拓扑相似性垃圾邮件行为识别技术研究*

张新有，李新洁
（西南交通大学 信息科学与技术学院，成都 610031）

摘　要：对基于流的垃圾邮件行为识别技术进行了研究。根据垃圾邮件与正常邮件通信拓扑具有较大差异的特性，引入相似度的概念，提出了一种基于拓扑相似性的垃圾邮件行为识别方法。该方法以收发件人联系表来表征收发件人，计算用户相似度以此将邮件用户划分为多个邮件用户群，通过计算邮件收发件人归属判别邮件是否为垃圾邮件。采用一个辅助分类器方便对原始邮件用户进行判别和分组，最后用真实的邮件集进行实验，结果证明基于拓扑结构相似性分类方法有较好的分类能力。

关键词：垃圾邮件；行为识别；拓扑；相似度；邮件用户群
中图分类号：TP393　　**文献标志码**：A　　**文章编号**：1001-3695(2012)10-3805-04
doi:10.3969/j.issn.1001-3695.2012.10.052

Research on spam behavior recognition technology based on similarity of topology

ZHANG Xin-you, LI Xin-jie
(*School of Information Science & Technology, Southwest Jiaotong University, Chengdu* 610031, *China*)

Abstract: Spam behavior recognition technology, especially the behavior recognition technology based on the e-mail flow was studied. According to the communication topology of legitimate e-mail and spam with large differences, this paper introduced the concept of similarity, proposed a behavior recognition technology based on the similarity of topology. With this method, the e-mail receivers and senders were marked by the contact list. The e-mail users were divided into e-mail user clusters by calculating similarity. The senders and receivers of the coming e-mail were classified to clusters to judge whether the e-mail was spam. This paper used an auxiliary filter for classification and recognition of the original e-mail information. Simulation test with real e-mail set was conducted, which shows that the method based on the similarity of topology provides a better result for spam classification.

Key words: spam; behavior recognition; topology; similarity; e-mail user cluster

目前，反垃圾邮件技术研究主要集中在内容过滤与行为识别两个方面。内容过滤技术起步较早，其理论和应用的发展也较为成熟。行为识别技术是近几年发展起来的，并已成为当前反垃圾邮件技术的研究热点。

行为识别技术在 MTA 通信阶段就能判断出所接收邮件是否为垃圾邮件，无须检查邮件内容[1]。常见的技术有分布协作指纹分析、源头认证、邮件流量分析以及数据挖掘方式获取行为特征规则等。数据挖掘获取垃圾邮件行为规则是比较传统的行为识别技术，也是国内外垃圾邮件行为识别技术研究的主要方向。它是在分析与统计邮件头发送信息基础上，建立行为识别模型的方法来识别垃圾邮件。该方法有较好的查准率及处理效率，但召回率并不理想，不同邮件发送行为是变化的，且容易伪造，要随着垃圾邮件发送行为的改变更新分类器，造成时间和资源浪费。而邮件用户的通信关系相对其他行为特征是比较稳定的，而且不易造假，具有较高的分类价值，可以作为反垃圾邮件技术的研究方向之一。

基于拓扑结构相似性行为识别技术是以邮件用户通信关系为研究基础，分析垃圾邮件与正常邮件通信拓扑结构特征，通过邮件用户之间拓扑相似度计算、集合划分等来判别垃圾邮件。该技术不仅识别垃圾邮件准确率高，而且能够有效减少伪真问题，提高召回率。

1　邮件通信拓扑结构

从数据流特征入手研究邮件发送行为的邮件过滤技术还不是很成熟，大体分为基于统计的方法和基于拓扑的方法[2]。基于拓扑的方法是将邮件视为流数据，邮件的发件人、收件人为顶点，双方通信关系作为边构建邮件网络拓扑图来表示邮件流向，在此基础上进行统计与分析。合法邮件是以相互交换信息为目的双向通信，收件人和发件人存在某种社会关系[3]。垃圾邮件发送者是以暴利为目的的大范围扩散电子邮件的单向通信行为，邮件的发送者和接收者不存在社会关系。从这个层面讲，两者通信行为因其发送动机不同有明显差别。

