

文章编号:1001-9081(2008)03-0714-05

基于朴素贝叶斯和支持向量机的自适应垃圾短信过滤系统

金展¹, 范晶¹, 陈峰², 徐从富¹

(1. 浙江大学 计算机科学与技术学院, 杭州 310027; 2. 浙江省公安厅 科技处, 杭州 310009)

(jinzhan@ccpithz.org)

摘要:随着短信业务的不断发展,垃圾短信的特征和内容也在不断变化,传统垃圾短信过滤系统中存在的主要问题是,短信特征和内容未能得到及时更新而导致过滤性能降低。考虑朴素贝叶斯的快速统计分类及支持向量机(SVM)的增量训练等特点,将其应用于垃圾短信过滤中,并把分析结果及时反馈给在线过滤子系统,使得系统具有更好的自适应性。实验结果表明,该方法可有效地解决当前垃圾短信过滤系统中存在的问题。

关键词:垃圾短信过滤;朴素贝叶斯;支持向量机

中图分类号: TP18; TP302.1 **文献标志码:** A

Spam message self-adaptive filtering system based on Naïve Bayes and support vector machine

JIN Zhan¹, FAN Jing¹, CHEN Feng², XU Cong-fu¹

(1. College of Computer Science and Technology, Zhejiang University, Hangzhou Zhejiang 310027, China;

2. Department of Science & Technology, Zhejiang Provincial Bureau of Public Security, Hangzhou Zhejiang 310009, China)

Abstract: With the development of the short message services, the characteristics and contents of the spam short Message are also changing constantly, the main problems that exist in the traditional short message filtering systems are that the characteristics and contents fail to be updated in time, which reduced the filter capability. This paper mainly utilized Na ve Bays advantage of rapid statistics classification and Support Vector Machine (SVM) incremental training characteristic in Spam Short Message filtering, and provided feedbacks to the online filtering sub-system in time in order to enhance the system-s self-adaptability. The experimental results show that this new method effectively deals with the above problems in the traditional spam short Message filtering systems.

Key words: spam short message filter, Naïve Bays; Support Vector Machine (SVM)

0 引言

随着我国移动通信业务的发展以及网络科技的创新,被称为“第五媒体”的手机已成为人们日常生活中不可缺少的一种通讯工具。手机短信业务也因其价格便宜、形式新颖、方便快捷,赢得了广大手机用户的青睐,而其构成的产业和市场也被人称为“拇指经济”^[1]。据有关部门统计,截至 2007 年 6 月,我国手机用户达到 5 亿户,居世界第一。2006 年中国的短信已经达到 5000 亿条。一方面给广大使用者带来了各种方便和帮助;另一方面也出现了诸多的问题,如垃圾短信的泛滥、短信诈骗、短信病毒、短信谣言等。短信正被极少数不法分子或不良企业利用,不仅影响了人们的正常生活,也给正在蓬勃发展的移动通信业带来了很大的负面影响。

所谓垃圾短信:是指没有经过接受者同意的、其内容具有违法性或广告性的、侵害接受者合法权益的短信息^[2]。短信的安全监控过滤是一个较新的课题,加上短信业务流程复杂、接收方式灵活多变,因此,短信内容的监控和过滤非常困难。虽然垃圾短信和垃圾邮件在一定程度上有相似之处,但是短信的安全监控和过滤有其自身特点,且不像垃圾邮件监控及过滤那样成熟。当前主要有两种机制实现垃圾短信的监控和

过滤,即内容关键字过滤机制和号码黑名单机制^[3]。其中,内容关键字过滤机制中的内容(关键字)主要依靠人工添加的方法来实现,尚无法实现自动添加;而号码黑名单的生成可分为手工添加、实时自动生成和准实时自动生成等方法,主要通过手工添加的方法来实现。

当前,在快速的在线垃圾短信内容分析方法中,其关键词和黑名单大多通过人工添加,主要缺点是非常呆板且数量有限,事实上,由于垃圾短信在不断演化,垃圾短信的发送特征和内容也在不断变化,为适应这种变化,必须研发新的垃圾短信自适应过滤系统,以提高系统的智能化水平。本文设计并实现了一个基于贝叶斯和支持向量机(Support Vector Machine, SVM)的自适应垃圾短信过滤系统,将贝叶斯分类和 SVM 等技术引入垃圾短信过滤中,并将分析结果及时反馈给在线垃圾短信过滤系统,使系统具有更好的自适应性和较高的智能化水平。

