基于流式计算的垃圾短信治理关键技术研究

Research on Key Technologies for Spam SMS Management Based on Streaming Computing

王九九，狄秋燕，马永亮（中国联合网络通信集团有限公司，北京 100033）
Wang Jiujiu，Di Qiuyan，Ma Yongliang（China United Network Communications Group Co.，Ltd.，Beijing 100033，China）

摘　要：
某运营商在现网垃圾短信治理中，常采用关键字+规则的方法，难以在拦截成功率和误拦正常短信之间找到平衡。基于文本语义分析识别垃圾短信，则需要解决大数据挖掘算法、海量数据处理、响应时效等问题，因此在大业务量的集约化平台上应用并不广泛。通过算法研究、开发原型系统等工作，探索基于流式计算的垃圾短信治理技术方案，研发了一套基于Storm+Mahout架构的垃圾短信识别原型系统，完成了性能和准确率测试，取得了较好的效果。

关键词：
垃圾短信治理；自然语言处理；大数据；流式计算
doi：10.12045/j.issn.1007-3043.2024.05.010
文章编号：1007-3043（2024）05-0056-06
中图分类号：TN919
文献标识码：A
开放科学（资源服务）标识码（OSID）：

Abstract：
A certain operator often adopts the method of keyword+rule in the management of spam messages on the current network, which makes it difficult to strike a balance between the success rate of intercepting spam messages and the error rate of normal messages. Based on text semantic analysis to identify spam messages, it is necessary to solve problems such as big data mining algorithms, massive data processing, and response time. Therefore, it is less applied on intensive platforms with large business volumes. It explores a spam message management technology solution based on streaming computing through algorithm research and prototype system development. A spam message recognition prototype system based on Storm+Mahout architecture has been developed, and the performance and accuracy tests have been completed, achieving good results.

Keywords：
Spam SMS management；NLP；Big data；Stream computing

引用格式：王九九，狄秋燕，马永亮. 基于流式计算的垃圾短信治理关键技术研究[J]. 邮电设计技术，2024(5)：56-61.

0 前言

通信技术的进步、移动终端的普及和移动通信网络的能力提升为移动信息服务在中国的推广带来了机遇。根据工信部统计数据，2022年我国移动短信业务总量为118 748亿条，比上年同期增长6.4%，移动短信业务收入为401亿元，比上年同期增长2.7%。其中，个人短信市场收入为22.7亿元，占比5.7%；验证码服务市场收入为131.39亿元，占比32.8%；行业应用短信服务市场收入为102.73亿元，占比25.6%；其他企业短信市场收入为143.98亿元，占比35.9%[1]。近年来，由于微信、line等社交通信技术的快速发展，国内个人短信业务呈现快速下滑态势，而企业短信业务整体维持良好的增长态势。

企业短信是当前国内行业移动信息服务的主要产品形式，它充分满足了金融、交通运输、电子商务、零售商贸、文化传媒、公共服务等行业集团客户的移动信息应用需求。企业短信在为用户提供便捷消息服务的同时，也为信息垃圾的传播提供了一条方便的渠道。垃圾短信不仅影响用户正常通信业务的使用，严重时还可能被犯罪分子利用进行违法活动。因此，垃圾短信作为热点问题一直受到社会各界的广泛关

收稿日期：2024-04-10

注。运营商作为短信的运营和监管者，承受着巨大的社会压力。同时，垃圾短信也严重影响了用户对运营商的信任，破坏了运营商长期建立的良好声誉。

1 垃圾短信治理现状及存在问题

当前垃圾短信拦截的精度和效率之所以不高，主要是由于原有的垃圾短信识别算法本身不足，具体表现在以下2个方面。

a）现网垃圾短信识别主要基于关键字和频次的组合来判断，如果关键字和频次的设定严格，则拦截成功率提高，但同时也会增加用户正常短信的误拦率；反之，若设定宽松，误拦率可以降低，但相应地，垃圾短信的拦截成功率也会下降。这2个关键指标之间存在天然的矛盾。

b）关键字及频次的规则过于固定，容易被不法分子探测并规避。在现网系统中，关键字规则制定的质量很大程度上取决于运维人员的水准，因此系统难以迅速并有效地根据用户投诉自动生成相应的拦截规则。

