基于多维数字化方法的智能垃圾短信检测与实现

Detection and Implementation of Intelligent Spam Short Message Based on Multi-dimensional Digital Method

王玉玲[1]，刘晓鸣[1]，王尧永[2]（1. 中国联通济南分公司，山东 济南 250002；2. 中国联通山东分公司，山东 济南 250001）
Wang Yuling[1], Liu Xiaoming[1], Wang Yaoyong[2]（1. China Unicom Jinan Branch, Jinan 250002, China; 2. China Unicom Shandong Branch, Jinan 250001, China）

摘　要：
随着垃圾短信发送模型不断变化，传统的基于发送频次与内容的检测方法已经不能满足新型垃圾短信检测的需要。在治理垃圾短信的实践过程中，创新性使用了基于短信发送位置（城市）不易变化的特征作为垃圾短信的检测依据，并使用Simhash算法、改进的朴素贝叶斯算法等新方法对待检短信进行智能判断，有效提高了垃圾短信检测查全率、查准率，实现对垃圾短信的精准拦截，降低了垃圾短信举报率。

关键词：
垃圾短信；数字签名；Simhash算法；朴素贝叶斯算法
doi：10.12045/j.issn.1007-3043.2023.01.004
文章编号：1007-3043（2023）01-0015-06
中图分类号：TN929.5
文献标识码：A
开放科学（资源服务）标识码（OSID）：

Abstract：
With the continuous change of spam message sending model, the traditional detection methods based on sending frequency and content can not meet the needs of new spam message detection. In the practice of managing spam messages, the features that sending location (city) of SMS is not easy to change is innovatively used as the detection basis of spam messages, and new methods such as simhash algorithm and improved naive Bayesian algorithm are used for intelligent judgment of spam messages, which effectively improves the recall and accuracy of spam message detection, realizes the accurate interception of spam messages and reduces the reporting rate of spam messages.

Keywords：
Spam SMS; Mathematical signature; Simhash algorithm; Naive bayesian algorithm

引用格式：王玉玲，刘晓鸣，王尧永. 基于多维数字化方法的智能垃圾短信检测与实现[J]. 邮电设计技术，2023（1）：15-20.

0 引言

随着移动互联网的蓬勃发展，行业短信被广泛应用于网站、APP验证码、物流快递、订单通知等领域，为产品宣传、服务维系提供了有效手段，但因部分商家群发广告短信导致行业垃圾短信投诉量激增。同时各类违规催收、暴力催收问题呈野蛮发展态势，其中暴力短信催债成为网贷债务催收的重要手段，在催收过程中，催债人或催债公司使用手机号码，对贷款人及其关系人（亲属、同学、同事、朋友）实施短信轰炸，以此向贷款人施压，迫使其还款。该行为给贷款人及其关系人的身心、工作、生活造成了恶劣影响，严重破坏了正常的经济、社会生活秩序。

