基于复杂网络的垃圾短信过滤算法

黄文良 [1,2,3] 刘 勇 [1,2] 钟志强 [3] 沈仲明 [3]

摘 要 对垃圾短信发送用户的识别和过滤具有十分重要的研究价值和社会意义. 随着新形式和内容的垃圾短信出现, 传统的关键字匹配和发送速度频率过滤方法无法有效地处理这一问题. 在对短信发送/接收网络形式化表达的基础上, 以真实短信发送和接收以及通话关系数据为例, 统计和分析了短信发送网络的网络特性. 进一步分析和挖掘了垃圾短信用户在网络上发送接收的异常模式和行为, 并以此提出了一个基于语音关联程度和短信回复比率的过滤算法 (NASFA 算法). 通过实验和分析表明, 本文的算法能够高效地识别垃圾短信发送用户, 同时能够有效地控制将正常用户误识别为垃圾短信用户的比率.

关键词 复杂网络, 无标度网络, 垃圾短信过滤, 幂律, 出入度比
中图分类号 TP391.4

Complex Network Based SMS Filtering Algorithm

HUANG Wen-Liang[1, 2, 3] LIU Yong[1, 2] ZHONG Zhi-Qiang[3] SHEN Zhong-Ming[3]

Abstract It is very important to recognize and filter the spam short messages (SMS). As the contents and formats of spam messages are diverse, the ordinary filtering methods based on keyword matching and sending speed can not tackle this problem effectively. This paper first presents a formalized representation of the SMS network. On the basis of real short message samples, the social characteristics of the SMS network are analyzed and studied. Further analysis and statistical work are carried out to discover the un-normal patterns of spam senders in SMS network. An N-degree association spam filter algorithm (NASFA) based on the un-normal patterns of spam senders is presented. Experiments and analysis show that the algorithm can efficiently recognize spam senders, and the wrong recognition rate is reduced significantly.

Key words Complex network, scale-free network, spam short messages (SMS) filter, power law, out-in degree ratio

手机短信这种先进的通信手段给人们带来了很多便利, 但与此同时, 垃圾短信也开始泛滥, 给我们的生活带来了很多负面影响, 成了社会一大公害. 统计数据表明, 数目庞大的发送短信总量中, 30 % 以上是垃圾短信. 垃圾短信中的绝大部分都是属于恶意欠费, 给运营商造成了极大的经济损失; 同时垃圾短信中存在着大量的谣言、诽谤和诈骗等影响社会安定团结的不良信息, 因而如何有效地过滤掉这些垃圾短信不仅仅是经济问题, 还是当前的一个重大社会问题. 为了解决垃圾短信泛滥问题出现了一系列的应对方法, 如关键字过滤方法 (即通过设置屏蔽的关键字辨识垃圾短信发送用户) 和发送速度频率过滤法 (通过短信发送速度频率来识别垃圾短信发送用户) 等, 但是使用效果都不明显. 从图论角度看短信发送网络在网络结构上与邮件发送网络具有许多的相同点. 目前在国际上已经出现了一些利用复杂网络理论技术来实现垃圾邮件的发现和过滤的工作 [1−2], 并且取得了一系列成功的应用. 因而可从中借鉴, 利用复杂网络的方法和理论来研究短信发送网络, 尤其是研究其中的垃圾短信发送现象的网络模式, 并提出相应的过滤方法. 文章接下来的内容首先给出短信发送网络的形式化的定义和描述, 通过真实短信数据证明短信网络满足小世界特性和无标度网络特性. 在此基础上进一步通过统计实验对比垃圾短信发送用户和正常用户的网络特征模式, 主要从其与通话网络的关联性和短信发送回复比率的角度进行对比实验, 并基于以上两个特征提出了一个垃圾短信过滤算法; 最后通过详细的实验以及与其他过滤方法的对比, 分析并总结各种垃圾短信过滤算法的优劣.

收稿日期 2008-02-29 收修改稿日期 2008-12-16
Received February 29, 2008; in revised form December 16, 2008
国家自然科学基金 (60803053), 国家博士后科学基金 (20070420231), 国家博士科学基金 (20081459) 资助
Supported by National Natural Science Foundation of China (60803053), China Postdoctoral Science Foundation (20070420231), and Excellent Postdoctoral Science Foundation of China (20081459)

1. 浙江大学工业控制技术国家重点实验室 杭州 310027 2. 浙江大学智能系统与控制研究所 杭州 310027 3. 中国联合网络通信有限公司浙江分公司 杭州 310006
1. State Key Laboratory of Industrial Control Technology, Zhejiang University, Hangzhou 310027 2. Institute of Cyber-Systems and Control, Zhejiang University, Hangzhou 310027 3. China United Network Communication Corporation, Zhejiang Branch, Hangzhou 310006

DOI: 10.3724/SP.J.1004.2009.00990

1 短信网络特性分析

下面首先给出短信网络的一些形式化的概念和定义.

