拼写纠错和Query类目预测

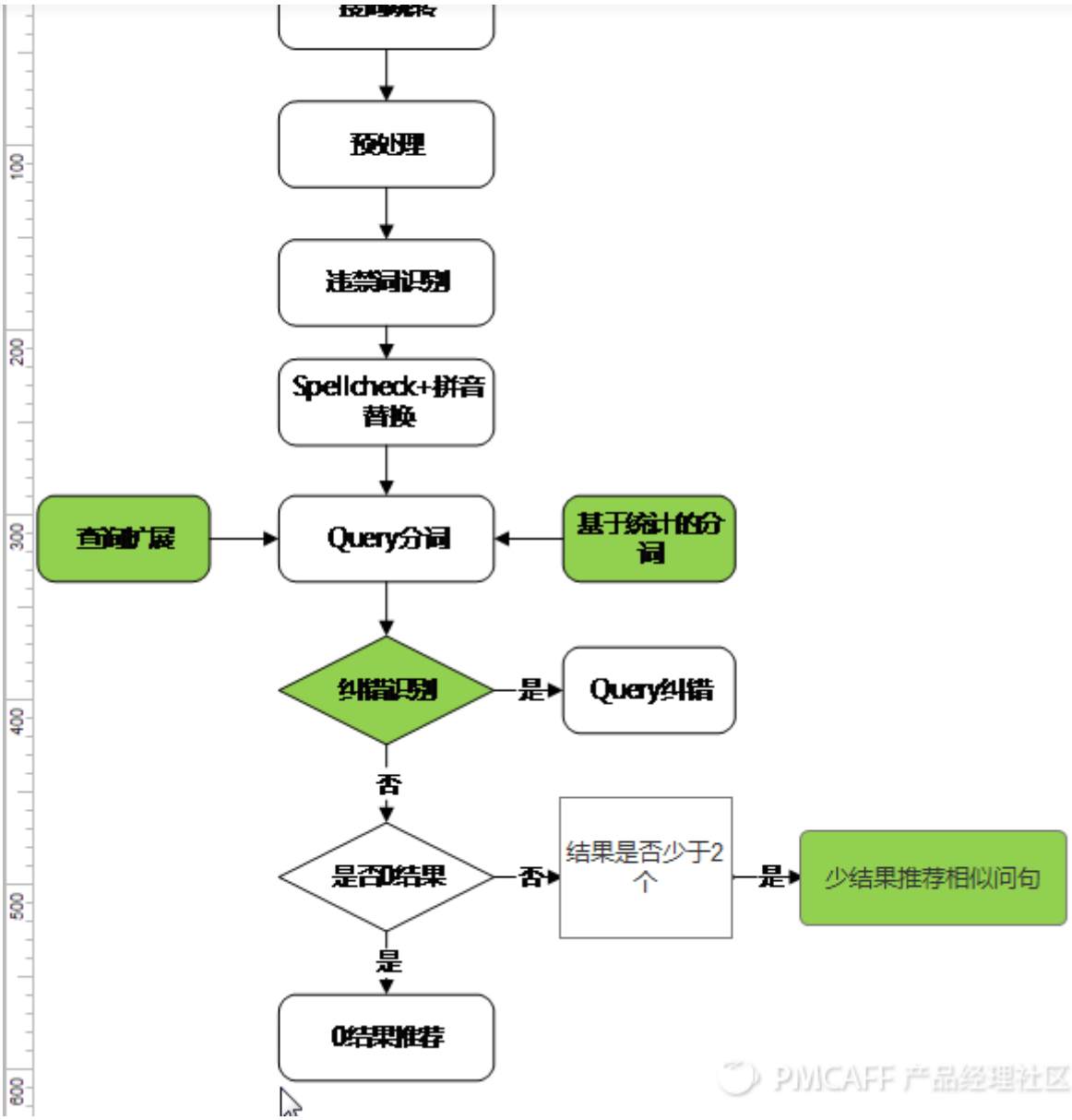
鲍佳 2601阅读 2017-10-18

下面我分享下在搜索项目呆了几个周期后的总结（指问题总结），有不足的请多指教，谢谢！

主要分为以下2个模块

1. 拼写纠错
2. Query类目预测

顺便简单画了个搜索过程



Contents-----拼写纠错

方法:

1. 错词表
2. 距离方法
3. 全拼音查找
4. 首字母查找
5. 编辑距离
6. 概率方法
7. 基于字的语言模型 + 易错字词典
8. 基于词的语言模型 + 拼音分词 + 易错词词典

距离方法:

$$Edit(i,j)=\begin{cases} i & \text{if } j=0 \\ j & \text{if } i=0 \\ \min\begin{cases} Edit(i-1,j)+1, \\ Edit(i,j-1)+1, \\ Edit(i-1,j-1)+[A[i]\neq B[j]] \end{cases} & \text{otherwise} \end{cases}$$

Jaccard系数

$$J(A,B)=\frac{|A\cap B|}{|A\cup B|}.$$

其他距离函数……

编辑距离+k-grams举例

ipone与iphone编辑距离为1

k-grams分词:

ipone:

grams1: $ i p o n e $

grams2: $i ip po on ne e$

grams3: $ip ipo pon one ne$

iphone:

grams1: $ i p h o n e $

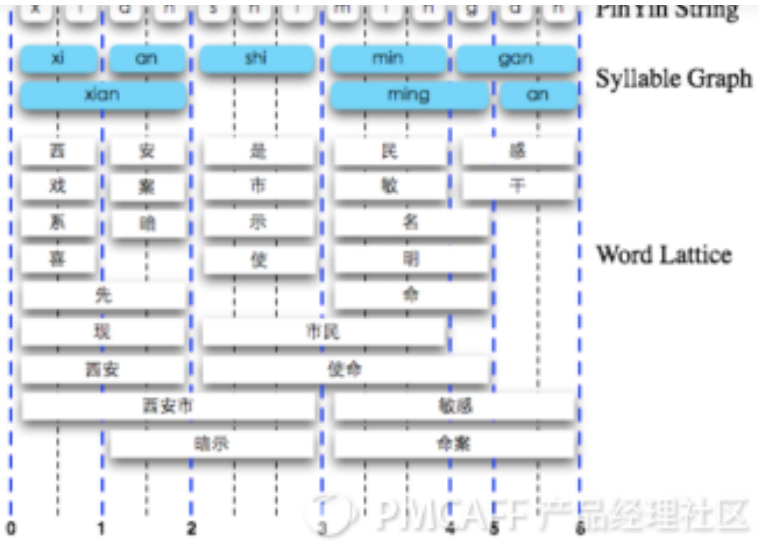
grams2: $i ip ph ho on ne e$

grams3: $ip iph pho hon one ne$

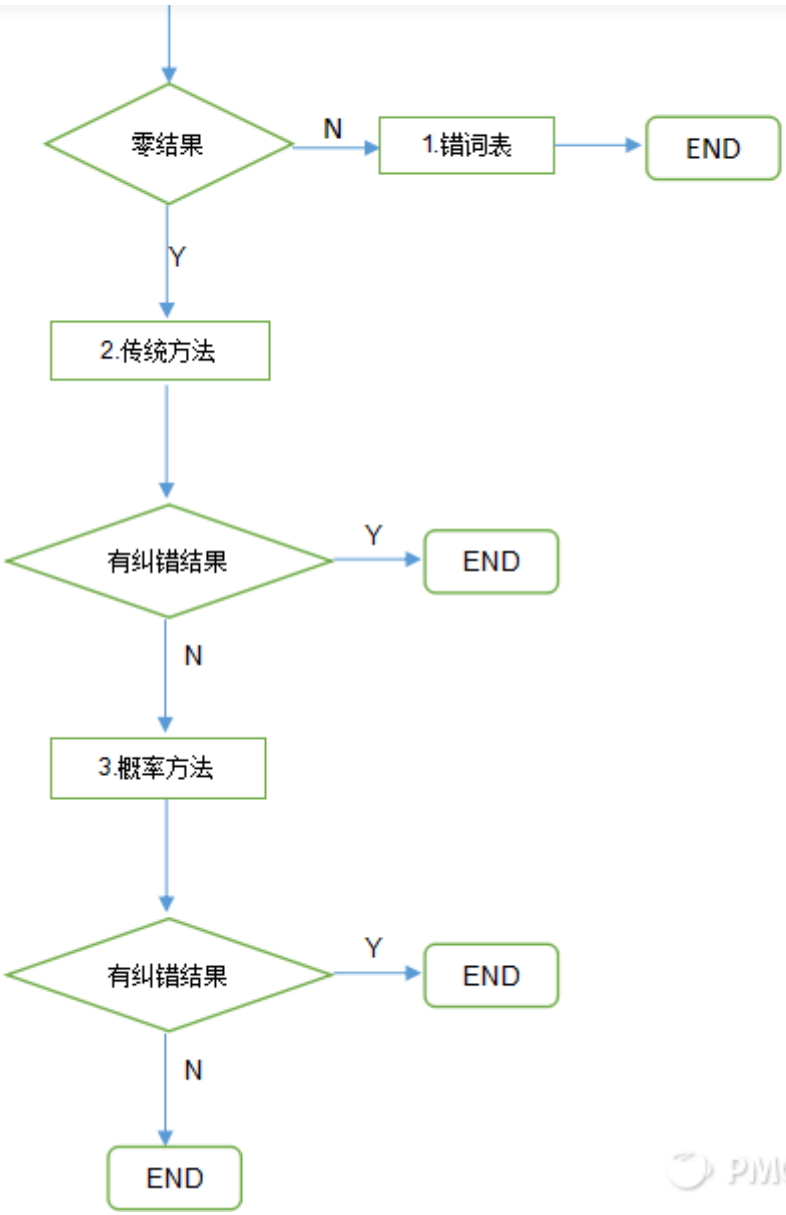
计算Jaccard，相似度最高为最优候选词

$表示词的开始或结束

实现：lucene spellchecker模块

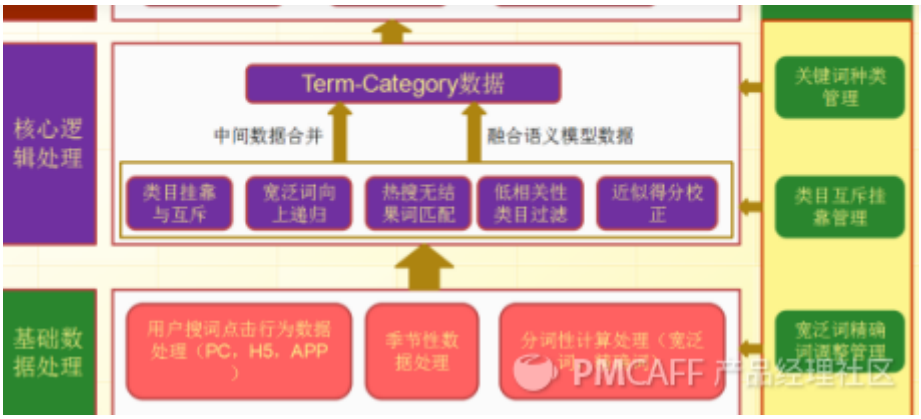

Pin Yin String
xi an shi min gan
xian ming an
Syllable Graph
西 安 是 民 感
戏 案 市 敏 干
系 暗 示 名
喜 使 明
先 命
现 市民
西安 使命
西安市 敏感
暗示 命案
Word Lattice
0 1 2 3 4 5 6

