基于广义数字的智能垃圾短信拦截系统设计

Design of Intelligent Spam Short Message Interception System Based on Generalized Digital

王金栋[1]，向前兰[2]，李　岳[1]（1. 中国电信股份有限公司陕西分公司，陕西 西安 710035；2. 咸阳师范学院，陕西 咸阳 712000）
Wang Jindong[1]，Xiang Qianlan[2]，Li Yue[1]（1. China Telecom Corportation Shaanxi Branch，Xi'an 710035，China；2. Xianyang Normal University，Xianyang 712000，China）

摘　要：
垃圾短信问题困扰已久，存在拦截规则呆板、拦截效率低、误拦截、错拦截等问题。采用基于广义数字的垃圾短信过滤系统方案，实现基于广义数字拦截规则的智能识别，减少人工判断、提高拦截效率。

关键词：
广义数字；垃圾短信；拦截
doi：10.12045/j.issn.1007-3043.2021.03.012
文章编号：1007-3043（2021）03-0055-03
中图分类号：TN918
文献标识码：A
开放科学（资源服务）标识码（OSID）：

Abstract：
The problem of spam messages has been plagued for a long time，and there are problems such as stiff interception rules，low interception efficiency，false interception，and false interception. It adopts a spam SMS filtering system scheme based on generalized numbers to realize intelligent recognition based on generalized digital interception rules，reduce manual judgment and improve interception efficiency.

Keywords：
Generalized number；Spam SMS；Intecept

引用格式：王金栋，向前兰，李岳. 基于广义数字的智能垃圾短信拦截系统设计[J]. 邮电设计技术，2021（3）：55-57.

0 引言

随着移动通信业务的飞速发展，手机给人们带来便利的同时，也带来了许多危害。与微信、QQ等社交工具相比，短信具有被叫号码不受限制、快捷、高效等优势，导致不法分子趁机以短信形式实施诈骗、广告宣传甚至传播手机病毒，轻则给用户带来骚扰，重则造成经济损失。目前，大部分运营商垃圾短信监控系统主要利用关键字策略、流量策略和被叫行为分析等方法进行变相组合监控和拦截[1-2]，同时配以人工审核对监控结果进行二次确认提高查准率。但随着近年来垃圾短信发送方法的不断升级，运维运营的治理成本大幅上升，治理效果却在下降。

收稿日期：2021-01-04

基于此，本文重点研究基于内容的智能垃圾短信分析系统，在原有垃圾短信网元中增加智能分析模块，及时发现并更新拦截策略；采用广义数字识别方法识别短信中的电话号码、银行卡等数字信息，提升拦截效率和效果。