短信发送网络可以形式化地表示为由点集 V 和边集 $\vec{\boldsymbol{E}}$ 组成的有向图 $G = (V, \vec{\boldsymbol{E}})$, 结点数记为

$N=|V|$, 边数记为 $M=|\overrightarrow{\boldsymbol{E}}|$. 其中点集 V 中的节点表示短信网络中的手机发送或接收号码, 边集中的边 e_{ij} 表示号码 v_i 发送短信给 v_j. 节点 v_i 的度 k_i 定义为与该节点的出度与入度之和. 其中节点的出度和入度分别记为 k_i^{out} 和 K_i^{in}, 表示与节点 v_i 存在短信联络的节点中, v_i 作为发送方或接收方的相邻节点数. 节点 v_i 的出度和入度的边集可记为 $\overrightarrow{\boldsymbol{E}}(v_i)_{\text{out}}$ 和 $\overrightarrow{\boldsymbol{E}}(v_i)_{\text{in}}$.

$$\overrightarrow{\boldsymbol{E}}(v_i)_{\text{out}}=\{\cup e_{ij}|\forall v_j\in V, e_{ij}\in\overrightarrow{\boldsymbol{E}}\}, k_i^{\text{out}}=|\overrightarrow{\boldsymbol{E}}(v_i)_{\text{out}}|$$

$$\overrightarrow{\boldsymbol{E}}(v_i)_{\text{in}}=\{\cup e_{ji}|\forall v_j\in V, e_{ji}\in\overrightarrow{\boldsymbol{E}}\}, k_i^{\text{in}}=|\overrightarrow{\boldsymbol{E}}(v_i)_{\text{in}}|$$

令节点 v_i 度所包含的边集为 $\overrightarrow{\boldsymbol{E}}(v_i)$, 则显然有 $\overrightarrow{\boldsymbol{E}}(v_i)=\overrightarrow{\boldsymbol{E}}(v_i)_{\text{in}}\cup\overrightarrow{\boldsymbol{E}}(v_i)_{\text{out}}$, 且 $k_i=k_i^{\text{out}}+k_i^{\text{in}}$. 为了后续计算方便, 这里也给出了短信发送网络的无向图 $G=(V,E)$ 定义, 其中结点集 V 的定义与有向图中的定义相同, 边集中的边定义为只要节点间存在短信发送或接收关系, 则此对节点之间就存在一条边. 因而节点的度定义为与该节点连接的其他节点数目.

为了有效地获得短信发送网络的具体特性, 本文中采用了浙江联通公司一个月时间段内某一段号码区间内的手机间短信发送和通话记录数据作为样本进行测量和统计, 分析其网络特性. 本文后续的实验和分析都是建立在此数据集之上. 该数据集包含的短信和通话网络的基本参数如表 1 所示.

表 1　短信发送网络特征统计

Table 1　Statistical characteristics of SMS network

特征	短信网络	通话网络
网络结点数目 N	53 509	53 509
非零出度结点数 (即存在短信发送记录或通话记录的结点数)	43 668	47 830
短信发送和通话记录总数	7 310 643	18 526 147
网络边数 (包含出度和入度)	584 853	2 366 628
平均度 $\langle k\rangle$	10.93	49.48
聚类系数 C	0.71	0.75
平均距离 L	5.4	3.2

表 1 中的平均度是有向图的出度入度的平均值, 计算公式为 $\langle k\rangle=|\overrightarrow{\boldsymbol{E}}|/|\boldsymbol{V}|$, 聚类系数[3] 定义在无向图 G 上, 节点 v_i 的聚类系数定义为节点实际存在的边数 $|E_i|$ 和总的可能的边数 $k_i\ (k_i-1)/2$ 之比, 整个网络的聚类系数 C 就是所有度数大于 2 的节点 (度为 1 的节点对计算无影响故去掉) 的聚类系数的平均值[1], 即

$$C=\frac{1}{N_2}\sum_i\frac{2|E_i|}{k_i\,(k_i-1)}$$

式中 N_2 表示度数大于 2 的节点数. 类似地, 网络结点 v_i 和 v_j 间的距离 d_{ij} 定义在无向图上, 表示这两个节点间的最短路径上的边数. 因此可以得出网络平均距离的 L 的定义为任意两个节点间的距离平均值, 即

$$L=\frac{2}{N\,(N+1)}\sum_{i\geq j}d_{ij},\ \text{且若}\ i=j,\ \text{则有}\ d_{ij}=0$$

从表 1 可知, 短信网络具有较低的平均距离 L 和较高的聚类系数 C, 表明该网络符合 Watts 和 Strogtz 提出的小世界模型[4−6], 即网络节点在局部呈献出非常高的连接程度 (高聚类系数), 同时联通到任意点的距离较短 (即非常低的网络平均距离).