社会关系网存在这样的自然属性，相互认识的两个人，他们有朋友或其他信任关系的人之间也很有可能相互认识，所以他们之间的通信概率很大。例如，A 认识 B、C、D、E、F，那么 B、C、D、E、F 相互认识的可能性很大。类似地，在邮件系统中，B、C、D、E、F 都是 A 的朋友，A 向 B、C、D、E、F 发过电子邮件，B、C、D、E、F 之间相互通信的可能性就很大，这样形成一个紧密联系的集合。通信关系拓扑结构如图 1 所示。

收稿日期：2012-03-13；**修回日期**：2012-04-26　　**基金项目**：国家自然科学基金—联合资助基金资助项目(U0970122)；中央高校基本科研业务费专项资金科技创新项目(SWJTU09CX040)

作者简介：张新有(1971-)，男，河南人，副教授，博士研究生，主要研究方向为分布式网络计算、网络安全(zhang.xyou@gmail.com)；李新洁(1984-)，河南人，工程师，硕士，主要研究方向为网络安全、网络测试.

垃圾邮件是 spammer 大范围广播发送的不请自来的邮件。发送者与接收者以及各个接收者之间一般相互不认识,不存在任何社会关系,只是单向的通信。如 A 是垃圾邮件发送者,将邮件发给 B、C、D、E、F,它们之间的通信关系形成一个单向的、稀疏的网络拓扑图,如图 2 所示。

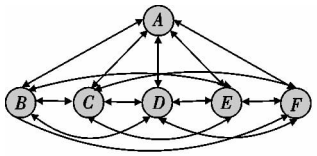


图1 正常邮件通信拓扑图

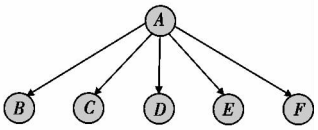


图2 垃圾邮件通信拓扑图

2 拓扑结构相似性

2.1 邮件用户向量表示

任何一个邮件用户都维护着一个收信人和发信人的联系列表。在邮件用户的多维向量空间中,采用发送者列表、接收者列表形成的多维向量表示邮件发送者和接收者。C_r 表示一个 e-mail 用户联系列表的收件人个数,邮件发送用户 s_x 的发送列表表示为向量 $\boldsymbol{s}_x$,为 C_r 维向量。$\boldsymbol{s}_x$ 的第 n 维向量值定义为

$$\boldsymbol{s}_x[n]=\begin{cases}1 & \text{if } s_x\rightarrow r_n\\ 0 & \text{otherwise}\end{cases} \tag{1}$$

其中:$s_x\rightarrow r_n$ 表示发件人 s_x 向收件人 r_n 至少发送过一次邮件。根据图 2,设在六维向量空间,A 向 B、C、D、E、F 发过邮件,那么 $\boldsymbol{s}_A=(0,1,1,1,1,1)$,0 表示没有向本地地址发送过邮件。

同理,C_s 表示 e-mail 用户的列表中发信人个数,邮件接收用户 r_x 的接收列表表示为向量 $\boldsymbol{r}_x$,为 C_s 维向量。$\boldsymbol{r}_x$ 的第 n 维向量定义为

$$\boldsymbol{r}_x[n]=\begin{cases}1 & \text{if } s_n\rightarrow r_x\\ 0 & \text{otherwise}\end{cases} \tag{2}$$

当 $\boldsymbol{r}_x[n]$ 为 1 时,表示至少有一封邮件来自 s_n。

通过上述方法将复杂的通信关系数据化、向量化,为拓扑相似性的进一步研究打下了基础。

2.2 相似度

相似度是两类事物或模式之间的相似程度,通常用距离和相关系数来衡量,其计算方法有多种[4]。通过夹角余弦求文本相似度已在很多领域中得到应用并取得良好效果,如论文检测系统[5]、入侵检测系统、购物网站商品推送系统等。本文将相似度的概念引入到垃圾邮件识别领域,通过比较邮件收件人或发件人通信关系相似度来识别垃圾邮件。采用夹角余弦计算相似度,$\boldsymbol{s}_x$、$\boldsymbol{s}_y$ 表示两个发件人的发送列表向量,如下所示

$$\mathrm{sim}(s_x,s_y)=\frac{\boldsymbol{s}_x\circ\boldsymbol{s}_y}{|\boldsymbol{s}_x\|\boldsymbol{s}_y|}=\cos(\boldsymbol{s}_x,\boldsymbol{s}_y) \tag{3}$$

