基于抽样的垃圾短信过滤方法*

钟延辉, 傅 彦, 陈安龙, 关 娜
(电子科技大学 计算机科学与工程学院, 成都 610054)

摘 要: 现有垃圾短信过滤系统主要采用对短信进行逐条分析判断的技术,因此处理的效率比较低。针对这一过滤技术的不足,提出了一个基于抽样的垃圾短信过滤方法,该方法引入用户信任度的概念,根据用户的信任度对用户发送的短信进行抽样过滤,极大地提高了处理效率。同时该方法整合了多项垃圾短信过滤技术(黑白名单、关键词及内容过滤技术),较之单一的过滤方法在准确率和效率上有很大的提高。

关键词: 垃圾短信; 用户信任度; 抽样过滤; 文本分类
中图分类号: TP393 文献标志码: A 文章编号: 1001-3695(2009)03-0933-03

Filtering algorithm of junk SMS based on sample

ZHONG Yan-hui, FU Yan, CHEN An-long, GUAN Na
(School of Computer Science & Engineering, University of Electronic Science & Technology of China, Chengdu 610054, China)

Abstract: The existing filter system of junk SMS use the technology which judge SMS one by one, therefore its efficiency is quite low. To overcome the shortcomings of existing filtering technologies, this paper proposed a filtering algorithm of junk SMS based on sample. Introduced the concept of user's confidence, and filtered messages by SMS center according to user's confidence. Implemented three kinds of filtering technology (black/white list based, key words based, content based) on junk short message filtering method, which increase the efficiency very significantly.

Key words: junk short message; user's confidence; sample filtering; text categorization

近几年来,由于移动通信技术的快速发展,催化了诸多增值服务的产生。短信作为移动通信的增值服务之一,在为人们提供价格低廉和便捷的通信服务的同时,滋生了大量以传播淫秽色情、商业欺诈以及商业广告等不良信息为目的垃圾短信[1],并严重干扰人们生活、妨害社会安全以及造成网络拥塞,垃圾短信的监管问题已经受到社会各界的广泛重视[2]。除了从立法层面加强对信息发布进行监管外,更重要的是要从技术层面探索有效的垃圾短信防范技术。因此,研究短信智能监管技术和系统,建立一个高效、准确、可靠的短信监管平台,实现短信内容的监管,不仅有利于国家安全和社会稳定,保护人民财产和正常生活,而且有重要的社会价值。目前,针对垃圾短信泛滥问题主要的过滤技术有基于黑 3白名单的垃圾短信过滤、基于规则的垃圾短信过滤和基于内容的垃圾短信过滤[3]。黑 3白名单技术有很大的局限性,而基于规则的垃圾术也有一些不尽人意的地方,如规则需要人工指定,没有经验的用户会影响规则的有效性和准确率。鉴于此,目前对垃圾短信的过滤研究很多都集中到内容分析方面。但是只有对短信进行逐条分析,才能判断是否为垃圾短信,因此处理的效率比较低。不管是基于关键词还是基于内容的垃圾短信过滤方法,都存在大量运算,这样会造成短信服务中心网络堵塞[1,2]。

对以上提到的各种技术在垃圾短信过滤中的不足,本文融合黑名单、关键字和内容过滤技术,提出了一个基于抽样过滤的垃圾短信过滤方法,在一定程度上弥补了采用单一处理方法所存在的问题。同时,实验表明该方法提高了垃圾短信过滤的效率和准确性。

基于抽样的垃圾短信过滤系统的架构

传统的垃圾短信过滤方法需要对短信进行逐条分析,才能判断是否是垃圾短信,因此处理的效率较低。但实际情况是大部分用户所发送的短信都不是垃圾短信,没有必要逐条分析。在兼顾垃圾短信过滤的准确率和效率前提下,本文在传统垃圾短信过滤方法的基础上,创新地提出给每个用户设定一个信任度,并根据用户信任度的不同,对发送到短信服务中心的短信进行不同强度的抽样,抽样到的短信根据构成短信的文字信息进行分类,而不必对每一条短信进行逐条分析,提高了垃圾短信处理效率,一定程度上解决了短信服务中心网络堵塞问题。本文设计的基于抽样的垃圾短信过滤系统主要包括以下步骤:

a)为用户设置一个用户信任度。

b)根据短信用户的信任度,对经过短信服务中心的短信进行不同强度的抽样过滤,用户信任度越低,抽样强度越高;反之,亦然。未被抽样到的短信作为正常短信予以发送,被抽样到的短信作为可疑短信由步骤 c)处理。

c)被抽样到的可疑短信采用依据构成短信的文字信息确定短信是否是垃圾短信,如是,直接过滤掉;如果不是,认为是正常短信,予以发送。

d)依据用户短信发送情况,更新该用户的信任度。正常短信的发送量越大,垃圾短信的发送量越小,用户信任度就会提高,反之则降低。

依据构成短信的文字信息判定是否为垃圾短信的过滤方法,主要有前述的基于关键词或基于内容的垃圾短信过滤方法。

收稿日期: 2008-06-10; 修回日期: 2008-08-21 基金项目: 国家"863"计划资助项目(2006AA01Z414)