垃圾短信的识别本质上属于文本处理技术，现有的互联网和大数据技术完全可以应用于垃圾短信的识别，与网页文本处理相比，垃圾短信识别有其特殊之处，例如在进行垃圾短信识别时，需要特别考虑实时过滤的响应及时性等因素。

2 国内外相关工作

基于文本分类的垃圾短信识别，目前国内外通用的做法为构建合理的语料库、正确分词、文本预处理、提取最具统计意义和代表性特征及建立科学高效的过滤模型[5]。国外相关工作的主要发展历程如下：Luhn在1950年提出了词频概念，并将其应用于文本分类，开启了该领域的研究[6]。Maron验证了概率索引与信息检索的相关性，提出了概率模型，极大地推动了早期文本分类技术的发展[7]。Salton等人于1975年提出了文本的空间向量表示模型，文本中具有区分度的关键词表示特征项，文本向量的分量值则表示特征项的权重[8]。Blei等人于2003年提出了隐含狄利克雷模型，应用贝叶斯方法构建了一种基于主题的文本表示模型，将文档集内各子集的主题以概率分布的形式给出[9]。

国内文本分类研究起始于八十年代初，侯汉清首先介绍了国外的分类技术，从而开启了国内中文文本分类的研究[10]。张培颖等人提出了一种基于语义距离的分类方法，该方法将语义信息考虑在内，提高了文本分类的有效性[11]。陈功平等提出了一种基于改进贝叶斯算法的过滤方法，该方法结合黑白名单机制有效减少了误判，其识别率普遍高于基于文本特征的方法[12]。李根等人提出了基于距离特征的自分类簇和自学习算法，该算法能学习新的诈骗信息样本的特征实现自我更新，具有持续识别新加信息的能力[13]。

3 基于流式计算的垃圾短信治理技术方案

3.1 整体思路

基于流式计算的垃圾短信治理技术方案如图1所示。

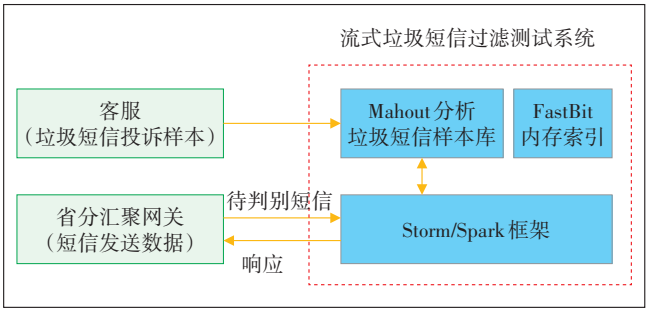

图1 基于流式计算的垃圾短信治理技术方案

本文将现有基于关键词+规则的垃圾短信治理算法，用以下算法模型来代替实现，以提高垃圾短信识别准确率。

a）以文本相似性检测算法来代替纯粹的关键字比对算法，以提高短信内容识别精度。

b）基于客服投诉垃圾短信样本，自动训练用于文本相似性检测算法的垃圾短信库，以确保垃圾短信库更新及时，内容准确，避免了关键字规则制定时可能遇到的质量问题和及时性问题。

本文通过算法创新和应用创新，解决了存在难题，并在模拟测试中取得了很好的效果。

a）在文本相似性检测算法方面，根据短信文本的特点，本文提出并实现基于“分层统计”的朴素贝叶斯算法，通过分层统计短信的词频-逆文档频率（Term Frequency-Inverse Document Frequency，TF-IDF）特征向量，解决了传统贝叶斯分类无法用于由特征向量TF-IDF表示的短信分类过程的难题，该算法不仅显著提高了垃圾短信识别精度，其复杂度也大幅降低。

b）在垃圾短信样本库的训练方面，本文开发了基

于Mahout分布式数据挖掘的短信样本库训练方法，该方法结合文本聚类算法，解决了大数据量的垃圾短信样本库训练的瓶颈问题，为基于文本语义分析的垃圾短信识别提供了高质量样本库。

c）由于上述各种算法均属于高复杂度的算法，本文设计了基于流式计算Storm的分布式计算架构，通过任务分解和分发，弹性调整垃圾短信识别过程中各任务执行的数量和比例，实现了文本语义分析等复杂算法的高并发实时处理。该架构支持集群节点的平滑扩展，能够实现处理能力的动态调整。

