基于子语义空间的挖掘短文本策略方法

孙洋，粟栗，张星，王峰生，杜海涛
（中国移动通信有限公司研究院，北京 100032）

摘　要：为解决精准识别短文本数据的问题，提出一种基于子语义空间的短文本策略挖掘方法。该方法首先采用语义空间技术，解决短文本在分析过程中存在的“词汇鸿沟”与“数据稀疏”问题；然后基于聚类算法将语义空间划分为多个子语义空间，在各子语义空间并行挖掘关联规则，提高了策略生成的效率与质量；最后利用二叉树进行策略归并，生成最简策略集。实验证明，与传统的分类模型相比，该方案生成的策略集在误报率为 6.5%的情况下，准确率可达 88%。在违规短信的发现处理中，使用该技术挖掘的策略集，覆盖能力强、准确率高，具有很强的实用性。

关键词：子语义空间；策略提取；短文本；关联规则挖掘；聚类
中图分类号：TP391.4
文献标识码：A
doi: 10.11959/j.issn.1000−0801.2020061

Method of short text strategy mining based on sub-semantic space

SUN Yang, SU Li, ZHANG Xing, WANG Fengsheng, DU Haitao
China Mobile Research Institute, Beijing 100032, China

Abstract: To solve the problem of identifying short text data accurately, a method of short text strategy mining based on sub-semantic space was proposed. Firstly, semantic space technology was used to solve the problem of “vocabularygap” and “data sparseness” in short text analysis. Then, based on clustering algorithm, the semantic space was divided into several sub-semantic spaces, and association rules were mined in the sub-semantic space, which improved the efficiency and quality of strategy generation. Finally, binary tree was used to merge strategies and generate the simplest strategy set. Experiments show that compared with the traditional classification model, the accuracy rate of the strategy set generated by the proposed scheme can achieve 85% when the false positive rate is 6.5%. In the processing of illegal short messages, using this technology to mine potential policy sets has strong coverage ability, high accuracy and strong practicability.

Key words: sub-semantic space, strategy extraction, short text, association rule mining, clustering

收稿日期：2019−12−12；修回日期：2020−03−06
基金项目：教育部-中国移动科研基金项目（No.MCM201805-2）
Foundation Item: Ministry of Education-China Mobile Research Fund (No.MCM201805-2)

1 引言

短文本策略提取算法是从海量短文本数据提取关键词并生成关联关系，目前针对策略提取的技术研究主要分为 3 类：基于统计的方法，基于词语网络的方法和基于语义、本体的自然语言理解的方法。

- 基于统计的方法，如参考文献[1]中提出的基于 TF-IDF 特征等计算方法，该类方法具有模型泛化性，但忽略了词汇语义的内在联系，如 *N*-gram 模型，限于计算能力，目前使用最多的为二元（Bi-gram）和三元（Tri-gram）两种方式，参考文献[2]在 *N* 元模型基础上引入了马尔可夫模型，增加了缓存部分的改进模型，短文本自身的特点导致应用效果比较差。
- 基于词语网络的方法，根据一定规则将文档映射为词语网络，利用词语网络计算词语的关键度，参考文献[3]中基于共现网络的关键词提取算法 TextRank，通过词语共现窗口构建共现网络，并根据共现网络计算词语的得分；参考文献[4]中的词语共现图模型，利用词间的共现强度作为图划分的依据，此时每个簇对应某个基本观点的连通子图，从而构成了某个特定主题，这样簇内的词汇形成几组策略词组，此种方法的弊端是在海量数据的情况下，词语网络无法获得清晰有效的簇集。
- 基于语义、本体的自然语言理解的方法，利用概念间的同义词、近义词关系，如基于“知网”[5]和哈尔滨工业大学信息检索实验室修订的《同义词词林扩展版》获得同义词，这种方法受限于字典本身的完善程度。

本文提出一种基于子语义空间挖掘短文本策略的方法，应用于违规短信识别过程中，构建语义空间并基于此关联规则挖掘生成策略集，此方法生成的策略集从语义层面出发，不仅考虑了词汇间的共现关系，能够挖掘出词汇间潜在的关联关系，具有很好的策略扩展能力，同时基于机器学习的算法可以自动学习和调优，提高了判定的准确率，降低了漏拦率。

