



(12) 发明专利

(10) 授权公告号 CN 101299729 B

(45) 授权公告日 2011.05.11

(21) 申请号 200810064806.1

(22) 申请日 2008.06.25

(73) 专利权人 哈尔滨工程大学

地址 150001 黑龙江省哈尔滨市南岗区南通  
大街 145 号 1 号楼哈尔滨工程大学科技  
处知识产权办公室

(72) 发明人 杨武 张乐君 王巍

(51) Int. Cl.

HO4L 12/58 (2006.01)

HO4L 9/36 (2006.01)

(56) 对比文件

CN 1564167 A, 2005.01.12, 全文 .

US 6560600 B1, 2003.05.06, 全文 .

EP 1139236 A1, 2001.10.04, 全文 .

审查员 王伦杰

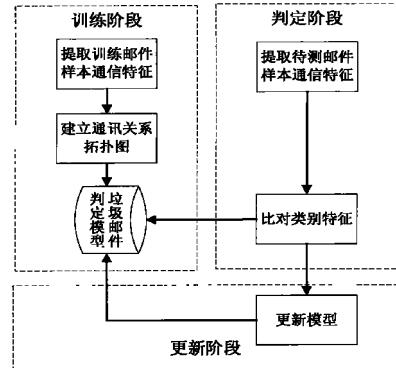
权利要求书 1 页 说明书 6 页 附图 5 页

(54) 发明名称

一种基于拓扑行为的垃圾邮件判定方法

(57) 摘要

本发明提供的是一个垃圾邮件的判定方法。根据电子邮件的通讯关系，建立一个通讯关系拓扑图；对通讯关系拓扑图中，具有双向通讯关系的用户归并为一个类，建立垃圾邮件的判定模型；通过待检测的邮件中提取 from 邮箱地址和 to 邮箱地址，并判断其是否为垃圾邮件；对垃圾邮件判定模型进行更新。本发明的优点在于：只需要获取邮件的少量信息就可以快速、准确的对垃圾邮件进行判定，并且可以根据不同的情况部署到不同的位置：如：邮件服务器、网关、骨干网出入口等等。由于其处理速度快，因此可以在源头上遏制垃圾邮件的传播。



1. 一种垃圾邮件的判定方法,其特征是:

1) 根据电子邮件的通讯关系,建立一个通讯关系拓扑图;

2) 对通讯关系拓扑图中,具有双向通讯关系的用户归并为一个类,建立垃圾邮件的判定模型;

3) 通过待检测的邮件中提取 from 邮箱地址和 to 邮箱地址,并判断其是否为垃圾邮件;所述的判断其是否为垃圾邮件的方法是:(1)首先提取出 from 邮箱地址和 to 邮箱地址,检查它们的类号,如果两个邮箱中至少有一个没有类号,暂判为正常邮件,把没有类号的邮箱的类号记为 1 并记录下通信关系;然后根据其以后的通信情况,再做相应判断和处理;否则,向下继续进行;(2)检查两个邮箱所属类号的最大公约数,如果最大公约数大于 1,则这封邮件被判为正常邮件;如果最大公约数为 1,向下继续进行;(3)根据发送者是否在接收者已发送但并未回复的地址中,来判断这封邮件是否是一封回复邮件,如果是,则这封邮件被判为正常邮件,同时,还要更新类的信息,否则向下继续进行;(4)统计这个发送者向这个接收者已发送但并没有得到回复的邮件数目,并将其与设定的阈值相比较,如果小于阈值,就判为正常邮件,如果大于等于阈值,则判为垃圾邮件;

4) 对垃圾邮件判定模型进行更新。

2. 根据权利要求 1 所述的一种垃圾邮件的判定方法,其特征是:所述的建立通讯关系拓扑图的方法为:1) 从每一封电子邮件中抽取出 from 邮箱地址和 to 邮箱地址;2) 建立一条从 from 邮箱地址到 to 邮箱地址有向图。