1 系统架构

本文系统的基本思路是,在特定的时间段内提取短信的特征内容,并将其反馈给在线过滤子系统,增加或更新短信的特征,以达到准确和智能过滤垃圾短信的目的。离线过滤子

收稿日期:2007-10-16;修回日期:2007-12-19。

基金项目:国家 863 计划项目(2007AA01Z197);国家自然科学基金资助项目(60402010)。

作者简介:金展(1980-),男,浙江杭州人,硕士,主要研究方向:人工智能、数据挖掘; 范晶(1980-),女,浙江衢州人,硕士,主要研究方向:人工智能、数据挖掘; 陈峰(1970-),男,浙江杭州人,高级工程师,主要研究方向:计算机应用; 徐从富(1969-),男,浙江杭州人,副教授,主要研究方向:人工智能、数据挖掘、信息融合。

系统包括三个模块:短信预处理模块、SVM 模块和统计样本模块。系统架构如图 1 所示。

在图 1 中,离线过滤子系统有三个接口,接口 1 是离线统计模块如何得到短信样本,它通过文本文件将嫌疑短信实时保存到文本文件中;接口 2 是离线统计分析后把信息传递给综合评价函数模块,通过配置文件传递,定义配置文件中, class 记录正常短信和垃圾短信的概率, replaynum 记录是否含有回复电话在正常和垃圾短信中的概率, keyword 记录每个词在正常和垃圾短信中的概率, length 记录不同短信长度在正常和垃圾短信中的概率四个文件,在线综合评价函数根据这些信息计算结果;接口 3 是垃圾短信数据库的接口,通过数据交换把新的数据导入数据库中。

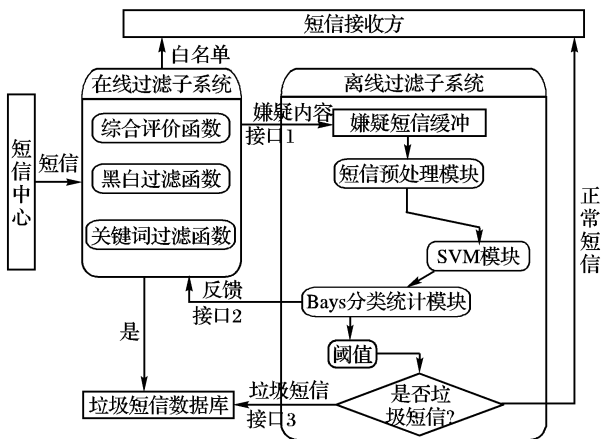


图 1 垃圾短信过滤系统框架

离线过滤子系统的处理流程如下:1) 短信文本预处理,其主要任务是剔除文档中所有与分类任务无关的内容,并将文本(即由字、词、数字、标点符号等构成的文字串)转化为由其包含的基本语义单位组成的表列。对于进入短信监控中心的短信,预处理模块将进行短信清洗、自动分词等预处理,并送 SVM 过滤模块。2) SVM 分类,不同于朴素贝叶斯所使用的生成模型,SVM 采用的是区别模型。基于 SVM 的分类方案为:对于每条分词后的短信,根据 $TF \times IDF$ 提取特征,并且把每条短信映射到向量空间中;运用 SVM 对短信进行分类;把分类结果存入统计样本中。3) 特征选择采用 IG (Information Gain) 选择最合适的特征,统计出词频高的词,反馈到综合函数,其核心思想是:选取部分信息,经过 SVM 学习训练后,把正常和垃圾短信分离,并进行关键词的学习;利用贝叶斯将统计出的正常短信和垃圾短信特征概率,并将关键词的分类,词频高的反馈更新在线过滤系统关键词库,系统设置一个阈值,综合评价函数根据计算得到垃圾短信特征的值,跟阈值比较,如果大于这个阈值说明是正常短信,反之小于则是垃圾短信。

2 关键技术

本文系统主要用到了两个关键技术,即朴素贝叶斯方法和支持向量机。朴素贝叶斯方法是一种系统的统计推断方法,具有较高的精确度。在统计上利用贝叶斯公式整合了短信的长度和短信内容,计算得到一个概率值,若超过设

定的阈值,即判断为垃圾短信;而 SVM 分类不仅具有良好的分类性能,同时具有增量训练特征,这是一般模型所不具备的。由于垃圾短信的发送特征和内容也在不断变化,故利用 SVM 较强的自适应性能,通过学习不断地调整综合评价函数(模块包含三个子模块:基于频率过滤的子模块、快速内容过滤子模块、基于朴素贝叶斯方法的概率计算子模块)以达到快速准确地过滤的目的。