作者简介: 钟延辉(1984-),男,硕士研究生,主要研究方向为数据挖掘(yanhui1984@gmail.com);傅彦,女,教授,博导,主要研究方向为数据挖掘;陈安龙,男,博士,主要研究方向为数据挖掘;关娜,女,硕士研究生,主要研究方向为数据挖掘.

基于抽样的垃圾短信过滤系统设计如图 1所示。它创新性地采用了基于用户信任度抽样的思想,同时融合了多种传统的垃圾短信过滤技术,构建了一个有效垃圾短信过滤系统。

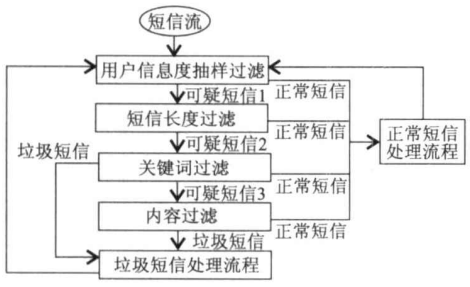

图1 基于抽样的垃圾短信过滤系统架构

系统中的主要过滤模块分析

系统的过滤部分主要包括四个模块,即用户信任度抽样过滤、短信长度过滤、短信关键词过滤和短信内容过滤。以下分别对四个模块进行详细描述。

1)过滤模块 1:用户信任度抽样过滤

所谓用户信任度抽样,是指根据用户的信任度对用户所发送的短信进行相应频度(强度)的抽样过滤。用户信任度是指对用户发送正常短信的信任程度,由用户垃圾短信的发送量、正常短信的发送量和总的短信发送量计算而成。如果用户信任度不同,那么用户抽样监测的频率(强度)也就不一样。被抽样到的垃圾短信作为可疑短信 1,到短信长度的过滤模块进行确认;未被抽样到的短信作为正常短信,并予以发送。

此外,本文结合传统的黑/白名单过滤方法[4]。如果用户的信任度为 0,那么表示该用户为黑名单用户,则该用户所发送的短信是垃圾短信,将直接过滤掉;当用户的信任度为 1,则表示该用户为白名单用户,即特权用户,用户所发送的短信是正常短信,直接给予通过。这样就将本文提出的抽样过滤方法与传统的黑/白名单过滤方法很好地结合起来了。

2)过滤模块 2:短信长度的过滤

接收用户信任度抽样过滤模块来的可疑短信 1,计算短信内容的长度,判断短信内容的长度是否超过设定的阈值。当短信长度超过设定阈值时,该短信为可疑短信 2,到关键词过滤模块进行确认;短信的长度小于设定阈值时,该短信为正常短信,并予以发送。

基于短信长度的过滤方法主要根据设定的垃圾短信最小长度阈值来判断用户发送的短信是否为正常短信。由于短信的长度小于 70个中文字符,可以根据短信的长度对短信进行初步判定,一般长度比较短的短信是垃圾短信的可能性相对比较小。在该模块中的短信长度可以通过分析历史垃圾短信的长度分布模式而动态设定。

3)过滤模块 3:短信关键词过滤

接收短信长度过滤模块来的可疑短信 2。根据设定的关键词列表,监测用户发送的短信中是否含有设定关键词[4]。若含有,则认为是垃圾短信,予以过滤;反之,则该短信作为可疑短信 3到短信内容过滤模块进行确认。

4)过滤模块 4:短信内容过滤

短信内容过滤主要根据短信的内容进行分类过滤。在本文中,该内容过滤主要采用贝叶斯分类算法[5~7],接收关键词过滤模块来的可疑短信 3,对其进行自动分类。分类为正常短信的将直接予以发送;若判断为垃圾短信,则拦截。同时根据新扩充的正常短信和垃圾短信更新分类器。

根据用户的信任度确定对不同短信用户的过滤强度,这样使得大部分用户所发送的短信都能直接通过,只有小部分用户所发送的短信接受相应的检测,大大提高了过滤系统的处理效率;此外,本文还整合了传统的垃圾短信过滤技术,构成了一个多层的垃圾短信过滤方法,对抽样到的短信进行有效的判别,较之采用单一过滤技术,在准确性和效率上都有了很大提高。