1 智能分析系统组网及业务流程设计

增加了智能分析系统后，需要对现有的短消息业务流程进行调整，在原有系统基础上增加智能分析系统，具体业务流程如图1所示。

a）MO为用户提交消息到短消息中心（SMSC）。

b）deliver_req为SMSC提交消息到监控平台进行监控处理。

c）deliver_rsp为监控平台根据现有监控策略对消息进行相关监控处理，并将结果反馈给短消息中心。

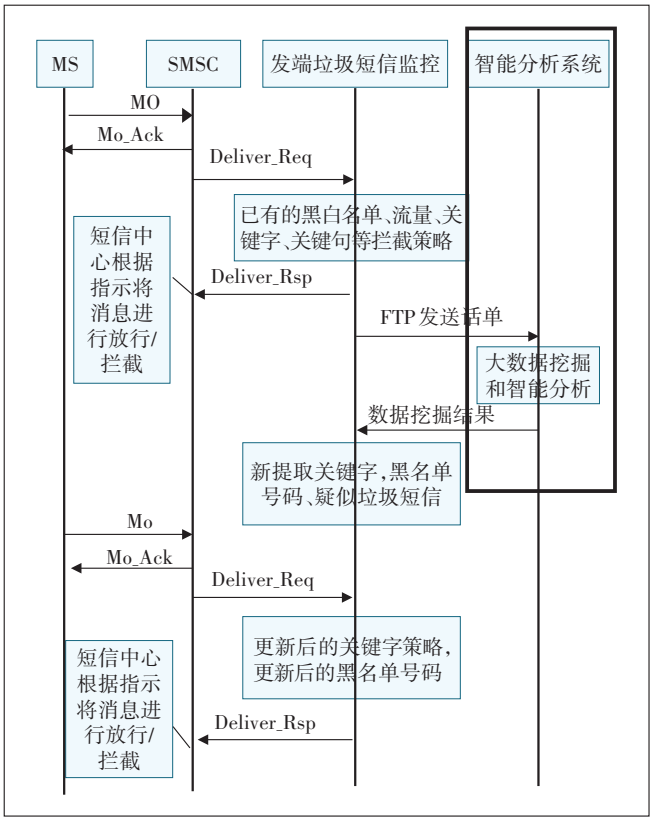

图1　短消息业务流程设计

d）监控平台将消息话单同步到智能分析系统。

e）智能分析系统将此条话单入库，并进行大数据挖掘和分析处理。

f）智能分析系统将相关数据挖掘结果同步到监控平台（新提取关键字、黑名单号码、疑似垃圾短信）。

g）后续SMSC提交到监控平台的消息，将根据更新后的监控策略进行处理。

2 广义数字识别

从垃圾短信产生的根源分析，诈骗或者宣传类的垃圾短信通常会包含联系电话或银行账号等重要信息，而由于成本及更换困难等多种原因，这些联系电话或账号相对比较固定，通过之前的垃圾短信内容分析，很多内容是经常变化的，但银行卡账号或者电话号码一般更换的较少。因此，数字特征是大量垃圾短信中具有明显特征且比较固定的特征信息。如果根据垃圾短信内容中的数字相关信息进行拦截，拦截效率大大提升的同时，拦截效果也会非常显著。

2.1 定义广义数字库

目前垃圾短信中所包含的电话号码、账号等数字已不是简单的阿拉伯数字，不法分子为了避免被拦截，往往在其中穿插了各种各样的“数字”：阿拉伯数字、中文简体数字、繁体数字、谐音数字、带符号的数字，这些表现形式多样的“数字”称为广义数字。

广义数字库可配置，包括阿拉伯数字、中文简体数字（如一、二、三）、繁体数字（如壹、贰、叁）、谐音数字、带符号的数字（①、❶）、以全角或上下标表示的数字（如１、$_1$、^{4}）等[3]。

通过智能垃圾短信拦截系统对大量话单的挖掘统计分析，会提取一份广义数字黑名单送往实时监控系统，经启用同步后用于垃圾短信的实时判断，当多个不同的主叫发送的短信中都含有上述广义数字时，实时短信垃圾监控系统会判断击中广义数字黑名单规则并直接实时拦截，从而减少短信下发。

2.2 广义数字特征向量提取流程

广义数字特征向量是从该条短信内容中提取的若干个连续广义数字组成的集合。

a）短信内容预处理。首先对短信内容进行特殊字符过滤，即去除掉空格和标点符号后，接着以广义数字库为基础，对短信内容中的广义数字进行检测，统一替换为阿拉伯数字。

b）单个连续数字段的最小长度（字符数）判断。连续K个或K个以上（K可以自定义，如K=3，即表示连续3个或3个以上的广义数字段才会被抽样出来）的广义数字才会被抽样出来，设某条短信内容中抽样出的数字各段组合集合：$\{a_1a_2\cdots a_i, b_1b_2\cdots b_j, \cdots\}$，其中$a_1a_2\cdots a_i$和$b_1b_2\cdots b_j$是抽样出的2个数字段，则必须满足$i,j\geqslant 3$。

c）2个连续数字段的最小距离（字符数）判断。广义数字特征向量应是由该条短信内容中相对集中的一段广义数字组成的集合，设某条短信抽样出的数字各段组合成集合：$\{a_1a_2\cdots a_i, b_1b_2\cdots b_j, \cdots\}$，则必须满足：$a_1a_2\cdots a_i$和$b_1b_2\cdots b_j$ 2段广义数字段之间的距离小于等于J个字符。其中J可以自定义，如J=4，则表示如果2组数字之间的其他字符超过4个或4个以上时$a_1a_2\cdots a_i$不会被抽样出来；继续检查$b_1b_2\cdots b_j$和下一段之间的距离。

d）有效的数字特征向量长度范围（字符数）判断。抽样出的广义数字段组成一个广义数字特征向量：$V=\{a_1a_2\cdots a_ib_1b_2\cdots b_j\cdots\}$。检查该向量的长度$x$需要满足长度范围$m\leqslant x\leqslant n$，考虑到目前手机号码为11位，固定电话号码不含区号一般为7~8位，含区号一般为11~12位，银行账号一般为16位，故可设定m=7，n=

16。

2.3 可疑广义数字特征向量判断原则

广义数字特征向量判断首先对短信内容进行预处理，将短信格式进行规整，然后抽样出广义数字特征，具体流程如图2所示。

为每个新提取出的广义数字特征向量设置一个计数器Q_i，当发现另一条包含该特征向量的短信时，该计数器Q_i累加；同时比较该特征向量的主叫号码是否相同，若不相同，则其相应的主叫号码离散度D_i加1。

