搜索扩招回之query 改写

张云
算法工程师

前言： 翻出一篇压箱底的旧文，2018年做搜索rewrite的一些方法心得，文章在2018年做搜索业务时写的，如有不对，欢迎指出，一起学习。

背景

召回和排序是搜索的两个重要模块。召回的功能就是根据用户的query，尽可能的找出更多query相关的结果。query改写(query rewrite)是query扩召回的重要组成部分。通过对原始query进行改写，扩展出多个相关query，作为原始query的补充，与原始query一起参与搜索，从而返回更加丰富和准确的搜索结果

query rewrite 的方法 query rewrite方法很多， 关键在于理解业务场景，构造合理的特征。

- 1 基于日志挖掘的方法 基于用户的行为数据，挖掘query和query之间的关系。 query点击日志和query session日志是比较容易想到的数据。对于点击同一文档的query建立query共现关系，以及基于同一个session下的query建立共现关系。 点击日志和session日志的数据各有优缺点，session 的数据不过度依赖搜索结果。点击的数据普适性更强。可以结合使用
- 基于session的优点是不依赖现有的搜索结果。如果用户搜q1搜不到结果，可以继续搜q2，发现了想要的。如此，q1和q2就建立了联系。以后用户搜q1，就可以把q2的结果返回回来。和session数据相比，点击的数据 缺点就是依赖现有的搜索返回的结果。如果搜索q1在现有搜索引擎下召不回任何结果，q1就没法和其他任何的query建立联系
- 点击的数据是多个不同用户搜索不同query点击的同一文档的数据。把不同用户之间的搜素建立关系。往往泛化能力更强。session数据建立的是个体内部的query关系。存在知识偏差。 举个例子：之前的一部电影叫做 "西虹市首富" ,大部分用户以为电影名为"西红柿首富"，很少有用户会先搜"西红柿首富"，再搜"西虹市首富"。session中 共线关系很强的都是 西红首富 西红柿 首富 等词，基于session很难挖掘到 "西红柿首富"扩展到"西虹市首富"。 而点击就会弱化这个关系，用户X搜索 "西虹市首富"点击了西虹市首富电影相关的doc，用户Y搜索"西红柿首富"，通过翻了很多页也点击了这个doc,基于点击的数据就能更容易的挖掘到 "西红柿首富"扩展到"西虹市首富" 基于挖掘的效果如下：

```
query    rewrite
小芳妹娘    小爱妹娘
```

中国邮政快递单号查询　　中国邮政
阿拉蕾表情　阿拉蕾的表情包
毛线拖鞋编织　　毛线编织
清马　　清远马拉松
圣诞节招募　圣诞节活动招募
关牧村歌曲　关牧村歌曲大全
tfbays　tfboys
中国水利网　中国水利
美女情趣　　美女小视频
卫生　　卫生巾

基于挖掘的算法计算简单，能够快速更新数据，数据量足够大的情况下，效果也基本够用。缺点是需要大量的日志，而且简单粗暴，不能挖掘更深层次的关系，对于低频query效果不是很好。

- 2 基于知识体系替换，通过query中词的同义词、上下位词替换改写query.

苹果6手机多少钱 -> iphone6多少钱
新鲜水果->新鲜苹果

上下位词关系一般牵涉到知识图谱挖掘，有兴趣可以去网上看看知识图谱的相关知识，这里不做过多介绍。 同义词挖掘方法很多，有语料对齐挖掘，上下文挖掘等 语料对齐法：可以拿点击日志、session日志、anchor语料 通过上述1.1方法挖掘到对齐语料。再使用机器翻译模型(比如 IBM Model1)等从对齐语料中挖掘同义词。 语料对齐+机器翻译：原理就是把原语言翻译成目标语言概率最大化。这里面用到了单词对齐。参考（宗成庆:《自然语言理解》讲义 11章）

上下文挖掘：简单来说就是同义词往往有着相似的上下文。通过计算词的上下文相似程度来挖掘同义词

折抵换购 iPhone XR 仅 RMB 176/月起
折抵换购 苹果 XR 仅 RMB 176/月起

挖掘的同义词效果如下：

面包机　厨师机
尿布湿　尿片
醒酒　　解酒
工行　　icbc
握力圈　握力器

```
ems 邮政速递
文胸    bra
文胸  内衣
文胸 胸罩
```

- 3 基于深度学习的方法 基于深度学习方法做query rewrite的方法很多。这里讲解一些笔者用到方法
- 3.1 基于上下文的query2vec

和word2vec doc2vec原理一样，利用语言模型的原理，上下文相关的query具有相似性。将共现的query看做句子。比如将用一个session的所有query 或者点击同一个doc的所有query 看做句子。把query看做token,训练 query向量。详细方法可参照相关论文

- 3.2 deepwalk deepwalk和3.1类似，区别在构造sentence上下文采用随机游走的方法。 随机游走的原理：将query之间的关系建立成图。通过从一个点随机游走，建立起多条条路径，每条路径上的query组成一个句子。再使用上下文相关原理(3.1)训练query的embedding. 随机游走的优点就是关系具有传递性，和query共现不同,可以将间接关系的query建立联系。少量的数据经过游走能够产生够多的训练数据。例如session1:q1->q2 session2:q2->q3. 共线的方法无法直接建立q1->q3的关系。而随机游走能够很好的解决。
- 3.3 bert 笔者将数据构造成 q1 + SEP + q2 label的方式。训练样本的构造可按参照上面挖掘的方法，把挖到的数据做正样本，随机的query 对为负样本。 将q1 ， q2作为输入，得到sentence embedding，再根据label来计算loss，这样就可以学习到 q1和q2的相似性。

```
苹果 iphone 1
苹果 电脑 0
```

- 3.4 翻译的思想 query 作为源语言，rewrite作为目标语言，训练翻译模型，将query通过推理翻译达到rewrite目的

编辑于 2020-09-03

「真诚赞赏，手留余香」

赞赏