BDTC 2017 中国大数据技术大会
Big Data Technology Conference 2017

研究垃圾短信大数据自动识别的新方法

目录

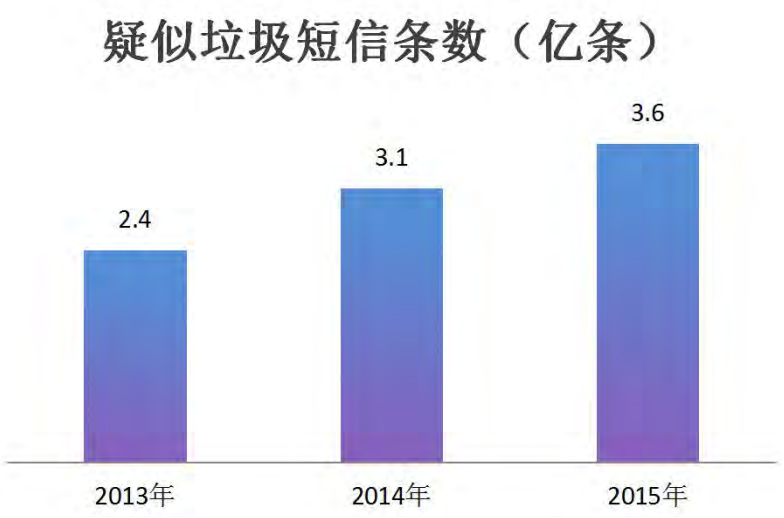
一 项目背景

二 技术方案

三 项目详细内容

四 应用及效果

为履行企业社会责任，保障广大用户通信权益，自2012年以来，中国移动持续开展垃圾短信治理工作。依托垃圾短信集中管控平台，治理主要采用“**系统监测**”模式。随着中奖诈骗类、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，年均疑似垃圾短信高达3亿余条，垃圾短信治理形式严峻。



系统监测：主要指通过“频次+关键词”等过滤方式，筛选得到现网疑似垃圾短信。如：内容含“发票&代开”，1分钟内发送20次等。

若想提升垃圾短信治理效率，亟需引入新的方法。

面临的问题以及现有方案

技术难点：

- **表示稀疏问题**：单条短信内容短小，传统的BOW模型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。
- **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。

短文本分类方面，终端安全公司大多使用以贝叶斯算法为代表的机器学习方法，在终端侧对用户接收到的短信进行识别，将疑似垃圾短信拦截在垃圾箱内。

目前已得到广泛应用。

过滤垃圾短信！

google公司前期提出Simhash算法，将长文本转化为64位的哈希码进行计算、比对。

方法在业界广受好评，但目前仅限于在长文本方面（如网页）应用。

查找重复网页！

目录

一 项目背景

二 技术方案

三 项目详细内容

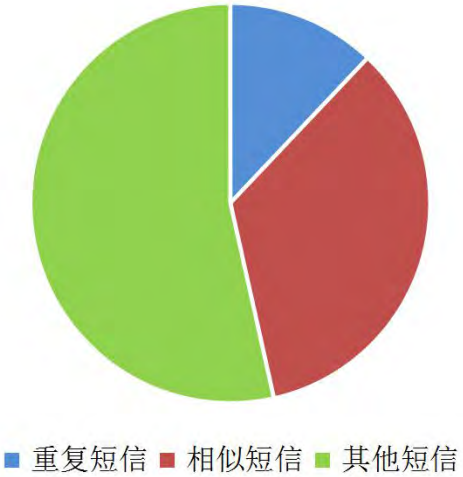
四 应用及效果

技术方案选择

将垃圾短信自动化识别系统成功应用到现网中，关键在于保障接入识别算法的准确率与查全率。

通过对2016年3、4月份历史数据进行抽样分析，约有12%的样本内容完全一致，有34%的样本内容相似，合计比例达46% ！

序号	样本总量	重复样本数量	相似样本数量	重复占比	相似占比
1	753383	73465	249232	9.8%	33.1%
2	946749	155991	322144	16.5%	34.0%
3	967990	118395	291414	12.2%	30.1%
4	790758	91571	291687	11.6%	36.9%
5	1047358	118226	387277	11.3%	37.0%
6	935610	109332	357542	11.7%	38.2%
7	1051106	126231	430548	12.0%	41.0%
8	535883	59017	209265	11.0%	39.1%
9	589327	69271	208354	11.8%	35.4%
10	642357	79160	240023	12.3%	37.4%
均值	898571	108143	305954	12.0%	34.1%



