CCF全国青年大数据创新大赛

垃圾短信基于文本内容识别

cloudComputing团队

报告人：王帅　　指导教师：徐睿峰

哈尔滨工业大学深圳研究生院智能计算研究中心HLT研究组

目录 / CONTENTS

团队介绍

团队：cloudComputing

队长：王帅

QQ：916794076（笔岸书塘）

队员：石锋、祝方泽、徐锋、蔡文举

指导教师：徐睿峰

学校：哈尔滨工业大学

实验室：智能计算研究中心HLT研究组

主要研究方向：自然语言处理\信息检索\机器学习

校招内推：www.myofferbus.com（ offer直通车）

团队战绩

团队：cloudComputing

队长：王帅

QQ：916794076（笔岸书塘）

队员：石锋、祝方泽、徐锋、蔡文举

指导教师：徐睿峰

学校：哈尔滨工业大学

主要研究方向：自然语言处理\信息检索\机器学习

校招内推：www.myofferbus.com（ offer直通车）

里程碑回顾

F值：0.99734

结合信息检索方法的
错误纠正策略

F值：0.99703

加入词组合特征+信息量特征

F值：0.99694

加入字节级N元文法特征

F值：0.99688

改进online-LR：增加迭代次数，
引入学习速率衰减，引入正则化

F值：0.99673

改进online-LR：加强错误边界学习，TONE策略

F值：0.99361

词特征+online-LR

算法介绍

整体思路

任务

实际任务映射为什么样的数据挖掘任务？监督学习型还是无监督型？二类分类多类分类？文本分类还是结构化数据的分类？短文本分类or长文本分类？

数据

样本如何定义？什么样的数据作为特征？样本的label怎么确定？如何划分训练集、验证集及测试集？

特征

如何从原始数据中提取机器学习模型适用的特征？业务理解和模型的原理如何有效结合？如何验证特征是否有效？样本特征缺失怎么办？

模型

选择合适的模型；根据具体的任务优化模型；模型调优；多模型融合

赛题分析

任务分析

基于短信文本内容，准确地、完整地识别出垃圾短信、正常短信

01 短信样本有类别标签，属于监督学习型

02 二类分类：垃圾短信（1），正常短信（0）

03 非结构化短信文本数据，属于短文本分类问题

数据

- 样本定义
  - 一条短信标识一条样本，利用带有审核结果标签的短信数据建立模型，识别未知标签短信
- 数据集划分
  - 离线学习模型：训练集+验证集+测试集
  - 增量式在线学习模型：训练集+测试集（训练集尽可能大）
  - 训练集越大越好
  - 验证集尽可能逼近测试集
- 正负样本平衡
  - 离线学习模型中正负样本比例控制1:10左右
  - 离线学习模型中常用随机上采样+随机下采样
  - 增量式在线学习模型：加强错误边界学习（TONE策略）