3. 根据权利要求 2 所述的一种垃圾邮件的判定方法,其特征是:所述的建立垃圾邮件判定模型的方法是:1) 设网络中所有节点的集合为 U, from 表里存放可直达 U 中每一个节点的节点,to 表里存放 U 中每一个节点可直达的节点;2) 在 U 中任取一个节点 a, 把 a 放进集合 T1 中;3) 在 from 表中查找出 a 可到达的所有节点 ak, 并加入到 T1 中;4) 在 from 表中查找 ak 可达的所有节点,并加入到 T1 中且已有的不再加入,重复这种查找直到 T1 不再发生变化; 5) 采用 3)、4) 中 from 表查找节点的方法,在 to 表中进行查找,得到另一集合 T2, 取 T1 和 T2 的交集 T 为节点 a 的类;6) 在 U 中去掉 T 中元素,再选择一个节点,重复 3)、4)、5) 过程得出新的类,如此下去直至 U 为空;7) 对每一个分出来的类,若其内部的元素个数大于等于 2 则为正常类,给其分配一个奇素数类号,其它所有节点都归为一个奇异类,为其分配类号为 1。

4. 根据权利要求 3 所述的一种垃圾邮件的判定方法,其特征是:所述的对垃圾邮件判定模型进行更新是选择如下方法之一:1) 新节点和所有类节点进行单向通信,把新节点加入到奇异类中;2) 新节点与奇异类中节点进行双向通信,则它们生成新的正常类;3) 新节点与正常类中节点进行双向通信,则把新节点加入该正常类;4) 奇异类中的节点之间进行双向通信,则它们生成新的正常类;5) 奇异类中的节点 a 和某正常类中节点进行双向通信后,把节点 a 也归为与 a 进行通信的节点的类中;6) 若两个不同的正常类节点进行双向通信,则这两个节点生成一个新的包含这两个类的聚类,它们的类号均新设为原先两个类号的积,但这两个节点都可正常与以前所属类别中的节点通信;7) 如果某个正常类里面已没有节点,则撤销该类,并把该类的聚类 B 并入 B 的另一个子类中,撤销 B。

## 一种基于拓扑行为的垃圾邮件判定方法

### (一) 技术领域

[0001] 本发明涉及的是一种垃圾邮件的判定方法。

### (二) 背景技术

[0002] 电子邮件凭借低廉、简单、快捷的优势已经成为人们工作和生活中的重要通信方式,但人们在享受电子邮件提供诸多便捷之时,也在忍受着它的副产品所带来了的痛苦,即互联网上垃圾邮件泛滥成灾,并且近几年有愈演愈烈的趋势。

[0003] 目前对垃圾邮件的治理还是集中在依靠垃圾邮件过滤技术。而以内容识别为主的邮件过滤系统,在使用过程中渐渐发现它们也存在着一些缺陷。内容过滤需要训练、分类、计算,过滤过程需要耗费大量系统资源,所以处理速度比较慢, CPU 和内存占用较高,效率低。准确性依赖大量的历史数据,故生命周期短。对于经常变换内容的垃圾邮件,效果也不是很好。因为它始终没有逃离关键词匹配的思想,所以关键词库需要不断更新维护,是一种被动的处理过程。另外,该技术需要将邮件全部接收下来再进一步处理,虽然判断出垃圾邮件,但并没有节省网络流量开销。

[0004] 对于拓扑行为的垃圾邮件判定还处于起步阶段,如 Scale-free topology of e-mail network[J], 2002, 偏重于建立邮件网络模型,以用户为节点,以通信关系为边,从邮件服务器日志中截取一定信息来建立网络模型,并通过试验证明邮件世界同样有 scale free 和 small world 属性; Comparative graphTheoretical Characterization of Networks of Spam and RegularEmail [EB/OL]. <http://arxiv.org/abs/cond-mat/0503725>, 通过邮件发送者和接收者产生的边界流图。作者通过用户图表和域图表在各个指标如网络聚合度、出入度差异等方面差异来分析垃圾邮件和正常邮件的特征,使用 HIS 算法来分析流量图的演化结构,并提出如何动态地调整图的关系结构的方法。上面典型的垃圾邮件判定方法还属于概念性的表述,如果没有大量后续工作的展开,是难以在垃圾邮件判定中得到应用。

### (三) 发明内容

[0005] 本发明的目的在于提供一种通过分析邮件之间的通讯拓扑关系来对垃圾邮件进行快速判定的方法。

[0006] 本发明的目的是这样实现的:

[0007] 1) 根据电子邮件的通讯关系,建立一个通讯关系拓扑图;

[0008] 2) 对通讯关系拓扑图中,具有双向通讯关系的用户归并为一个类,建立垃圾邮件的判定模型;