2 技术基础

2.1 语义空间

为了计算词汇间的关系，传统的方法采用“词袋”将词汇映射为向量空间（VSM）中的点，其中 VSM 的维度等于词汇的总数 *N*，动辄会达到上万或者百万维度。而且短文本由于字数较少、表达随意等自身特点，在挖掘策略过程中凸显出两个严重的问题：“词汇鸿沟”问题，文本在分析过程中，每个词汇转换成“词袋”映射到空间向量的点，表达很随意，无法很好地刻画词与词之间的相似性；“数据稀疏”问题，短文本的字数少而且使用关键词表示后，每条数据只有几个关键词，计算过程中会引起严重的数据稀疏问题。比如“苹果手机”和“iPhone”两个意义相近的词汇，映射到向量空间后词汇关联性变弱：

- “苹果手机”表示为 $[0\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,\cdots 0]_{1\times N}$；
- “iPhone”表示为 $[0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,0\,0\,0\,0\,0\,\cdots 0]_{1\times N}$。

为了解决以上问题，引入了语义空间的概念，也叫词向量空间[6]，其中语义空间可以设置为固定的 *M* 维度，词汇转化成 *M* 维的词向量表示，这样不仅大大缩减了向量空间维度，减少了计算量，而且从语义层面计算词汇关联性，提高了分析的准确率。语义空间[7-8]的本质为：将向量每一个元素由整数改为浮点数，变为整个实数范围的表示；将原来稀疏、巨大维度的向量空间压缩嵌入一个稠密、低维的语义空间。上述“苹果手机”和“iPhone”两个词，在 *M*（如 *M*=200）维语义空间内词向量表示如下：

- “苹果手机”表示为 [0.01 0.07 0.96 0.32 0.24 …]$_{1\times M}$；
- “iPhone”表示为 [0.02 0.24 0.84 0.45 0.01…]$_{1\times M}$。

2.2　关联规则挖掘

关联规则挖掘[9]是数据挖掘的一个经典无监督学习算法，可以发现事务数据集中项集之间的相互依存和关联关系，如果两项或者多项属性之间存在关联，那么其中一项的属性可以依靠其他属性值进行预测。比如经典案例“啤酒与尿布”，购买啤酒的人在一定概率下会购买尿布，{啤酒}→{尿布}就是一条关联规则。

- 项集：数据库中不可分割的最小单位信息称为项，用符号 i 表示；项的集合称为项集，用 I 表示。
- 事务：所有的数据统称为事务，每个事务是一个非空项集。
- 关联规则：关联规则 $x\rightarrow y$ 的蕴含式。其中 x、y 都是 I 的真子集，并且 $x\cap y=\phi$，x 称为前提，y 称为结果。关联规则反映 x 中的项目出现时，y 中项目也跟着出现的规律。
- 项集的支持度：包括项集的事务个数称为项集的支持度计数（项集的频数）。
- 最小支持度 min_support：项集在所有训练元组中同时出现的最小次数。
- 关联规则的支持度（support）：关联规则的支持度是交易集中包含 x 和 y 的事务的并集和所有事务之比：

$$\text{support}(x\rightarrow y)=p(x\cup y) \tag{1}$$

- 关联规则的置信度（confidence）：关联规则的置信度是交易集同时包含 x 和 y 的事务与所有包含 x 的事务之比：

$$\text{confidence}(x\rightarrow y)=\frac{\text{support}(x\cup y)}{\text{support}(x)}=p(y\mid x) \tag{2}$$

关联规则挖掘问题可以分为两个子问题：

- 找出数据库中所有大于或等于指定的最小支持度的频繁项集；
- 利用频繁项集生成所需要的关联规则，根据设定的最小可信度筛选出强关联规则。

3　基于子语义空间挖掘违规短信策略词

基于子语义空间挖掘策略的方法，针对短文本策略提取由两个步骤组成：关键词提取和关联关系组合。首先从短文本样本中提取关键词，然后确定这些关键词的关联关系（与、或、非关系）形成策略，利用这些挖掘的策略识别更多的目标短文本，本文将其使用于违规短信的识别过程中。