为了配合工信部关于垃圾短信的专项治理工作，某运营商制定了防范打击通信诈骗电话、骚扰电话治理、垃圾短信治理3项考核指标。

1 现状分析

首先调查现网垃圾短信拦截情况，垃圾短信检测拦截系统建于2012年，主要以短信发送频次和短信内容中的关键字作为垃圾短信的检测手段（见图1）。

发送频次检测：以单位时间内手机发送短信条数

收稿日期：2022-11-30

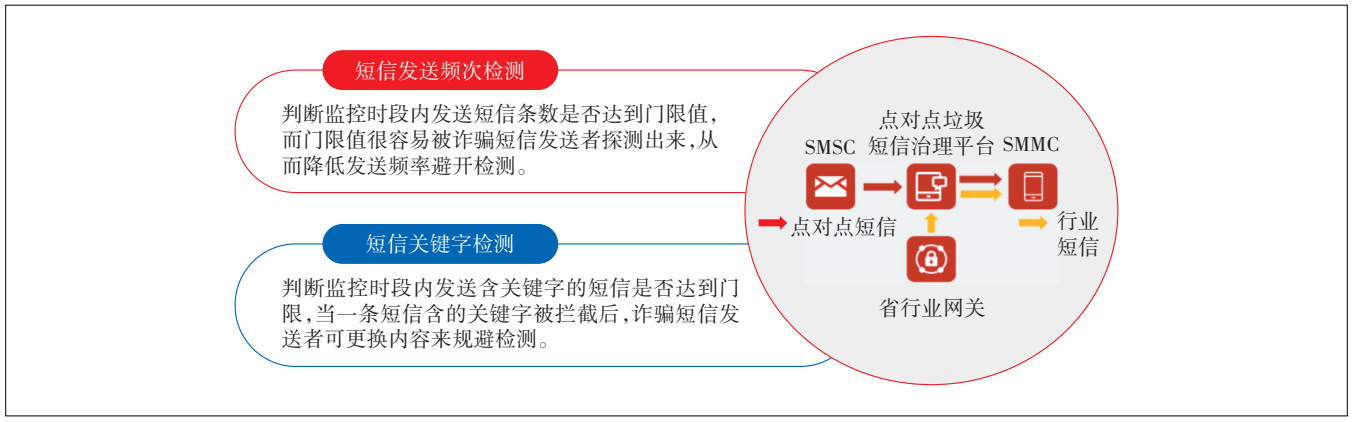

图1　垃圾短信检测手段

作为检测条件，如单位时间内发送的条数达到门限值，则作为疑似垃圾短信进行处理，但门限值很容易被发送者探测出来，从而采取低于门限值的发送频率避开检测。统计表明，当门限值设为30条/h时，短信发送频次检测出的垃圾短信准确率只有8%，需人工进行仲裁。如降低判断门限，准确率则会大幅降低，需要人工仲裁的短信呈指数型上升，受拦截准确率及人工仲裁工作量的限制，检测门限无法设置过低。

关键字检测：以监控时段内发送含敏感关键字短信的数量是否达到检测门限，来判断是否为垃圾短信，但当含有关键字的短信被拦截后，垃圾短信发送者可更换短信内容规避检测。另外，对于内容不断变化垃圾短信，关键字检测方法发现垃圾短信能力较弱，需依靠用户举报被动发现。

2 解决方案

为了弥补现有垃圾短信检测手段的不足，顺应人工智能时代机器学习技术潮流，本文集中进行前瞻性技术的应用研究，着力解决垃圾短信检测难的技术难题，并提出了3种解决方案。

2.1 基于发送位置检测

根据垃圾短信投诉在短信中心反查举报号码注册的MSC地址，确定垃圾短信发送者所在城市，并定时提取该城市所有MSC发出的全部短信进行检测（见图2）。经过统计分析，共确定9个垃圾短信易发城市，定时提取从这9个城市发送的短信进行重点检测。

统计漫游到全国各MSC的短信提交量，若MSC提交量大于正常值则判断为垃圾短信，并根据该MSC确定所在城市。据此，又将5个城市确定为垃圾短信易发地（市），定时提取其MSC发送的短信（见图3）。