数据清洗

切忌过分清洗

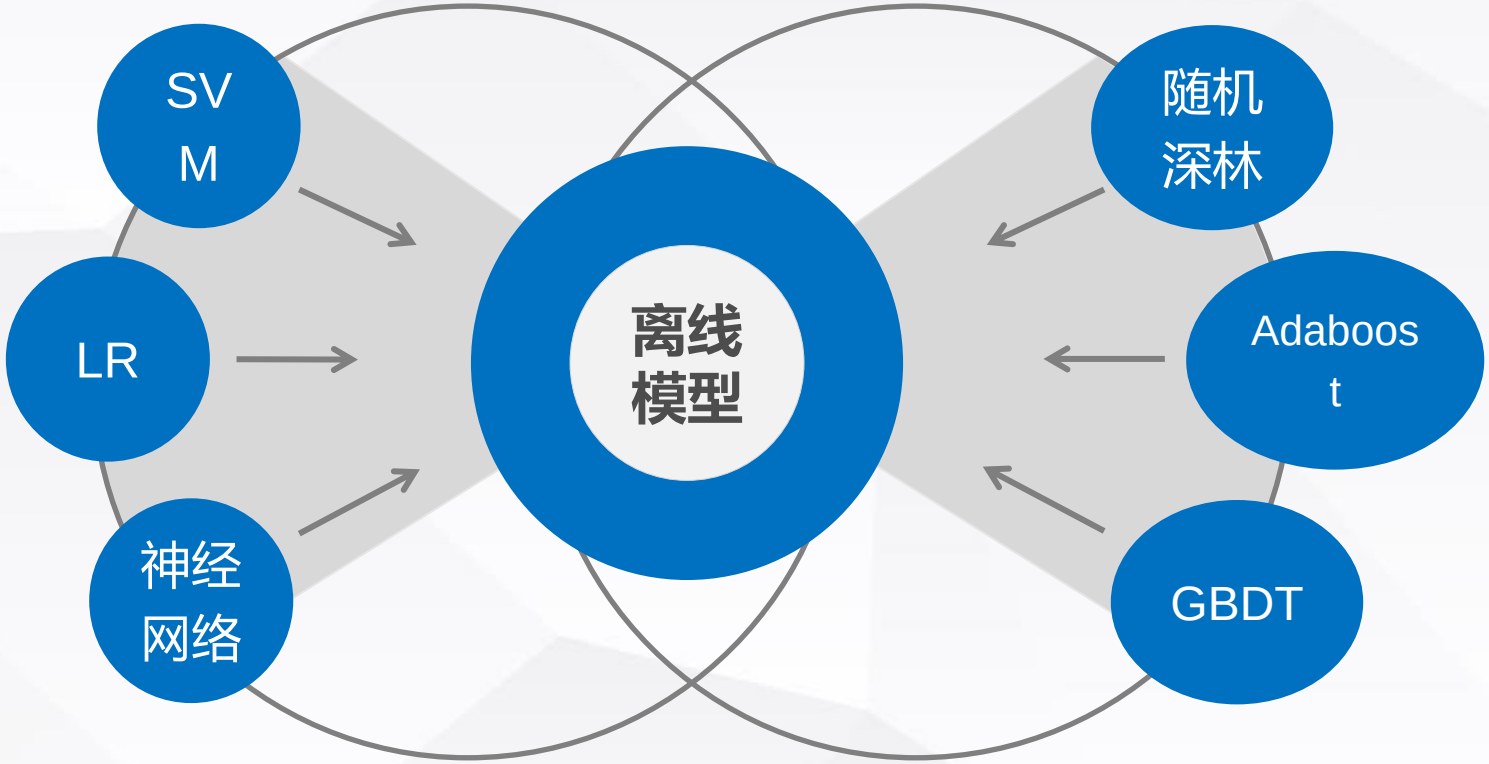
•短信文本包含中文和英文，简体和繁体形式，及含特殊字符等

•考虑到短信特定的不规则表达在很大程度上是识别垃圾短信一个重要的特征，勿过分清洗，草率的清洗数据很可能导致重要信息的丢失

短信文本统一转简体表述

号码等脱敏字符串转单字符

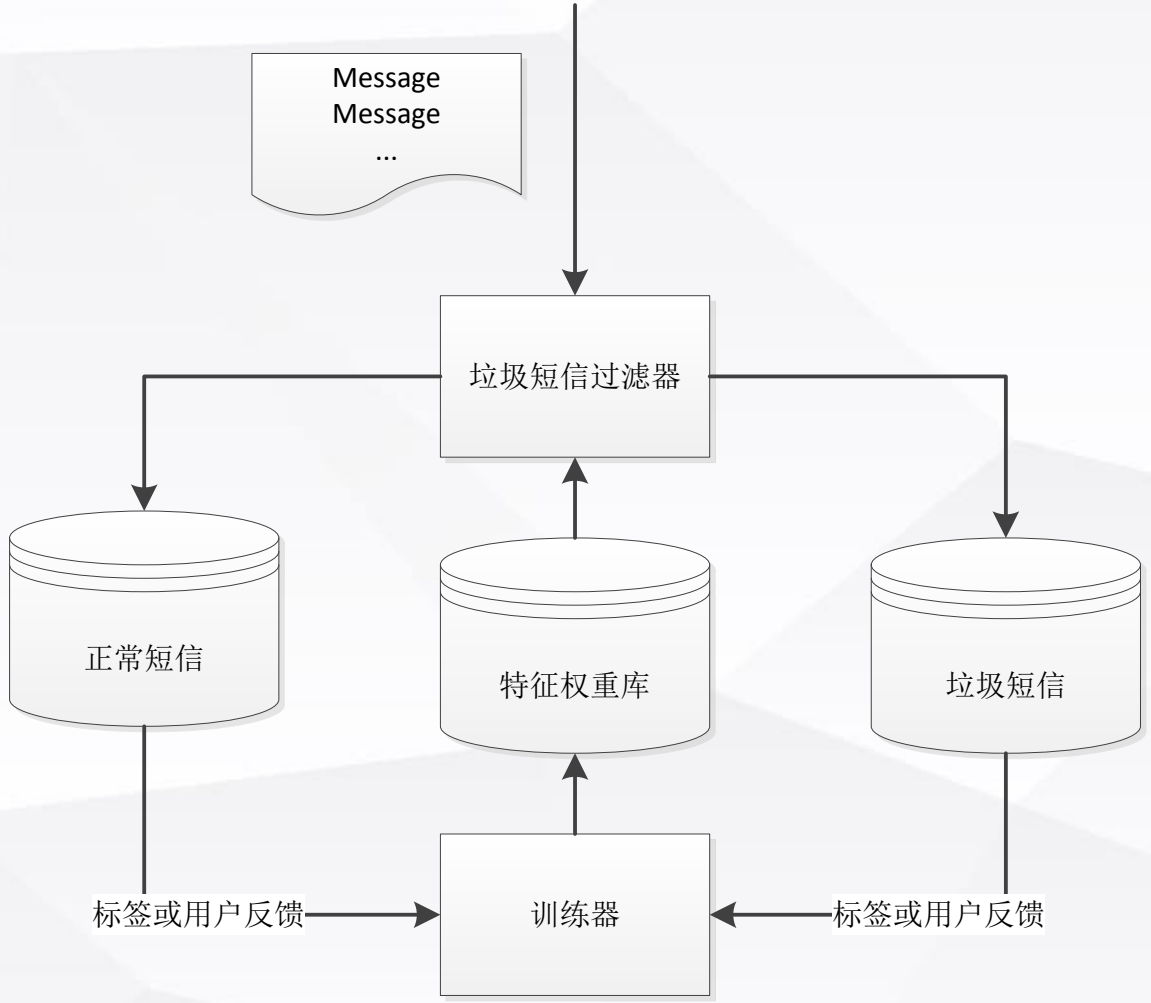
模型选择



- 垃圾短信随时间推移发生演进，导致固有模型失效
- 需要重新生成模型，而新模型训练需要重新整合数据，资源耗费大
- 无法及时感知垃圾短信环境的变换，无法实时动态更新模型

模型选择

•在线学习（online-learning，增量式学习算法）



•根据用户的反馈或标注不断自动更新系统模型参数，适应不断变化的应用环境

•可大大减小更新模型的资源消耗，提升线上应用效率

•对在线更新学习算法要求严格，模型参数的更新算法复杂度要低，要模型快速收敛或近似收敛，以适应实际应用需求

online-LR算法

•LR常用逻辑方程：

$$f(x_i) = P(Y = \text{垃圾短信} \mid \bar{x}) = \frac{\exp(\bar{w} * \bar{x})}{1 + \exp(\bar{w} * \bar{x})}$$