纠错流程



再来讲讲Query类目预测

Term-Category2.0架构



当然遗留的问题也多，比如：

类别多，末级类目8000多个，且sku分布不均衡

Query偏短，平均长度5，词个数1~3

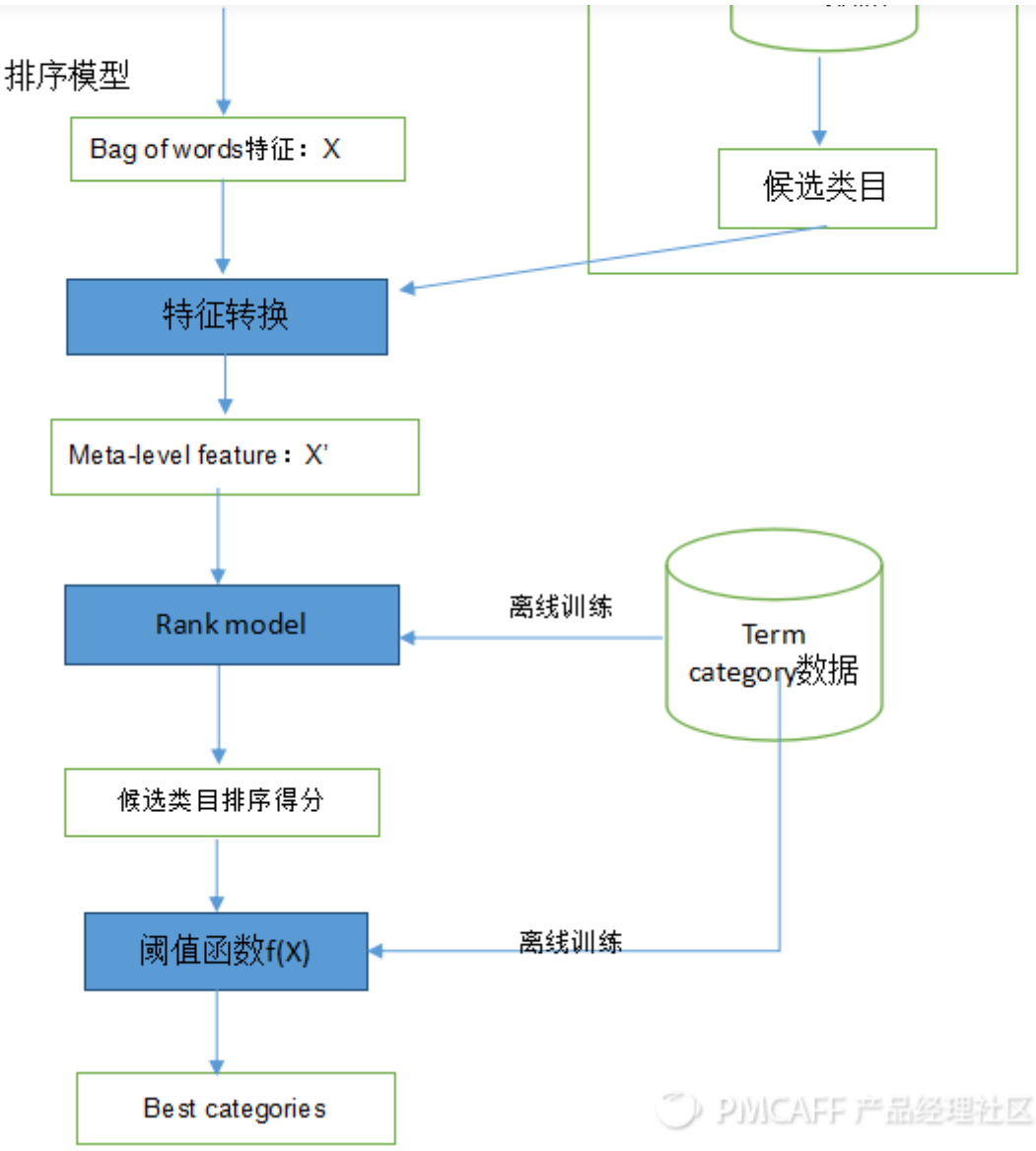
Query多类别，多类占比20%

方法

f(X) → Y，直接预测Y

f(X,Y) → {+1,-1}，Y作为特征

分类流程：



排序特征

Query检索指定类目下top K文档得分

Query与指定类目下top K文档 L2距离得分

Query与指定类目下top K文档L1距离得分

Query与指定类目下top K文档cos得分

实验比较

Baseline：MultinomialNB

检索模型

检索+排序模型

实验结果

baseline	0.1636	0.0538	0.0669	0.2949
检索模型_query	0.3506	0.2008	0.2319	0.4364
检索-排序模型_query	0.1850	0.1107	0.1243	0.1909
检索模型_product	0.4766	0.2652	0.3076	0.3304
检索-排序模型_product	0.5089	0.2765	0.3271	0.3461

baseline：移除多标签数据，转为多分类，用 MultinomialNB
检索模型：query 检索，得到候选类目，按照类目-商品数倒序，返回 top 1，细分索引数据为 query 和 product
检索-排序模型：对候选类目用排序模型排序，返回 top 1

参考

纠错

[lucene扩展] spellChecker原理分析

SunPinyin代码导读 - 输入法引擎 - A statistical language model based Chinese input method - Google Project Hosting

分类

Multilabel Classification with Meta-level Features

Deep Classification in Large-scale Text Hierarchies

我靠，梳理到这里发现这个流程和一号店有点相似。。。最后只能说明一点，类似的搜索产品拼音纠错和query类目预测都比较相似，推测和淘宝的也不会差很多。。。

后续补充。。。。。

写下你的评论

写下评论，分享你的见识…