评语：该方案通过动态库（解决容量扩展问题）、多指纹库（解决数据噪声影响）、预备库的提前过滤（减少噪声）以及多算法模型融合方法，真正解决了垃圾短信的识别难问题。

垃圾短信大数据自动识别技术

林华生[1]，钱岭[2]，周莹[1]，杨希[2]，李婧[2]
1. 中国移动通信集团公司信息安全管理与运行中心，北京 100031
2. 中移（苏州）软件技术有限公司，江苏 苏州 215163

摘要
近年来，垃圾短信问题层出不穷，严重侵害了广大用户的通信权益。详细分析了垃圾短信治理面临的挑战、存在的问题及解决思路，利用人工智能手段，设计了一套基于指纹算法的垃圾短信大数据自动识别系统，对新系统的整体架构、工作流程和关键技术进行了详细的阐述。该新系统可以高效治理垃圾短信，全力对抗非法短信侵袭。

关键词
垃圾短信；大数据；人工智能；指纹算法

中图分类号：TP393　　文献标识码：A　　doi: 10.11959/j.issn.2096-0271.2018024

Automatic identification technology for spam message

LIN Huasheng[1], QIAN Ling[2], ZHOU Ying[1], YANG Xi[2], LI Jing[2]
1. China Mobile Information Security Center, Beijing 100031, China
2. China Mobile (Suzhou) Software Technology Co., Ltd., Suzhou 215163, China

Abstract
In recent years, the problem of spam messages is endless, which seriously infringes on the communication rights and interests of the majority of users. The management challenges facing the spam messages, the exisiting problems and the respording solutions were analyzed in detail. Based on the use of artificial intelligence methods and big data, an automatic identification system with fingerprint algorithm of spam messages was designed. The architecture, workflow and key technoligies of the system were elaborated. And the system can deal with spam messages effectively.

Key words
spam message, big data, artificial intelligence, fingerprint algorithm

1 引言

手机用户的不断增加，特别是智能手机使用量的增长，使得人们获得信息的效率大大提高。但是手机渗透率的快速增长却伴随着垃圾短信的快速扩散，不仅对人们的日常工作和生活产生影响，而且存在着极大的安全隐患（利用短信进行诈骗、勒索等犯罪活动）。治理垃圾短信不仅需要工业和信息化部的监督和手机安全厂商的屏蔽，更需要运营商主动出击，利用先进的技术从源头上拒绝垃圾短信的发送。

2 面临的挑战

2.1 垃圾短信发送成本低，社会影响大

中奖诈骗类、政治违法类、涉黄涉黑类、病毒诱导类、商业广告类等违规短信层出不穷，垃圾短信数量居高不下，中国移动通信集团有限公司（以下简称中国移动）年均处理疑似垃圾短信高达3亿余条。

垃圾短信不仅给手机用户造成了不可避免的骚扰，更对社会造成了不良的影响，主要体现在以下几个方面。

- 利用短信进行勒索、诈骗的违法犯罪活动日渐猖獗（如以中奖、敲诈等方式出现）。
- 少数不法分子利用短信传播黄色信息，毒化社会风气。
- 短信营销泛滥，某些商家企图通过短信传销商品，给用户带来一定的骚扰。
- 不法分子通过短信中嵌入的恶意链接，诱导用户点击，远程操控用户手机，窃取用户隐私。

2.2 现有治理平台识别效率不足

2012年以来，中国移动持续开展不良信息集中治理工作，依托不良信息集中管控平台，对监测发现的不良信息进行处理。但是监测策略准确率存在瓶颈，导致整个系统的有效识别率仍然存在不足。另外，传统的分类模式不能及时、准确地对垃圾短信进行分类，影响后续关键词的生产以及策略的精细化管理工作。随着垃圾短信的日益增长，管控平台存在较长的处理时延。

要解决上述问题，就必须考虑在现有平台中接入新的自动识别系统，对管控平台监测发现的疑似垃圾短信进行自动识别，提高垃圾短信治理效率。同时不断训练扩充新的关键词库，提高识别的有效性。

3 存在的问题及解决思路

随着大数据技术和人工智能技术的日趋成熟，短文本识别技术在机器翻译、网页查重等方面均有广泛的应用。一些互联网安全公司也开发了“手机安全卫士”等产品，应用贝叶斯学习（Bayesian learning）、支持向量机（support vector machine，SVM））[1,2]等机器学习算法识别垃圾短信，并在用户终端侧进行拦截和提醒。此外，Google公司提出的指纹算法①[3]将长文本转化为64位的散列码进行计算、比对。但由于短信的特殊性，仍存在以下问题。