•特征权重更新：

$$w = \begin{cases} w + (1 - p) * x_i * Train\_rate, y_i = \text{垃圾短信}1 \\ w - p * x_i * Train\_rate, y_i = \text{正常短信}0 \end{cases}$$

online-LR 之 算法改进一

•加强错误边界学习（TONE：Train On or Near Error）

•避免随机抽样，有效解决正负样本比例极度不平衡问题

•可有效解决样本的过学习和欠学习问题

$\vec{w} = 0$; //initialize weights to 0

for each $\vec{x}_i, y_i$

$$p = \frac{\exp(\vec{x}_i \cdot \vec{w})}{1+\exp(\vec{x}_i \cdot \vec{w})}$$

if (p>0.5)

predict spam;

else

predict ham;

if(abs(p−0.5)<θ or prediction error) //TONE

if ($y_i == 1$)

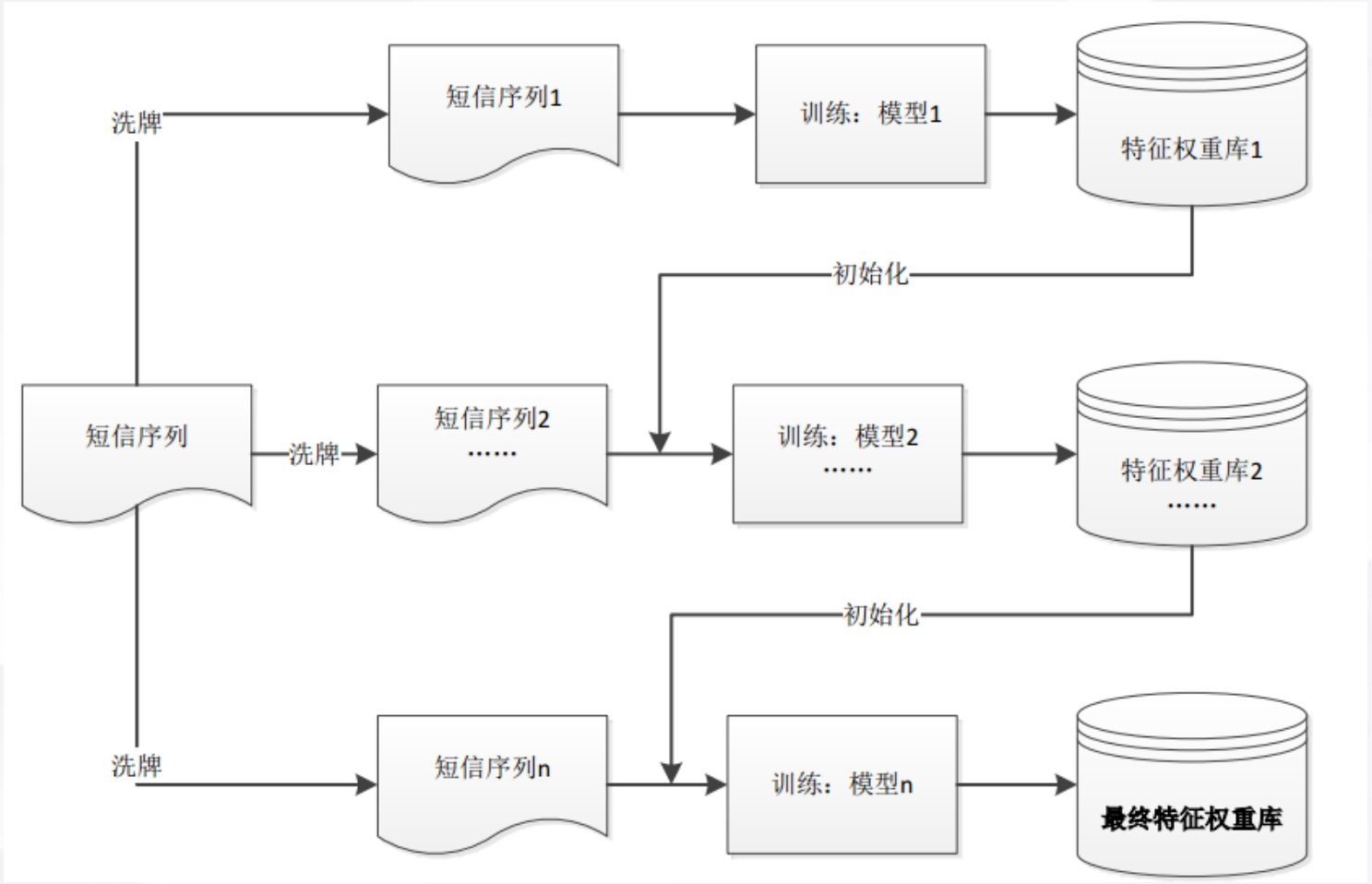
$\vec{w} = \vec{w} + (1-p) * \vec{x}_i \times$ rate

