Query意图方法（2）- 基于文本分类

原创 XG数据 WePlayData 2019-10-22

在Query意图方法（1）- 基于片段意图一文中介绍了基于片段的意图计算。本文进一步介绍一种采用文本分类方法对query进行意图分类，目前是将query分类到预定义意图系统中的一个或多个类中。

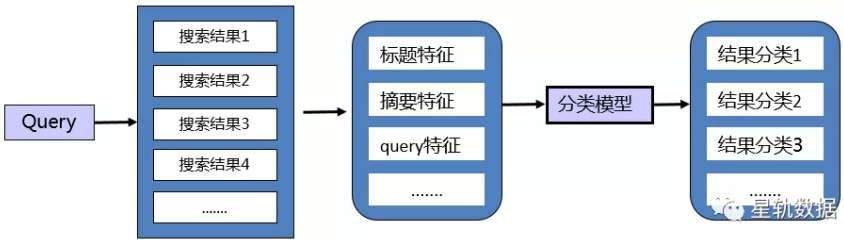
相比于文章，80%的query通常都是短文本，同时意图的类别比较多都对准确的query意图分类提出了挑战：

- 缺少标注数据集，同时人工标注样本成本较高，需要尽可能自动化构造样本并减少人的参与；
- 短文本造成特征比较稀疏，需要引入外部资源来补充和丰富query的特征表示；
- query存在强实体词误导分类准确性，比如包含“宝宝”的query大概率被分成母婴类，即使是非母婴意图的query“亲亲我的宝贝”、“天线宝宝”等；

介绍具体优化思路前，先给出一些较难的query及误识别类别：

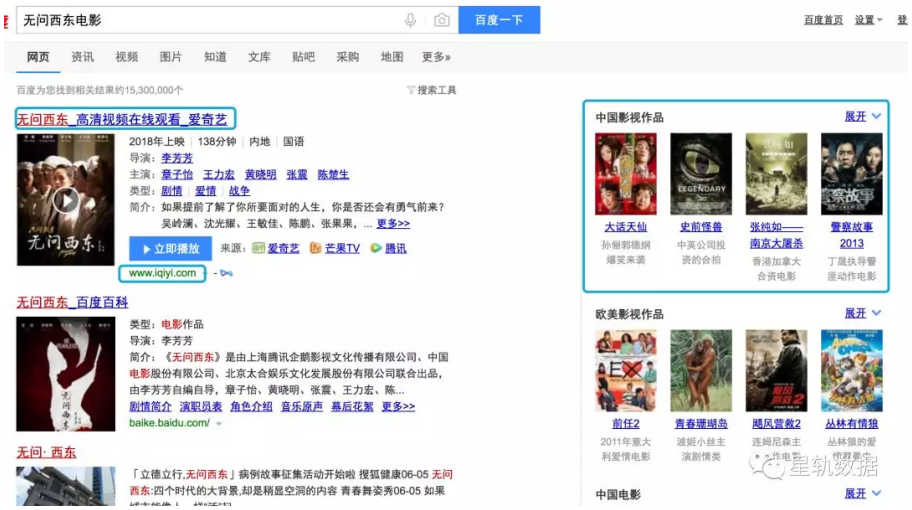
- 面朝大海、春暖花开 - 旅游
- 医生表情包 - 医疗
- 怎么恢复聊天记录 - 健康
- 华侨城集团招聘 - 旅游
- 烧烤店爆炸 - 美食

下图给出了query意图分类的一般流程，相比于模型的选择，意图分类的瓶颈更在于如何获取标注样本和进行特征扩展。因此本文重点从样本获取、特征扩展这两点介绍一些优化意图分类的方法。



一、样本获取

a. **百度元搜**：抓取query在百度的搜索结果，根据结果项的内容（url、站点名）可对query意图自动标注，这种方法不依赖于私有资源（如点击日志），同时百度展示结果已经融合很多先验信息，对有歧义的短实体query有自动消歧的作用。



如上图，从query“无问西东电影”结果中的标题和url可看出其意图是影视，同时位于右侧的知识图谱展示（“大话天仙”、“前任2”）也可以看成影视意图的正例。

b. **点击日志**：搜索点击日志记录用户搜索的展示和点击结果，其点击行为相比于百度元搜中的展示行为更能反映query的意图，因此可类似的分析点击doc的url和标题来生成一些有标注的query意图样本。这种方法需要能获取到用户点击日志，在冷启动阶段采用百度元搜的方式是个更好的选择。

c. **数据增强**：数据增强在深度学习模型中经常被采用的方法，有利于提高样本的多样性、不仅仅是重采样达到正负例的平衡、针对性构造一些难区分的样本。本文介绍一种基于类别特征词数据增强方法，首先挖掘一批类别特征词（比如医疗中的“医生、增生、疾病、预防”，然后找出包含特征词的一些query同时基于元搜结果获取query的意图结果，此时可有偏的选取一批非医疗类意图的query来增强医疗意图的负例，比如“医生表情包”是表情包/图片意图、“城野医生”是美容意图，“吉格斯医生”是影视意图，这样更能丰富样本的多样性和特征词的上下文模式，缓解强实体词对意图分类的影响。实际增强中，可针对性的选取一些易混淆的冲突类别，比如“游戏”和“旅游”，“游戏”和“军事”。

二、特征扩展

a. **短文本**：query大部分都是短文本，尤其是头部query，其信息量少并且存在较多的歧义性，使得很难只从文本本身信息推断其主要意图。通常做法是对query去扩充文本，扩充文本的方式有很多种，常用的做法是基于元搜的结果或点击日志作为query的文本扩展，此外还有查询知识图谱的相关节点来扩展，比如搜“城野医生”可获取相关节点“森田药妆”、“资生堂”等节点。最后基于扩充的文本进行特征抽取来表示query。

b. **内容型实体**：query中包含大量的内容实体，比如“爱情公寓资源”中的“公寓”，这种内容实体对意图识别无意义，并且会干扰模型的学习。常用的做法是采用知识嵌入的方法，首先根据知识图谱来识别“爱情公寓”，将“爱情公寓”替换成_ENTITY_特殊标记，然后再进行模型学习。

相关阅读

1. Query理解 - 搜索引擎“更懂你”
2. 搜索引擎新的战场 - 百度、头条、微信
3. 当我们关注舆情系统时，我们忽略了什么？