2.1 贝叶斯分类统计

朴素贝叶斯分类^[4,6]可预测类成员关系的可能性,如给定样本属于某个特定类别的概率。朴素贝叶斯分类算法可以和判定树、神经网络等分类算法相媲美。本文以短信频率、长度和内容作为短信的主要特征,其中,以短信关键词为特征概率的算法主要由朴素贝叶斯完成。

朴素贝叶斯分类假定一个属性值对给定类的影响独立于其他属性值。朴素贝叶斯分类可描述如下:

- 1) 设短信样本可用一个 n 维特征向量 $X = (X_1, X_2, \dots, X_n)$ 表示,其中, X_1, X_2, \dots, X_n 分别表示对 n 个属性 A_1, A_2, \dots, A_n 的度量。
- 2) 本系统有两个类 C_1, C_2 , 代表正常短信和垃圾短信, 给定一个未知的数据样本 X (即没有类标号), 分类法将预测 X 属于具有最高后验概率(条件 X 下)的类, 即朴素贝叶斯分类将未知的样本分配给类 C_j , 当前仅当 $P(C_j | X) > P(C_i | X)$, $1 \leq j \leq m; j \neq i$ 。根据贝叶斯公式, 样本属于某个类 C_j 的概率为:

$$P(C_j | X) = \frac{P(X | C_j)P(C_j)}{P(X)}$$

由于 $P(X)$ 对于所有类为常数, 只需要 $P(X | C_j)P(C_j)$ 最大即可。如果类的先验概率未知, 则通常假定这些类是等概率的, 即 $P(C_1) = P(C_2)$, 并据此只对 $P(C_j | X)$ 最大化。否则, 最大化 $P(X | C_j)P(C_j)$ 。其中, 类的先验概率可用 $P(C_j) = S_j/S$ 估计, S_j 是类 C_j 中训练样本数, S 是样本总数。

3) 当特征项很多时, 计算 $P(X | C_j)$ 很复杂。为了简化计算, 可作特征项间条件独立的朴素假定, 即属性间不存在依赖关系。可得:

$$P(X | C_j) = \prod_{k=1}^n P(X_k | C_j)$$

为对未知样本 X 分类, 对每个类 C_j , 计算 $P(X | C_j)P(C_j)$ 。样本 X 被指派到类 C_j , 当且仅当:

$$P(X | C_j)P(C_j) > P(X | C_i)P(C_i); 1 \leq j \leq 2, j \neq i$$

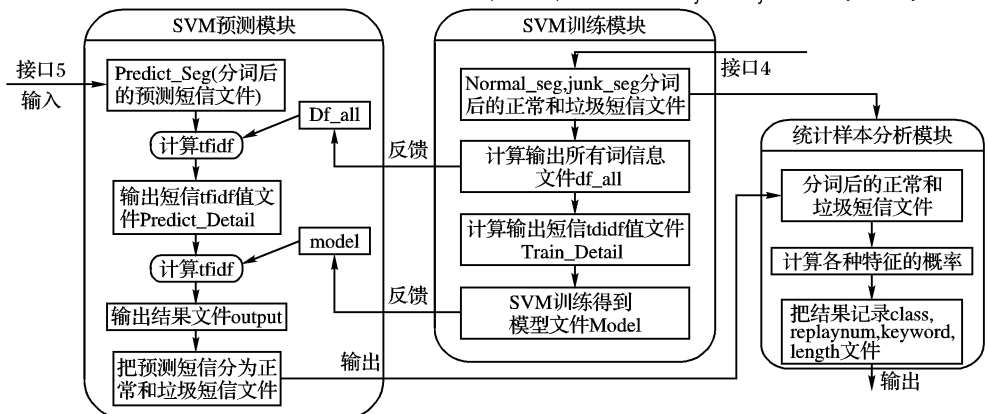


图 2 SVM 模块的流程

短信过滤是两类分类问题, 且存在两种分类错误: 将垃圾短信判断为非垃圾短信以及将非垃圾短信判断为垃圾短信。

一般地,第二种错误更严重,故定义一个变量 $a = \frac{P(C_{junk} | X)}{P(C_{norma} | X)}$,并给定一个常量 ϵ ,若 $a > \epsilon$,则分类器预测 X 是一个垃圾短信,否则,判定为合法短信。