系统主要处理流程

. 用户信任度抽样流程

首先给每个用户设定一个信任度。以下是用户信任度抽样过滤模块的具体算法描述:

a)初始化用户信息。统计用户总的短信发送量 sum_i,正常短信的发送量 m_i。用户的信任度为

$$credit_i = \begin{cases} credit\_MIN & m_i / sum_i \le credit\_MIN \\ m_i / sum_i & 其他 \\ credit\_MAX & m_i / sum_i \ge credit\_MAX \end{cases} \quad (1)$$

其中:i是用户编号;最小信任度 credit_MIN和最大信任度 credit_MAX可以根据情况自适应地调整以符合实际要求。为了防止某些普通用户的信任度过低导致用户发送的短信直接被认为是垃圾短信,或者信任度过高导致对使用用户疏于监控。黑名单用户信任度为 0,特权用户信任度为 1,直接过滤或予以发送,这类用户不需要进行抽样分析短信的文字内容,即可判定短信是否是垃圾短信。此时,用户发送垃圾短信的频率为

$$junk\_p_i = (sum_i - m_i) / sum_i = 1 - credit_i \quad (2)$$

当用户为没有历史信息的新用户时,则用户信任度设为 credit = credit_MIN,用户短信的发送量 sum_i 和正常短信的发送量 m_i 都为 0,此时需要连续对新用户发送的短信进行抽样,以获取用户发送短信的情况。设置有连续抽样标志 $flag_i$。$flag_i$ = true表示需要连续抽样;$flag_i$ = false表示信任度概率连续抽样。当用户发送的短信需要连续抽样时,n表示连续为正常短信的最大条数。此时设有一个计数器 $n1_i$,表示连续检测时连续为正常短信的数量,初始值为 0。

b)接收用户短信,检索用户信任度,判断是否为 0、1或其他值。如果为 0,则为黑名单用户,短信予以拦截;如果为 1,则为特权用户,短信予以直接通过,并发送;如果为其他值,则转到步骤 c)。

c)用户短信发送量 $sum_i = sum_i + 1$,当连续抽样标志 $flag_i$ = true时,转到短信长度过滤模块;否则按信任度概率选择当前短信是否需要抽样检测。若抽样到则转到短信长度过滤模块;否则进入正常短信处理流程。

. 用户短信长度过滤流程

从用户信任度抽样过滤模块中接收可疑短信 1,进行短信长度过滤。设垃圾短信的长度大于等于 x,因此,当短信长度小于 x时,则认为该短信是正常短信,予以直接通过并发送;对短信长度大于等于 x的短信作为可疑短信 2,需进入关键词过滤模块确认。其具体的过滤步骤为:

a)接收用户信任度抽样过滤模块来的可疑短信 1;

b)计算每条短信的长度 x;

c)判断短信的长度是否大于设置的阈值 x,如果小于设定阈值转到步骤 d),否则转到步骤 e);

d)短信为正常短信,进入正常短信处理流程;

e)该短信为可疑短信 2,转到关键词过滤模块进行处理。

. 关键词过滤流程

根据设定的关键词列表,监测用户发送的短信中是否含有设定关键词。

a)接收短信长度过滤模块来的可疑短信 2。

b)根据设定的关键词列表,判断短信中是否含有列表中的关键词。若无,则将该短信作为可疑短信 3,转到短信内容过滤模块;若有,则认为是垃圾短信,进入垃圾短信处理流程。

. 内容过滤流程

短信内容过滤主要根据短信的内容进行分类过滤。具体的过滤步骤如下:

a)接收短信关键词过滤模块来的可疑短信 3作为待分类短信。

b)对训练样本和可疑短信进行分词。

c)基于几率比的特征提取。

d)采用贝叶斯分类算法,对可疑短信 3进行分类:

(a)若是垃圾短信,则进入垃圾短信处理流程。

(b)若是正常短信,则进入正常短信处理流程。

. 正常短信处理流程

当用户短信确定为正常短信后,正常短信的发送量 $m_i = m_i + 1$,并判断连续抽样标志 $flag_i$ 是否为 true:如果 $flag_i$ = false,则该短信予以发送,返回到用户信任度过滤模块;如果 $flag_i$ = true,则连续正常短信数 $nl_i = nl_i + 1$。判断 nl_i 是否小于 $(1 - credit_i) \times n$:如果 $nl_i < (1 - credit_i) \times n$,则发送该短信,返回到抽样过滤模块;若 $nl_i \geq (1 - credit_i) \times n$,则 $flag_i$ = false, $nl_i = 0$,并发送该短信,返回到过用户信任度抽样过滤模块。