短信样本分析饼图

算法原理对比

鉴于此，以算法准确率与查全率为核心参考指标，我们重点考虑准确率极高的“基于短信内容精确匹配的识别算法”与“基于指纹技术的大数据识别算法”，以及在垃圾邮件处理上已成功获得广泛应用的“基于贝叶斯学习的大数据识别算法”。

	基于短信内容精确匹配的识别方法	基于指纹技术的大数据识别方法	基于贝叶斯学习的大数据识别方法
自动识别率	**仅限完全一致短信，识别率低**	具有匹配相近文本的能力，识别率较高	具有语义识别能力，可以“举一反三”，识别率较高
识别准确率	准确率极高	准确率较高	**准确率较低**
数据库/模型维护复杂度	数据库添加增量，简单	数据库添加增量，简单	模型必须重新生成，复杂
运行效率	**数据库过大，匹配效率低**	数据库可控，运行速度快	数据库可控，运行速度快

关键指标：

自动识别率：等于自动识别出的短信数量在总短信数量中的占比。

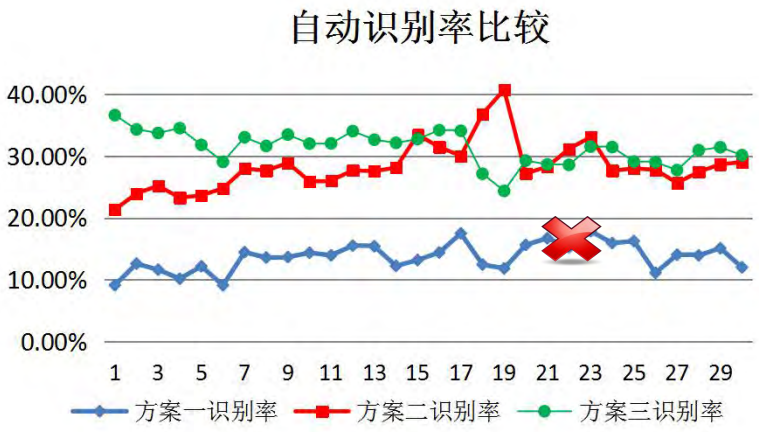
识别准确率：等于新方法识别正确的短信条数与识别的短信总条数之比。

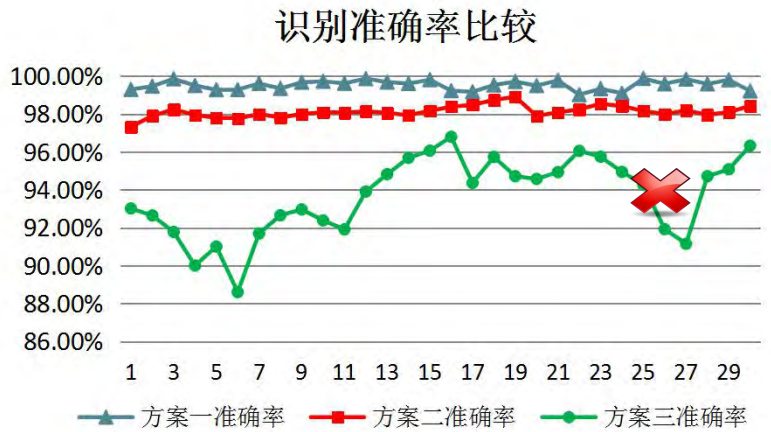
识别率达到30%以上

准确率必须高于99%

算法效果比对

我们使用2016年4月的数据，共计2836万条样本，对三个方案的核心算法进行**模拟测试**，效果对比如下：





三种方案实验比对效果图

实验发现，方案一的自动识别率仅为11.5%，方案三的识别准确率仅为93%，与指标要求差别大。方案二的指纹算法同时具有较好的自动识别率和识别准确率，与原理比对结果一致。项目最终将系统算法锁定为基于指纹技术的大数据识别算法。