2.2 基于 SVM 的分类及反馈学习

SVM 对垃圾短信进行训练,提取垃圾短信的关键词,反馈给综合函数,实现关键词库的动态变化更新,以提高在线垃圾过滤子模块的自适应性。其实现流程图如图 2 所示。

在图 2 中,接口 4 是短信样本训练模块的数据入口,接口 5 是短信预处理模块数据的入口。在基于 SVM 的文本分类中,进行关键词的训练,通过反馈可以动态的更新关键词,提高了在线垃圾短信的过滤效果,且大大优于不进行关键词学习的方法。

2.2.1 短信预处理

本文系统以短信文本为主体,在对短信进行基于语义的过滤之前,需要对文本进行预处理,使文本更适合后继分析的需要。文本预处理的主要任务是剔除文档中所有与分类任务无关的内容,并将文本(即由字、词、数字、标点符号等构成的文字串)转化为由其包含的基本语义单位组成的表列。分词这项预处理工作与文本的表示有着十分密切的联系,甚至可看成是文本表示的一个重要组成部分。分词的关键在于如何选择恰当的基本语义单位。对于进入短信监控中心的短信,预处理模块将进行短信清洗、自动分词等预处理,并送到朴素贝叶斯过滤模块。它包括短信清洗、去除停用词、自动分词、词性标注、特征项提取、特征项权重计算和生成短信向量一系列过程,其处理流程如图 3 所示。

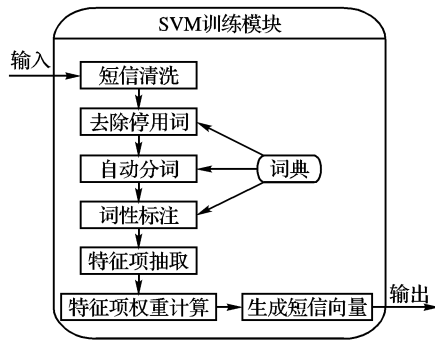


图 3 数据预处理流程

2.2.2 SVM 分类反馈学习

1) 问题描述

定义 1 短信集合分为三类:训练短信集合 E , 测试短信集合 T , 预测短信集合 U ; Θ 为训练后产生的分类器, S 为分类器 Θ 中的支持向量集; R 为经分类器 Θ 分类处理后的结果集, 并用 R^E, R^T, R^U 分别表示集合 E, T, U 的分类结果集; F 为反馈学习的短信集合, $d = \{(t_1, w_1), (t_2, w_2), \dots, (t_m, w_m)\}$ 为短信向量, 用 d^v 和 d^h 分别代表 S 中的支持向量和 F 中的反馈短信向量; $C = \{C_i | i = 1, 2\}$ 为预定义类别, 其中, C_1, C_2 分别表示正常短信和垃圾短信, 并用 $C_i^{init}, C_i^{auto}, C_i^h$ 分别表示短信的原始分类类别、自动分类类别、反馈分类类别; $P = \{P_1, P_2, P_3\}$ 作为短信集合类型标记, 分别表示短信为训练样本、测试样本、预测样本, 则:

$$E = \{\langle d_i, c_i^{init}, p_1 \rangle | i = 1, \dots, n^E\}$$

$$T = \{\langle d_i, c_i^{init}, p_2 \rangle | i = 1, \dots, n^T\}$$

$$U = \{\langle d_i, p_3 \rangle | i = 1, \dots, n^U\}$$

$$S = \{d_i^s | i = 1, \dots, n^S\}$$

其中, n^E, n^T, n^U, n^S 分别为集合 E, T, U, S 的大小; $R = \{R^E, R^T, R^U\}$, $R^E = \langle E, C_i^{auto} \rangle$, $R^T = \langle T, C_i^{auto} \rangle$, $R^U = \langle U, C_i^{auto} \rangle$; $F = \{\langle d_i^h, c_i^h, p_j \rangle | d_i^h \in p; i = 1, \dots, n^F, j = 1, 2, 3\}$, n^F 为集合 F 的大小。