当某个广义数字特征向量满足：计数器Q_i达到阈值且主叫号码离散度D_i达到阈值，则该条特征向量判定为可疑广义数字特征向量。

2.4 广义数字结果输出

根据可疑广义数字特征向量挖掘规则，将短信内容及主叫号码提取出来，并根据人工判断是否启用规则，通过对样本挖掘，结果举例如图3所示。

图3中每行第1列为提取的数字，第2列为其权重，权重越高垃圾短信嫌疑越大，一般权重为0为垃圾短信和正常短信的临界点。

经过对内容进行人工判断，如图3所示训练结果均可被认定为垃圾短信，认定判断准确。

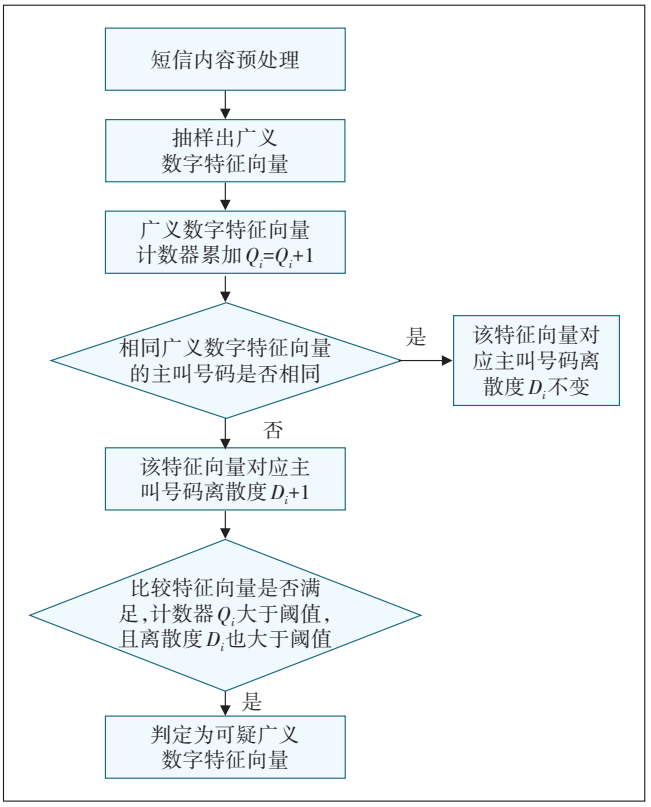

图2　广义数字判断流程

66026222,3609,49,现买现租！新街口商业圈[精锐“SOHO”]精装现房酒店式公寓,40-62平,投资30万稳定年赚3万,买到即是赚到！电话:66026222
84713763,1691,47,速办企业贷款,最高一千万,民间融资首选专业高效,有房产即可办理;月综合成本1.8%;三百万一日得！江苏邦成:84713763金轮大厦24A
31905777,856,54,让您久等了,金地长青湾[天阅]147-167平产品,浑河脉唯一墅区高层,给您双河一湖顶级亲水享受,赠双层挑空卧室。31905777
31886666,634,199,就差20万？世茂五里河帮你补齐！150平金廊稀缺准现房现在购买立减20万！T6精装酒店公寓2万抵5万;抢到就赚了！31886666
11513111111118021403448,541,15,急-用-款,5千-30万无-抵-押-正-规-安-全,电-话:18021403448新街口新世纪-投-资,如有打扰敬请原谅
88725555,491,188,浑南核心臧品！五层电梯洋房独立入户,悦享8万平商街,尽在咫尺的超市、影院。143平洋房起价7100元/平限时限量8872555

图3　广义数字训练结果

3 结论

本文重点对基于内容的智能垃圾短信拦截系统进行了分析，与其他垃圾短信鉴定系统不同的是，此次主要以广义数字样本识别对垃圾短信的内容进行了判断，并且通过文本实验进行抽样，可行性强，判断准确率高，可以为运营商垃圾短信治理提供强有力的支撑手段。

参考文献：

[1] 张尼，张智江，宋建，等. 垃圾短消息过滤技术综述[J]. 移动通信，2009,33(6):17-21.
[2] 徐英慧，刘梅彦. 基于内容的手机端垃圾短信过滤策略研究[J]. 北京信息科技大学学报(自然科学版),2013,28(1):51-55.
[3] 钱庆锋，万博文. 基于广义数字的垃圾短信拦截策略的研究[J]. 中国新通信,2015,17(4):42-43.
[4] 易阳锋. 垃圾短信监控的原理与实现[J]. 中兴通讯技术,2005(6):49-54.
[5] 张燕，傅建明. 垃圾短信的识别与追踪研究[J]. 计算机应用研究,2006(3):245-247.

作者简介：
王金栋，工程师，硕士，主要从事短信系统及反诈系统维护工作；向前兰，博士，主要从事本科教学工作；李岳，工程师，硕士，主要从事陕西电信销售渠道建设工作。