基于贝叶斯及多模式串模糊匹配算法的不良短消息甄别混合模型

张文波[1]，蒋春华[2]，姚天昉[1]

[1]上海交通大学 计算机科学与工程系，上海 200240

[2]上海交通大学 软件学院，上海 200240

E-mail: hellowenniu@gmail. com; jiang.chunhua2008@163.com; yao-tf@cs.sjtu.edu.cn

摘 要：手机短信息业务一方面给人们带来诸多便利，另一方面一些不法分子利用手机短信息进行违法犯罪活动也日益猖狂，如何防范和打击此类犯罪活动对执法机关来说都是一个新的挑战。本文针对不良短消息的识别和分类问题，提出了一个基于贝叶斯分类算法和改进的多模式串模糊匹配算法的不良短消息甄别混合模型，以实现对不良短消息的识别和分类。短消息文本经由朴素贝叶斯分类器进行是否不良的判断，对确认为不良的短消息再经过多模式串的模糊匹配进行不良类别的分类。实验表明该方法提高了不良短消息识别的准确率，具有良好的应用前景和实际效益。本文重点分析不良短消息识别和分类过程。

关键词：不良短消息；朴素贝叶斯；多关键词匹配；WM 算法

A Hybrid Model for Filtering Illegal Message Based on Bayes Classification and Fuzzy Matching Algorithm

Zhang Wenbo[1], Jiang Chunhua[2], Yao Tianfang[1]

[1] Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240

[2] Software School, Shanghai Jiao Tong University, Shanghai 200240

E-mail: hellowenniu@gmail.com; jiang.chunhua2008@163.com; yao-tf@cs.sjtu.edu.cn

Abstract: Mobile phone short message service brings a lot of convenience in modern world, while it also leads more and more criminal acts via mobile phone network, which has become a new problem and challenge for law enforcement agencies. In this paper, we propose a hybrid model for filtering illegal short message based on Naïve Bayes classification and fuzzy matching of multi-pattern string algorithm to deal with identification and classification of illegal message. A short message will firstly be identified as illegal/legal by Naïve Bayes classifier, and then those messages labeled illegal will be classified to different kinds of illegal messages by fuzzy matching of multi-pattern string algorithm. The experiments show that this mode could improve performance for identifying and classifying illegal short messages. This paper will focus on how to identify illegal short messages and classify them.

Keywords: illegal message; Naïve Bayes; multi-keyword matching; WM algorithm

1 引言

近年来，随着通信产业的发展，传统的信息传播方式已经无法满足人们互通信息的需求，通过手机传播短消息已经成为人们又一主要的通信方式，人们对短消息的认同感也正在逐渐增强，短消息这一新兴的信息载体和传播方式也逐渐影响人类的生活方式。

短消息作为人类日常生活交流工具给人们的生活提供便利的同时，一些不法分子利用手机发送虚假信息实施诈骗活动；或散布谣言、传播色情、反动信息；或发送大量广告信息影响他人的正常生活；这些不良短消息不仅影响人们的正常生活，对公共安全也是一大隐患。

据统计，2006 年，我国手机用户全年收到的垃圾短信总量达 1836 亿条，手机用户人均每周收到的垃圾短信数量已达 6.46 条；2007 年，我国手机用户全年收到的垃圾短信总量达 3538 亿条，与 2006 年同期相比增加了 1702 亿条，增幅达 92.7%。手机用户人均每周收到的垃圾短信数量已达 12.44 条；2008 年，我国手机用户全年收到的垃圾短信总量达 2944 亿条，手机用户人均每周收到

的垃圾短信数量达 10.35 条。

利用短消息的不良活动严重干扰了正常的社会秩序，侵害了广大人民群众的合法权益，社会各界对此反映强烈，如何有效地治理不良短消息、保护用户个人隐私、打击不良犯罪活动等问题引起了人们的格外关注。本文提出一种基于最小风险的朴素贝叶斯分类器和多模式串的模糊匹配算法的短消息甄别混合模型。通过实验证明该模型能够有效的对短消息文本进行不良短消息的甄别。

2 短消息甄别混合模型

本文提出的混合模型对短消息的甄别，包括对不良短消息的识别及分类。混合模型包含两大模块：基于最小风险的朴素贝叶斯分类器[1]的不良短信识别模块和基于多模式串模糊匹配的不良短信分类模块。模型总体框架如图：