3.2 整体技术方案

基于流式计算的垃圾短信治理技术方案，采用Storm+Mahout的架构，实现对垃圾短信实时判别。具体包括以下2个部分内容。

a）基于Mahout的大数据量垃圾短信样本库训练。实现了短信文本向量可视化，即将短信文本转化为以特征向量数字表示的特征向量。采用Mahout分布式Kmeans算法，训练短信样本库，并完成“分层统计”的短信样本库训练。

b）基于流式计算框架Storm实现高并发实时处理。采用Storm框架装载中文词库、短信样本库，并基于Storm实时流处理能力，完成“分层统计”的朴素贝叶斯短信分类。

3.3 关键技术原理和实现

3.3.1 基于“分层统计”的朴素贝叶斯分类算法

朴素贝叶斯算法是文本文档分类算法中较为有效的算法之一，其特点是速度快、效率高、耗费少、应用广泛。由于其稳定性较好、实现简单，且易于开发维护，该算法能够满足手机短信过滤要求[2]。但传统的朴素贝叶斯分类算法在垃圾短信文本相似性方面存在一定的局限性。针对这一问题，本文根据短信文本的特点，提出并实现基于“分层统计”的朴素贝叶斯算法，通过分层统计短信的TF-IDF特征向量，解决了传统贝叶斯分类不足的问题。

3.3.1.1 朴素贝叶斯分类算法的局限性

朴素贝叶斯分类算法相比于KNN（K-近邻）算法，在复杂度上有了很大的降低，但是朴素贝叶斯算法是基于有限集合中的特征进行计算的。在计算短信文本特征向量中，TF-IDF值是一个连续性数据，没法计算不同值出现的概率，因此不能直接使用朴素贝叶斯分类方法。

3.3.1.2 基于“分层统计”的朴素贝叶斯分类算法

针对朴素贝叶斯分类问题，本文提出了“分层统计”方法，该方法将连续的TF-IDF值变为有限的区间集合，通过统计落在不同区间内样本短信的条数，从而统计特征向量不同元素的概率。具体实现方法如下：

a）设计“层数”。层数是本算法的关键点。对短信样本的所有特征向量的TF-IDF值进行统计，根据统计结果的范围划分合理的层数。本文设计了12层，包括（0，0-1，1-2，…，9-10，>10）。

b）统计不同层占比，即短信样本库中“类别内部各层占比”，其内部的格式如图2所示。

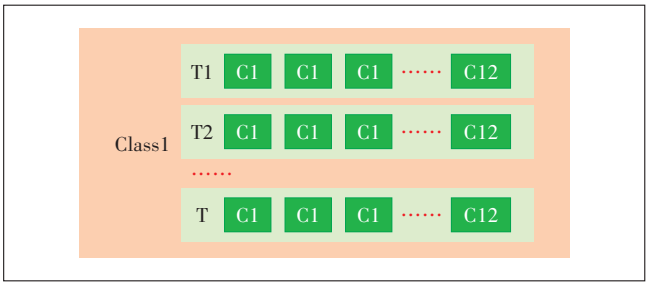

图2 短信样本库某一类别内部组成

对于类别1，T1表示的是该类别中第1个特征元素在各个层级的概率分布。其中T1内部的C1表示的是第1个词（也称为“元素”）的TF-IDF值为0的概率，依次类推。Class1表示的是类别1的整体概率，这个概率是相对于其他类别组成的短信样本总体而言的。

c）对于待分类短信的特征向量，按照之前朴素贝叶斯的计算公式。根据每个元素的TF-IDF值，判断其所属的层，然后获得该元素（词）在该类别的概率。当所有的元素查询完成后，乘以类别总体概率，最终获得该短信属于该类别的概率。

3.3.2 基于Mahout的大数据量垃圾短信样本库训练

在垃圾短信样本库的训练方面，本文设计并开发了一种基于Mahout分布式数据挖掘技术的短信样本库训练方法。该方法结合文本聚类算法，解决了大数据量垃圾短信样本库训练时存在的瓶颈问题。为基于文本语义分析的垃圾短信识别提供了高质量样本库。