对于基于 SVM 的分类^[13], 判断位于正常和垃圾短信边界处的那些可疑短信是分类模型学习的主要任务。模型的分类处理主要与分类边界处的支持向量有关, 而其他短信对分类处理不起作用。对基于 SVM 分类的反馈学习而言, 若经分类后证明是正常的短信, 表明原分类模型已包含该短信的相关分类信息, 因此它对反馈学习没有价值; 而那些分类错误或可疑的短信, 则说明含有原模型中所不包含的分类信息, 是进行反馈学习的重点。因此, 基于 SVM 分类反馈学习的重点是针对误分类短信或可疑的短信, 而这些短信往往是整个分类结果集合中的一小部分。事实上, 正常短信大大超过垃圾短信, 从理论上讲, 该学习方法具有反馈样本少的优点。

2) 分类反馈学习的模式类型

对于分类处理, 根据短信集合的不同可分为训练集、测试集和预测集。不同的集合在反馈处理时有一定的不同。另外, 根据反馈短信与分类模型中支持向量关系的不同, 也可以将反馈学习分为封闭反馈学习和开放反馈学习, 其中, 封闭反馈学习是指对训练集的反馈, 而对测试集、预测集的反馈则为开放反馈学习, 相关处理方法如下:

(1) 训练集的反馈: 在训练集 E 训练完成后, 通过获得的分类模型 Θ 对 E 进行封闭测试, 产生分类结果集 R^E , 对其中的错误分类文档集 R_{error}^E 进行反馈处理, $R_{error}^E \subset R^E$, 一般 $R_{error}^E \ll R^E$, 若 $R_{error}^E \neq \emptyset$, 则结束; 否则, 对于 $(d_i, c_i^{init}, c_i^{auto}) \in R_{error}^E$, 通过人工标记, 若原短信分类类别 C_i^{init} 不准确, 且其类别相关度高, 则由人工给出其正确的人工分类类别 c_i^h , 并保存到反馈集合 F 中。对于训练短信集的反馈处理, 可用于发现训练短信中的噪声数据(原始分类不正确的短信), 从而纠正训练短信中的错误数据信息, 并提高模型分类质量。

(2) 测试集的反馈: 测试集的反馈处理流程与训练集基本一致, 测试集可认为是具有人工分类信息的预测类短信。对于由训练集合 E 产生的分类器 Θ , 选取测试集合 T , 由分类器 Θ 进行分类处理, 产生分类结果集 R^T , 其中错误分类集为 R_{error}^T 。对 R_{error}^T 进行反馈, 对于 $(d_i, c_i^{init}, c_i^{auto}) \in R_{error}^T$, 因为 $c_i^{init} \neq c_i^{auto}$, 若经过反馈处理后, 发现 c_i^{init} 正确而 c_i^{auto} 错误, 同时该短信 d 具有较高的类别相关性, 则将其保存到反馈集合 F 中, 并将 c_i^{init} 赋值给 c_i^h 。通过对测试集的分类反馈处理, 可弥补训练短信所产生分类模型的不足, 进一步提高模型分类性能。

(3) 预测集的反馈: 预测集是实际分类应用中需要处理的大量无类别信息的短信, 是分类模型在应用中的主要对象。对于已有的分类器 Θ , 预测集 U 经分类处理后获得的结果集为 R^U , 反馈处理需要对整个 R^U 进行。对于结果集中自动分类 c_i^{auto} 错误, 且类别相关度高的短信, 给出其相应反馈类别 c_i^h , 并将其存入反馈集 F 中, 用于反馈训练学习。若系统可以给出不能分类短信集, 不妨记作 R_{un}^T , 则也可重点对 R_{un}^T 进行反馈处理。

3) 训练短信反馈的支持向量消重

在对训练短信集 E 进行反馈处理时, 由于反馈短信集需要与原模型的支持向量集一起进行优化训练, 而训练短信的反馈集主要由具有不正确的原始分类的训练短信组成, 这些短信有可能已成为原分类模型 Θ 的支持向量, 若仍留在原支

持向量集合 S 中,则会影响模型的效果,因此,在进行训练短信的反馈学习过程中,反馈训练前需要首先对该类反馈短信进行支持向量的检测,进行消重处理,以消除对分类性能的影响。为更好地进行支持向量的相似消重处理,定义 θ_{sim} 为短信相似度系数,作为进行支持向量消重的阈值。短信相似度的计算步骤如下:

a) 对反馈短信集合 F 中的任一反馈短信进行短信类型判断,若该短信为训练短信,则其反馈短信向量为 $\mathbf{d}_j^b(t_i, w_i)$; $j = 1, \dots, n^F$;

b) 读取支持向量集 S 的支持向量 $\mathbf{d}_k^s(t_i, w_i)$; $k = 1, \dots, l$, l 为向量集合 S 中支持向量数;

c) 对于任一训练反馈短信向量 \mathbf{d}_j^b 和支持向量 \mathbf{d}_k^s , 计算其向量相似度:

$$sim(\mathbf{d}_j^b, \mathbf{d}_k^s) = \frac{\mathbf{d}_j^b \cdot \mathbf{d}_k^s}{\|\mathbf{d}_j^b\| \times \|\mathbf{d}_k^s\|}$$