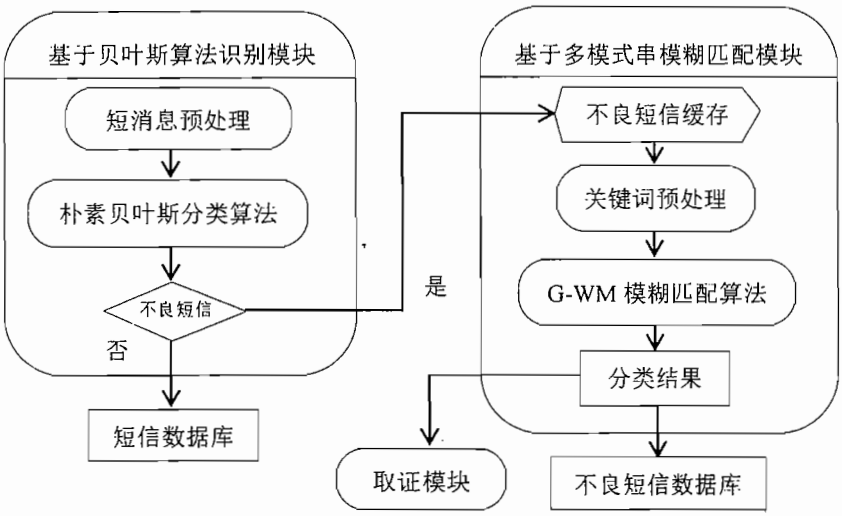

图 1 短消息甄别混合模型总体框架

基于朴素贝叶斯分类器的不良短消息识别模块由短消息预处理子模块和贝叶斯实时分类子模块两部分组成，完成对短消息是否不良的过滤；基于关键词的模糊匹配模块由关键词预处理子模块和基于 G-WM 算法的模糊匹配子模块组成，根据预先设定的一系列关键词在收到一条不良短消息时，在不良短消息中找出所有匹配的关键词，并根据关键词的权重累加权值得到最终值，判断短信所属类别。

3 基于朴素贝叶斯分类算法的不良短信识别模块

3.1 贝叶斯分类算法

文本分类有很多比较经典的方法，包括 K-NN[2]、决策树[3]、SVM[4]、贝叶斯等。朴素贝叶斯分类方法可以高效的对短文本信息进行分类识别，可以很好的解决本文所需要的二元（是/否不良短消息）分类的问题。作为一种简单实用的分类算法，朴素贝叶斯分类法能够很好的胜任本模块的分类问题。

3.2 特征提取及其算法

考虑到算法的速度和效率，本文采用 TFIDF 方法[5]进行特征提取[6]，具体采用下面几项来表示特征：特征名称、在所有类别的训练例中包含该特征的文档个数、在本类别的训练例中包含该

特征的文档个数、该特征的权重等。其中，特征名称表示对短消息正文或主题部分进行挖掘所得到的一些词或词组。

作为半结构化的文本[7]，短消息由短消息头(结构化信息)和短消息体(无结构化信息)组成。对每一封训练例子不仅从头部信息中提取特征，同时也从短消息体中提取特征，共同构成这封短消息的特征模式。

3.3 基于贝叶斯分类算法的不良短信识别

考虑各种误判造成的损失不同，本文采用基于最小风险的贝叶斯决策来对短消息进行过滤与分类[8]。将短消息决策为不良短消息或者合法短信息的条件风险为：

$$R(c=2\mid d_x)=0\times P(c=2\mid d_x)+k\times P(c=1\mid d_x)=k\times(1.0-P(c=2\mid d_x)) \tag{1}$$

$$R(c=1\mid d_x)=1\times P(c=2\mid d_x)+0\times P(c=1\mid d_x)=P(c=2\mid d_x) \tag{2}$$

其中，$P(c_j\mid d_x)=\dfrac{P(c_j)P(d_x\mid c_j)}{P(d_x)}$，$P(d_x\mid c_j)=\prod_{i=1}^{n}P(w_i\mid c_j)$

$P(w_i|c_j)$为 c_j 个文本中出现特征 w_i 的文本数。

决策为不良短消息时，要满足 $R(c=2|d_x)<R(c=1|d_x)$，即：

$$k(1.0-P(c=2\mid d_x))<P(c=2\mid d_x) \tag{3}$$

令 $\theta=\dfrac{k}{1+k}$，整理得：$P(c=2\mid d_x)>\theta$ (4)