根据发送位置定位方式所提取的短信，根据发送条数、短信长度、离散度及号码入网时间4个常规条件排除明显不是垃圾短信的短信，以减少智能算法工作量（见图4）。

该方法对特定漫游城市重点检测，提高了检测的针对性，因此检测门限、检测时间粒度均可较传统方

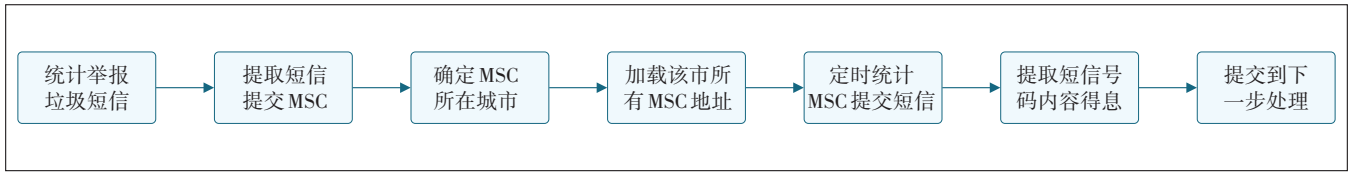

图2　查找发送位置流程图

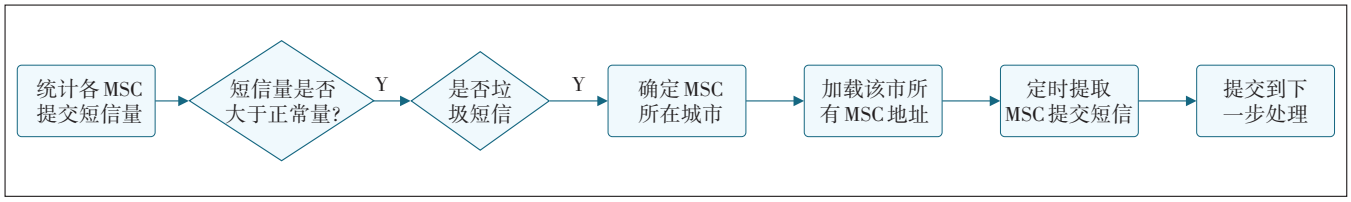

图3　查找MSC流程图

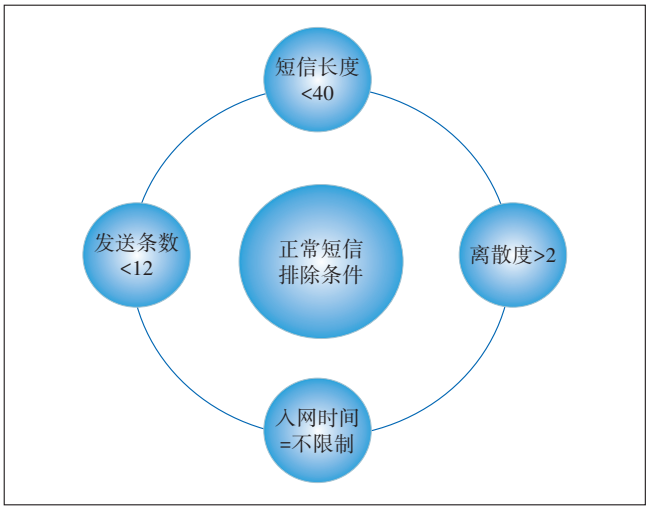

图4　垃圾短信检测条件

法降低，如检测时间粒度可低至1 min，检测门限可降至2条，再辅助套餐、被叫号码离散度等判断手段，垃圾短信检测准确率可达99%以上，较传统检测方法提高30倍以上。

2.2 基于SimHash算法的垃圾短信匹配算法

2.2.1 垃圾短信样本库检测的基本思想

利用已知的垃圾短信样本检测待检短信是发现垃圾短信有效方法，如果采用字符串匹配方法对短信进行比较，需要两重循环来遍历待检短信和垃圾短信样本中的所有字符串，进而统计这2个集合中相同字符串的个数。对内存和时间的消耗都非常大，检测效率低，样本库只能维持在几百条左右。

2.2.2 Hash算法的优势与不足

借鉴互联网网页的去重方法，使用Hash算法将短信进行数字化，从而实现垃圾短信的快速检测。Hash算法实现原理：将不同长度规则的短信内容通过Hash算法转换为一个相同长度的字符串（数字签名），用这些数字签名来表示原文本。当某条短信的数字签名与垃圾短信样本库数字签名一致时，则可认定为垃圾短信。这样就将字符串比较转换成了数字运算，从而提高检测速度，样本库也可以达到百万级。