进行短信向量相似度判断:若 $sim(\mathbf{d}_j^b, \mathbf{d}_k^s) \geq \theta_{sim}$, 则判定这两条短信相同,将 \mathbf{d}_k^s 从支持向量集合 S 中删除;若不同,则存入集合 S 中,并提取短信的关键词。

针对以上处理,具体的反馈处理算法描述如下:

a) 对分类结果集 R 中任一条短信 $d_i \in R$ 进行反馈处理,若短信属于训练短信或测试短信,则可对 R 中的错误分类结果集 R_{error} 进行处理,根据短信的内容特征对需要进行反馈处理的短信,给相应的反馈类别 c_i^b 存入反馈短信集 F ;

b) 所有短信反馈处理完毕后,开始进行反馈训练学习,从反馈集 F 中读取任一短信,生成相应短信向量 $\mathbf{d}_j^b(t_i, w_i)$; $j = 1, \dots, n^F$;

c) 从原分类模型 Θ 的支持向量集合 S 中读取支持向量 $\mathbf{d}_k^s(t_i, w_i)$; $k = 1, \dots, l$;

d) 对集合 F 的反馈短信进行短信类型判断,若短信 $\mathbf{d}_j^b(t_i, w_i)$ 属于训练短信的反馈短信,则对其与 $\mathbf{d}_k^s(t_i, w_i)$ 进行支持向量消重处理;

e) 重复 d) 直到所有训练反馈短信支持向量消重完毕,获得消重后的支持向量集 S' ;

f) 取反馈学习短信集 F 与支持向量 S' 的并集作为训练短信集,进行支持向量机的重新优化训练。

由上述可知,SVM 模块主要实现垃圾短信和正常短信的分类,通过基于关键词的学习,反馈到预测模块中,使在线关键词库能够动态更新变化,以提高在线垃圾短信过滤系统的过滤准确率。短信训练主要是基于关键词的学习,融合关键词学习策略的 SVM 增量分类学习过程。

3 实验数据

本文系统的测试数据由浙江联通公司提供三段数据,时间从 2007/05/01 00:07:56 至 2007/05/02 20:12:22(五一劳动节)假日一段数据,时间从 2007/07/01 00:07:56 至 2007/07/03 18:12:22 平常的一段数据和 2007/09/24 00:07:56 至 2007/09/24 24:00:00 中秋节(节日)的一段数据,共计 36 万条,主要测试在不同时段过滤系统的效果比较。系统测试环境为: Intel Pentium 4 CPU, 3.00 GHz, 1 GB 内存, Windows XP 操作系统。

为了评价垃圾短信过滤效果,本文使用三个评价指标:SP(垃圾短信准确率)、SR(垃圾短信识别查全率)和 SE(垃圾短信识别错误率), $N_{junkt \rightarrow junk}$ 为正确识别出垃圾短信数, $N_{normal \rightarrow junk}$ 为正常短信被误判为垃圾短信数, $N_{junk} = N_{junkt \rightarrow junk} + N_{normal \rightarrow junk}$ 为垃圾短信总数。垃圾短信过滤的性能评价采用文本分类的相关指标。具体地,本文的垃圾短信过滤系统的判定结果及其比较分别如表 1 和表 2 所示。

在表 2 中,SP = 1 说明实验数据几乎没有丢,SR 全查率同过训练后效果近提高了近 1.7%,SE 错误率提高了近 0.1%,错误率跟训练的次数有关,但 4 次以后的效果将不明显,系统的过滤效率跟训练样本的存在一定关系。

表 1 垃圾短信过滤系统性能评价表(取 7 月的一段数据共 18 万条)