else

$\vec{w} = \vec{w} - p * \vec{x}_i \times$ rate

online-LR 之 改进算法二

•增量式迭代多次，每次迭代对短信序列随机洗牌，模拟短信环境的随机变化过程，使学习更充分



•线上应用时可直接转在线模式

online-LR 之 算法改进三

•学习速率衰减，引入衰减速率

$$Train\_rate_k = Start\_rate * \varphi^k$$

其中，k代表第k次迭代，Start_rate代表起始学习速率，φ代表延迟学习速率

•则融入学习速率衰减的特征权重的更新策略：

$$w = \begin{cases} w + (1-p) * x_i * Start\_rate * \varphi^k, y_i = \text{垃圾短信1} \\ w - p * x_i * Start\_rate * \varphi^k, y_i = \text{正常短信0} \end{cases}$$

online-LR 之 算法改进四

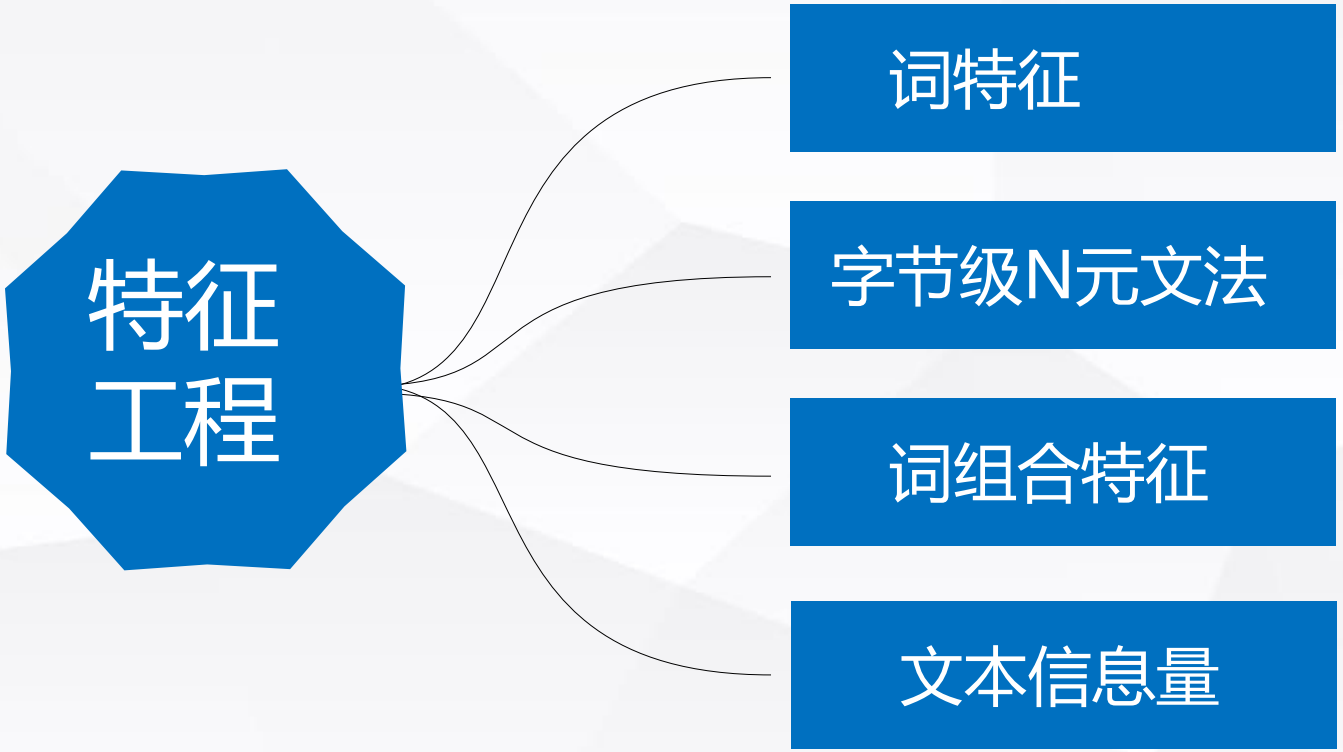
•借鉴离线LR中的风险最小化原则，同样引入正则化：

$$w=\begin{cases} w*(1-\varphi^{k}*lambda)+(1-p)*x_i*Start\_rate*\varphi^{k}, y_i=\text{垃圾短信}1 \\ w(1-\varphi^{k}*lambda)-p*x_i*Start\_rate*\varphi^{k}, y_i=\text{正常短信}0 \end{cases}$$