3.3.2.1 Mahout样本训练模块

a）原始短信文本和转换程序。原始短信样本文件以txt文本的形式存储在HDFS上，可以用Mahout进行TF-IDF计算和Kmeans聚类训练。先对txt类型的短信样本文件进行转换，并按照<Key，Value>键值对

的形式组合成需要的Sequencefile文件格式。

b）Mahout训练过程。常用的文本特征提取方法包括TF-IDF、互信息（MutualInformation，MI）和信息增益（Information Gain，IG）等[4]。通常采用TF-IDF算法，它是一种文本统计的方法，能够反映单个词在文档集合中的区分度[3]。

所有短信文本先经过文本向量化，表示为特征向量。全部过程如图3所示。

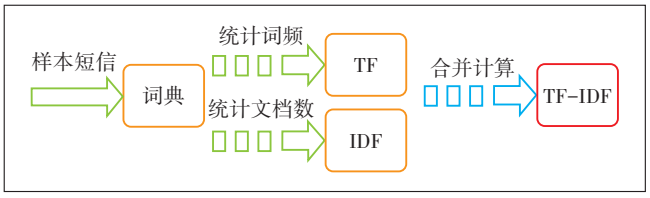

图3　TF-IDF计算过程

词典是通过Mahout训练后转换获得的。在处理每条短信文本时，首先通过中文分词，以获得文本关键词；其次，通过与“词典”对照，将其转换为序列号；再次，合并计算获得该关键词的TF-IDF值，短信所有关键词的TF-IDF值组成的向量即为短信的特征向量。

在短信样本库训练和生成过程中，通常分为2种情况：预先已分类和预先未分类。本文利用Mahout提供的算法，对2种情况都做了支持。在预先未分类情况下，采用Kmeans聚类算法，其中K值由人工设置完成。

Mahout训练过程如图4所示，原始短信文本经过中文分词处理，并被转为Sequencefile文件后，这些文件被输入到Mahout中进行训练，最终生成包括Terms（词典）、IDF（逆文档频率）、TF（词频）、TF-IDF（词频-逆文档频率）的Sequencefile文件。

3.3.2.2　垃圾短信样本库生成模块

采用朴素贝叶斯分类算法，经过“分层统计”将TF和TF-IDF转换为“类别整体占比”和“类别内部各层占比”。垃圾短信样本库的各部分组成如图5所示。

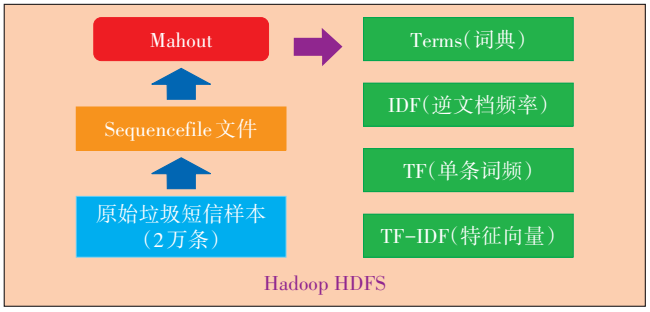

图4　Mahout训练过程及输出结果

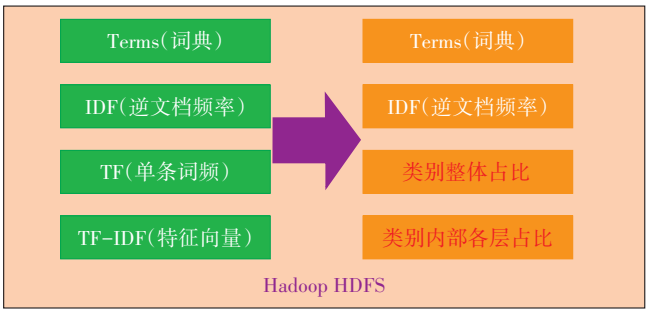

图5　垃圾短信样本库的各部分组成

3.3.3　基于流式计算框架Storm实现高并发实时处理

流式框架Storm技术拥有低延迟、高性能、分布式、可扩展、容错等特性[18]，可以保证消息不丢失，消息处理严格有序[19]。本文设计了一种基于流式计算Storm的分布式计算架构，通过任务分解和分发机制，实现了文本语义分析等复杂算法的高并发实时处理。