主要过程分为 5 步，系统流程如图 1 所示。

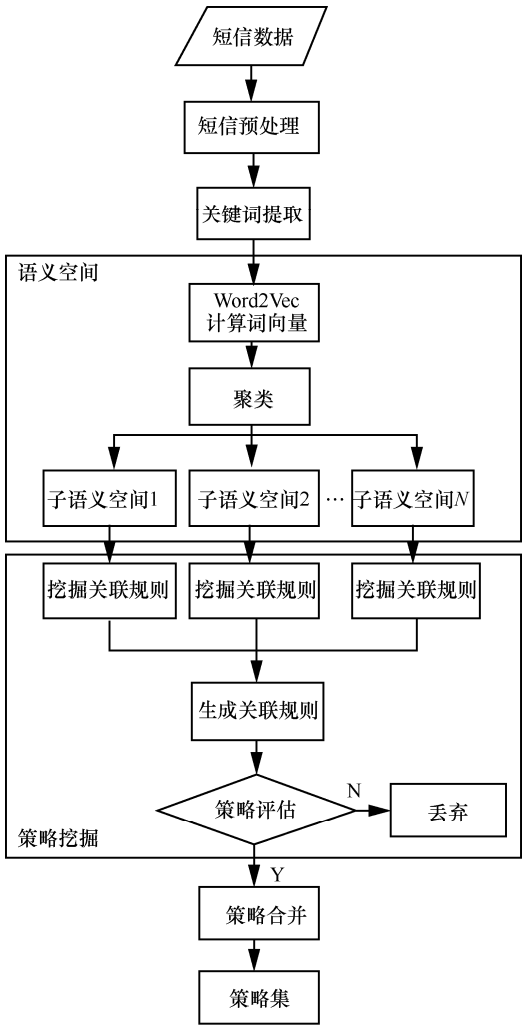


图 1　策略集生成系统流程

步骤 1 对短信样本进行预处理，提取关键词；

步骤 2 构建关键词的语义空间，并基于语义空间进行聚类，生成子语义空间；

步骤 3 针对各子语义空间进行并行挖掘，生成关系明显及具有潜在关系的关联规则；

步骤 4 利用训练语料评估关联规则生成策略；

步骤 5 构建策略二叉树压缩合并策略，形成精简策略集。

3.1 步骤 1

（1）短信预处理

短信预处理[10]的主要功能是对含有干扰项的短信样本按规则进行归一化处理，如繁体字、数字序列、干扰符号等，并按匹配转换规则将原始信息样本转换为统一的文本序列。短信样本为：

- 本#公司*代开各种 Fa 票，代刻公章，电联 13② 88 0 57&898；
- 你的手机号码已选为《淘宝周年慶》获奖用户，获得 1 陆万奖金登陆 tkeux.c0m 验证码：7768 领取。

处理方法如下：

- 处理数字：按照数字对照表进行转换，然后根据上下文将相对连续的数字序列提取出来，如陆→6、②→2 等；
- 处理干扰字符：将待测短信中的#、&、%、α、β、я、ы、щ、■、●、★和※等特殊符号剔除；
- 处理繁简体：提供繁简体的对照表，将繁体转换成简体，如周年慶→周年庆。

（2）分词并去除停用词和无用词

对预处理后的短信进行如下操作：首先采用分词工具，生成分词和词性标注结果；然后按照停用词表删除停用词、无用词，一个句子中能够真正表达句子含义的词为名词、名词性短语、动词和动词性短语等，电话号码、银行卡号等在策略提取中的意义较小，可以通过正则等方式识别并使用统一的词汇进行替换，最后只保留上述 4 种词性的词汇。

（3）关键词提取

此方法的目的剔除超高频和超低频的词，对关键词的提取精度要求不必太高，可以利用 TF-IDF 计算词汇权重。TF-IDF 是一种用以评估词汇对于语料库的重要程度的统计方法，词汇在一篇文章中出现的频率高，并且在其他文章中很少出现，则认为此词或者短语具有很好的类别区分能力，适合用来分类。按照词汇权重由高到低排序，根据需求的关键词个数或者设置阈值获得指定的关键词，其中 TF-IDF 词汇权重计算式如下：

$$W_i = \mathrm{TF}_i \times \mathrm{IDF}_i \tag{3}$$

其中，TF_i 为词汇 W_i 出现的频次，$\mathrm{TF}_i = \frac{\text{词汇}W_i\text{出现的频次}}{\text{样本的总词汇数}}$。$\mathrm{IDF}_i$ 为倒排文档频次，$\mathrm{IDF}_i = \ln\frac{N}{N(W_i)+1}$，$N$ 为样本总数，$N(W_i)$ 为包含 W_i 的文章总数。

