搜索系统算法解析-OAA

Term改写

改写是QP中的重要组成部分，在搜索系统中，用户输入的查询词（Query）和文档之间可能存在用词不一致的情况，或者同一意思可以通过不同的表达方式传达。通过同义改写，检索系统可以更好地理解查询和文档的语义匹配，从而提高检索的相关性和结果召回量。

工业界常见的改写方式主要有两种，基于分词的Term粒度匹配改写和全词Query粒度改写。Term粒度改写是指将Query分词为多个Term词序列，通过Term同义词替换实现语义泛化，如：宝宝辅食 --> [宝宝, 辅食] --> [婴幼儿, 辅食]。Query粒度改写指直接对Query做语义泛化，如：瘦身饮食计划 --> 减肥食谱。

本章节将详细介绍Term改写，对其下游应用、技术方案两块进行展开。

布尔检索应用

Term改写主要应用在搜索召回检索模块，如基于倒排索引的Term文本匹配召回通路。

假设有`Query`可以被分词为`[A, B, C]`，基于布尔检索的召回会根据分词构建查询串表达式：`A and B and C`，即检索结果`Doc`的文本内容需要同时包含Term词`A`、`B`、`C`，若`A`有同义Term词`A1`、`A2`、`A3`，`B`有同义Term词`B1`，则查询串表达式可以构建为：`(A or A1 or A2) and (B or B1) and C`。

当以结合同义词的布尔查询串进行检索时，由于原Term词和同义Term词是`OR`的关系，即检索命中同义词的笔记也会被召回，由此可以较大的提高相关结果召回量。此外，改写词可以作为特征参与相关性排序模块中。

数据挖掘

Term改写的数据来源主要有两类：同义词典和PT表（Phrase Table），其中同义词典中的数据格式一般为有向二元组`<w1, w2>`，表示`w2`是`w1`的同义词；而PT表存储的是两个词和两者的关联信息，并不显式表示两者是同义关系，格式为`<w1, w2, feat>`，其中`feat`表示`w1`到`w2`的转移特征（对齐特征）。

同义词典

同义词典的语料一般来自于现有的辞海、百科、知识图谱等知识库数据，比如很容易可以从辞海中挖掘出：<中国, China>。此外也可以通过人工标注、模型判别对隐式匹配（向量召回）语料进行筛选和过滤后写入词典。

PT表

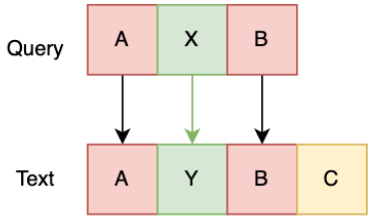
数据来源

PT表的数据来源于用户搜索行为数据，通过历史搜索行为构建句对，具体如下：

- Query-Title：通过用户点击Doc的Title和搜索词Query构建平行句对
- Query-Doc-Query：若不同Query历史点击同一篇Doc，那么这些共现Query之间构建平行句对
- Session：短时间内，同一用户多次输入不同Query进行检索，那么这些Query之间构建平行句对

词对齐

根据数据源挖掘出平行句对后，则需要通过词对齐的方式构建词表，对齐方式如下：



假设有`Query=AXB`，`Text=AYBC`，以共现词`A`和`B`作为锚点，则`X`和`Y`构成1-gram词对齐，`AX`和`AY`构成2-gram词对齐。另外可以采用文本翻译领域中的词对齐工具实现词对齐（如FastAlign、GIZA++）。

对齐特征

根据平行语料挖掘出大量词对齐数据后，即可统计出对齐特征，常见的特征如下：

特征
Term长度
Term是否包含数字/英文
Term左右熵
Term在Query中出现的频次
Term在对齐语料Text中出现的频次
Term Pair在Query-Text中出现的频次
Term Pair出现的比率
改写Term在当前原Term的所有改写Term中按出现频次排名的位置
Term Pair的相似度

有了词和对齐特征即可构建PT表作为线上Term改写的候选召回。

模型预测

当离线挖掘构建出PT表后，对于输入的查询词Query将通过PT表获取改写Term和相关特征后，则可通过树模型（GBDT）预测判别Query中的Term是否可以替换成召回的改写Term，流程如下：

1. 对查询词Query进行分词获得Term序列
2. 对Term序列构建N-Gram短语（1-3 gram）
3. 对N-Gram短语通过PT表查询得到改写候选短语和对齐特征
4. 对召回的PT短语进行对应Term替换，组成改写Query
5. 针对每个改写Query基于Term替换点构建特征
    1. 全局特征：Query/改写Query语言模型得分，Query/改写Query统计特征（检索量）等
    2. Term词粒度特征：PT表对齐特征
    3. 上下文特征：替换点前后组成的N-Gram对齐特征
6. 通过GBDT模型打分，判别当前Term改写是否可行（GBDT模型可通过人工标注数据集训练得到，学习目标可以是样本关系，如同义/近义/无关）

总结

综上，Term改写可以抽象为以下流程，并可分为离线数据挖掘和在线模型预测两个部分，一个好的改写系统可以有效降低长尾低频Query的零少结果率和Query换词率等指标。