3.3.3.1　Storm任务初始化模块

如图6所示，任务被分配在多个“逻辑节点”上，但是资源的调用会涉及到物理节点Supervisor的内存，因此需要在每个物理节点上都加载相关资源。需要加载的资源包括“中文词库”和训练完成的“垃圾短信样本库”。

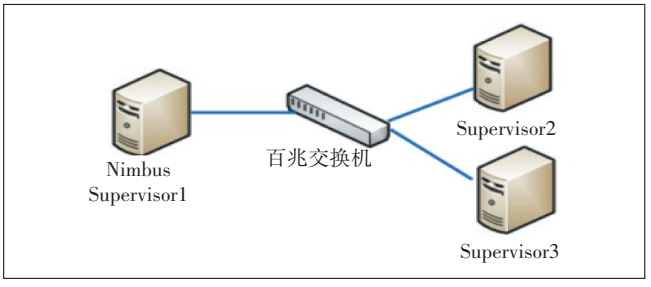

图6　Storm集群网络架构

3.3.3.2　Storm任务实时处理模块

Storm任务启动以后，接收各个短信中心发来的短信文本，对短信文本进行实时处理。流式短信文本经过的处理过程如图7所示。

短信文本实时处理依次经过：中文分词处理、特征向量计算、短信分类3个阶段。短信分类阶段采用基于“分层统计”的朴素贝叶斯分类处理算法。

图8所示为对一条短信的特征向量进行基于分层统计的朴素贝叶斯分类过程。图8中橙色部分表示的是类别1的类整体概率和内部各层概率。特征中的每个元素的TF-IDF值，到该类别中相应的位置查找其所

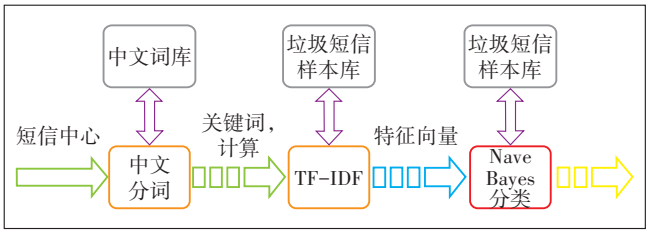

图7 流式短信文本经过的处理过程

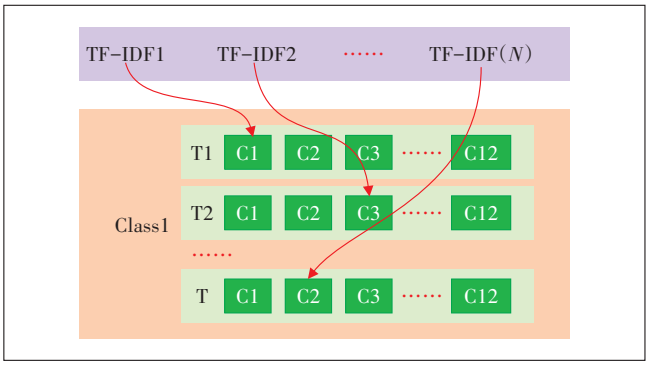

图8 基于分层统计的朴素贝叶斯分类处理过程

属的层并获得概率，将所有的概率与类别总概率相乘即为该短信属于该类别的概率。

3.4 测试结论

3.4.1 测试环境

由3台服务器组成一个Storm集群，其中1台服务器作为主节点运行管理进程Nimbus，同时也作为Supervisor节点；另外2台服务器单独作为Supervisor节点。该集群也运行Hadoop集群，节点分配与Storm相同。Hadoop集群主要用于Mahout训练和存储垃圾短信样本库。