将训练集内所有短信数据按照以上 3 个步骤处理，使用关键词进行特征化表示。短文本中词汇即使多次出现，对句子的影响也很小，所以只保留一次。

3.2 步骤 2

利用 Word2Vec 思想将步骤 1 提取出的关键词转化为词向量并构建语义空间，此时如果直接针对整个语义空间挖掘关联规则，不仅数据量大，而且空间内所有词向量并非都存在关联关系，容易造成挖掘速度慢或者内存溢出等问题。所以采取切割语义空间的思想，对大数据“分而治之”，本文利用聚类算法划分语义空间，将语义上相近的词向量聚类形成子语义空间，将子语义空间内相似度不小于阈值 θ（θ=0.98）的词汇生成“或”关系（如 $A_1|A_2|A_3\ldots|A_n$），即自动生成近义词词典。子语义空间部分词汇见表 1。

表 1　子语义空间部分词汇

类别	词汇
推广类	好礼|豪礼、大放送、赔钱|亏本|贱价、火热|火爆、喜悦、赔钱|贱卖|赔本|亏本、相思|思念、传达|传送|传递、便民、主流、粉丝
欺诈类	房租|租金|房费、卡上|卡里|卡内|账户|账上|账号、房东、幸运、评选|选中|抽选|抽中|中选、旺肖|平肖|特肖、给料|发料、办证、刻章、发票、积分
违法类	遇难、法轮、谎言、机缘、彷徨、劫难|灾难、佛法、罪恶|罪行、镇压、功学、保佑、中共|共产党|中央
涉黑类	流连忘返|乐不思蜀、排列、输钱、百家乐|六合采|六合彩|赌球|打麻将、随时随地|足不出户|不出门、赌博|博弈、间断、点子|运到|运气、先知、赌王、分金、储直|充值、http|www.|.com|.c0m|.cn|.cc|.hk|.net|.apk|m.|.lt

表 2　子语义空间内策略词顺序

策略词	特征 ID	支持度
百家乐|六合采|六合彩|赌球|打麻将	T1	60
赌博|博弈	T2	60
http|www.|.com|.c0m|.cn|.cc|.hk|.net|.apk|m.|.lt	T3	60
点子|运到|运气	T4	40
分金	T5	20
排列	T6	20
输钱	T7	20
先知	T8	20
随时随地|足不出户|不出门	T9	10

3.3　步骤 3

对各子语义空间使用 FP-Growth 关联规则算法，并行挖掘频繁项集生成策略。

（1）统计一级频繁项集

扫描违规短信训练数据，统计每个子语义空间内的关键词出现的频次，计算一级频繁项集。违规和正常短信区别之一为：在指定的时间窗口内违规短信重复发送量至少是正常短信的十几倍，所以设置一级频繁项集最小支持度 min_support=10、置信度 confidence=80%可以减小误命中正常短信的概率，其中大于最小支持度的关键词作为策略词，按照词频进行排序，顺序见表 2。

（2）转换训练短信数据形成数据记录表

每条短信转换成策略词及其相应的特征 ID，每条记录的支持度为短信内各策略词支持度的最小值，见表 3。

$$\text{Supportsentence} = \min_{W_i \in \text{sentence}} \{\text{Support}(W_i)\} \quad (4)$$

（3）利用 FP_Growth 挖掘关联规则

扫描数据记录表，生成 FP-tree 多叉树，结构如图 2 所示，应用 FP-Growth 算法挖掘该树的频繁项集，关联规则见表 4。

3.4　步骤 4

（1）评估关联规则

生成的关联规则包含 2~N 个策略词，如下所示：

“<百家乐|六合采|六合彩|赌球|打麻将，排列>、

<百家乐|六合采|六合彩|赌球|打麻将，赌博|博弈，排列>、

…

<百家乐|六合采|六合彩|赌球|打麻将，赌博|博弈，http|www.|.com|.c0m| .cn|.cc|.hk|.net|.apk|m.|.lt，点子|运到|运气>”