当垃圾短信发送者发现发送完全相同内容的垃圾短信易被拦截时，会尝试用在短信中增加虚假称呼等方法来规避垃圾短信的检测。

如表1所示，2条短信内容只有称呼不同，但生成的数字签名完全不同，因此普通的Hash算法无法检测与样本短信内容不完全相同的短信。

2.2.3 SimHash算法原理与实现

表1　数字签名对比

短信内容	数字签名
李经理你好，高新管委会单位学区房，城市广场168平，送车位地下室，低于市场价10万	0001000001100110100111011011110
张先生你好，高新管委会单位学区房，城市广场168平，送车位地下室，低于市场价10万	1010010001111111110010110011101

SimHash算法是一种局部敏感Hash。所谓局部敏感，是假定A、B具有一定的相似性，在Hash之后，仍然能保持这种相似性。SimHash的基本原理是对于2个给定的变量 x,y，哈希函数 h 总是满足：

$$P_{rh\in F}\left[h(x)=h(y)\right]=\mathrm{sim}(x,y)$$

其中，P_r 表示 $h(x)=h(y)$ 的可能性，$\mathrm{sim}(x,y)\in[0,1]$ 是相似度函数，一般也用雅可比函数 $\mathrm{Jac}(x,y)$ 来表示 x,y 的相似度，$\mathrm{sim}(x,y)$ 表示如下：

$$\mathrm{sim}(x,y)=\mathrm{Jac}(x,y)=\frac{|x\cap y|}{|x\cup y|}$$

h 属于哈希函数簇 F，需要满足以下条件：

a）如果 $d(x,y)\leqslant d_1$，则 $P_{rh\in F}\left[h(x)=h(y)\right]\geqslant P_1$）。

b）如果 $d(x,y)\geqslant d_2$，则 $P_{rh\in F}\left[h(x)=h(y)\right]\leqslant P_2$）。

称 F 为 (d_1,d_2,p_1,p_2) 上的敏感哈希簇函数。其中 $d(x,y)$ 表示 x,y 变量之间的距离，通俗而言，如果 x,y 足够相似时，那么它们映射为同一Hash函数。

其基本做法是通过将原始的文本映射为64位的二进制数字串，然后通过海明距离（Hamming Distance）来度量2个串（通常是二进制串）的差异，2个二进制串对应的位有几个不一样，那么海明距离就是几，例如 x=1010，y=1011，那么 x 和 y 的海明距离就是1，值越小越相似。

2.2.4 SimHash算法实现流程

a）分词：判断短信内容分词，形成这条短信的特征单词，最后形成去掉噪声词的单词序列，并为每个词加上权重，权重一般分为5个级别（见表2）。

表2　SimHash算法分词

短信内容	分词后
李经理你好，高新管委会单位学区房，城市广场168平，送车位地下室，低于市场价10万	李经理(3)你好(1)高新(3)管委会(4)单位(1)，学区房(2)，城市(1)广场(1)，168平(5)送车位(3)地下室(2)，低于(1)市场价(1)10万(5)

注：括号里数字代表单词在整个句子里重要程度，数字越大越重要。

b）Hash：通过Hash算法把每个词变成Hash值，将字符串变成了一串串数字。

c）加权：通过 步骤b）的Hash生成结果，按照单词的权重形成加权数字串。

d）合并：把上面各个单词算出来的序列值累加，变成只有一个序列串。

e）降维：把步骤d）算出来的序列数值变成01串，形成最终的SimHash数字签名，如果每一位大于0记为1，小0记为0。

图5所示为SimHash算法流程示意。

对于待检测短信，首先转化为64位的数字签名，与样本库中已存在的数字签名逐一进行对比，当海明距离小于阈值N时，则认为其与垃圾短信样本库的内容相似，可认定为垃圾短信（见图6）。

在实践过程中，当阈值N取值为5时，判断得出的垃圾短信准确率可达95%，可以将其直接加入垃圾短信黑名单；当阈值N取值为10时，判断得出的垃圾短信准确率约为60%，需要人工复核其是否为垃圾短信。