判断类型	第一段 6 万条短信(反馈前)		第二段 6 万条短信(训练样本)		第一段 6 万条短信(4 次反馈后)	
	实际为垃圾短信	实际为合法短信	实际为垃圾短信	实际为合法短信	实际为垃圾短信	实际为合法短信
系统判定为垃圾短信	$N_{junkt \rightarrow junk} = 3065$	$N_{normal \rightarrow junk} = 308$	$N_{junkt \rightarrow junk} = 8564$	$N_{normal \rightarrow junk} = 523$	$N_{junkt \rightarrow junk} = 3009$	$N_{normal \rightarrow junk} = 251$
系统判定为合法短信	$N_{junkt \rightarrow normal} = 0$	$N_{normal \rightarrow normal} = 56627$	$N_{junkt \rightarrow normal} = 0$	$N_{normal \rightarrow normal} = 50913$	$N_{junkt \rightarrow normal} = 0$	$N_{normal \rightarrow normal} = 56740$

表 2 7 月的某一段数据结果比较

测试项目	6W 垃圾短信				
	反馈前	1 次反馈后	2 次反馈后	3 次反馈后	4 次反馈后
$SP = \frac{N_{junkt \rightarrow junk}}{N_{junkt \rightarrow junk} + N_{junkt \rightarrow normal}}$	1	1	1	1	1
$SR = \frac{N_{junkt \rightarrow junk}}{N_{junk}}$	90.6%	91.1%	91.7%	92.1%	92.3%
$SE = \frac{N_{normal \rightarrow junk} + N_{junkt \rightarrow normal}}{N_{junk}}$	0.513%	0.495%	0.458%	0.423%	0.419%

本系统在节假日峰值时采集的样本数据比较如图 4、5 所示。图 4 表明:本系统在短信特征变化较大的环境中,垃圾短信的过滤效果比平常的要提高 2% ~ 3%;图 5 表示垃圾短信的错误率跟识别垃圾短信特征有关,系统判断为垃圾短信或可疑的短信越多,系统的漏判越高。实验结果表明,本文提出的

反馈学习训练的方法在短信过滤上获得了较好效果,提高了系统的自适应性和智能化水平,改善了系统的过滤效果。

4 结语

本文针对传统垃圾短信过滤系统中存在的主要问题,特别

是垃圾短信的关键词库等特征未能得到更新而导致过滤性能降低的问题,将朴素贝叶斯分类、文本分类和支持向量机等方法引入垃圾短信过滤中,利用离线子系统的训练和学习分析结果反馈到在线子系统的快速过滤模块中,不但满足了短信的准确性和自适应性强的要求,而且根据短信的内容及时地更新在线过滤子系统的特征和内容,使得系统具有更好的自适应性和较高的智能化水平。

未来工作主要有:进一步的优化本文算法;考虑引入控制理论中的相关方法,以实现更好的自适应调节功能;将本文方法应用于下一代垃圾彩铃的过滤应用中。

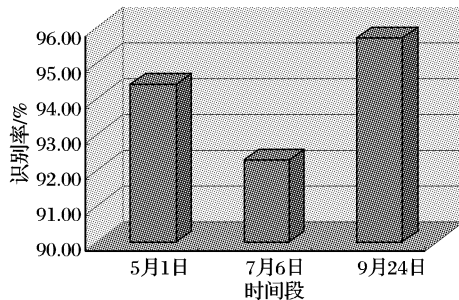


图4 过滤系统在不同时间的过滤效果比较

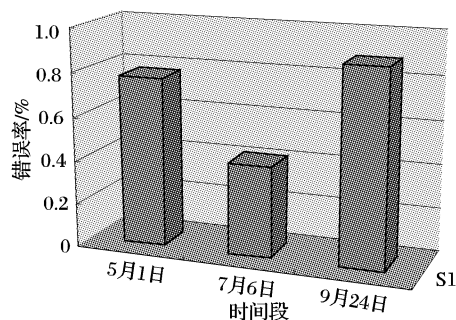


图5 过滤系统的错误率比较

参考文献:

- [1] 王丽萍. 我国垃圾短信的法律分析及防范策略[J]. 湖南税务高等专科学校学报, 2006, 19(3): 42-44.
- [2] 孙晋文, 肖建国. 基于SVM文本分类中的关键词学习研究[J].

计算机科学, 2006, 33(1): 182-184.