表 3　数据记录表

短信 ID	策略词表示	特征 ID 表示	支持度
M1	百家乐|六合采|六合彩|赌球|打麻将,赌博|博弈,http|www.|.com|.c0m|.cn|.cc|.hk|.net| .apk|m.|.lt, 随时随地|足不出户|不出门	T1,T2,T3,T4	40
M2	百家乐|六合采|六合彩|赌球|打麻将, 分金, 排列	T1,T5,T6	10
M3	百家乐|六合采|六合彩|赌球|打麻将, http|www.|.com|.c0m|.cn|.cc|.hk|.net|.apk|m.|.lt,分金, 排列	T1,T3,T5,T6	10
M4	赌博|博弈,输钱,先知,随时随地|足不出户|不出门	T2,T7,T8,T9	10
M5	赌博|博弈, http|www.|.com|.c0m|.cn|.cc|.hk|.net|.apk|m.|.lt ,输钱,先知	T2,T3,T7,T8	10

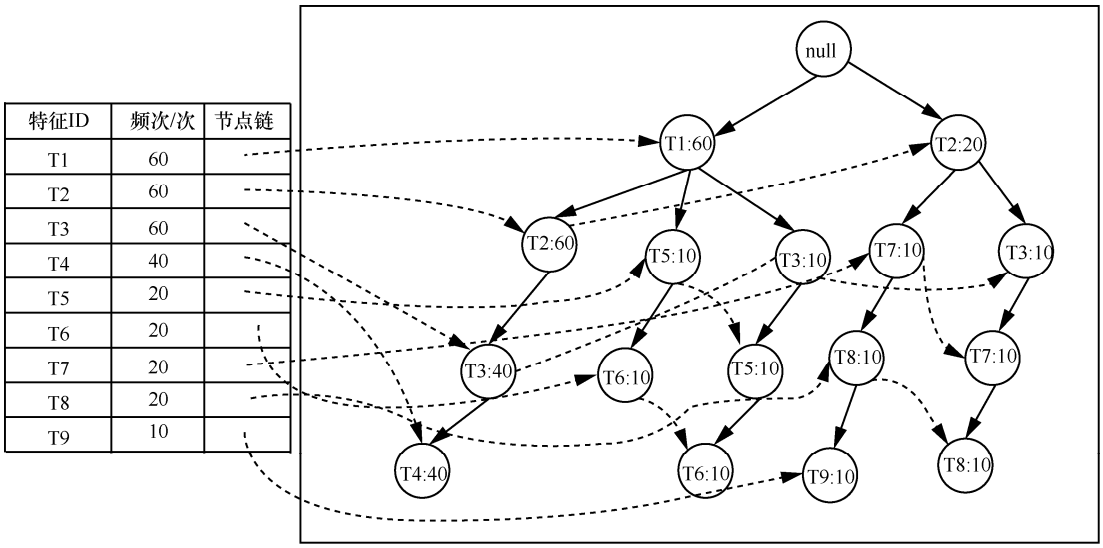


图 2　FP-tree 树结构

表 4　关联规则

ITEM_ID	模式基	条件 FP_TREE	关联规则
T9	{(T2 T7 T8:10)}	<T2:10 ,T7:10 ,T8:10>	T2T9:10,T7T9:10,T8T9:10,T2T7T9:10,T2T8T9:10,T2T7T8 T9:10
T8	{ (T2 T7:10)	<T2:10, T7:10>	T2T8:20,T7T8:10,T2T7T8:20,T3T8:10,T2T3T7T8:10
	(T2 T3 T7:10)}	<T2:10, T3:10, T7:10>	
T7	{(T2:20)}	< T2:20 >	T2T7:20
T6	{(T1 T2 T5:60)	<T1:60, T2:60 ,T5:60>	T1T6:70,T2T6:70,T5T6:70,T1T2T6:70,T1T5T6:70,T2T5T6:70,T1T2T5T6:70,
	(T1 T2 T3 T5:10)}	<T1:10,T2:10,T3:10,	T1T2T3T6:10,T2T3T5:10,T1T3T5:10,T1T2T3T5T6:10
		T5:10>	
T5	{(T1 T2:10)	<T1:20,T2:20,T3:10>	T1T5:20,T2T5:20,T3T5:10,T1T2T5:20,T1T3T5:10,T2T3T5:10,T1T2T3T5:10
	(T1 T2 T3:10)}		
T4	{(T1 T2 T3:40)}	<T1:40,T2:40,T3:40>	T1T4:40,T2T4:40,T3T4:40,T1T2T4:40,T1T3T4:40,T2T3T4:40,T1T2T3T4:40
T3	{(T1 T2:40)	<T1:50,T2:50><T2:20>	T1T3:50,T2T3:70,T1T2T3:50
	(T1 T2:10)(T2:20)}		
T2	{(T1:60)}	<T1:60>	T2T1:60