目录

一 项目背景

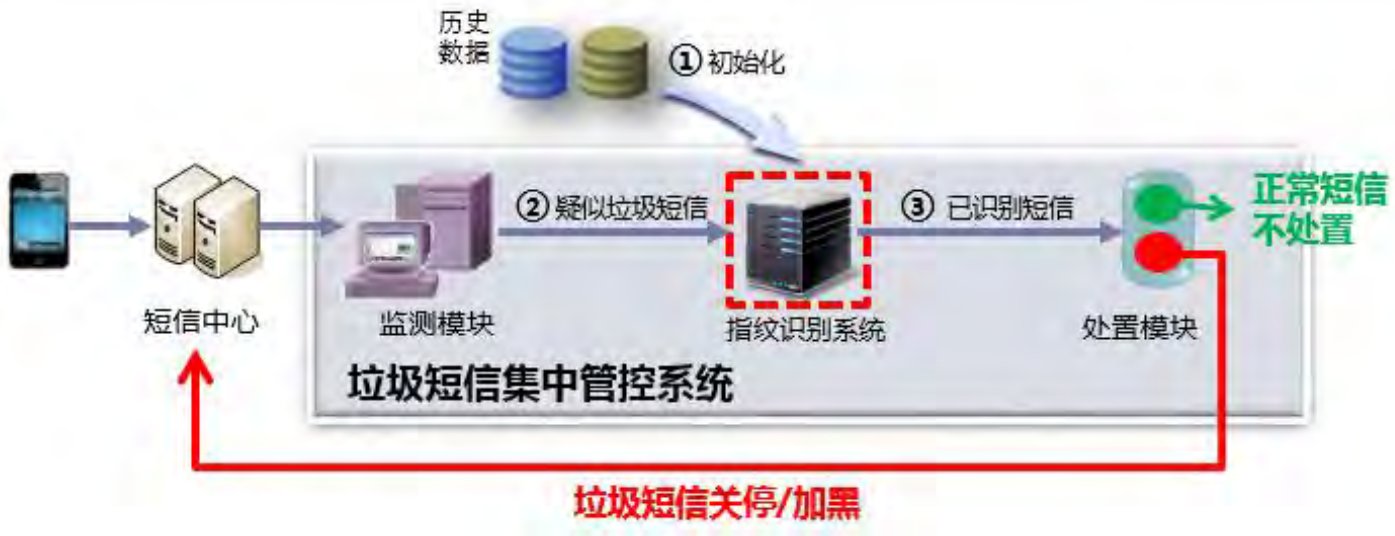
二 技术方案

三 项目详细内容

四 应用及效果

总体流程

在二次过滤模块中央平台中，引入垃圾短信自动化识别系统，用以提升垃圾短信识别率。具体垃圾短信大数据识别应用方案，如下图：



- 利用中移信安中心前期积累的海量短信样本对人工智能分类器和指纹数据库进行初始化；
- 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短信，并进行自动识别；
- 得到识别结果的短信，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；

系统整体架构

基于上述核心算法，项目组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。

- 应用创新指纹算法对待识别短信进行处理；**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
- 创新采用**分布式多机多核系统架构**，通过**kafka**实现内部服务之间的通讯，有效保障了现网的实时运行需求