[0009] 3) 通过待检测的邮件中提取 from 邮箱地址和 to 邮箱地址,并判断其是否为垃圾邮件;

[0010] 4) 对垃圾邮件判定模型进行更新。

[0011] 所述的建立邮件通讯关系拓扑图:1) 从每一封电子邮件中抽取出 from 邮箱地址

和 to 邮箱地址 ;2) 建立一条从 from 邮箱地址到 to 邮箱地址有向图。

[0012] 所述的建立垃圾邮件判定模型是 :1) 将网络中所有节点集合为 U, from 表里存放可直达该节点的节点, to 表里存放该节点可直达的节点。2) 在 U 中任取一个节点 a, 把 a 放进集合 T1 中 ;3) 在 from 表中查找出 a 可到达的所有节点 ak, 并加入到 T1 中 ;4) 在 from 表中查找 ak 可达的所有节点, 并加入到 T1 中 ( 已有的不再加入 ), 重复这种查找直到 T1 不再发生变化 ;5) 同样的方法在 to 表中进行查找, 得到另一个集合 T2, 取 T1 和 T2 的交集 T 为节点 a 的类 ( 当然也是 T 中任意一个元素的类 ) ;6) 在 U 中去掉 T 中元素, 再选择一个节点, 重复 3.4.5 过程得出新的类, 如此下去直至 U 为空 ;7) 对每一个分出来的类, 若其内元素个数大于等于 2 则为正常类, 给其分配一个奇素数类号, 其它所有节点都归为一个奇异类, 为其分配类号为 1 。

[0013] 所述的垃圾邮件判定方法是 :1) 首先要提取出 from 邮箱地址和 to 邮箱地址, 检查它们的类号, 如果两个邮箱中至少有一个没有类号, 说明是新邮箱之间通信或已有的类与新邮箱通信, 这时暂判为正常邮件, 把没有类号的邮箱的类号记为 1 ( 奇异类 ) 并记录下通信关系。然后根据其以后的通信情况, 再做相应判断和处理。否则, 向下继续进行 ;2) 检查两个邮箱所属类号的最大公约数, 如果最大公约数大于 1, 则这封邮件被判为正常邮件。如果最大公约数为 1, 向下继续进行 ;3) 看发送者是否在接受者已发送但并未回复的地址中, 既判断这封邮件是否是一封回复邮件。如果是, 则说明发送者和接收者在互相通信, 则这封邮件被判为正常邮件。同时, 还要更新类的信息。否则向下继续进行 ;4) 统计这个发送者向这个接收者已发送但并没有得到回复的邮件数目, 并将其与我们设定的阈值相比较。如果小于阈值, 就判为正常邮件。如果大于等于阈值, 则判为垃圾邮件。

[0014] 所述的模型更新具体包括以下几种情形 :1) 新节点和所有类节点进行单向通信, 把新节点加入到奇异类中 ;2) 新节点与奇异类中节点进行双向通信, 则它们生成新的正常类 ;3) 新节点与正常类中节点进行双向通信, 则把新节点加入该正常类。4) 奇异类中的节点之间进行双向通信, 则它们生成新的正常类 ;5) 奇异类中的节点和某正常类中节点进行双向通信后, 把奇异类的节点也归为与其进行通信的节点的类中 ;6) 若两个不同的正常类节点进行双向通信, 则这两个节点生成一个新的包含这两个类的聚类 ( 原先的两个类称为该类的子类 ), 它们的类号均新设为原先两个类号的积, 但这两个节点都可正常与以前所属类别中的节点通信 ;7) 如果某个正常类里面已没有节点, 则撤销该类, 并把该类的聚类并入聚类的另一个子类中, 撤销聚类。

[0015] 针对以上情况, 本发明从邮件的拓扑行为出发, 提出了基于拓扑行为的垃圾邮件判定方法。经实验验证, 这个技术能够很好地解决已有的垃圾邮件判定技术的不足。

[0016] 本发明是基于如下问题而设计的 :

[0017] 由于基于邮件内容的垃圾邮件识别技术, 具有识别速度慢的特点, 并且无法从源头上有效遏制垃圾邮件的转播和蔓延, 因此需要一种可以快速、有效的垃圾邮件判定方法。