对于用户，合法短消息比不良短消息更为重要，因此合法短消息被误判为不良短消息将可能给用户带来更大的损失，所以 k 要远大于 1。当 $P(c=2|d_x)$满足公式（4）时，可保证决策为不良短消息的风险比决策为合法短消息的风险小。而 $k\geqslant 1$，当 $k=1$ 时，$\theta=0.5$；$k=9$ 时，$\theta=0.9$；$k=99$ 时，$\theta=0.99$。因此，将不良短消息的过滤阈值设为 0.5, 0.9 和 0.99 分别能保证 $k=1, k=9$ 和 $k=99$ 时的决策风险最小。

4 基于改进的 WM 关键字模糊匹配算法的不良短信分类算法

短消息经过朴素贝叶斯分类之后，被判定为是否不良，接下来采用内容关键词分类算法来实现不良短消息的详细分类。

4.1 不良短信的分类

本文将不良短信分为以下六类：

一、假冒银行、银联或其他单位名义发送短消息进行诈骗或者敲诈勒索公私财物的；

二、散布淫秽、色情、赌博、暴力、凶杀、恐怖内容或者教唆犯罪、传授犯罪方法的；

三、非法销售枪支、弹药、爆炸物、走私车、毒品、迷魂药、淫秽物品、假钞、假发票或者明知是犯罪所得赃物的；

四、发布假婚介、假招聘、快速致富等传销内容，或者引诱、介绍他人卖淫嫖娼的；

五、多次发送干扰他人正常生活的，以及含有其他违反宪法、法律、行政法规禁止性规定的内容的，如代开发票、代办证件或小额贷款等；

六、未分类或其他短消息举报；

4.2 G-WM 算法的预处理

作为多模式串精确匹配算法，WM 算法精确度和效率都比较高，但容易遗漏与模式串相似的文本串。短消息发送者经常采用反过滤手段对某些关键词进行处理，若直接应用 WM 算法容易将

不良误判为合法，因此，本文采用模糊匹配算法进行不良短消息的分类。

模糊匹配需要寻找与关键词相似的字符串，目前已有许多有效的模糊匹配算法，但他们都需要寻找输入模式经过一切可能变换后的结果，需要进行大量的计算，性能比较差。既要提高系统匹配精确度，又要提高性能，本文所采用的关键词模糊匹配分类算法就是对 WM 算法进行改进，实现精确匹配和模糊匹配的转换，以满足实际需求。

不良短消息发送者会采用各种手段变换关键词，主要表现如下：

一、在关键词中插入一些无意义的干扰符号，如将“彩票中奖”写成“彩？票中/奖”；

二、利用关键词的组合构造特定的信息。如：“话费赠送”只有在和“中奖”同时出现时才有可能属于诈骗类短消息；

三、对关键词进行同音词或拼音的转换，如将“彩票”写成“采票”或“cai 票”等变换形式。

针对上述三种可能出现的情况，在关键词匹配之前需要对短消息文本和关键词进行预处理。预处理包含下列步骤：

（1）消除噪音

消除噪音是通过噪音字典删除噪音字符，噪音字典通常包含一些标点符号、助词、代词等，如 “法？轮功” 变成 “法轮功”，短信息文本以其原始形态出现。

（2）关键词分配权重

短信息属于哪一类别违法短信息通常不是只由某个关键词决定，对关键词分配权重可以解决多个关键词的组合决定短信息文本的类别的问题。根据各个关键词在每个类别中起到的作用不同为每个关键词分配一个权重，最后计算一个累加权值，如果该值大于预先设定的阈值，则认为该短信息是属于某类违法短信息。

（3）同音字拼音化替换处理

同音字拼音化替换处理可以消除短信息中的同音误差、或以拼音替代汉字的情况。本文利用字典文件将一个拼音和属于这个拼音的所有汉字组合成一个类，对这些类进行编号，以一个整数来表示该类。

表 1 显示对短消息文本和关键词列表进行简单的预处理示例。

表 1　不良短消息预处理示例

原始内容	去除噪音	拼音转换
尊敬的朋友你好！想要测听对方的通^话与短~信吗？本公司能为你配*这类手 机与卡!市区可送货。详询:13755563011 王经理	尊敬的朋友你好想要测听对方的通话与短信吗本公司能为你配这类手机与卡市区可送货详询 13755563011 王经理	zun jing de peng you ni hao xiang yao ce ting dui fang de tong hua yu duan xin ma ben gong si neng wei ni pei zhe lei shou ji yu ka shi qu ke song huo xiang xun 13755563011 wang jing li