- [3] 易阳峰. 垃圾短信的监控与原理实现[J]. 中兴通讯技术, 2005, 11(6): 49-54.
- [4] Di NUNZIO G M. A bidimensional view of documents for text categorization[C]// Proceedings of the 26th European Conference on Information Retrieval (ECIR 2004), Lecture Notes in Computer Science (LNCS) 2997. Germany: Springer Heidelberg, 2004: 112-126.
- [5] FORMAN G. An extensive empirical study of feature selection metrics for text classification[J]. Journal of Machine Learning Research 3, 2003(3): 1289-1305.
- [6] EYHERAMENDY S, LEWIS D D, MADIGAN D. On the naive bayes model for text categorization[C]// Proceedings of the 9-th International Workshop Artificial Intelligence and Statistics. Florida, US: IEEE Press, 2003.
- [7] ANDROUTSOPOULOS I, KOUTSIAS J, CHANDRINOS K V, et al. An evaluation of naive bayesian anti-spam filtering[C/OL]. [2007-09-15]. http://www.ics.forth.gr/~potamias/mlnia/paper_2.pdf.
- [8] ANDROUTSOPOULOS I, PALIOURAS G, KARKALETSIS V, et al. Learning to filter spam e-mail: A comparison of a naive bayesian and a memory-based approach[C]// Proceedings of the workshop: Machine Learning and Textual Information Access. Washington, DC: IEEE Press, 2000: 1-13.
- [9] SEBASTIANI F. Machine learning in automated text categorization [J]. ACM Computing Surveys, 2002, 34(1): 1-47.
- [10] CHANG C C, LIN C J. LIBSVM-A library for support vector machines (version 2.33) [EB/OL]. [2007-09-02]. <http://www.csie.ntu.edu.tw/~cjlin/libsvm/index.html>.
- [11] ANDROUTSOPOULOS I, KOUTSIAS J, CHANDRINOS K V. An evaluation of naive bayesian anti-spam filtering[C]// Proceedings of 2000 Workshop on Machine Learning in the New Information Age. Barcelona: IEEE Press, 2005: 9-17.
- [12] SEBASTIANI F. Machine learning in automated text categorization [J]. ACM Computing Surveys, 2002, 34(1): 1-47.
- [13] 孙晋文, 肖建国. 基于SVM的中文文本分类反馈学习技术的研究[J]. 决策与控制, 2004, 19(8): 927-930.

(上接第 713 页)

2) 我们通过大量实验发现 $\hat{H}_n(p)^h$ (“高兴”时的熵序列统计积分值)与 $\hat{H}_n(p)^n$ (“平常”时的熵序列统计积分值)之间有非常好的规律性。在算法中我们利用了这一规律,并取得了良好的效果。如何更好的诠释 $\hat{H}_n(p)^h$ 与 $\hat{H}_n(p)^n$ 二者之间的规律,提高算法的正确识别率以及算法的鲁棒性是后继的工作之一。

3) 如何将识别算法应用到其他不同的情感当中,以进行更多情感的区别判断。

参考文献:

- [1] 马义德, 袁敏, 齐春亮, 等. 基于PCNN的语谱图特征提取在说话人识别中的应用[J]. 计算机工程与应用, 2005, 41(2): 81-84.
- [2] ECKHORN R, REITBOECK H J, ARNDT M, et al. A neural network for feature linking via synchronous activity: results from cat visual cortex and from simulations[M]// Cambridge: Cambridge University Press, 1989: 255-272.
- [3] 刘勍, 马义德, 钱志柏. 一种基于交叉熵的改进型PCNN图像自动分割新方法[J]. 中国图象图形学报, 2005, 10(5): 579-584.
- [4] DENG L, KHEIRALLAH I. Dynamic formant tracking of noisy

speech using temporal analysis on outputs from a nonlinear cochlear model[J]. IEEE Transactions On Biomedical Engineering, 1993, 40(5): 456-467.

- [5] ECKHORN R, REITBOECK H J, ARNDT M, et al. Feature linking via synchronization among distributed assemblies: simulation of results from cat cortex [J]. Neural Computing, 1990, 2(3): 293-307.
- [6] SKOURIKHINE A N. A pulse couple neural network for image smoothing and segmentation[C]// International Symposium on Computational Intelligence. Kosice: IEEE Press, 2000.
- [7] JOHNSON J L. Pulse-coupled neural networks[C]// Adaptive Computing: Mathematical and Physical Methods for Complex Environments, Proceedings of SPIE. 1996, 2824(11). Proceedings of Adaptive Computing: Mathematics, Electronics and Optics. Orlando FL.
- [8] 张颖, 罗森林. 情感建模与情感识别[J]. 计算机工程与应用, 2003, 33: 98-102.
- [9] 周洁, 赵力, 邹采荣. 情感语音合成的研究[J]. 电声技术, 2005, 10: 57-59.