评估规则过程中应注意如下问题：

- 规则内的策略词个数过少会导致误命中很多正常短信；
- 规则内的策略词个数过多，会导致违规短信虽然命中精准但是命中量少，抑制了规则的作用；
- 有些策略词组是中性词汇，如“<先知，随时随地|足不出户|不出门>”，容易误命中正常短信，不适合作为过滤垃圾短信的策略。

（2）关联规则转换为策略

一条策略由 *N* 个策略词及 3 种关联关系{“与

（&）”、“或（|）”、“非（!）”}组成，表现形式为<（A_1|A_2）&（B_1|B_2|B_3）&…!(F_1|F_2|F_3)>，其中转换后的形式如下：

“<(赌博|博弈)&分金>

…

<(百家乐|六合采|六合彩|赌球|打麻将)&(赌博|博弈)&排列>”

3.5　步骤 5

评估策略后发现存在大量的策略包含、交叉等情况，如：“0001 T1&T6、0002 T2&T6、0003 T1&T2&T6、0004 T1&T5&T6、0005 T1&T2&T5&T6”。

通过构建策略二叉树的形式，将策略进行有效的合并、压缩形成精简策略集，构建二叉树的过程如下。

（1）统计元素频次：首先统计每条策略内的每个元素的频次，按照由高到低排序{ T6：5；T1：4；T2：3；T5：2}。

（2）策略按照元素频次调整顺序

各个元素按照频次排序后的策略顺序进行调整对比，见表 5。

表 5　策略按照频次调整顺序对比结果

策略 ID	排序前	排序后
0001	T1&T6	T6&T1
0002	T2&T6	T6&T2
0003	T1&T2&T6	T6&T1&T2
0004	T1&T5&T6	T6&T1&T5
0005	T1&T2&T5&T6	T6&T1&T2&T5

（3）构建二叉树

再次扫描按照频次调整顺序的策略表，构建一棵有序二叉树。构建原则为所有策略都从 root 节点出发，其中孩子节点在树的左侧，兄弟节点在树的右侧，各条策略的最后一个元素附带策略 ID，每个节点为三元组<策略词，ID，状态>，其中作为策略的最后一个节点带有 ID 和状态，状态有两种：{0：未提取、1：已提取}，已经进行了策略合并的元组，则节点状态标记为 1，避免二次生成策略，策略二叉树结构如图 3 所示。

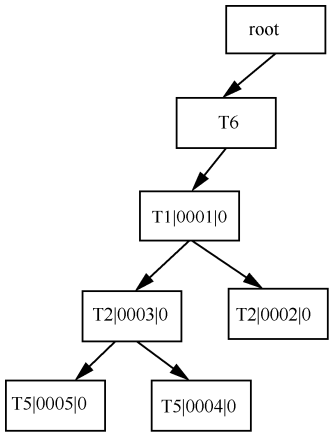


图 3　策略二叉树结构

（4）遍历二叉树合并策略，合并原则为：

- 对二叉树进行递归的中序遍历；
- 遍历发现最后一个节点包含 ID，则从根节点到此路径上所有节点形成一条“和”关系策略；
- 一旦发现该节点有右子树，则一直对右子树进行扫描，如果包含 ID 则成为“或”关系，同时将该节点标记为 1。

生成后的策略结果如下，基于二叉树的合并方案，在数量级上达到了 1/3~1/2 压缩，获得很好的压缩、合并的效果，效果如下：

0003　T6&T1& (T2| T5)

0001　T6& (T1|T2)

4　实验结果与分析

本系统的软硬件测试环境为：Intel Core CPU、2 GHz 主频、4 GB 内存、120 GB 硬盘、Windows 7 操作系统。实验采用 4 组真实数据，包括两组违规短信数据和两组正常短信数据，包括欺诈、涉黄、涉黑、违法、恶意 URL、推广、正常 7 类数据，全部数据经过了人工审核。训练和测试数据集见表 6。

表 6　训练和测试数据集

测试类型	数据集	样本数量/条
封闭（违规）	SMS_train	280 000
封闭（正常）	normal_train	265 000
开放测试（违规）	SMS_test	28 000
开放测试（正常）	normal_test	26 800