4.3　G-WM 算法的搜索与匹配

短消息文本和关键词列表经过简单的预处理之后并不能马上进行匹配，在预处理基础之上还需要进行编码的转换，在编码转换的过程中，本文进行了压缩编码。

（1）压缩编码

压缩编码是 Sun Kim 和 Yannggon Kim 设计的一种多模式匹配算法中采用的一种技术[9]。通过压缩编码，可以得到一个 PCODE 值或 TCODE，对于每个待匹配模式串 P 还需要设置一个掩码串 PMASK，对文本串 T 和 PMASK 进行逻辑与操作，将其结果再和模式串 P 进行逻辑异或操作，如果为 0，则匹配成功，否则视为模式串没有出现在文本中。

（2）构建哈希表

在对短消息内容和关键词列表进行一系列的变换过程之后，还需要构建一个哈希表 PHASH，

通过 PHASH 表来过滤一批候选的可能匹配的关键词。

（3）构建 SHIFT 表

SHIFT 表是用来存储扫描短消息文本时可以移动的最大且安全的距离。在短消息中搜索关键词，一般至少可以移动 1 位，取最小关键词长度为 2，本文考虑当前待匹配短消息文本串（两个字节）紧邻的下一个字符所取的作用，将当前待匹配短消息文本串的最后一个字符和其紧邻的字符组合为一个新的字符块 K_B，用此字符块来确定可以移动的最大距离，这样最大移动距离可以达到最小关键词长度+1。

（4）搜索与匹配

在搜索与匹配阶段，使用 K_B 计算哈希值，查找 SHIFT 表计算移动的距离。如算法 1。

算法 1：关键词模式串 P 与短消息文本串 T 匹配算法

输入：关键词模式串 P，短消息文本串 T；

Step 1：从短消息文本字符串 T 左端开始，依次取两个字节，计算二进制编码，赋值给变量 n；

Step 2：搜索 PHASH 表，检查 PHASH[n]和 PHASH[n+1]，若 PHASH[n]= PHASH[n+1]，转到步骤 5，否则，转到步骤 3；

Step 3：取当前 PHASH[n]值赋给变量 INDEX，从短消息文本串的当前位置自左向右取第 INDEX 个关键词模式串对应的字节数 i，计算其二进制编码 TCODE；

Step 4：取当前关键词模式串 P 中第 INDEX 个关键词的 PMASK 与 PCODE，将 TCODE 与 PMASK 进行逻辑与操作，结果为 0 时转步骤 5，不为 0 时，再将 TCODE 与 PCODE 进行异或操作，结果为 0 则表示匹配成功，后移 i 位转步骤 1；否则转 5；

Step 5：计算 K_B 的散列值 h，根据 SHIFT[h]向后移动 SHIFT[h]个字符，转步骤 1。

通过算法 1 发现，匹配后该关键词只是候选关键词中的一个，在编码层匹配的基础上还需要进行汉字层匹配，在进行汉字层匹配时使用相似度[10]来权衡关键词与在待检测文本中获取的子串的相似程度。

$$\text{相似度}=\frac{\text{关键词和子串对应位置汉字相同的个数}}{\text{关键词中汉字的个数}} \tag{5}$$

通过设定一个相似度阈值，不同长度模式的相似度阈值不同，我们根据计算的相似度与相似度阈值比较，如果大于该阈值，则确定子串与模式匹配。

在短消息文本中搜索到属于某类不良短消息的关键词后，还要根据预先设定的关键词权重计算匹配度，如果匹配度大于某个阈值，则此短消息即可认定为该类别的不良短消息。

5 实验结果及分析

5.1 实验数据

由于短消息涉及到个人的隐私问题，因此公开的短消息语料库目前还没有。在实际工作中，本文抽取来自公安部门接收的群众报警短消息共 20000 条，其中不良短消息 15000 条，合法短消息 5000 条，每一条短消息均不含任何个人隐私信息。

5.2 贝叶斯分类器

本文将收集的短消息随机分为 10 组，抽取 9 组作为训练样本，另外一组作为测试样本，在实验过程中，对 20000 条短消息随机组合，反复测试，取平均值作为测试的结果。结果如表 2。