- 表示稀疏问题：单条短信内容短小，传统的词袋（bag of words，BOW）模型[4]无法获取足够的特征信息用来区分垃圾短信和非垃圾短信。
- 数据噪音问题：存在大量的非正规语言的使用现象，传统的基于词汇的文档表示模型无法处理该问题。

① http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.473.7179&rep=rep1&type=pdf

● 动态演化问题：短信内容和语言使用随时间高速演化，固定的特征集合和分类模型无法应对该问题。

针对算法识别的问题，中国移动创新提出“指纹+语义”交叉融合算法模型。Simhash算法是一种经典的模糊匹配算法，通过匹配指纹编码实现分类，适合识别常见的群发类垃圾短信，精准性强；语义识别算法通过学习语料特征，能够“举一反三”，适合发现、识别新的样本，扩展性强。两种大数据识别技术互补，交叉融合可实现自动识别率的提升，具体如下。

（1）准确率的提升

引入大数据技术后，需实时对违规号码进行关停，这对大数据识别算法来说，准确率要求极高。通过对指纹库进行多库分解、引入动态数据库等机制，识别准确率高达99.8%，实现技术应用。

（2）大数据的实时处理

由于大数据识别算法需要存储大量的历史知识，为实现对疑似垃圾短信的实时处理，系统采用分布式多机多核系统架构，通过Kafka技术，有效打破了固有指纹算法的性能瓶颈。

4 系统解决方案

4.1 整体架构

基于Simhash算法，结合应用场景和线上持续运营要求，完善系统功能设计，满足以指纹识别算法为核心算法的线上识别功能和运营功能。系统整体架构如**图1**所示。

该系统具有以下几个特点。

● 应用创新指纹算法对待识别短信进

图1 系统整体架构

行处理；核心算法可扩展，支持引入新算法交叉融合识别。

● 在基础运营功能的基础上，打造稽核质检、投诉回溯核查等针对指纹算法特点研发出的持续运营功能。

● 采用金库管理模式，对数据安全进行双重保障。

● 采用分布式多机多核系统架构，通过Kafka实现内部服务之间的通信，有效保障了现网的实时运行需求。

系统架构主要采用分布式多机多核的方案，通过将系统模块服务化，完成系统模块间的解耦，进一步提高了系统算法的扩展能力。分布式多机多核具有速率高、性能稳定的优点。

4.2 工作流程

在垃圾短信集中管控平台中，引入垃圾短信大数据自动化识别系统，用以提升垃圾短信识别率。具体的垃圾短信大数据识别应用方案如**图2**所示。

大数据自动化识别系统工作流程如下。

● 建立垃圾短信大数据自动识别系统，利用前期积累的海量短信样本对数据库进行初始化。

● 将系统与现有垃圾短信治理模块对接，接收监测模块发来的全量疑似垃圾短信，并进行自动识别。

● 得到识别结果的垃圾短信，直接送至处置模块实时处置；未识别的短信按照原有流程进行处理。

4.3 关键技术

结合技术应用场景和短信文本短的特点，考虑到中国移动线上治理的极高准确性要求，团队研发出具有自主知识产权的指纹识别算法。

短信文本通过文本预处理完成噪声的处理。文本预处理主要包括简繁转换、大小写归一化、半角全角归一化以及拼音转文本等；利用自研的分词算法对文本进行切分，然后使用CityHash算法实现分词文本到指纹的转换，最后再利用FNV-1算法对散列指纹进行再散列，以减少指纹冲突。基于生成的指纹，创新提出针对指纹的动态数据库、基于多指纹库识别、汉明距离动态调优等技术。

（1）动态数据库机制

如**图3**所示，在入库方面，为指纹算法设计二次入库技术，降低训练数据中误判造成的影响；在出库方面，动态剔除入库早、不常使用的指纹，解决指纹库膨胀的问题，保障指纹库的容量可持续高效运营，并进一步提升算法识别准确率。