该方法对使用多个号码大量群发重复性垃圾短信的情形有良好的检测效果，上线后，长期发送重复垃圾短信的情况基本消失，取得了良好的治理效果。

2.3 基于改进朴素贝叶斯算法的智能检测

朴素贝叶斯算法具备稳定性较好、实现简单且易于开发维护的特点，是文本文档分类算法中较为有效的算法。

2.3.1 朴素贝叶斯算法原理

朴素贝叶斯分类方法是在条件独立性假设的前提下，计算该文本所属类别的概率，是建立在贝叶斯定理之上的一种分类算法（Dreiseitletal.，2002）。贝叶斯定理是用来计算随机事件A和B的条件概率之间的关系，其计算方法如下：

$$P(B|A)=\frac{P(A|B)P(B)}{P(A)}$$

朴素贝叶斯分类算法通过计算文本类别与词分布的联合概率，进而对文本进行分类。具体计算方法如下：

$$P\left(C_j|D\right)=\frac{P\left(D|C_j\right)}{P(D)},j=1,2,3\cdots m$$

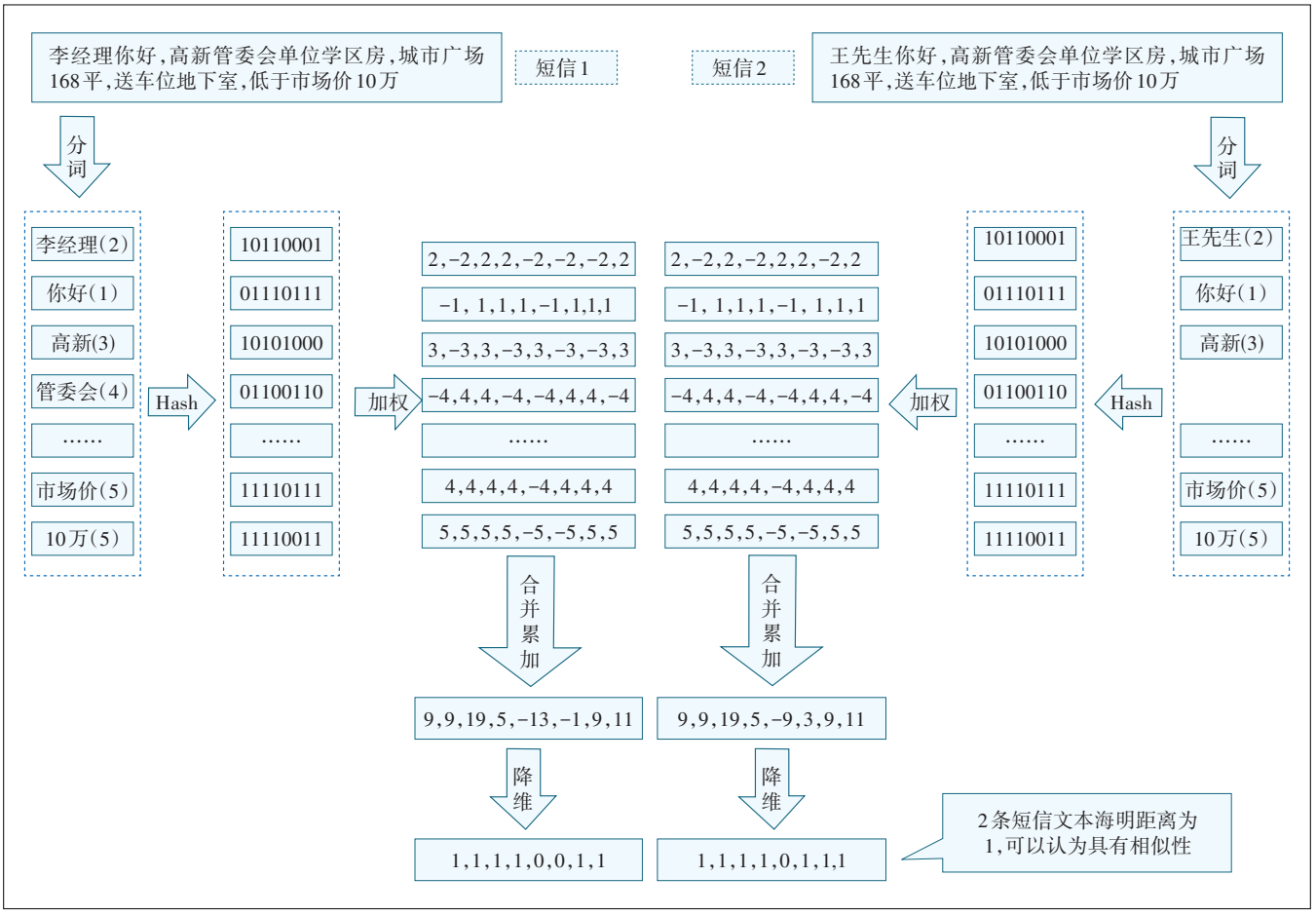

图5 SimHash算法流程示意图

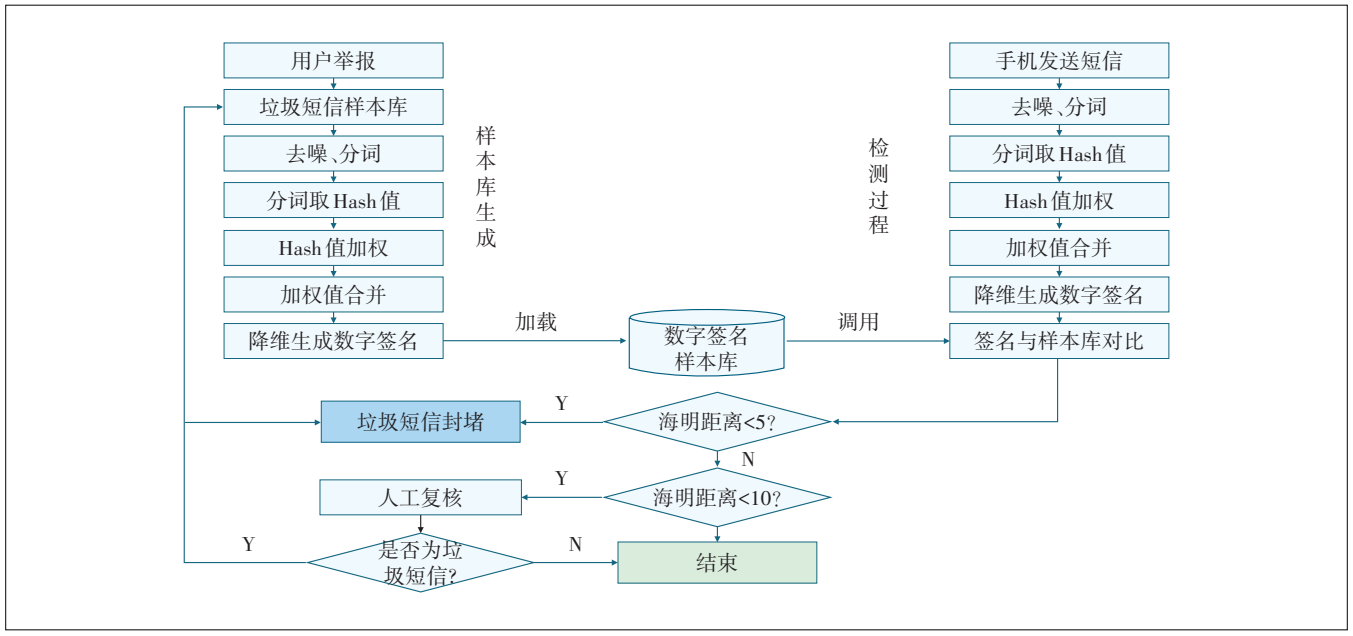

图6　判断流程

$P(C_j|D)$是给定的文本D属于类别C_j的概率，$P(D|C_j)$是类别C_j包含文本D的概率，最后把$P(C_j|D)$中值最大的一个作为给定文本D所属的类别。朴素贝叶斯算法是在词之间相互独立的假设下计算的，对于式中$P(D)$对总体的计算结果没有影响，故而求解$P(C_j|D)$可转换成求$P(D|C_j)\cdot P(C)$的值。计算公式可转换为：