Strom在不同的Supervisor节点，对短信处理的3个过程进行任务封装：

a）短信文本流任务（Spout）：属于Spout类型任务，负责短信实时流入。

b）中文分词任务（Bolt1）：属于Bolt任务，提取短信文本的关键词。

c）短信分类任务（Bolt2）：属于Bolt任务，内部具体完成短信文本向量化，特征向量TF-IDF计算、朴素贝叶斯分类等处理过程。

根据不同任务计算量和处理时间不同，为了避免短板效应，合理分配任务比，Spout∶Bolt1∶Bolt2=1∶4∶2。

3.4.2 测试方法

采用2万条垃圾短信样本用于Mahout训练，获得垃圾短信样本库，包含Terms（词典）、IDF（逆文档频率）、类别总体占比、类别内部各层占比。

短信中心的1 508万条短信作为测试短信，并作为Spout任务短信流输入。Spout任务采用并行模式，每个Spout任务单独的读取测试短信，形成并行流计算。

3.4.3 测试结果

3.4.3.1 垃圾短信样本库结果

Mahout训练并获得的2W条垃圾短信样本、中文词库部分截图如图9所示。

图9第1部分为中文词库，用于中文分词处理；第2部分为Terms（词典），关键词与序号对应关系；第3部分为IDF（逆文档频率），词典中所有词的逆文档频率；第4部分为类别1的内部各层占比。

3.4.3.2 物理节点负载情况

按照之前任务分配方式运行本文的垃圾短信治理软件，通过监控各物理节点得到各物理节点资源占用情况，具体如表1所示。

从表1可以看出，Supervisor3节点在CPU、内存、网络等资源的使用上都比其他节点要高。尤其CPU

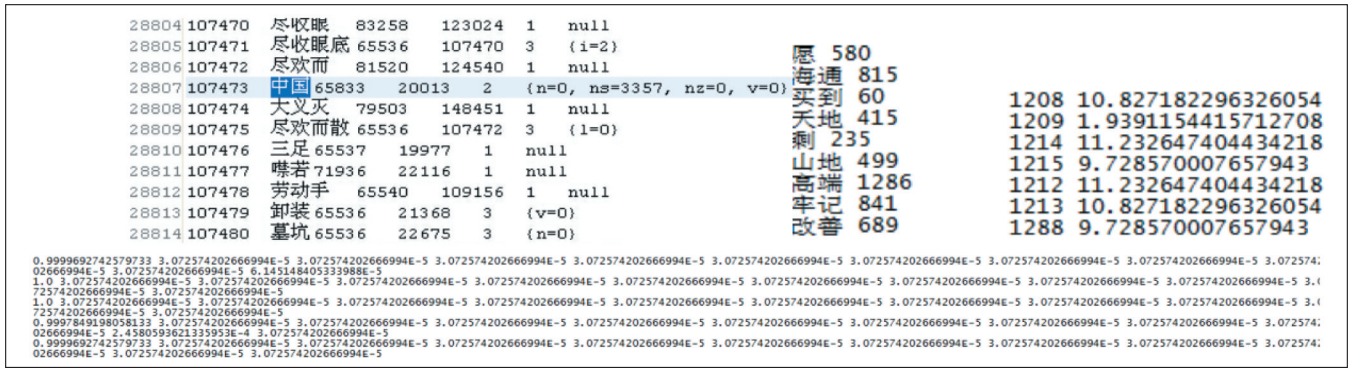

图9 训练完成的短信样本库各组成部分截图

表1　Supervisor各物理节点负载情况指标

Node	CPU/%	Memory/G	Net/M
Supervisor1	86.70	22.23	8.30
Supervisor2	81.44	21.91	5.80
Supervisor3	97.10	24.86	12.33

和网络带宽都基本已经达到资源上限。从任务分配情况来看，Supervisor3上执行了3个任务。包括2个计算复杂度较高的Bolt1和Bolt2，因此CPU资源使用率较高。单个Spout任务运行在Supervisor节点上，所有的短信文本流入都通过Supervisor3，造成带宽占用较高，已经达到百兆资源上限。

从物理节点负载监控结果来看与预期结果相一致，从3个节点的状态来看，Storm7个任务已经基本达到了物理节点的处理峰值，也基本上避免了单个节点负载超限的问题。

3.4.3.3　短信处理结果

对比实验，设置不同任务总数和分配方式，测试系统所能支持的短信最大处理能力。实验结果如表2所示。

表2　测试结果对比

实验	Spout/个	Bolt1/个	Bolt2/个	Time/min	Total/条	Speed/(条/s/台)
提前过滤(12)	3	6	3	26.7	40 476 680	8 422
不提前过滤(12)	3	6	3	29.0	40 476 680	7 754
提前过滤(7)	1	4	2	6.8	15 079 520	12 319