复杂网络的另外一个特性是无标度性. 短信网络是一种典型的无标度网络, 具有 BA 模型[7] 的一些特征, 即增长 (Growth) 特性和优先连接 (Preferential attachment) 特性. 但是短信网络的增长模型要稍微不同于经典的 BA 模型, 文献 [8] 指出短信网络中, 一个联系人因为认识网络中的某个人而加入到网络中, 同时随着时间推移, 他会以一定的概率认识该人的朋友, 然后以较小的概率认识该人朋友的朋友. 对表 1 中数据的度分布分析可得短信网络中分布符合 Stretched exponential 分布[9], 即短信网络具有无标度性[10−13].

2　垃圾短信发送模式

为了有效地过滤掉短信发送过程中的垃圾短信, 就必须先了解对应的垃圾短信发送模式, 分析和获取其发送的普遍模式和规律, 并据此制定相应的过滤算法. 本节中从两个方面对垃圾短信发送用户的网络特性进行分析, 分别是垃圾短信发送用户与其对应的语音通话网络之间的关系和垃圾短信发送用户的短信回复比例.

2.1　短信发送与通话网络的相关性分析

从垃圾短信发送的目的性分析, 其发送对象多为匿名发送, 即发送者与接收者之间不认识. 故在网络特性中表现为, 短信发送者与接收者之间几乎很少存在语音通话记录. 首先给出短信网络在通话网络上的相关性定义.

定义 1 (短信网络的语音相关性). 对无向图的短信网络 G_s 和通话网络 G_p, 若节点 v_i 和 v_j 是语音相关, 则短信网络的边 $e_{ij}\in E_s$ 且在通话网络中 $d_{ij}\neq\infty$.

上述定义表明, 若短信发送网络中存在联系的两个节点是语音相关的, 则该节点对至少在通话网络中存在一条通路.

基于以上的相关性定义, 我们可以给出 “短信 –

语音” 的 N-度相关的定义.

定义 2 (N-度相关). 对于同一组节点构成的短信网络 G_s 和通话网络 G_p, 若节点 v_i 和 v_j 是 N-度相关仅当 $e_{ij} \in E_s$ 且在通话网络中 $d_{ij} = N$, 记为 $(v_i, v_j)_N$.

在表 1 数据的基础上, 我们分别选取了一个典型的普通短信发送用户和垃圾短信发送用户作为观测对象, 分析其短信发送的语音相关性. 对比正常短信发送用户和垃圾短信发送用户, 可以发现正常用户的语音相关性要显著地高于垃圾短信发送用户. 普通短信用户经过三次语音相关性关联之后基本上都与源节点建立关联, 而垃圾短信发送者在 3-度相关下绝大部分节点仍然与源节点没有关联.

对于整个短信网络, 需要定义一个能够反映所有短信发送节点联络关系与通话关系之间的相关性. 因此, 我们给出网络 N-度相关概率的定义.

定义 3 (网络 N-度相关概率). 对于同一组节点构成的短信网络 G_s 和通话网络 G_p, N-度相关概率为

$$P_N = P[(v_i, v_j)_1, \cdots, (v_i, v_j)_N | e_{ij} \in E_s]$$

且有 $v_i, v_j \in V$, $i \neq j$.

由定义 3 可知, 网络 N-度相关概率是短信网络无向图中 i-度相关的边所占总边数的比例 ($1 \leq i \leq N$), 因此可以得到

$$P_N = \frac{\sum\limits_{e_{ij} \in E_s} H(d_{ij}^p \leq N)}{|E_s|}$$

式中, d_{ij}^p 是节点 v_i 和 v_j 在语音网络中的距离, $H(x)$ 是假设检验函数

$$H(x) = \begin{cases} 1, & x = \text{true} \\ 0, & x = \text{false} \end{cases}$$

根据 N-度相关概率的定义, 表 2 给出了短信发送网络与语音通话网络之间的相关性统计.

表 2 短信发送网络的语音相关性统计

Table 2 Statistical characteristics of SMS network corresponding to phone call network

序号	N-度相关概率
P_1	0.6007
P_2	0.8612
P_3	0.9872

上面统计数据表明, 绝大多数短信网络中的发送关系 (边) 都能够在通话网络中找到一条在三跳之内的路径. 由于统计采用的数据中包含的垃圾短信发送节点数目非常低, 不到整个网络的 1 %, 因而其对统计结果的影响较小, 故以上特性可以认为是普通用户的短信网络所具有的.

2.2 短信网络回复分析

由于垃圾短信发送的多为广告或反动色情内容, 一般正常用户收到垃圾短信后基本不会回复 (不排除仍然有极小部分用户回复的可能). 因而可以从用户的短信发送回复情况上予以考虑垃圾短信发送者的行为模式.

一般正常用户发送短信对象基本上都会有回复行为, 即源节点和目的节点间建立的是双向链接, 而仅仅有少部分发送对象没有回复, 分析后发现, 此部分所占比例极低的节点没有回复的原因可能是用户发错或者是回复的延迟 (本文使用数据中仅包括一个月的数据). 而垃圾短信的发送用户则可以明显地发现图中双向边的比例非常低, 绝大部分是单向边.