其中:$\boldsymbol{s}_x\circ\boldsymbol{s}_y$ 表示向量的内积,$|\boldsymbol{s}_x|$ 与 $|\boldsymbol{s}_y|$ 为向量的模。若两个向量完全没有共同的收件人时,当 s_x 的某个值为 1 时,s_y 必为 0,反之亦然,由向量内积的计算公式可知 $\boldsymbol{s}_x\circ\boldsymbol{s}_y=0$,得到 $\mathrm{sim}(s_x,s_y)=0$。相反地,当收件人完全相同时,s_x 与 s_y 相等,$\boldsymbol{s}_x\circ\boldsymbol{s}_y=1$,$\mathrm{sim}(s_x,s_y)=1$。相似度取值范围为[0,1]。同理,计算收件人 r_x、r_y 相似度 $\mathrm{sim}(r_x,r_y)$。

合法邮件通信流的相似度与垃圾邮件通信流相似度代表的含义是不同的。合法邮件的相似度代表一组邮件用户之间的社会关系;而在垃圾邮件中,较大的相似度代表不同的 spammer 共享同一组的垃圾邮件攻击目标或同一个 spammer 使用不同的账号发送垃圾邮件。

2.3 邮件用户分类

相似度是收件人或发件人之间通信关系相关程度的表示,根据相似度的值将收件人、发件人划分成不同的集合。用向量模型方式表示发件人集合 senders cluster 与收件人集合 receivers cluster,即发件群与收件群。如存在一个 senders cluster 为 sc_x,sc_x 为其集合内所有元素的向量和,如下:

$$\boldsymbol{sc}_x=\sum_{s_y\in sc_x}\boldsymbol{s}_y \tag{4}$$

其中 $\boldsymbol{s}_y$ 为属于集合 sc_x 内的收件人的向量表示。

通过计算一个发件人与发件群 senders cluster 之间的相似度,当相似度达到阈值 ε 时,将收发件人划分到相应集合当中。发件人 s 与 sc_x 相似度由式(5)得到,该式是式(3)的扩展。

$$\mathrm{sim}(sc_x,s)=\begin{cases}\cos(\boldsymbol{sc}_x-\boldsymbol{s},\boldsymbol{s}) & \text{if } s\in sc_x\\ \cos(\boldsymbol{sc}_x,\boldsymbol{s}) & \text{otherwise}\end{cases} \tag{5}$$

当 $\boldsymbol{s}$ 已经包含在 $\boldsymbol{sc}_x$ 中,为了相似度计算的准确性,首先从 $\boldsymbol{sc}_x$ 中删除 $\boldsymbol{s}$,再计算 $\boldsymbol{sc}_x-\boldsymbol{s}$,$\boldsymbol{s}$ 的余弦值,这样新的分类的结果不受原有分类的影响,保证了分类的不断更新。可以看出,阈值 ε 的选择直接影响发件群分类的个数和分类精度,ε 越小,分类群数越少,分类的精度较低。

类似地,可将收件人划分成多个收件群。

3 拓扑结构相似性分类算法

3.1 分类原理

拓扑结构相似性分类算法需结合辅助分类器对邮件进行分类。辅助分类器可以使用现有的垃圾邮件分类技术实现。当新邮件到达时,首先经相似性分类算法,将邮件传送到辅助分类器作初始判断。基于初始判别的结果,再通过相似性判断邮件收件人和发件人归属群组以及之前获得的历史信息,对邮件是否为垃圾邮件作出新的分类判断,该判断有可能与辅助分类器的分类结果有所不同。过程简述如图 3 所示。

3.2 拓扑结构相似性分类算法设计

任何新的电子邮件 m 首先通过相似度计算,找到 m 的发件人和收件人所属的集合,分别表示为 sc_m、r_{cm}。当相似度达到阈值 ε 时,将其划分为同一群组,否则将该用户单独分类。