分类结果采用混淆矩阵，见表 7。

表 7　混淆矩阵

真实	预测		
	违规短信	正常短信	总计
违规短信	TP	FN	TP+FN
正常短信	FP	TN	FP+FN
总计	TP+FP	TN+FN	

- 违规短信的评测指标：准确率 $p=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$，召回率 $R=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}$，$F$ 值：$\frac{2\times P\times R}{P+R}$。

评估关联规则的实质是通过实验确定挖掘的关联规则中包含的策略词个数，再一次扫描训练语料统计关联规则命中情况，计算命中违规短信的准确率、召回率最后获得 F 值。如图 4 所示，实验发现当策略内的策略词个数小于两个时，容易引起误判，超出 5 个时，F 值迅速下降到 0.4 以下，所以策略内的策略词个数在[3,4]时效果最佳。

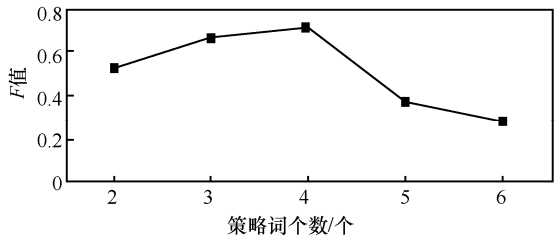


图 4　判断策略词个数

其中对各条策略要求其命中违规短信准确率 $P_b \geqslant 0.95$，为了避免对正常短信的误命中，增加约束条件“命中正常短信的个数≤100”。

过滤过程中存在两种类型的错误：将违规短信错分为正常短信；将正常短信被错分为违规短信。第 2 种错误带来的影响更大，因此一个好的违规短信分类算法应该在满足将第 2 类错误率控制在一个较小的范围前提下，尽量提高整个分类模型的准确率。

表 8　SVM 分类测试结果（单位：条）

类别	欺诈	涉黄	涉黑	违法	恶意 URL	推广	正常	均值
训练数据	60 000	35 000	45 000	40 000	50 000	50 000	280 000	
测试数据	5 000	4 000	4 000	5 000	5 000	5 000	26 800	
正确数据	4 328	3 598	3 079	4 275	3 243	4 309	26 509	
误判数据	443	458	125	627	367	2032	4 973	
准确率	0.907	0.887	0.961	0.872	0.898	0.680	0.842	0.864
召回率	0.866	0.900	0.770	0.855	0.649	0.862	0.989	0.841
误判中的正常类数据量	330	26	42	243	300	1 750		
误判率	0.069	0.006	0.013	0.050	0.083	0.276		0.083

- 模型评测指标：正常短信误判率 $N=\frac{\mathrm{FP}}{\text{全部短信数量}}$，模型的准确率 $A=\frac{\mathrm{TP}+\mathrm{TN}}{\text{全部短信数量}}$。

SVM 分类测试结果见表 8。

SVM 对恶意 URL 的识别不好，由于恶意 URL

表 9　基于子语义空间的策略集分类测试结果（单位：条）

类别	欺诈	涉黄	涉黑	违法	恶意 URL	推广	正常	均值
训练数据	60 000	35 000	45 000	40 000	50 000	50 000	280 000	
开放测试数据	5 000	4 000	4 000	5 000	5 000	5 000	26 800	
正确	4 128	3 729	2 902	4 345	3 446	4 209	24 935	
错误	265	429	139	593	207	1 132	6 082	
判定为该类	4 393	4 158	3 041	4 938	3 653	5 341	31 017	
准确率	0.940	0.897	0.954	0.880	0.943	0.788	0.804	0.887
召回率	0.826	0.932	0.726	0.869	0.689	0.842	0.930	0.831
错误中的正确数量	189	32	78	329	287	950		
误判率	0.043	0.008	0.026	0.066	0.079	0.178		0.066

表 10　两种模型对比

算法	正常短信误判率 N	模型的准确率 A	模型召回率 R	模型的 F 值
SVM	8.3%	86.4%	84.1%	85.25%
基于子语义空间的策略集	6.6%	88.7%	83.1%	85.76%

的明显特征是包含一个 URL，但是文本很短小，含有的信息比较少，所以识别准确率和召回率比较低。基于子语义空间的策略集分类测试结果见表 9。

针对上述样本集，将本文所述的方法与 SVM 分类算法进行对比，结果见表 10。

研究、分析与实验证明，此算法从语义层面出发，尤其适合对于短文本的处理工作，提取的违规短信策略能在保证低误判率的情况下，准确发现约 88%的垃圾短信。该能力应用在现网中，能在基本不增加人工审核的前提下，准确有效地发现大量垃圾信息，具有很强的实用价值。