. 垃圾短信的处理流程

当用户短信确定为垃圾短信后,判断连续抽样标志 $flag_i$ 是否是 true:如果 $flag_i$ = true,则连续正常短信数 $nl_i = 0$;若 $flag_i$ = false,则连续正常短信数 $nl_i = 0$, $flag_i$ = true。将该短信予以拦截,返回到用户信任度抽样过滤模块。

实验与分析

. 实验数据

目前对垃圾短信过滤的研究还不多见,由于短信涉及个人隐私问题,公开的短信语料暂时还没有。为此,本文自建了 2 000条短信的语料库。该短信语料库含有垃圾短信和正常短信各 1 000条。将这 2 000条短信各自均分成 5份,随机各自选取 3份合成训练集(1 200条),其余合成测试集(800条)。共做 5组实验,即 5次交叉验证,取其平均值作为参评结果。

. 实验评价标准

目前对垃圾短信过滤性能进行评价的指标主要有以下两个方面:

首先,假设测试集合中共有 N条短信,定义几个变量(表 1),显然有 N = A + B + C + D。

定义如下评价指标:

a)召回率(recall): recall = A/(A + C) ×100%

这个指标反映了垃圾短信过滤的能力,召回率越高,误否认就越少。

b)正确率(precision): precision = A/(A + B) ×100%

这个指标反映了过滤系统辨别垃圾短信的能力,正确率越大,将非垃圾短信判定为垃圾短信的可能性就越小。

表 1 变量定义表

判　断	垃圾短信	合法短信
判为垃圾短信	A	B
判为正常短信	C	D

. 实验结果与分析

实验在 Windows XP平台上进行,采用基于 VC ++平台的 C ++编程语言实现。其中分词算法采用了中国科学院计算技术研究所的分词系统 ICTCLAS(Institute of Computing Technology, Chinese Lexical Analysis System);特征选择采用几率比(odds-ratio, OR)方法。特征词库的存储采用哈希表存储结构。

表 2和 3表明:随着特征数量的增加,两个方法中的召回率和正确率都在相应增加,说明在内容过滤部分恰当地选择特征数量对分类的召回率和正确率有很大的影响。此外,本文所采用的方法比单一的贝叶斯方法在垃圾短信过滤上的召回率和准确率都有很大的提高,而且处理时间也大大降低(表 2和 3)。但由于目前中文分词技术及贝叶斯分类算法还存在一些不足,本文中的方法在准确率方面必然受到一定的影响。

表2 单一贝叶斯过滤方法实验结果

特征数量	召回率/%	准确率/%	处理速度/s
1 000	80.11	81.21	4.01
2 000	84.31	85.24	4.07
3 000	89.96	90.31	4.12

表3 抽样过滤方法实验结果

特征数量	召回率/%	准确率/%	处理速度/s
1 000	85.21	85.86	0.93
2 000	90.02	90.75	1.96
3 000	96.11	96.28	0.99

结束语

本文提出了一个基于抽样的垃圾短信过滤方法,创新性地引入了用户信任度的概念,有效地对每个用户进行不同强度的监管,同时引入可疑短信库,弥补了把正常短信判为垃圾短信的损失。本文的实验结果是在程序初始运行时得出,由于该方法具有自动更新的功能,随着程序的持续运行,其效率和准确性将会有所提高。下一步工作是寻找一种更有效的适合于贝叶斯分类的特征提取方法以及从垃圾短信库中自动有效提取关键字的方法。

参考文献:

[1] 人民网. 数字中国[EB/OL]. (2005-05). http://www.people.com.cn.

[2] 中国社会科学院. 手机托起"第五媒体"[EB/OL]. (2005). http://www.cass.net.cn

[3] 易阳锋. 垃圾短信监控的原理与实现[J]. 中兴通讯技术, 2005, 12(6): 49-54.

[4] 胡于进,周小铃,凌铃,等. 基于向量空间模型的贝叶斯文本分类方法[J]. 计算机与数字工程, 2004, 32(6): 28-30, 77.

[5] CRISTIANNINI N, SHAWE T J. An introduction to support vector machines and other kernel-based learning methods[M]. Cambridge: Cambridge University Press, 2000.

[6] LIU Tao, LIU Shengping, CHEN Zheng, et al. An evaluation on feature selection for text clustering[C]//Proc of the 20th International Conference on Machine Learning (ICML-03). 2003: 488-495.

[7] TAN A H, YU P. A comparative study on Chinese text categorization methods[C]//PRICAI 2000 Workshop on Text and Web Mining. 2000: 24-35.