新的电子邮件是否为垃圾邮件由其发件人或收件人所属的群决定。邮件 m 的发件群为垃圾邮件发送者集合的概率用 PS_m 表示;PR_m 是邮件 m 的收件群为垃圾邮件接收者集合的概率。PS_m、PR_m 的值为集合内所有用户发送或接收垃圾邮件概率的平均值。为了计算方便,此处的概率可以用垃圾邮件的频数来代替。

此处用 Sprank 表示邮件 m 为垃圾邮件的概率。Sprank 是邮件分类最终判决结果,可由 PS_m、PR_m 计算得到,如图 4(a)所示。

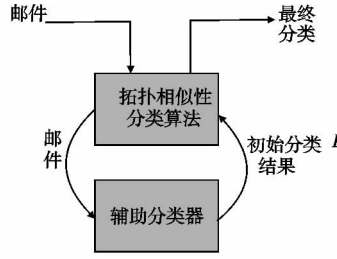
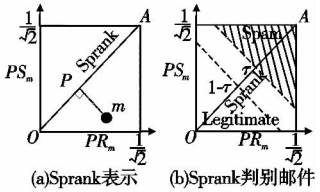


图3 拓扑相似性分类原理



(a)Sprank表示 (b)Sprank判别邮件

图4 Sprank与PS_m、PR_m

其中,邮件 m 为垃圾邮件的概率 Sprank 可用 PS_m、PR_m 坐标表示,m 为 PS_m、PR_m 组成的坐标系中的一个点。Sprank 为从原点 o 到 m 在对角线上的投影点 P 的距离。PS_m、PR_m 的取值范围变为$[0, 1/\sqrt{2}]$,这是因为当采用原始范围$[0,1]$时 Sprank 最大值即对角线上从 o 到 A 的长度为$\sqrt{2}$。为了清晰合理地表示垃圾邮件概率,用$\sqrt{2}$对 PS_m、PR_m 标准化,即 $PS_m = PS_m/\sqrt{2}$,$PR_m = PR_m/\sqrt{2}$。PS_m、PR_m 的取值范围变为$[0, 1/\sqrt{2}]$,则 Sprank 取值范围为$[0,1]$。Sprank 为

$$\text{Sprank} = \begin{cases} \sqrt{PS_m^2 + PR_m^2} \times \sin\left(\frac{\pi}{4} - \arctan\frac{PS_m}{PR_m}\right) & \text{if } PR_m > PS_m \\ \sqrt{PS_m^2 + PR_m^2} \times \sin\left(\frac{\pi}{4} - \arctan\frac{PR_m}{PS_m}\right) & \text{if } PR_m < PS_m \\ \sqrt{PS_m^2 + PR_m^2} & \text{if } PR_m = PS_m \end{cases} \quad (6)$$

当 PS_m、PR_m 的值较大时,Sprank 的值也随之增大,说明该邮件为垃圾邮件的概率越高。当两者的值较小时,Sprank 的值也较小,表明邮件为合法邮件的可能性较大。如图 4(b)所示,当 Sprank 大于给定的阈值 τ 时,m 所在位置为右上方阴影部分 Spam 区,邮件判定为垃圾邮件;当 Sprank $< 1 - \tau$,邮件 m 在左下方 Legitimate 区,判定为合法邮件。否则对邮件不作判定,采用原始分类器判定结果。由此可见,参数 τ 的值是否合适决定了分类的准确率的高低。

3.3　算法描述

基于拓扑结构相似性的分类算法描述如下:

```
for 到达邮件 m do
    辅助分类器对邮件作初始判别;
    计算相似度确定 m. sender 所在发件群 sc_m;
    计算 PS_m;
    PR_m = 0;
    for r ∈ m. receivers do
      为 m. receivers 确定所在的发件人集合 rc_m;
      计算 PR_m;
      PR_m + = PR_m;
    end for
    PR_m = PR_m/ size (m. receivers);
    Sprank = f(PR_m, PS_m),函数 f 求 Sprank;
    if Sprank > τ then
      邮件判定为 spam;
    else if Sprank < 1 - τ
      邮件判定为 Legitimate;
    else
      保持辅助分类器初始判定结果;
    end if
end for
```