5　结束语

本文针对识别短文本类的违规短信存在的主要问题，设计并实现了基于子语义空间的短文本策略提取算法，实验结果证明了该算法的有效性。目前大数据的建设对运营商提出了更多的挑战[11]，其中违规短信在不断演化，尤其在敏感词的变形方面，预处理只能解决一部分问题，而且其发送特征和内容也在不断变化，为适应变化，除了应用上述基于策略的内容分析方法外，还可以进行短信内容的智能化离线分析，将文本分类和机器学习中的新方法引入垃圾短信过滤中，并把分析结果及时反馈给在线过滤系统，大大提高了违规短信拦截率，降低了误拦率。

参考文献：

[1] YIH W, GOODMAN J, CARVALHO V R. Finding advertising keywords on Web pages[C]//Proceedings of the 15th International Conference on World Wide Web. New York: ACM Press, 2006: 213-222.

[2] KUHN R, DE MORI R. A cache-based natural language model for speech recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1990, 12(6): 570-583.

[3] MIHALCEA R, TARAU P. TextRank: bringing order into texts[Z]. 2004.

[4] 王庆, 陈泽亚, 郭静, 等. 基于词共现矩阵的项目关键词词库和关键词语义网络[J]. 计算机应用, 2015, 35(6): 1649-1653.
WANG Q, CHEN Z Y, GUO J, et al. Project keyword lexicon and keyword semantic network based on word co-occurrence matrix[J]. Journal of Computer Applications, 2015, 35(6): 1649-1653.

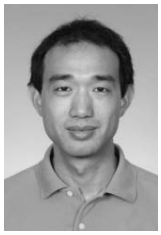
[5] 董振东, 董强, 郝长伶. 知网的理论发现[J]. 中文信息学报, 2007, 21(4): 3-9.
DONG Z D, DONG Q, HAO C L. Theoretical findings of HowNet[J]. Journal of Chinese Information Processing, 2007, 21(4): 3-9.

[6] BENGIO Y, DUCHARME R, VINCENT P, et al. A neural

probabilistic language model[J]. Journal of Machine Learning Research, 2003(3): 1137-1155.
[7] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[J]. Computer Science, 2013.
[8] BENGIO Y, DUCHARME R, WINCENT P. Neural probabilistic language model neural probabilistic language model[Z]. 2003.
[9] HANM J W, KAMBER M, PEI J. 数据挖掘概念与技术[M]. 范明, 孟小峰, 译. 北京: 机械工业出版社, 2012: 157-179.
HANM J W, KAMBER M, PEI J. Data mining concepts and techniques[M]. Translated by FAN M, MENG X F. Beijing: Machinery Industry Press, 2012: 157-179.
[10] 朱龙珠, 徐宏, 刘莉莉. 基于深度学习的 95598 重大服务事件识别研究[J]. 电力信息与通信技术, 2018, 16(11): 19-23.
ZHU L Z, XU H, LIU L L. Research on recognition of 95598 significant service events based on deep learning[J]. Electric Power Information and Communication Technology, 2018, 16(11): 19-23.
[11] 陈涛, 鲁萌, 陈彦名. 运营商大数据技术应用研究[J]. 电信科学, 2017, 33(1): 130-134.
CHEN T, LU M, CHEN Y M. Research on operators' big data technologies and applications[J]. Telecommunications Science, 2017, 33(1): 130-134.

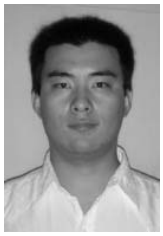
[作者简介]

孙洋（1983− ），女，中国移动通信有限公司研究院工程师，主要研究方向为自然语言处理、机器学习和大数据安全。



粟栗（1981− ），男，博士，中国移动通信有限公司研究院教授级高级工程师，主要研究方向为大数据安全和密码学。

张星（1980− ），男，中国移动通信有限公司研究院工程师、技术经理，主要研究方向为数据安全、数据存储和虚拟化技术。



王峰生（1979− ），男，现就职于中国移动通信有限公司研究院，主要研究方向为移动通信网络安全。

杜海涛（1979—），男，博士，中国移动通信有限公司研究院高级工程师，主要研究方向为大数据安全和移动通信网络安全。