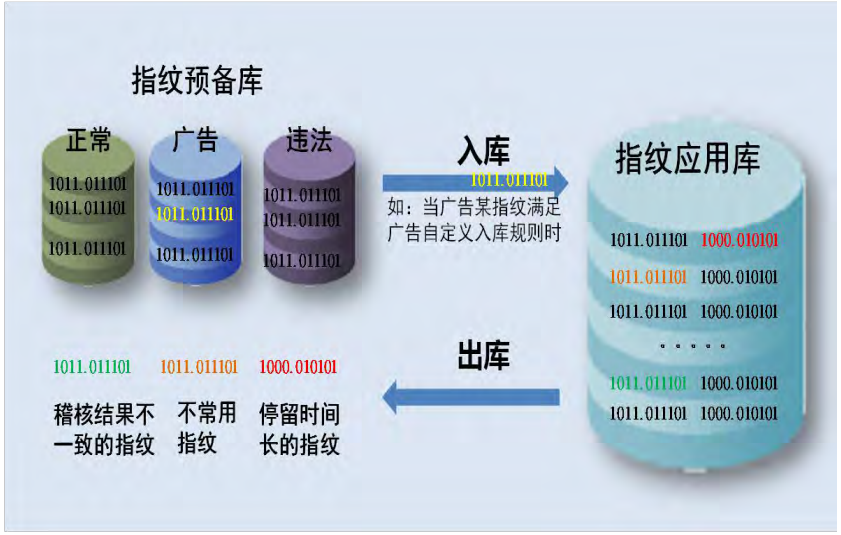
研究团队基于Google公司Simhash开源算法，结合技术应用场景和短信文本短的特点进行重构开发，突破算法准确率等方面的局限性，研发出具有自主知识产权的指纹识别算法。

考虑到中国移动线上治理的极高准确性和性能要求，我们创新提出动态数据库、基于多指纹库识别、指纹筛选与指纹比对分离技术，以达到满足现网应用的要求。

1. 动态数据库机制

在**入库方面**，为指纹算法设计二次入库技术，并采用哈希再散列技术（FNV-1），来降低训练数据冲突造成的影响；

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。



2. 研发多指纹库存储

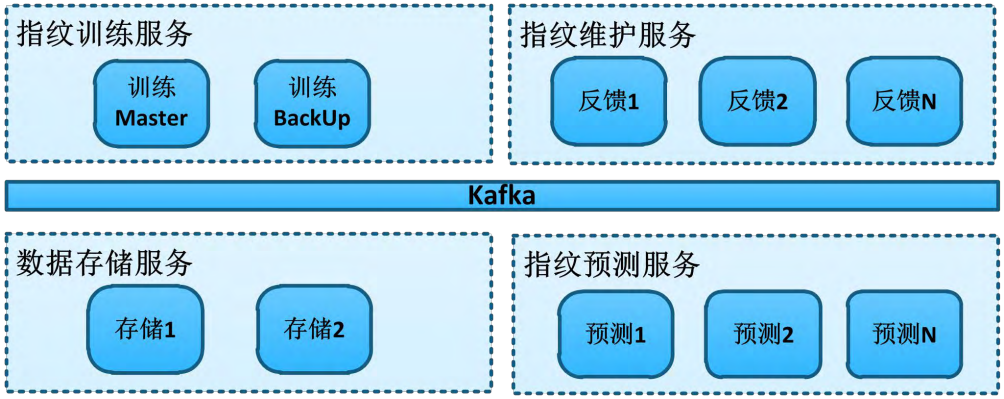
- 搭建**多指纹库**，根据处置方式不同，分为正常短信指纹库、违法诈骗短信指纹库、商业广告指纹库；
- 对违规类型指纹库采用更为严格的**校验入库机制**和优先级更高的**识别反馈机制**；
- 根据考察各指纹库相互冲突指纹，实现对数据库的进一步去噪，降低算法误识别比例。

3. 指纹筛选与指纹比对分离

为了保证分布式模块中预测指纹库的一致性，算法对指纹库的筛选入库（训练）和指纹比对（预测）进行了分离。通过统一的指纹筛选库完成对入库指纹进行筛选，以保证在分布式系统中，所有指纹比对（预测）模块使用的指纹库是相同的。