在运用贝叶斯分类方法进行不良短消息的判断时，采用最小风险策略的朴素贝叶斯分类方法

在分类效果上有明显的提高，使得短消息判为正常短消息的机率加大，查全率增大；而短消息判断为不良短消息的机率减少，查准率有所增加；同时，不同的阈值对查全率、查准率和 f-measure 的影响也有所不同，阈值越高，不良短消息的查全率越低，查准率越高，f-measure 也越低，这说明阈值越高，加大对短消息判断为不良短消息的风险，漏掉一部分不良短消息，但判断的却更为准确，减少了用户的损失。

表 2　基于最小风险贝叶斯分类的实验结果

阈值	短消息类别	Recall	Precision	F-measure
0.5	正常短消息	88.56%	83.4%	86.18%
	不良短消息	80.73%	90.72%	85.43%
0.9	正常短消息	90.17%	81.21%	85.45%
	不良短消息	73.89%	91.48%	81.76%
0.99	正常短消息	92.38%	78.43%	84.83%
	不良短消息	71.56%	93.24%	80.97%

5.3　混合模型

基于混和模型的短消息甄别模块的设计是建立在先过滤后分类的基础之上，从数据库中选取 2000 条短消息，其中 1500 条为不良短消息，500 条为正常短消息，设置关键词 200 个，获得数据如表 3 所示：

表 3　短消息甄别混合模型实验结果

类别	Recall	Precision	F-measure
第一类	89.64%	97.45%	93.38%
第二类	79.19%	86.86%	82.85%
第三类	95.59%	98%	96.74%
第四类	83.26%	90.72%	86.83%
第五类	73.7%	85.37%	79.11%
第六类	87.53%	98.67%	92.43%

从表 3 中可以看出，由于在认定短消息的类别时关键词的设定非常重要，对于“第二、四、五类”的关键词的界定比较模糊，对查准率造成一定的影响，而“第一、三类”关键词意思比较明确，歧义少，查准率相对来说比较高，因此，要想提高本模块的性能，在日常工作中不断积累充实关键词库非常重要。

6　结论

针对目前社会上存在的利用手机短信息进行违法活动的现象，本文提出并设计一个基于贝叶斯分类算法和多模式串模糊匹配算法的不良短消息甄别混合模型。通过对实验数据的分析知道，对短消息的分类识别具有较高的正确率。

在贝叶斯分类算法模块，本文采用 TFIDF 方法进行特征提取，降低了算法的复杂性，使得算法的速度和效率都能达到不错的效果。贝叶斯分类器在是/否不良短信的分类上，达到了较高的准确率。模型总体效果上，改进的多模式串模糊匹配算法能够有效地应用于不良短信分类模块。总体实验证明了该模型的有效性。

参 考 文 献

[1] A. H. Tan, P. YU. A comparative Study on Chinese Text Categorization Methods. PRICAI 2000 Workshop on Text and Web Mining, Melbourne, 2000, 24-35.

[2] Apt. Text Mining with Decision Rules and Decision Trees [C] Proceedings of the Conference on Automated Learning and Discovery, CMU,1998: 302-317.

[3] Malcolm Corney, Gender-Preferential Text Mining of E-mail Discourse[J/OL], http://www.acsac.org/2002/papers.

[4] Yang Y, Pedersen J P. A Comparative Study on Feature Selection in Text Categorization[C]. In: Proc of the 14th In Conf on Machine Learning (ICML'97), 1997: 412-420.

[5] Y. Yang and X. Liu. A re-examination of text categorization methods. In *22nd A CM InternationalConference on Research and Development in Information Retrieval (SIGIR-99),* pages 42-49, Berkeley, US, 1999. Osmar R. Zgiane, Jiawei Han, and Hua Zhu.Mining recurrent items in multimedia with progressive resolution refinement.

[6] Hooman Katirai: Filtering Junk E-Mail, A Performance Comparison between Genetic Programming & Naïve Bayes, paper of Department of Electrical & Computer Engineering University of Waterloo.

[7] C. Apte, F. J. Damerau, and S. M. Weiss: Automated learning of decision rules for text categorization. *A CM Transactions on InformationSystems,* 12(3): 233-251, 1994.

[8] W. Lam and C. Y. Ho. Using a generalized instance set for automatic text categorization. In *21st ACM SIGIR International Conference on Information Retrieval,* pages 81-89, Melbourne, Australia, 1998.

[9] Kim S, Kim Y. A Fast Multiple String-Pattern Matching Algorithm. In: Proceedings of the 17th AOM/IAOM International Conference on Computer Science. May 1999. 44.

[10] 刘钦东, 王倩, 黄新波. 面向中英文混合环境的多模式匹配算法. 软件学报, V01.19, 2008, 674-686 页.