为了进一步分析整个网络中的用户短信发送/回复的情况, 本文中给出一个短信收发比率的定义.

定义 4 (短信收发比率). 短信发送网络有向图 $G = (V, \overrightarrow{\boldsymbol{E}})$ 中, 结点 v_i 的短信收发比率定义为节点的入度与出度之比, 记为 λ, 计算公式为

$$\lambda = \frac{k_i^{\text{out}}}{k_i^{\text{in}}}$$

短信回复比率能够较好地表征某一节点的短信发送和接收的比例, 从而反映出用户短信发送和回复的情况, 若 $\lambda = 1$ 表明用户和其联系的用户间都存在双向联系, 而 $\lambda < 1$ 表明用户被联系多于其联系他人, $\lambda > 1$ 表明用户联系他人多于被其他人联系. 因此, 我们可以类似图 3 中的方法, 统计网络的收发比率分布, 即 $P(\lambda)$, 实验结果如图 1 和图 2 所示. 图中的纵坐标为收发比率为 λ 的节点数 ($P(\lambda) \propto \text{Frequency}(\lambda)$), 且为了避免引入不必要的

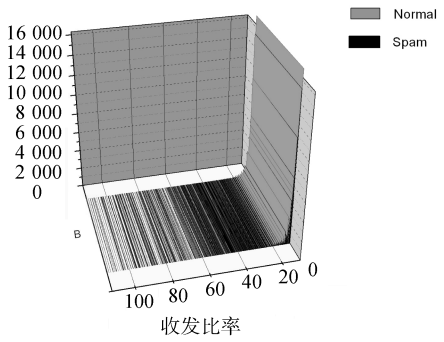


图 1 垃圾短信与正常短信的回复统计柱状图

Fig. 1 Replying statistics between the spam SMS network and normal SMS network

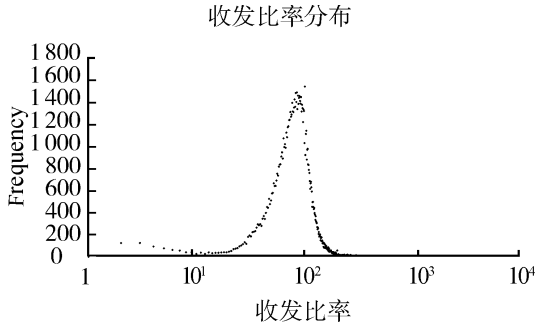


图 2　短信收发比率分布 (横坐标为对数坐标)

Fig. 2　Distribution of the receiving and replying ratio (x-coordinate is logarithmic)

误差, 仅仅统计满足 $|\overrightarrow{\boldsymbol{E}}_i| \geq 20$ 的节点 v_i. 图 1 中浅色柱表示所有的节点都是正常的短信发送用户, 而深色柱则表示在此短信收发比率中包含有垃圾短信发送用户. 经统计可以发现垃圾短信用户基本上分布于 $\lambda > 30$ 一端. 图 2 中采用描点方法给出 λ 的分布情况 (λ 取对数坐标), 可以看出呈泊松分布的态势.

3　过滤算法

上文对短信网络的特征以及垃圾短信发送者的行为模式的分析统计, 表明垃圾短信发送行为在短信回复比率和语音关联程度上存在较大的异常. 因此, 本节给出了一个综合 N-度相关和短信回复比率的过滤算法 (N-degree association spam filter algorithm, NASFA). 该算法对短信发送节点的发送短信进行监控运算, 并根据节点发送短信的网络特性和模式, 判断该节点是否为垃圾短信发送用户.

算法 1 (N-degree association spam filter algorithm, NASFA).

输入. N, λ_0, ω_0, ϖ, v_i

步骤 1. 令 $\omega_i \leftarrow 0$;

步骤 2. 对于以节点 v_i 为发送者的每一条短信, v_j 是短信接收节点;

步骤 3. 若 $(v_i, v_j)_N$ = false, 则 $\omega_i \leftarrow \omega_i + \varpi$, 否则 $\omega_i \leftarrow 0$;

步骤 4. 若 $\omega_i > \omega_0$, 转步骤 5, 否则转步骤 2;

步骤 5. 计算发送接收比率 $\lambda_i = k_i^{\text{out}}/k_i^{\text{in}}$;

步骤 6. 若 $\lambda_i \geq \lambda_0$, 将 v_i 加入阻止名单 (垃圾短信名单), 否则转步骤 1.

