一、旧特征存在的问题

1.1TitleRank:

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

Python L1 (default) ···

```
            意大利面条   怎么   做
            60          20    10
80 意大利面条    100         10   10
10 的                      10
10 做法                         50

原来的方法：query中被命中的个数/query的总个数=1/3
（1/3+0.6+0.6+...+0.3）*title中重要度占比

现在的方法：query命中的权重/query的词权重和=60/100

'-----1.新方法理论'
1.query和title的交集，A集合
A集合元素占query的比重 * A集合元素占title中的比重

'-----1.1.新方法例子'
相似度分数=(60*1+20*0.1+10*0.5/60+20+10) *
(80*1+10*0.1+10*0.1/80+10+10)

```

```
'-----2.位置编码的引用'
好处：
（1）transformer，平安，QQ浏览器
（2）线性函数。[2,0] y=-kx+b
第一个词是最重要的，最后是一个词不重要的
[0,20]由2递减到1

'----3.tile分为主题域和非主题域-'
title域：意大利面条怎么做 —美食-小红书
旧的输入：意大利面条怎么做 —美食-小红书
新的方法：0.6 *(q,意大利面条怎么做） +0.3*（美食-小红书）


'----4.-主题域和非主题域的抽取--'
步骤：
1.特征：title长度；content——title（文章的标题）长度
（1）站点位置和内容标题域的长度一致
（2）站点位置大于内容标题域的长度
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
3.其他情况，通过长度来获取标题


'-----5.主题域中间重复主题的处理'
<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>
如果出现2个断句有交集，取max[（q,t1），（q,t2）]

'------6.实验结果------'
评价指标介绍：
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
            》2W,提给产品，人工标注（vs 百度）
实验结果：线上5%流量，无线感知增益连续3天8万，
类别其他：深度学习模型5W-12W（3个）；5
改进;同义词改进1W，3W（20来个人）
```

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。