$$P(C_j|D)=P(D|C_j)\cdot P(C_j)=P(C_j)\prod_{i=1}^{n}P(D_i|C_j)$$

后验概率的最大值所对应的类即为该未知样本的分类：

$$H_{NB}(D)=\arg\max_{1\leqslant j\leqslant k}P(C_j)\times\prod_{i=1}^{n}P(D_i|C_j)$$

面对每天1亿多条的海量短信检测，朴素贝叶斯算法的处理速度及准确率仍有提升空间，需采取改进措施，以提升处理效率及准确性。

2.3.2　文本表示改进，提升准确率

短信文本表示改进方面，在对垃圾短信预处理时，针对噪声数据大和jieba分词不能识别新词的问题，数据采用流程化处理，包括繁体字转换、数字和特殊符号替换、错别字纠正、文本转拼音4个部分。对未能识别的新词，引入了改进的新词识别工具，将获得的新词字典导入jieba自定义词库中。并且为了减少非垃圾短信预测为垃圾短信的概率，引入了“例外”一类。对“例外”这类使用固定阈值和差值阈值选择方法，用于获得科学的阈值，以提高检测准确性。

2.3.3　特征提取改进，提升效率

垃圾短信特征项的提取，改为以基本短语为单位的分词方法，结合基本短语构成算法，并根据基本短语的定义实现由词到基本短语的转换。

短语结构模型的界定是一个确定不同类型短语边界位置的过程，是以单词为构件形成短语的主要步骤。作为能代表文本主要特征的一般名词短语和动词短语，其界定规则对降低特征项空间的维度及提高准确性来说非常重要。基本短语模式特征项提取应当遵循以下2个规则[5]：

a）一般名词短语结构模型界定。汉语简单非嵌套式名词短语（baseNP）的结构有：

（a）baseNP+baseNP，如“公路里程”“高校教师”等。

（b）base NP+名词名动词，如“公路建设”“高校发展”等。

（c）限定性定语+baseNP，如“双核”“三好学生”等。

（d）限定性定语+名词名动词，如“中国人口调查”“三峡工程建设”。

b）一般动词短语结构模型界定。一般动词短语结构模型形式主要有：

（a）述补结构、如压马路、走路等。

（b）述宾结构，如修改论文、选角等。

（c）状中结构，如立刻动手、到黄山游玩等。

（d）连动结构，如去洗手、开动等。

（e）联合结构，如边走边唱、甲和乙等。

（f）其他动词短语，如“着、了、过”属性的动词，睡了、坐着、想了想、听说过等。

短信文本短语特征项提取过程：先把第一个分词词语与后面的词语分别进行组合，通过过程测试检验，则认为合格；如果不通过，则将该短信所有短语列为垃圾短信短语（词语）向量集，继续取下一个分词词语重复上述过程，直到最后一个词语完成组合测试为止（见图7）。

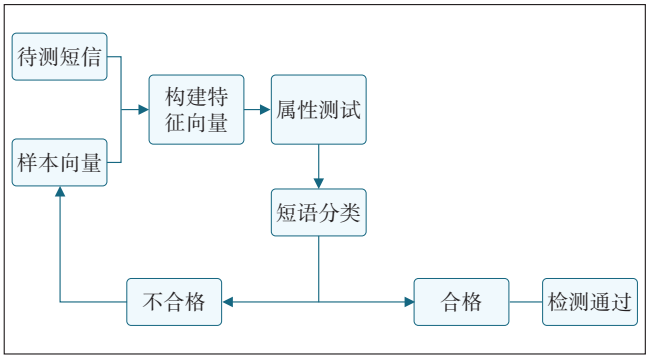

图7 垃圾短信短语特征项提取流程图

基于互信息方法，利用统计思想划分分词短语的边界。互信息是考察一个消息中两信号间的相互依赖度的度量，也是分词词语间结合的紧密程度的度量，通过短信文本相邻词性标记的互信息值大小来进行判断，其极小值的位置为短语的边界。互信息方法计算公式为：

