

一、旧特征存在的问题

1.1 TitleRank:

定义：query中词被全部命中为满分（按照个数来）。考虑命中位置（定格）

缺点：

- 1.没有考虑命中的词在title中比例（改进为命中的词占title词权重的比例）
- 2.query中用词的个数占比不合理（改进为：用权重）
- 3.没有从标题中抽取主题句（深度学习模型也如此）
- 4.原有的tiltlerank扩展性不好，通过线性加权的方式随意添加规则。且容易受到cache影响

1.2 新特征举例

```
1          意大利面条  怎么  做
2          60          20   10
3 80 意大利面条   100          10  10
4 10 的              10
5 10 做法              50
6
7 原来的方法：query中被命中的个数/query的总个数=1/3
8 (1/3+0.6+0.6+...+0.3) *title中重要度占比
9
10 现在的方法：query命中的权重/query的词权重和=60/100
11
12 '-----1.新方法理论'
13 1.query和title的交集，A集合
14 A集合元素占query的比重 * A集合元素占title中的比重
15
16 '-----1.1.新方法例子'
17 相似度分数=(60*1+20*0.1+10*0.5/60+20+10) *
18 (80*1+10*0.1+10*0.1/80+10+10)
19
20 '-----2.位置编码的引用'
21 好处：
22 (1) transformer, 平安, QQ浏览器
23 (2) 线性函数。[2,0] y=-kx+b
24 第一个词是最重要的，最后是一个词不重要的
25 [0,20]由2递减到1
26
27 '-----3.title分为主题域和非主题域-'
```

28 title域：意大利面条怎么做 -美食-小红书
29 旧的输入：意大利面条怎么做 -美食-小红书
30 新的方法： $0.6 * (q, \text{意大利面条怎么做}) + 0.3 * (\text{美食-小红书})$
31
32
33 '-----4.-主题域和非主题域的抽取--'
34 步骤：
35 1.特征：title长度；content—title（文章的标题）长度
36 （1）站点位置和内容标题域的长度一致
37 （2）站点位置大于内容标题域的长度
38 2.是否为站点：美团—吃喝玩乐-看电影，content_title长度为0，通过-美团
39 3.其他情况，通过长度来获取标题
40
41
42 '-----5.主题域中间重复主题的处理'
43 <title>意大利面的做法大全_意大利面怎么做好吃有营养_家常做法_下厨房</title>
44 如果出现2个断句有交集，取 $\max[(q, t1), (q, t2)]$
45
46 '-----6.实验结果-----'
47 评价指标介绍：
48 无线感知增益：单天点击top3，前3个，有1个我的实验，另外2个原来，点了我的+1
49 》2W,提给产品，人工标注（vs 百度）
50 实验结果：线上5%流量，无线感知增益连续3天8万，
51 类别其他：深度学习模型5W-12W（3个）；5
52 改进；同义词改进1W，3W（20来个人）
53
54 formart1 模型:web_diff.dat && wap_diff.dat t-test概率:0.864123 新旧ERR@3胜
 出率:0.505228 新旧DCG@1胜出率:0.500843 新旧DCG@3胜出率:0.502719 ERR@3胜出落败
 比:1.098814 DCG@1胜出落败比:1.030675 DCG@3胜出落败比:1.024779 折合千词ERR提
 升:0.290392 折合千词DCG@1提升:0.750000 折合千词DCG@3提升:3.544444
55 formart2 模型:web_diff.dat && wap_diff.dat
56 Total:3130 Top1标注覆盖率:0.947284 Top3标注覆盖率:0.763898
57 dcg1 胜出: 168 落败: 163 持平:2634 胜出率0.500843 胜出落败比:1.030675 千词分数提
 升0.750000
58 dcg2 胜出: 405 落败: 374 持平:1948 胜出率0.505684 胜出落败比:1.082888 千词分数提
 升3.604167
59 dcg3 胜出: 579 落败: 565 持平:1247 胜出率0.502719 胜出落败比:1.024779 千词分数提
 升3.544444
60 err1 胜出: 168 落败: 163 持平:2634 胜出率0.500843 胜出落败比:1.030675 千词分数提
 升0.118975
61 err2 胜出: 403 落败: 380 持平:1944 胜出率0.504217 胜出落败比:1.060526 千词分数提
 升0.324136
62 err3 胜出: 278 落败: 253 持平:1860 胜出率0.505228 胜出落败比:1.098814 千词分数提
 升0.290392
63 DCG@1:0.5008, DCG@3:0.5027, ERR@1:0.5008, ERR@3:0.5052
64 影响面:26.08%
65 平均指标增长:0.24%

1.3 .未来工作

语义计算：

①分析2个句子的句法树，利用tree-kernel来判断语法相似程度。

③同义词级别

主题抽取：

②针对新闻文本，抽出query实体事件和title实体事件

④query分为主题域和非主题域

⑤非主题域划分为标签域和站点域，也是2:1的比例

1.4.和其他模型比较：

问：NASM模型是啥?(2017)

答：

数据集：利用点击数据做标签，认为这2句是相似的，分2类，相关性强的为1类，不相关的为另一类

query与title每个词做Attention，加权求和，可以理解为找到和query类似的词，即命中的词，

缺点：

1.LTR标签集合，人为2个句子相似会给高分，如果title中含有query的词个数不同，分数是不一样的，（由早期的titlerank决定的）。但是在该模型中是一样的

2.query的词自身没有权重

3.title中是将站点和其他一起输入的。

优点：

用于Local query的召回

最新NASM模型：

title词向量添加docid和keyword，相当于明确站点信息。

缺点：词向量只有对海量的网站有用，对于频率较低的词作用不大，应当用规则。