能是指游戏、电影、电视剧、歌曲收听等多种可能。
- 时效性：同一个搜索关键词，在不同的时间搜索，意图不一致，比如小米8手机上市，相关的查询意图、购买意图的占比会随着时间占比发生变换。
- 数据冷启动的问题，用户行为数据较少时，很难准确获取用户的搜索意图。
- 没有固定的评估的标准，CTR、MAP、MRR、nDCG 这些可以量化的指标主要是针对搜索引擎的整体效果的，具体到用户意图的预测上并没有标准的指标。来获取用户的真实需求。

方案调研

目前调研的更多偏向于垂直搜索，以一特定类别为主题，只抓取与主题相关信息，根据主题特点有针对性的建立相应的索引检索方式，筛选方式，以及展现方式，如机票搜索，地图搜索，购物搜索。有如下几类方法：

词表穷举法

最简单直接的方法，通过词表的直接匹配来获取查询意图，也可加入适用于较为简单且查询较为集中的类别，比如电视台节目查询，节假日查询，餐馆查询等

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

查询模式：[brand]+[product]; [product]+[attr]; [brand]+[product]+[attr]

规则解析法

适用于一些查询不集中但非常符合规则的类别，通过规则解析查询来做意图识别和关键信息提取的，比如汇率查询，计算器，度量衡等。

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。

基于统计模型的意图分类

主要表现为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，如电商领域，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，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。

文本分类法

目前的多数研究中，都将意图识别归为分类问题，由此，多种文本分类模型可以用于实现意图识别。

1. 传统文本分类工作主要集中在三个方面：特征工程，特征选择和使用不同机器学习算法进行分类。对于特征工程，广泛使用是 bag of words 特征。 此外，还有一些更复杂的特征设计，如词性标签，名词短语和Query 的长度、Query 的频次、Title 的长度、Title 的频次、BM-25、Query 的首字、尾字等特征选择旨在删除嘈杂的功能和改善分类表现。 最常见的特征筛选方法是删除停用词（例如，“the”），也可以使用信息增益，互信息，或L1正则化选择有用的特征
2. 基于神经网络的文本分类在流程上更为简洁，不需要花费太多的时间在特征工程和特征选择上。常见的基于神经网络的文本分类方法有：1）TextCNN；2）Bi-GRU+FC；3）BTM-BiGRU，BTM为短文本主题模型；4）fastText；5）HAN(Hierarchical Attention Network) 等。

应用探讨

对于现阶段的以文搜图而言，实践性更强的为query理解方面，围绕**基于统计模型方面的意图分类**可能的细节有：

Term weight

常见的term重要程度多使用tf*idf 但term的重要程度并不是严格正比于term的tfidf。常见会根据 1）语料统计 2）点击日志 3）有监督学习进行优化。在语料统计方面的计算方法主要有：

imp

$$\begin{cases} Tmp_i = \dfrac{B_T}{\sum_{j=1}^{M_i} B_{T_{i,j}}} \\ B_T = \dfrac{N}{\sum_{i=1}^{N} \dfrac{1}{Tmp_i}} \end{cases}$$

其中BT为term的imp值，初始值可设为1，Tmp_i是query中的第i个term的重要性占比，N指所有包含第i个term的query数目。imp从在query中的重要性占比基础上，采用迭代的计算方式优化词的静态赋权

利用term weight ，根据分值给词进行打分，比如，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，可根据term weight来进行自定义方法排序降权。

query 理解

包括 query纠错，query扩展，query删除，query转换、query 建议。

query纠错

中文纠错以拼音为基础，编辑距离等其他方式为辅的策略。

1、基于编辑距离： 一般选择与候选词有相同的拼音的词，

2、基于噪声信道模型：

*p(x|w)*是正确的词编辑成为错误词x的转移概率，包括删除（deletion）、增加（insertion）、替换（substitution）和颠倒（transposition）四种转移矩阵,将转移矩阵计算公式代入公式的噪声信道模型公式中，根据不同候选词和纠错词之间的变换关系选择转移矩阵类型，就能得到概率最大的候选词。中文一般考虑的是替换即可。

$$\begin{aligned}\hat{w} &= \underset{w \in V}{\operatorname{argmax}} P(w \mid x) \\ &= \underset{w \in V}{\operatorname{argmax}} \frac{P(x \mid w) P(w)}{P(x)} \\ &= \underset{w \in V}{\operatorname{argmax}} P(x \mid w) P(w)\end{aligned}$$