4　实验

1)实验内容　为了验证行为识别技术中拓扑相似性算法的分类效果,采用传统行为识别技术构造辅助分类器。在真实的邮件数据上对拓扑相似性分类算法进行实验,对比辅助分类器分类结果,评价指标反映分类效果。

2)数据来源　用某大学校园邮件服务器一周的日志信息共 15 388 条,5 388 条用于收发信人群组划分和阈值计算,10 000条用来测试。

3)方法及工具　辅助分类器的构造方法是提取邮件头信息中的行为特征,采用新西兰怀卡托大学开发的数据挖掘软件 WEKA[6] 中的 LibSVM 支持向量机分类算法建立行为识别模型。

采用拓扑相似性分类算法判别邮件之前,首先要确定最佳参数 ε 和 τ,由训练集合实验得到。选定阈值后,对待测样本进行实验,统计分类结果。

4)评价指标　本文采用的评价指标有查准率(precision)、召回率(recall)、$F1$ 值[7]。查准率即垃圾邮件检对率 $P = TS/(TS + FL)$;召回率即垃圾邮件检出率 $R = TS/(TS + FS) = TS/N$;因查准率与召回率相互制约,故常采用二者的调和平均 $F1$ 值作为评价标准 $F1 = \frac{2PR}{P+R}$,$F1$ 值越大,对垃圾邮件的综合过滤能力越强。其中,N 为垃圾邮件总数,TS、FS、FL 及 TL 代表的含义如表 1 所示。

表 1　各参数含义

	实为垃圾邮件/封	实为合法邮件/封
判定为垃圾邮件	TS	FS
判定为合法邮件	FL	TL

5)实验过程与结果分析

a)确定邮件集分类阈值 ε。收发件群的划分对分类结果有较大的影响。为了选定最佳阈值 ε,首先计算邮件服务器中 5 388 封电子邮件涉及的收发件人相似度。改变 ε 值可以使得增加或减小的分类的紧密度。ε 越小,群组个数越少,分类的紧密度减小,垃圾邮件识别效果越差;ε 越大,分类的紧密度越大,处理效率就会降低。如图 5 所示,当 $0 \leqslant \varepsilon < 0.5$ 时,群组个数呈稳步上升趋势;当 $\varepsilon = 0.5$,群组数为 20;当 $\varepsilon \geqslant 0.5$ 群组个数呈平稳状态,说明 $\varepsilon = 0.5$ 相似度对集合的划分达到了最优,保证群组划分质量。

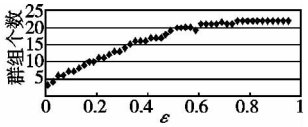


图5　阈值ε与群组个数的关系

b)确定最佳 Sprank 分类阈值 τ。基于辅助分类器的初始分类结果,改变 τ 的值,通过拓扑结构相似性分类的方法分别判别垃圾邮件、正常邮件,得出与初始结果一致的比率。如图 6 所示。

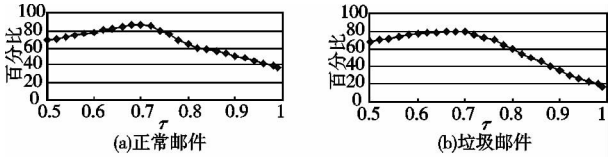


图6　改变τ，拓扑相似性与辅助分类器判定结果相同比率

从图 6 可以看出,变动 τ 值,拓扑相似性算法判定结果与辅助分类器的结果差异也随之改变。比较图(a)与(b),(a)中正常邮件的识别率均高于(b)中垃圾邮件的识别率。由此可见,拓扑相似度的方法对正常邮件的判定效果可达到 90%,可以有效地减少误报问题。当 $\tau = 0.68$ 时,无论是正常邮件还是垃圾邮件都达到了最高的识别率,分别为 86%、80%,选定 $\tau = 0.68$ 为最佳参数值。

c)采用最佳参数 $\varepsilon=0.5,\tau=0.68$ 对邮件数据集合按照拓扑相似性分类算法实验。结果如表 2 所示,表中分别为拓扑相似性分类算法与辅助分类器分类结果的查全率、查准率和 $F1$ 值。