NASFA 算法中的输入值 v_i 是待检验节点; N 是语音相关度参数, 由表 1 可知语音网络的平均距离为 3.2, 且从表 2 可知正常用户网络中的三度语音相关后就可覆盖绝大部分的正常用户, 因而一般取 $N = 3$; λ_0 是收发比率阀值, 由短信网络收发比率的统计实验可知, 一般取 $\lambda_0 > 30$ 可较好地区分垃圾短信发送用户和正常用户; ω_0 和 ϖ 分别是过滤敏感性阀值和增长步长, ω_0/ϖ 的取值可控制过滤的敏感程度, ω_0/ϖ 越大则敏感性越低, 反之则敏感性越高. NASFA 算法的工作原理可以用图 3 的状态转换关系描述. 初始节点 v_i 处于 "未标记号码" 状态, 此时 NASFA 算法监控节点 v_i 和其接收方节点的语音相关性, 若不相关则相应的控制阀值递增 $\omega_i \leftarrow \omega_i + \varpi$, 否则若超过敏感性阀值转入下一状态 "可疑号码" 状态. 此时再计算节点的收发比率, 若超出预期值则此号码未垃圾短信发送号码, 转入 "垃圾号码" 状态, 否则重置节点敏感性参数, 返回 "未标记号码" 状态.

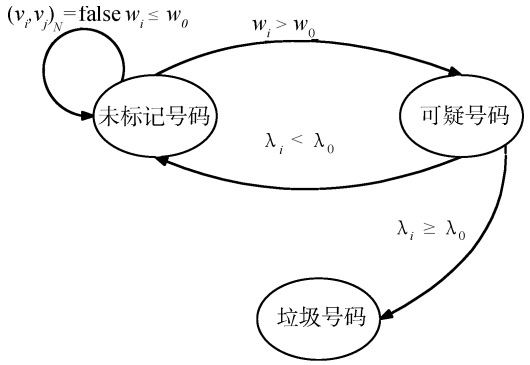


图 3　垃圾短信用户屏蔽算法状态转换关系

Fig. 3　Transition of the spam sender recognition algorithm

NASFA 算法是一个常驻后台运行的短信垃圾实时监控算法, 因而对其效率的要求非常高. 分析发现, 其中的主要计算复杂度耗费在节点 N-度相关的计算上. 当然, 也可以将 N-度相关问题看作是寻找节点间的最短路径问题, 使用 Dijkstra 算法[14] (计算该点到其他所有点间的最短距离) 或 Floyd 算法[15] (计算所有点之间的最短距离), 但是显然以上两种算法时间复杂度仍然过大, 时间复杂度分别为 $O(N^2)$ 和 $O(N^3)$ (N 是网络节点数) 且需要大量的空间存放中间结果. 基于通话网络和短信网络的小世界特性, 我们可以考虑一些专门针对小世界网络中的搜索算法[16], 降低计算复杂度.

由于过滤算法中一般最多只需要取到 4-度相关, 即只需要判断最短距离在 4 以内的节点. 而短信过滤中需要能够尽可能快地实现节点之间的相关性关系判断. 因此, 我们给出了一个支持计算节点之间是否存在 4-度相关以内关系的判别算法, 基于广度优先策略的 N-度相关邻域判定算法 (N-degree association neighborhoods algorithm, NANA), 在介绍算法之前先给出一些相关的定义.

定义 5 (节点 a 的邻域). 无向图 $G(V, E)$ 中, 节点 a 的邻域定义为与 a 有连接的所有节点构成的集合, 记为 A^*, 其计算式为

$$A^* = \{\cup b | \forall\, b \in V,\ e_{ab} \in E\}$$

类似地, 可以给出节点 a 的二级邻域 A^{**}:

$$A^{**} = \{\cup B^* | \forall\, b \in A^*\}$$

NANA 算法如下.

算法 2 (*N*-degree association neighborhoods algorithm, NANA).

输入. 节点 $a, b \in V$, N

输出. $H(d_{ab} \le N)$ //判断节点 a 和 b 间的关联度是否在 N 以内, $N = 1, 2, 3, 4$

步骤 1. 计算 A^*, B^*;

步骤 2. if $a \in B^*$, then $d_{ab} = 1$;

步骤 3. if $A^* \cap B^* \neq \phi$, then $d_{ab} = 2$;

步骤 4. if $A^{**} \cap B^* \neq \phi$, then $d_{ab} = 3$;

步骤 5. if $A^{**} \cap B^{**} \neq \phi$, then $d_{ab} = 4$;

步骤 6. return $H(d_{ab} \le N)$.

NANA 算法每次进行查询时仅需要搜索节点 a 和 b 的两个邻域 A^* 和 B^*. 若节点 a 与节点 b 的邻域 B^* 存在交集, 则表明此时 a 和 b 的距离为 1 (即 1-度相关); 若 A^* 和 B^* 相交不空, 则表示距离为 2; 对于 3-度相关只需要判断 A^{**} 与 B^* 相交是否为空; 类似地, A^{**} 与 B^{**} 相交不空, 则 4-度相关.

NANA 算法判断节点 1-度相关的时间复杂度为O($\langle k \rangle$), 其中 $\langle k \rangle$ 是图的平均度数. 采用 Hash 方法计算 2-度相关 (步骤 3) 的时间复杂度也为 O($\langle k \rangle$), 类似地, 采用 Hash 方法计算 3-度相关 (步骤 4) 的时间复杂度也是 O($\langle k \rangle$), 而步骤 5 的 4-度相关的时间复杂度为 O($\langle k \rangle^2$). 由于在 NASFA 算法中一般只需要计算 3-度相关, 应该可以保证是在 $\langle k \rangle$ 线性时间内完成相关度的判断.