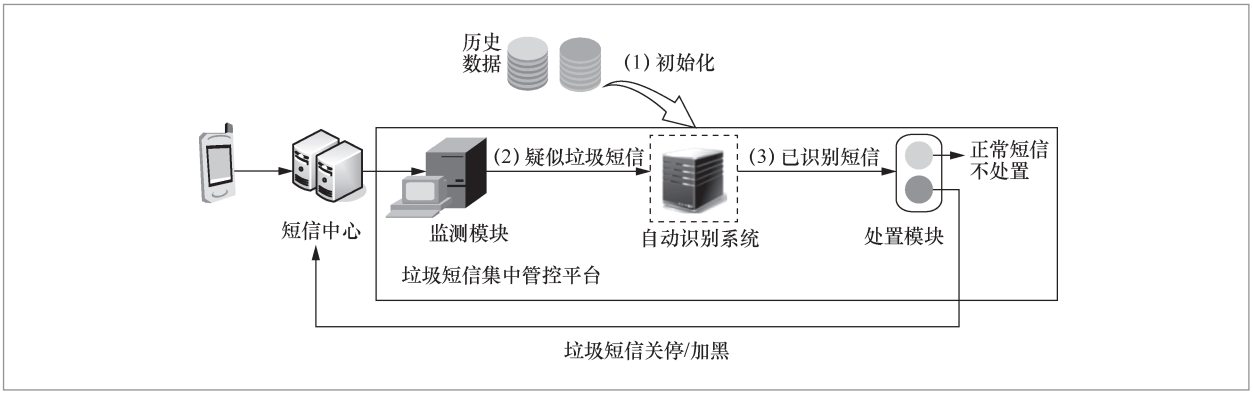


图 2 垃圾短信大数据识别应用方案

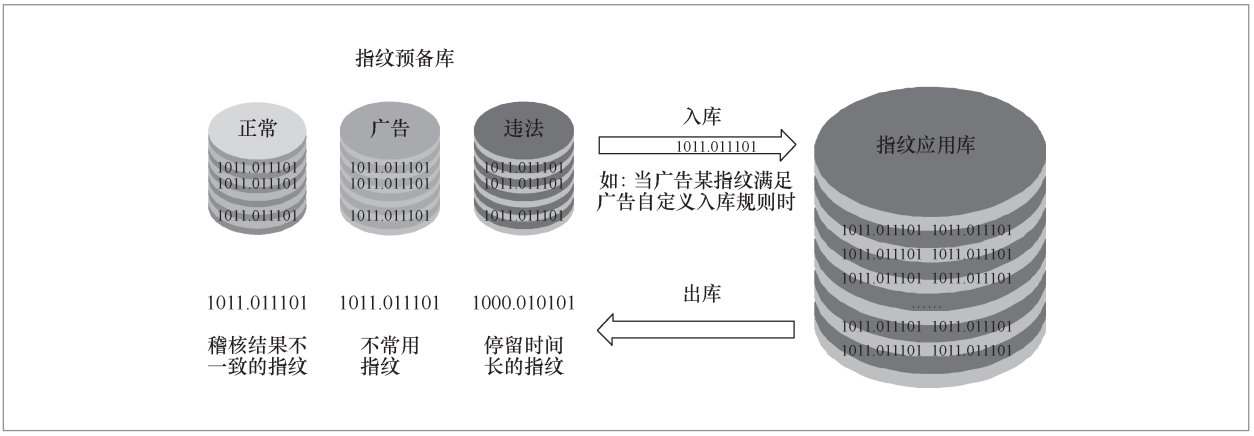


图 3　指纹库动态维护机制

（2）多数据库指纹存储机制

系统搭建多指纹库，根据处置方式不同，分为正常短信指纹库、违法诈骗短信指纹库、商业广告指纹库，并对后两者采用更为严格的校验入库机制和优先级更高的识别反馈机制。同时，根据考察各指纹库相互冲突的指纹，实现对数据库的进一步去噪，降低算法误识别比例。

（3）汉明距离动态调优

作为算法的核心参数，汉明距离表征不同文本之间的相似程度，即汉明距离越大，文本相似程度越低；反之，该距离越小，文本内容则越接近。算法前期通过调优测试明确初始汉明距离，在后续持续运营时，根据实时的自动识别率和识别准确率，动态实现汉明距离的调优。

5 应用效果

垃圾短信大数据自动识别系统于2017年1月在中国移动全网上线，覆盖31省。截至目前，累计接收垃圾短信系统全量疑似垃圾短信1.1亿余条，自动识别处理4 300万条，自动识别率达40.1%，识别准确率达99.8%。系统处理速率达14 000条/s，垃圾短信自动判定平均处理时长仅为0.07 ms。上线以来，系统运行情况良好。

由于垃圾短信自动判定平均处理时长仅为0.07 ms，违规号码的关停及时性大大提高，月均可减少不法分子发送的垃圾短信约8 500万条，有效地保障了广大用户的通信权益。

上线以来，垃圾短信治理效率大幅提升，通过系统的应用，垃圾短信治理团队有效降低133人，每年可节约人力成本1 452万元。

参考文献：

[1] 王斌，潘文锋. 基于内容的垃圾邮件过滤技术综述[J]. 中文信息学报，2005，19(5)：1-10. WANG B，PAN W F. A survey of content-based anti-spam Email filtering[J]. Journal of Chinese Information Processing，2005，19(5)：1-10.

[2] ANDROUTSOPOULOS I，KOUTSIAS J，CHANDRINOS K V，et al. An evaluation of naive Bayesian anti-spam filtering[J]. Tetsu-to-Hagane，2000(2)：9-17.