公式 5：噪声信道模型纠错公式

3、n元语法模型：取*wi*之前n-1个历史，根据马尔科夫假设，一个词只和他前面n-1个词相关性最高，这就是n元语法模型。

$$P(w_1 w_2 \ldots w_n) \approx \prod_i P(w_i \mid w_{i-k} \ldots w_{i-1})$$

注意，对于Real-Word Error问题（每个词都是正确的，但是组合起来意思不对）首先注重的是候选词集合的生成，（包括编辑距离n内的词，同音词，词本身）。

query扩展

可通过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。词典生成方法如下：

1、人工构建的同(近)义词词典

2、基于查询日志挖掘出的查询等价类(Web上很普遍)

3、机器方法找近义词，多基于词语的共现统计信息

query删除

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，query删除是需要用到实体识别的

query转换

会存在这样一种情况，query 查无结果，也无法通过query同义词扩展和query删除来对原query进行处理。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。

query 建议

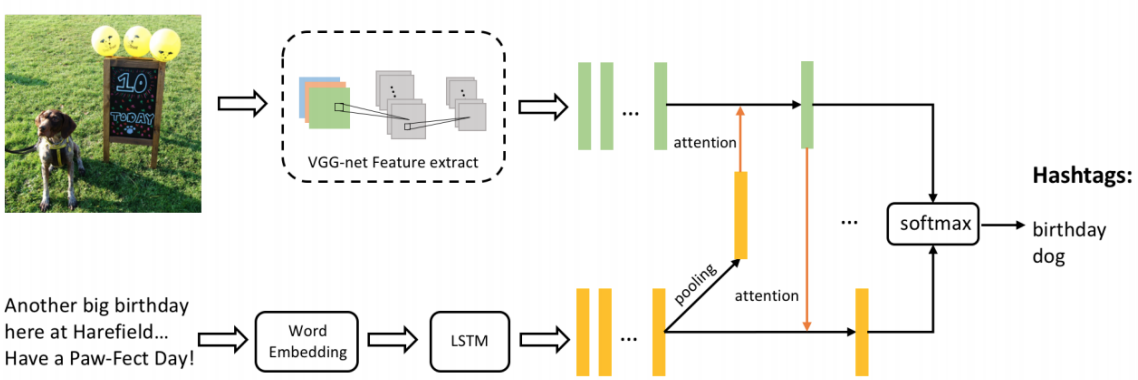
在用户键入时，在输入下拉框中进行提示，一般通过Trie树，返回可能的搜索项，并结合用户的query log，将热词的排序提前展示。

多模信息挖掘

以上是对传统以文搜文的意图识别调研，在以文搜图智能检索领域，image 往往只有一部分信息和 text 文本是相关的

而在社交平台上，更是充斥着纯图像微博或者短文本微博，目前以文搜图在两个模态上的处理是独立的

- Hashtag Recommendation for Multimodal Microblog Using Co-Attention Network 利用co-attention机制给图片和文本打一个tag



- Mention Recommendation for Multimodal Microblog 识别图像和文本相互关联的部分，见文档

Mention Recommendation for Multimodal Microblog调研文档V1.0.pdf
7MB

在图文pair的信息理解上，挖掘图片的出关键信息，可以提高意图识别的准确度。

思路

1）query层，可通过query的多个细节处理，识别用户隐藏意图，通过扩大召回的结果数量，达到提升用户意图理解的目的。

2）在数据层面，通过挖掘融合图文信息，使得图片和文本之间的关键联系更加明确，将匹配结果做更细致的精排，返回真正相关的图片，达到提升用户意图理解的目的。

参考资料

“搜你所想”之用户搜索意图识别

杨艺, 周元. 基于用户查询意图识别的Web搜索优化模型[J]. 计算机科学, 2012, 39(1):264-267.

伍大勇, 赵世奇, 刘挺, et al. 融合多类特征的Web查询意图识别[J]. 模式识别与人工智能, 2012, 25(3).

ACL 2018l 基于胶囊网络实现零样本意图识别

Query词权重方法（1） - 基于语料统计

爱奇艺视频场景下NLP应用与文本舆情分析（1）

5G 时代下：多模态理解做不到位注定要掉队