4 实验与讨论

4.1 实验

本节通过一系列实验来验证本文提出的过滤算法的性能. 实验采用表 1 中的短信发送数据集, 并且根据其发送的先后顺利依次经由 NASFA 算法处理. 为了对比 NASFA 算法的有效性, 还同时采用了短信过滤中常用的关键字过滤和发送速度频率过滤法作为对比.

为了定量分析算法, 假设正常短信用户为正样本 (Positive) 数为 P, 垃圾短信发送用户为负样本 (Negative) 数为 N, 过滤算法正确识别样本数为 T (True), 过滤算法错误识别的样本数为 F (False). 因而有, 真正 (True positive, TP) 被模型预测为正的正样本, 表示正常用户被正确识别数; 真负 (True negative, TN) 被模型预测为负的负样本, 表示垃圾用户被正确识别数; 假正 (False positive, FP) 被模型预测为正的负样本, 表示垃圾用户被错误识别为正常用户数; 假负 (False negative, FN) 被模型预测为负的正样本, 表示正常用户被错误识别为垃圾用户数. 对于垃圾短信过滤系统来说, 一个性能良好的过滤算法必须使得本来是垃圾发送用户的正确识别率 (TN) 要尽可能的高, 但是前提是必须保证正常用户不受到影响的情况下, 即要求正常用户被错误识别为垃圾用户的比率 (FN) 要尽量低 (对于运营商来说宁可放过部分垃圾短信, 也必须保障正常用户的使用). 因此, 本文中的实验主要衡量如下两个参数:

1) 假负率 (False negative rate, FNR), 计算如下:

FNR = FN /(TP + FN), 即 FNR 为被预测为负的正样本数/正样本实际数.

2) 真负率 (True negative rate, TNR), 计算如下:

TNR = TN /(TN + FP), 即 TNR 为负样本预测结果数/负样本实际数.

一般来说, FNR 越低则表明过滤算法的误识别越低, 同时 TNR 越高则表明过滤算法识别垃圾用户的性能越高.

我们选取了三组关键字作为基于关键字匹配过滤方法的匹配关键字, 分别记为 K_1, K_2 和 K_3. 其中 K_1 是目前基本上所发现的垃圾短信为样本分词后提取出的关键字列表, K_2 和 K_3 分别是从 K_1 中随机选取了 50 % 和 1 % 比率的关键字形成的过滤关键字集.

同时对于发送速度频率过滤法, 我们选取了三组速度作为测试, 分别是 V = 50 条/分, 300 条/分, 350 条/分. 实验结果如图 4 (见下页) 所示.

NASFA 算法采用的参数设置和实验结果如表 3 和图 5 (见下页) 所示.

表 3 NASFA 算法实验参数设置及实验结果 ($N = 3$, $\lambda_0 = 30$)

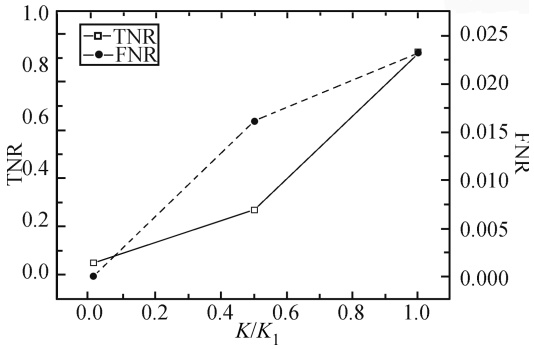
Table 3 Experimental parameters and results of NASFA

ω_0/ϖ	FNR (%)	TNR (%)
1	0.118	100
4	0.077	100
15	0.0075	98.33
20	0	90.79

从图 4 中的实验结果可以发现, 基于关键字匹配的过滤算法在关键字取得较全的时候, 能够较好地识别垃圾短信用户 (TNR = 84.04 %), 但是也相应地将大量的正常用户误识别为垃圾用户 (FNR = 2.32 %, 将近 1 230 个正常用户被错误识别), 这显然无法在实际中所接受. 而如果关键字匹配集设置过

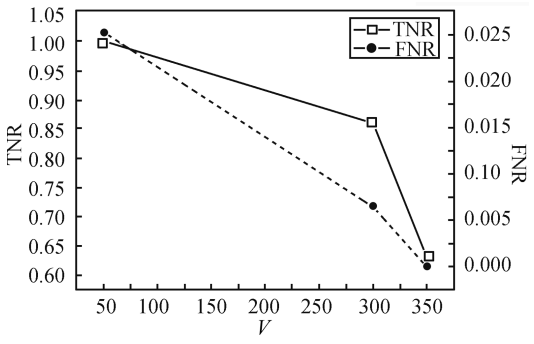
于简单又基本上起不到过滤的作用 (K_3 的 TNR 仅为 4.18 %). 同时, 当前的垃圾短信用户往往可以通过拆字的方法避开关键字过滤, 并且随着垃圾短信内容的更新, 需要人工去更新关键字匹配集, 因而此方法实用性不强, 无法有效地过滤短信.