表 2 分类实验查准率 P、查全率 R、$F1$ 值

邮件	辅助分类器			拓扑相似性分类算法		
个数	P	R	$F1$	P	R	$F1$
2 000	0.78	0.74	0.577	0.89	0.85	0.756
4 000	0.77	0.75	0.615	0.91	0.88	0.806
6 000	0.75	0.69	0.517	0.93	0.90	0.837
8 000	0.79	0.75	0.592	0.91	0.89	0.802
10 000	0.77	0.76	0.585	0.94	0.92	0.864
平均	0.772	0.738	0.577	0.916	0.888	0.811

由表 2 中的 $F1$ 值可得到图 7 所示的直观表示。

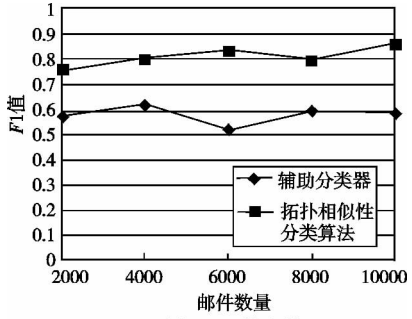


图7 F1值比较

从表 2 可以看出,邮件数量从 2 000 增加到 10 000,辅助分类器与拓扑结构相似性分类算法垃圾邮件查全率、查准率以及 $F1$ 值的变化较小,说明分类效果稳定。通过比较两种分类方法评价指标的平均值,可以看出本文提出的分类算法的召回率、查准率、$F1$ 值都有较大的提高。图 7 为两种分类方法 $F1$ 值比较,可见基于拓扑相似性分类算法的综合分类能力远远高于传统行为识别方法。由于本文是选用某学校邮件服务器日志数据进行实验,鉴于时间段选择、通信范围等因素,实验绝对结果可能受到一定的影响,但并不影响验证相对效果。

5 结束语

本文对邮件行为识别技术中的数据流行为进行了研究,具体分析了用户通信的拓扑结构,得出了垃圾邮件与正常邮件的通信流拓扑特征有较大差别。进一步通过用户联系列表表示邮件用户行为,提出计算相似度的方法将邮件进行分类,给出分类算法并在真实的邮件服务器日志数据上进行了实验验证。基于拓扑相似性的分类算法在一定程度上取得了较为理想的过滤效果,但是还有待改进的地方,比如对于新的“信任”关系,收发件人第一次单向通信、没来得及回信的邮件以及群发邮件等被误判为垃圾邮件。目前的反垃圾邮件技术各有优缺点,单一依赖某一种技术不能完美地解决所有垃圾邮件问题,多种技术融合的过滤技术将会是反垃圾邮件以后的研究方向。

参考文献:

[1] 赵治国,谭敏生,丁琳. 垃圾邮件行为识别技术的研究与实现[J]. 计算机应用研究,2007,24(11):229-231.

[2] 白秋颖,章睿,张耀龙. 基于网络会话层的垃圾邮件行为识别[J]. 计算机工程与应用,2007,43(1):167-169.

[3] OSCAR P, VWANI R. Personal e-mail networks: an effective anti-spam tool [J]. IEEE Computer, 2005, 38(4): 61-68.

[4] WANG Hui, LIN Zhi-wei, McCLEAN, S, *el al*. Measuring similarity for multidimensional sequences[C]//Proc of IEEE International Conference on Data Mining. 2010: 281-287.

[5] SONG Wen-he, MA Chun-xia. The study of thesis replica detecte method based on similarity of text [C]//Proc of the 3rd IEEE International Conference on Computer Science and Information Technology. 2010: 596-600.

[6] 陈慧萍,林莉莉,王建东,等. WEKA 数据挖掘平台及其二次开发[J]. 计算机工程与应用,2008,44(19):76-81.

[7] 潘洁珠,周晓,吴共庆,等. 基于小样本学习的垃圾邮件过滤方法[J]. 计算机工程,2010,36(21):245-247.