$$\mathrm{MI}(F)=\sum_{i}P(C_j)\times\lg\frac{P(D|C_j)}{P(D)}$$

基于短语朴素贝叶斯检测算法的主要改进在于利用互信息计算短信文本特征项提取算法，将特征项提取由以词为单位改为以短语为单位，降低样本空间规模，从而提升效率。

3 结束语

本文提出的基于人工智能的垃圾短信治理新方法，将垃圾短信特征加入朴素贝叶斯机器学习算法，实现垃圾短信精准画像，将垃圾短信管控重心前移，从事中被动拦截变为源头主动管控。

某运营商垃圾短信智能检测系统上线后，提高了垃圾短信检测查全率、查准率，实现对垃圾短信的精准拦截，有效降低了垃圾短信举报率。随着深度学习技术的不断进步以及数据处理能力的不断提升，为持续提高短信息质量和垃圾短信的治理效率，基于深度学习技术的垃圾短信治理是值得研究的方向。

参考文献：

[1] 袁婷婷. 基于人工神经网络的垃圾短信识别研究[D]. 长春：东北师范大学，2012.

[2] 李琼阳. 一种改进的朴素贝叶斯算法在垃圾短信用户识别中的应用[D]. 广州：华南理工大学，2017.

[3] 王行甫，杜婷. 基于属性选择的改进加权朴素贝叶斯分类算法[J]. 计算机系统应用，2015，24(8)：149-154.

[4] 金展，范晶，陈峰，等. 基于朴素贝叶斯和支持向量机的自适应垃圾短信过滤系统[J]. 计算机应用，2008，28(3)：714-718.

[5] 杨霞，董红斌，张海玉，等. 基于分布估计算法的朴素贝叶斯分类问题研究[J]. 电脑知识与技术，2010，6(11)：2704-2705，2731.

[6] DOMINGOS P，PAZZANI M. On the optimality of the simple bayesian classifier under zero-one loss[J]. Machine Learning，1997，29(2)：103-130.

[7] 张俊. 压降工信部12321平台垃圾短信被举报率浅谈[J]. 信息通信，2017(7)：255-257.

[8] 12321网络不良与垃圾信息举报受理中心. 2016年12月12321受理网络不良与垃圾信息举报数据分析[J]. 互联网天地，2017(1)：56-60.

[9] 李雪梅. 基于文本分类的多层次垃圾短信过滤系统研究[D]. 重庆：重庆理工大学，2012.

[10] 佚名. 12321举报中心正式开通不良和垃圾彩信举报通道[J]. 中国信息安全，2012(11)：21.

[11] 李润川，咎红英，申圣亚，等. 基于多特征融合的垃圾短信识别[J]. 山东大学学报(理学版)，2017，52(7)：73-79.

[12] KIM Y. Convolutional neural networks for sentence classification[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing(EMNLP). Doha：Association for Computational Linguistics，2014：1746-1751.

[13] 姜鹤. SVM文本分类中基于法向量的特征选择算法研究[D]. 上海：上海交通大学，2010.

[14] 孙茂松，陈新雄，张开旭，等. THULAC：一个高效的中文词法分析工具包[CP]. 北京：清华大学自然语言处理与社会人文计算实验室，2016.

[15] 梁桢，李禹生. 基于Hash结构词典的逆向回溯中文分词技术研究[J]. 计算机工程与设计，2010，31(23)：5158-5516.

作者简介：

王玉玲，高级工程师，硕士，主要从事短信平台的维护及建设工作；刘晓鸣，高级工程师，学士，主要从事短信平台的维护及建设工作；王尧永，高级工程师，硕士，主要从事IT平台的需求分析及建设工作。