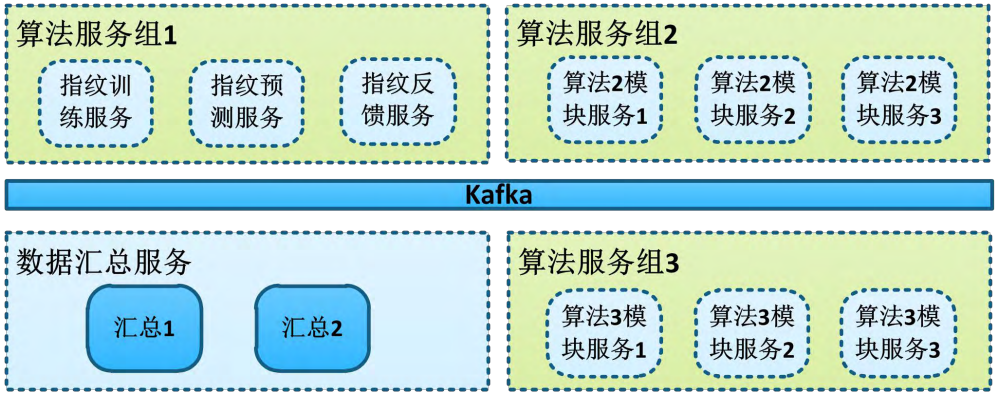
指纹比对与指纹入库的解耦，增强了算法识别部分的可扩展。

4. 在分布式结构上实现了所有模块的服务化



在系统中我们将所有模块进行服务化，模块之间无直接交互，全部通过中间件kafka进行间接的消息传递。这种模式不仅仅实现了服务的解耦，也间接实现了服务的负载均衡。

5. 后续规划与展望



在系统中可以引入多种算法，实现算法融合。通过算法模型的实时更新，能够及时识别现网中的新型垃圾短信（CNN，RNN，LSTM等深度学习算法，word2vec语义扩展等）。令识别系统对垃圾短信的识别更准更全。

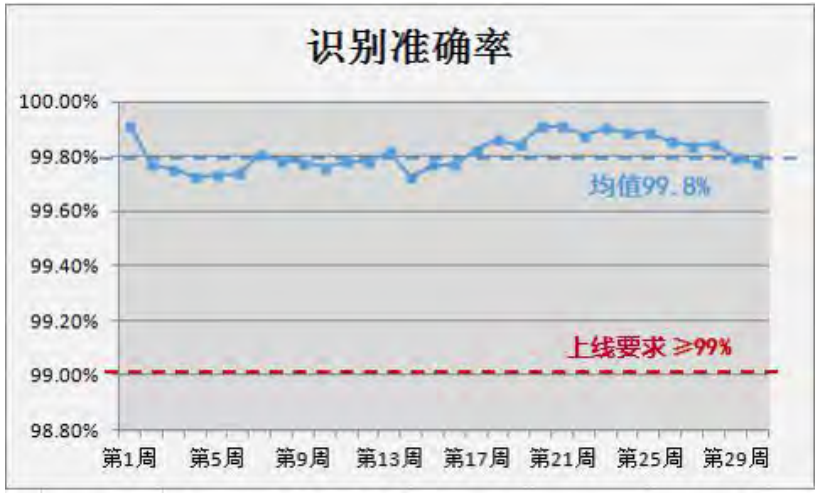
目录

一 项目背景
二 技术方案
三 项目详细内容
四 应用及效果

历时两年，经过6轮44组，累计分析现网数据68亿余件次，分析指标5万余项的大数据测试验证，中国移动垃圾短信大数据自动识别系统于**2017年1月上线**，**覆盖全网31省**。

运行效果

截止目前，系统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4300万条，自动识别率达到**40.1%**，识别准确率**99.8%**，运行效果良好。



社会效益

垃圾短信大数据自动识别系统上线以来，运行状态良好，对线上疑似垃圾短信开展持续治理。在此期间，服务支撑了“党的十九大”、“金砖国家领导人厦门会晤”、“一带一路高峰论坛”等多次重大保障，圆满完成任务，实现了垃圾短信的高效治理。

1. 月均减少垃圾短信近亿条

通过引入大数据识别技术，垃圾短信自动判定平均耗时仅为**0.07毫秒**，违规号码的**关停及时性大大提高。**

系统月均识别违规号码14.4万个，通过估算，可月均减少不法分子发送的垃圾短信约8500万条，有效的保障了广大用户的通信权益！

（ 50条/号码/分钟*6分钟*14.4万个号码/月*2个月=**8500万条垃圾短信** ）

2.垃圾短信投诉同比降低26%

系统上线后，中国移动10086999平台受理垃圾短信投诉同比下降26%，效果明显，保障中国移动垃圾短信治理持续处于行业领先水平。

Thanks!

BDTC 2017 中国大数据技术大会
Big Data Technology Conference 2017