图 4 中给出的基于发送速度的过滤方法实验结果表明, 此类方法采用短信发送速度过滤容易造成很强的误过滤, 在过滤速度设置过低时, 容易将正常用户误识别为垃圾短信发送者, 而速度设置过高则无法起到垃圾短信过滤的左右 (当前的垃圾短信用户已经采用限制垃圾短信发送速度的方式来避开发送速度过滤方法).



(a) 关键字匹配过滤算法

(a) Key words matching method



(b) 速度过滤方法

(b) Speed filtering method

图 4　传统方法实验结果

Fig. 4　The experimental results of traditional methods

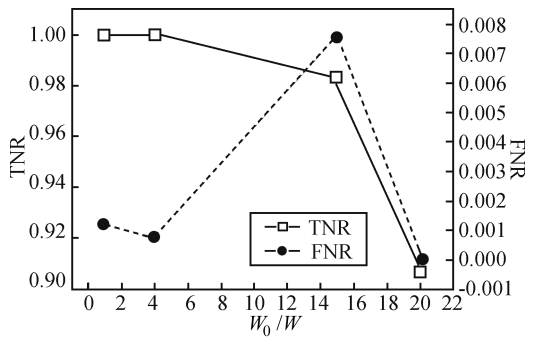


图 5　NASFA 算法实验结果

Fig. 5　The experimental results of NASFA

表 3 和图 5 所示的实验结果证明 NASFA 算法能够非常准确地识别垃圾发送用户同时控制正常用户的误识别率, 其误识别率 (FNR) 要显著低于采用关键字匹配和速度过滤方法.

4.2　讨论

通常垃圾短信发送者从社会网络的行为模式区别于普通发送用户的特性在于, 其希望尽可能发送给更多的用户, 同时出于费用的考虑又不会与发送的用户间发生语音通信关系, 而普通用户收到此类短信后一般也不会进行回复. NASFA 算法正是基于以上特征设计的.

由于 NASFA 算法是以短信发送者的社会网络最根本的行为模式作为过滤条件的, 因而该算法能够匹配垃圾短信发送的本质模式, 正常用户几乎不会出现类似垃圾短信的发送方式, 故可以高效准确地识别出垃圾短信发送者.

除了以上三种垃圾短信过滤方法外, 参考垃圾邮件过滤方法, 还存在其两种具有代表性的过滤垃圾短信的思路方法, 如文献 [1] 中采用的网络的聚类系数 C 作为判别垃圾邮件发送者, 文献 [17] 中采用的先训练分类器在通过分离器识别垃圾短信的方法.

对于聚类系数方法, 仔细分析发现 C 并不适合用于垃圾短信用户的过滤中, 主要原因有如下几点:

1) 聚类系数 C 的技术需要获得与当前被检验节点相连通的其他所有节点之间的连接关系, 因而给实时过滤算法的实现造成了技术困难;

2) 对于发送对象 (目的号码段) 为某一特定人群的垃圾短信, 目的号码段之间可能存在较多联系, 从而会获得比较高的聚类系数, 造成此类垃圾短信发送者无法被识别的问题.

对于训练分类器的方法, 虽然在邮件过滤中取得了比较满意的效果, 但是仍然无法较好地应用于垃圾短信过滤中, 主要有如下几点原因:

1) 训练过程需要有大量的训练样本, 造成识别有一定的滞后性. 而垃圾短信过滤中因为发送是需要计费的, 因而不能等待垃圾短信发送数量到达一定程度后才能识别垃圾短信发送者, 这样往往会给运营商带来大量的经济损失.

2) 训练分类器的方法需要事先人工标定大量的垃圾短信样本, 且分类的效果极其依赖于标定样本的数量, 因而在实际使用上不太可行.

3) 由于 SMS 发送文本存在字数限制, 因而分词之后可训练的特征非常少, 因此影响识别准确率, 不利于采用.

4) 垃圾短信的内容也会随时更新, 因而也要求重新训练分类器, 并且重新标定垃圾短信样本, 这就给后续过滤系统造成了额外的维护成本.

5 结束语

垃圾短信的过滤与防止, 不仅仅关系到运营商的经济利益问题, 同时也是一个重大社会问题. 本文提出了一个全新的垃圾短信过滤思路, 即从复杂网络的观点对短信发送网络和通话网络进行建模, 统计和分析短信网络的特性, 并通过统计实验分析和发现垃圾短信发送的网络模式, 据此提出相应的垃圾短信过滤算法. 文章中设计和提出综合 N-度相关和短信回复比率的过滤算法 (NASFA), 并通过实验分析表明该算法的不但能够有效地阻止垃圾短信的传播和发送, 最大限度地保障正常短信用户的使用, 同时具有良好的可扩展性, 能够快速可靠地适应新的垃圾短信发送手段和途径. 将来的工作将进一步细化研究不同种类垃圾短信的传播特征, 并着重研究其中的谣言类短信的传播范围和传播模式.