[3] HO P T，KIM H S，KIM S R. Application of sim-hash algorithm and big data analysis in spam email detection system[C]//The 2014 Conference on Research in Adaptive and Convergent Systems，October 5-8，2014，Towson，USA. New York：ACM

Press, 2014: 242-246.

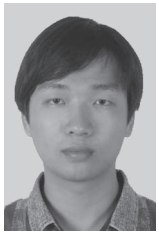
[4] SRIRAM B, FUHRY D, DEMIR E, et al. Short text classification in twitter to improve information filtering[C]// The 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, July 19-23, 2010, Geneva, Switzerland. New York: ACM Press, 2010: 841-842.

作者简介

林华生(1976-),男,中国移动通信集团公司信息安全管理与运行中心副总经理,负责大数据识别等新技术在中国移动不良信息治理应用场景下的方案设计和应用模型研究等。

钱岭(1974-),男,中移(苏州)软件技术有限公司大数据产品部高级工程师、总经理,中国敏捷开发十大贡献者之一,中国移动技术咨询委员会IT专家组成员。兼任工业和信息化部数据中心联盟大数据促进委员会副主任委员、中国计算机学会大数据专家委员会通信委员、苏州大数据产业副主任委员。主要研究方向为大数据产品规划、研发和应用全流程以及软件过程改进等。

周莹(1989-),女,中国移动通信集团公司信息安全管理与运行中心项目经理,主要研究方向为信息指纹技术的算法设计及应用系统开发。



杨希(1991-),男,中移(苏州)软件技术有限公司算法工程师,主要研究方向为中文文本领域的自然语言处理、数据挖掘等。

李婧(1988-),女,中移(苏州)软件技术有限公司项目经理,主要研究方向为通信领域的不良信息研究治理。

收稿日期:2018-01-12

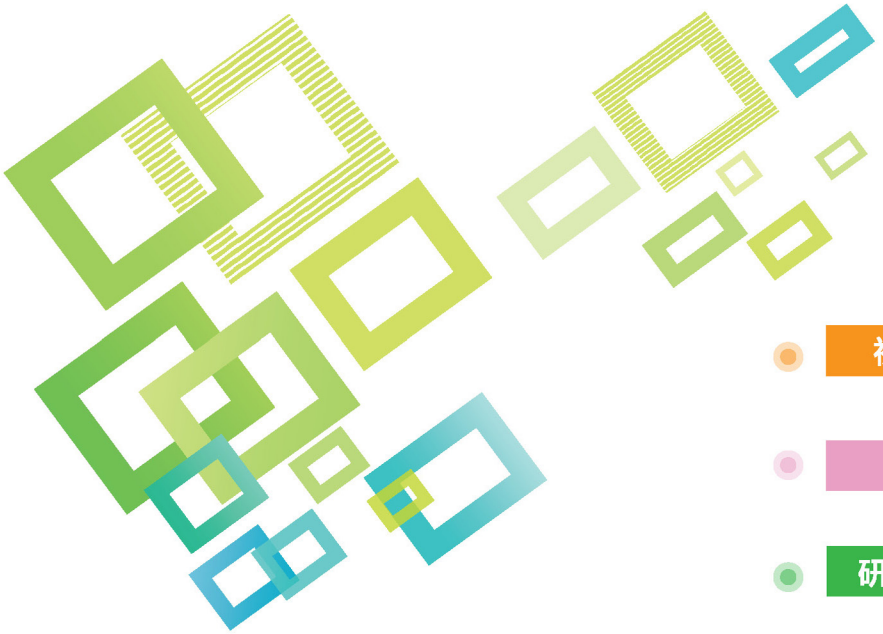
电信科学

Telecommunications Science

www.telecomsci.com

- 中国通信学会和人民邮电出版社携手打造
- 中文核心期刊
- 中国科技核心期刊
- 被国内外核心数据库和科技文摘期刊收录

聚焦自主创新
助推信息通信发展



- 视点聚焦
- 综述
- 专题
- 会议评述
- 研究与开发
- 运营技术广角

邮发代号：2-397 月刊 每月20号出版 2018年共12期 68元/期 全年价816元

编辑部电话：010-81055443/5459/5467

邮局订阅

凭邮发代号，在全国各地邮局（所）订阅

发行部订阅

将订阅信息以传真或邮件方式发送至发行部，也可致电发行部提交订阅信息

电话：010-81055598

传真：010-81055464

E-mail: faxing@bjxintong.com.cn

银行汇款

户名：北京信通传媒有限责任公司

开户行：中国工商银行北京体育馆路支行

账号：0200008109200044661

邮局汇款

北京信通传媒有限责任公司发行部

地址：北京市丰台区成寿寺路11号邮电出版大厦8层

邮编：100078

邮发代号：2-537 国外代号：C9118 定价：35.00元

ISSN 2096-0271

9 772096 027186 03>