其中，lambda代表正则化参数

特征工程

- 特征决定性能上限
- 单纯的分词特征远远不够

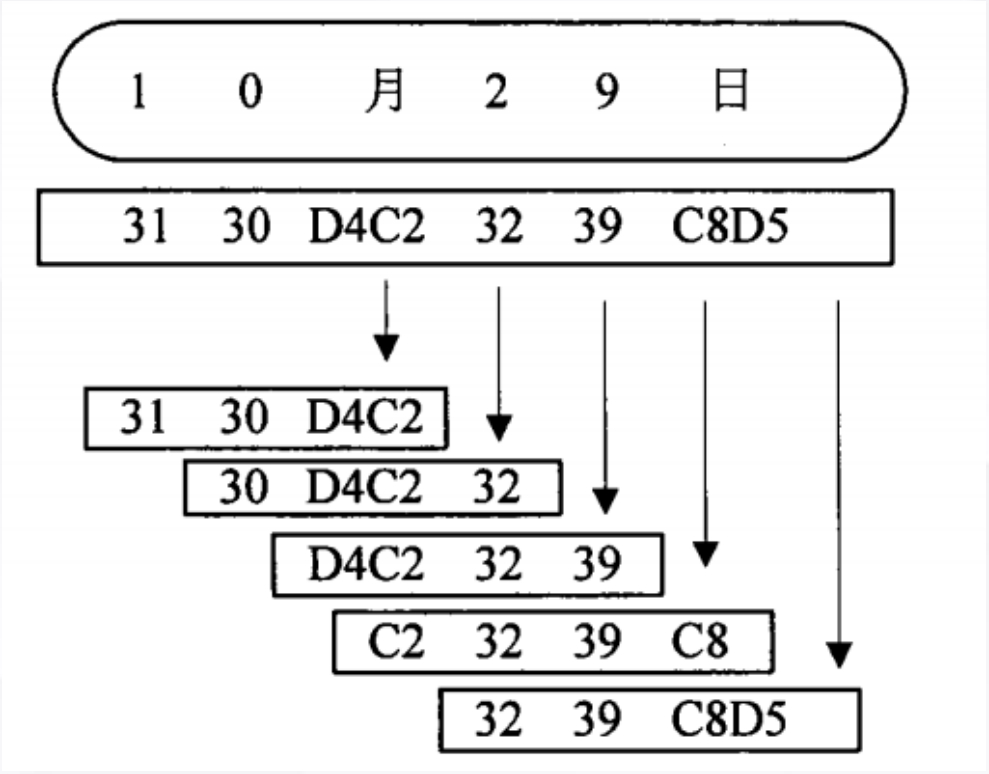


特征工程

- 中文分词
  - 利用开源分词工具ANSJ分词处理，并保留词性

本|r 公司|n 位于|v 茶山镇|n 现|v 招|v 焊工|v 铣工|v 钳工|n 有|v 兴趣|n 可|v 来电|v 询问|v

- 字节级N元文法
  - 有效提取垃圾短信隐藏形式的强特征
  - 避免繁杂的短信文本解析
  - 可以处理图像及病毒性的垃圾短信

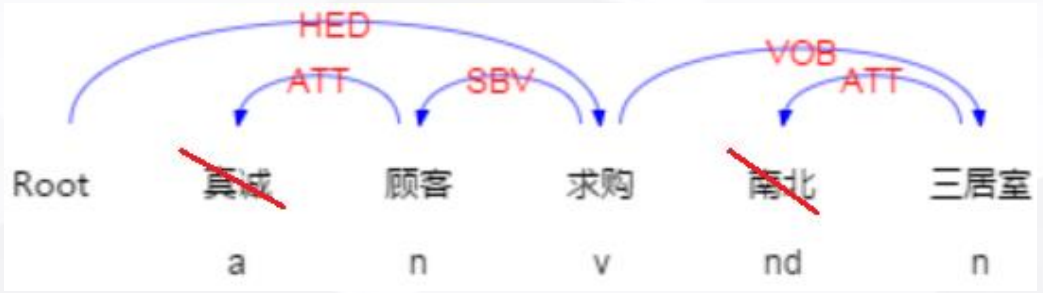


例如：

- product-->pro_duct，prod-uct；办证-->办\证
- prod-uct-->prod，rod-，od-u，d-uc，-uct

特征工程

- 词组合特征
  - 非修饰性实词组合成元组特征（名词+动词\形容词）
  - 转换成正则表达式模板，用于特征匹配



(.{2,10})客户求购(.{2,10}三居室(.*))