References

1 Boykin P O, Roychowdhury V P. Leveraging social networks to fight spam. *IEEE Computer*, 2005, **38**(4): 61−68

2 Kong J S, Rezaei B A, Sarshar N, Roychowdhury V P, Boykin P O. Collaborative spam filtering using e-mail networks. *IEEE Computer*, 2006, **39**(8): 67−73

3 Wang Xiao-Fan, Li Xiang, Chen Guan-Rong. *Complex Network Theory and Its Implmentation*. Beijing: Tsinghua University Press, 2006
(汪小帆, 李翔, 陈关荣. 复杂网络理论及其应用. 北京: 清华大学出版社, 2006)

4 Watts D J, Strogatz S H. Collective dynamics of small-world networks. *Nature*, 1998, **393**(6684): 440−442

5 Albert R, Barabasi A L. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 2002, **74**(1): 47−97

6 Kleinberg J M. Navigation in a small world. *Nature*, 2000, **406**(6798): 845

7 Barabasi A L, Albert R. Emergence of scaling in random networks. *Science*, 1999, **286**(5439): 509−512

8 Wu Ye, Xiao Jing-Hua, Wu Zhi-Yuan, Yang Jun-Zhong, Ma Bao-Jun. Research on the growing process of short message networks. *Acta Physica Sinica*, 2007, **56**(4): 2037−2041
(吴晔, 肖井华, 吴智远, 杨俊忠, 马宝军. 手机短信网络的生长过程研究. 物理学报, 2007, **56**(4): 2037−2041)

9 Laherrere1 J, Sornette D. Stretched exponential distributions in nature and economy: "fat tails" with characteristic scales. *The European Physical Journal B*, 1998, **2**(4): 525−539

10 Ebel H, Mielsch L I, Bornholdt S. Scale-free topology of e-mail networks. *Physical Review E*, 2002, **66**(3): 518−521

11 Newman M E, Forrest S, Balthrop J. Email networks and the spread of computer viruses. *Physical Review E*, 2002, **66**(3): 510−513

12 Adamic L A, Huberman B A, Barabasi A, Albert R, Jeong H, Bianconi G. Power-law distribution of the world wide web. *Science*, 2000, **287**(5461): 2113−2115

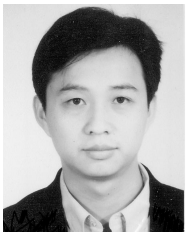
13 Gallos L K, Argyrakis P. Distribution of infected mass in disease spreading in scale-free networks. *Physica A*, 2003, **330**(1-2): 117−123

14 Dijkstra E W. A note on two problems in connection with graphs. *Numerische Mathematik*, 1959, **1**(1): 269−271

15 Floyd R W. Algorithm 97: short path. *Communications of the ACM*, 1962, **5**(6): 345

16 Kleinberg J. The small-world phenomenon: an algorithmic perspective. In: Proceedings of the 32nd ACM Symposium on Theory of Computing. Portland, Oregon, USA: ACM, 2000. 163−170

17 Sahami M, Dumais S, Heckerman D, Horvitz E. A Bayesian approach to filtering junk e-mail. In: Proceedings of AAAI Workshop on Learning for Text Categorization. Madison, Wisconsin: Springer, 1998. 55−62



黄文良 博士, 中国联合网络通信有限公司浙江分公司副总经理. 主要研究方向为信息处理和社会网络.
E-mail: huangwl0901@gmail.com
(**HUANG Wen-Liang** Ph. D., vice general manager of China United Network Communication Corporation, Zhejiang Branch. His research interest covers information processing and social network.)

刘 勇 博士, 浙江大学工业控制国家重点实验室讲师. 主要研究方向为智能信息处理, 机器学习和机器人视觉. 本文通信作者.
E-mail: yongliu@iipc.zju.edu.cn
(**LIU Yong** Ph. D., lecturer at the State Key Laboratory of Industrial Control Technology, Zhejiang University. His research interest covers intelligent information processing, machine learning, and robotics vision. Corresponding author of this paper.)

钟志强 硕士, 中国联合网络通信有限公司浙江分公司研发中心主任. 主要研究方向为软件工程和计算机网络技术.
E-mail: strongzhong@gmail.com
(**ZHONG Zhi-Qiang** Master, manager at the Research and Development Center, China United Network Communication Corporation, Zhejiang Branch. His research interest covers software engineering and computer network technology.)

沈仲明 博士, 中国联合网络通信有限公司浙江分公司总经理. 主要研究方向为数据挖掘和社会网络.
E-mail: shen9199@gmail.com
(**SHEN Zhong-Ming** Ph. D, general manager of China United Network Communication Corporation, Zhejiang Branch. His research interest covers data mining and social network.)