提前过滤是指对于其中一部分在中文分词词库中，却不在短信分类词典(Terms)中的关键词进行过滤。从实验结果可以看出，在Storm任务数为7，任务配比为1∶4∶2的情况下，本系统3个虚拟机集群达到最大处理能力。每个虚拟机节点平均处理能力为12 319条/s。按该运营商全国短信中心峰值10~20万条/s计算，仅需20台虚拟机集群即可满足需求。

4　基于流式计算的垃圾短信治理应用前景

本文提出一种基于流式计算的垃圾短信治理技术方案，并采用Storm+Mahout架构构建垃圾短信识别的原始系统。在实验室测试环境中，单节点处理能力达到1.23万条/s以上；识别精度达到99.5%以上，并支持平滑扩展。经该运营商集约化垃圾短信治理方案评估，相比于传统垃圾短信治理技术，预计将为该运营商节省80%以上的投入，同时将大幅提高该运营商垃圾短信识别的精度和速度，未来在现网投入使用后，将会极大改善现有垃圾短信现状，提升中国联通的品牌效益，保障用户的安宁权，体现运营商社会责任。同时，该技术方案的成功落地实施，也将为其他运营商集约化垃圾短信平台建设提供有益借鉴。

参考文献：

[1] 智研咨询. 2023-2029年中国企业短信服务行业市场专项调研及投资前景规划报告[EB/OL]. [2024-01-02]. https://www.sohu.com/a/611020222_120956897.
[2] 金小梅，毛本清. 基于改进贝叶斯算法的垃圾短信过滤研究[J]. 科技与创新，2019(6)：21-23.
[3] 刘鑫，王皓晨，黄宇煦. 基于朴素贝叶斯分类的电信诈骗信息的识别[J]. 计算机时代，2023(4)：29-32，38.
[4] 赵卫东，董亮. 机器学习[M]. 北京：人民邮电出版社，2018.
[5] 李围围. 垃圾短信识别的探索与研究[D]. 柳州：广西科技大学，2018.
[6] United States. National Bureau of Standards. Auto-encoding of documents for information retrieval systems[A]. Yorktown Heights：IBM Research Center，1958.
[7] MARON M E，KUHNS J L. On relevance，probabilistic indexing and information retrieval[J]. Journal of the ACM，1960，7(3)：216-244.
[8] SALTON G，WONG A，YANG C S. A vector space model for automatic indexing[J]. Communications of the ACM，1975，18(11)：613-620.
[9] BLEI D M，NG A Y，JORDAN M I. Latent dirichlet allocation[J]. The Journal of Machine Learning Research，2016，3：993-1022.
[10] 侯汉清. 分类法的发展趋势简论[J]. 情报科学，1981(1)：58-63，30.
[11] 张培颖，王雷全. 基于语义距离的文本分类方法[J]. 计算机技术与发展，2013，23(1)：128-130，134.
[12] 陈功平，沈明玉，王红，等. 基于内容的短信分类技术[J]. 华东理工大学学报(自然科学版)，2011，37(6)：770-774.
[13] 李根，王科峰，贲卫国，等. 基于自分簇自学习算法的垃圾短信识别[J]. 吉林大学学报(信息科学版)，2021，39(5)：583-588.
[14] 王继重. 基于Hadoop和Mahout的K-Means算法设计与实现[D]. 大连：大连海事大学，2016.
[15] 周煜敏，王鹏，汪卫. 基于Storm的实时大规模传感器监控平台的开发和实现[J]. 计算机应用与软件，2019，36(12)：7-11，28.
[16] 蔡宇，赵国锋，郭航. 实时流处理系统Storm的调度优化综述[J]. 计算机应用研究，2018，35(9)：2567-2573.

作者简介：
王九九，毕业于北京交通大学，硕士，主要从事网络技术、业务IT一体化相关工作；狄秋燕，毕业于北京邮电大学，硕士，主要从事数据管理相关工作；马永亮，毕业于北京邮电大学，硕士，主要从事业务IT支撑相关工作。