特征工程

- 信息量特征
  - 正常短信通常位于一定长度范围内，既不会太长也不会太短

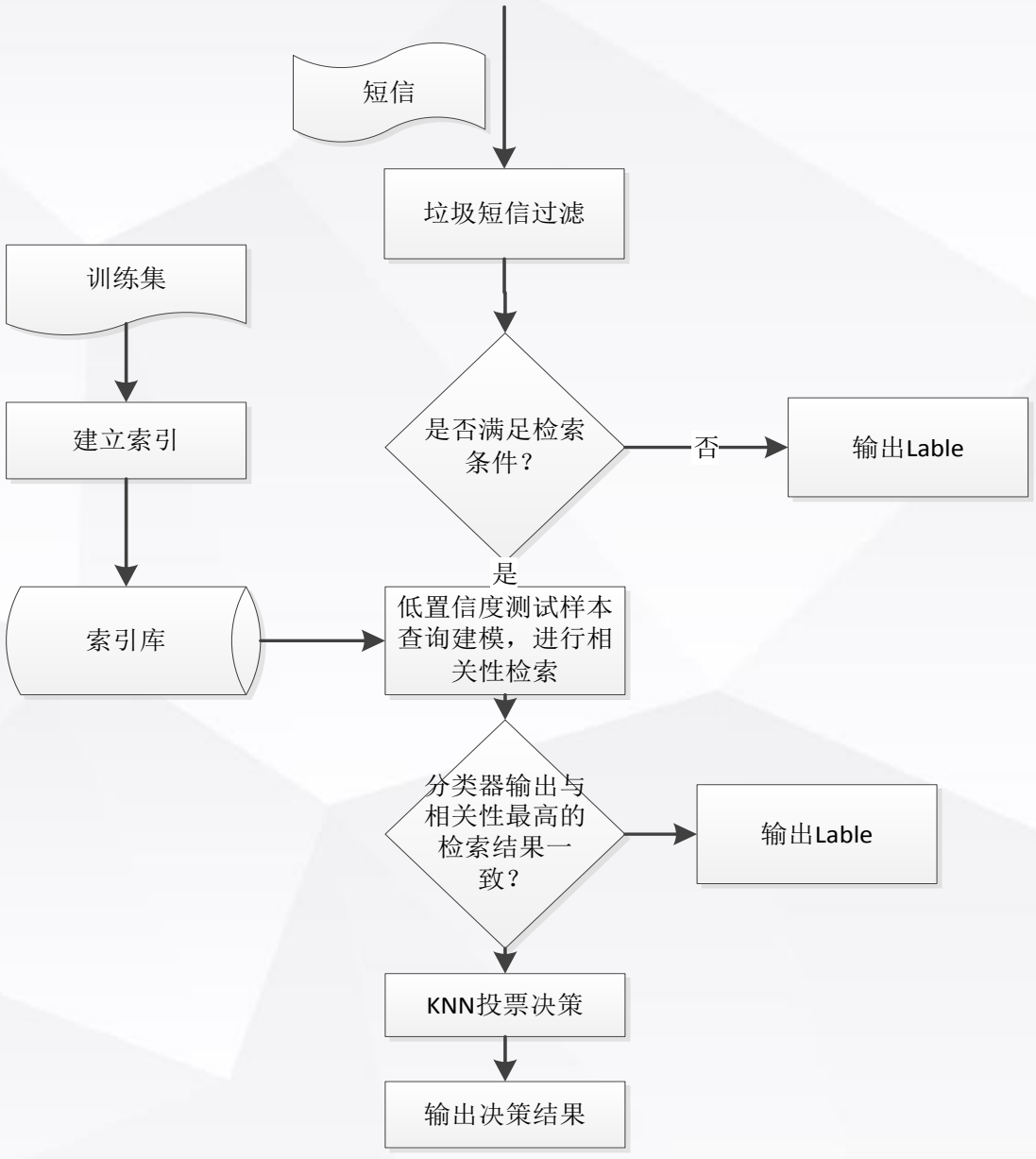
$$信息量 = \frac{1}{\exp(L-k)} + b$$

其中，L代表文本长度，k为调节因子，b为信息量平滑因子

- 文本越短，模糊性越大，特征缺失严重，文本长度特征相对重要

短文本特征缺失问题

- 短信内容较短，特征缺失严重
- 特征缺失导致分类器很难正确分类
- 结合信息检索方法的错误纠正策略
  - 将短信查询建模，进行相关性检索
  - 根据TopN结果，进行KNN投票决策
  - 易扩展其他检索技术来优化分类结果



线上排名

排名	队伍名称	最高得分	提交次数	最近提交时间
1	cloudComputing	0.99719	23	2015-12-18
2	overflow	0.99711	16	2015-12-18
3	NUDT_Yang	0.99674	7	2015-12-20
4	AIIIII	0.9965	7	2015-12-20
5	lifematrix	0.99646	14	2015-12-20
6	何处明心	0.99644	27	2015-12-19
7	以上成绩作废	0.99621	17	2015-12-20
8	CIKE_two	0.9958	57	2015-12-20
9	NUST	0.9957	40	2015-12-20
10	逍遥三老	0.99556	11	2015-12-14

算法特色与优势

经验总结与建议

尝试过，并验证有效

- 改进online-LR算法：四个优化
- 构建有效特征工程：词+词组+字节级N元文法+信息量
- 结合信息检索方法的分类错误纠正

尝试过，但验证无效

- 引入TF、TF-IDF特征不如0-1特征有效
- 加入主题模型并不有效
- SVM，GBRT等离线模型复杂度高，且性能并不如优化后的online-LR

未尝试，感觉有效

- 引入更加有效的检索模型，如语言模型
- 结合信息检索中查询反馈技术扩展样本特征，更加有效解决特征缺失

谢谢！欢迎批评指正！

中国好创意

http://www.wid.org.cn/

中国好创意

CCF TFBD