[0018] 为了能够快速判定垃圾邮件就必须采用一种需要信息量小且有效的方法, 首先获取邮件的消息头信息, 将 from 邮箱地址和 to 邮箱地址提取出来 ;其次建立邮箱地址之间的通讯关系图, 并建立识别模型 ;将待检测邮件的消息头部信息提取出来, 并放入已有的判定模型中进行判断 ;最后对判定模型进行更新。

[0019] 本发明的主要技术特征体现在 :

[0020] 1) 需要少量邮件信息,处理速度快

[0021] 判定邮件的属性往往需要获取邮件的内容,这样处理速度就比较慢,而且必须将邮件全部收下来以后才能进行,因此不能从源头上遏制邮件的传播。本发明之需要获取邮件的头部信息中的一部分,因此可以仅获取部分信息,就可以对邮件属性进行判断,可以从源头上对邮件进行判断。

[0022] 具体技术路线是 :1. 根据不同的部署情况,从 SMTP 协议中获取邮件信头部信息 : 从 Received 字段开始,到连续两个回车换行结束 ;2. 从获取数据中提取邮件地址信息,包括 from 字段和 to 字段 ;3. 将邮件地址字段信息对输入到判定模型中进行邮件属性判定。

[0023] 2) 垃圾邮件判定模型可以自动实时更新

[0024] 具体技术路线 :1. 根据待判定邮件的地址信息与已存在类节点之间通讯关系,更新识别模型 ;2. 根据奇异类之间的通讯关系和奇异类与正常类之间的通讯关系更新识别模型。3. 根据正常类之间的通讯关系更新识别模型。

[0025] 本发明的优点在于 :只需要获取邮件的少量信息就可以快速、准确的对垃圾邮件进行判定,并且可以根据不同的情况部署到不同的位置 :如 :邮件服务器、网关、骨干网出入口等等。由于其处理速度快,因此可以在源头上遏制垃圾邮件的传播。

#### (四) 附图说明

[0026] 图 1 基于拓扑行为的垃圾邮件判定步骤 ;

[0027] 图 2 A、B、C、D、E 形成的拓扑网络集合 1 ;

[0028] 图 3 A、B、C、D、E 形成的拓扑网络集合 2 ;

[0029] 图 4 邮件网络拓扑图 ;

[0030] 图 5 垃圾邮件子图 ;

[0031] 图 6 合法邮件子图 ;

[0032] 图 7 邮件拓扑示意图 ;

[0033] 图 8 基于拓扑行为的垃圾邮件判定技术测试结果 ;

[0034] 图 9 处理时间对比表 1。

#### (五) 具体实施方式

[0035] 下面结合附图举例对本发明做更详细地描述 :

[0036] 1) 拓扑行为识别模型原理

[0037] 合法邮件是在发信人和收信人存在社会关系前提下,以相互交换信息为目的,双向通信的结果 ;而垃圾邮件是在发送者利益驱动下,以大范围扩散为目的,单向通信的产物。两者本质上的不同必然导致其行为的显著差异,因此垃圾邮件和合法邮件在单 / 双向行为特征上是可以区分的。

[0038] 通过对大量合法邮件和垃圾邮件样本分析总结,发现合法邮件体现了通信双方之间一种亲戚、朋友、同事、上下级等社会关系,而且与合法的通信双方有“朋友”关系或“信任”关系的人之间也有很大概率通信的可能性。例如 A 认识 B、C、D、E,那么 B、C、D、E 相互认识的可能性就很大,这是社会关系网络的一个自然属性。同样对于邮件系统, A 给 B、C、D、E 都发过邮件,因为都是 A 的朋友,所以 B、C、D、E 之间通过 A 相互认识,他们之间也会相

互通信,就会形成一个小的紧密联系的集合,如图 2 所示。

[0039] 而垃圾邮件却是那些不请自来,希望能有更多的人获得发送者传递信息的邮件。接收与发送者之间并不认识,也没有任何社会关系。垃圾邮件属于滥发行为,发送者与接收者、多数接收者之间并不存在社会关系。比如 A 是垃圾邮件发送者,将邮件发给 B、C、D、E,它们形成的疏松网络如图 3 所示。

[0040] 用某大学校园邮件服务器上一周的日志信息,使用 Graphviz 绘图工具生成网络拓扑图验证上面分析的正确性。由日志信息建立的邮件网络拓扑图如图 4 所示。

[0041] 图中每个节点是邮箱地址的散列值,从图中可以拆分出一个垃圾邮件子图和一个合法邮件子图,分别见图 5 和图 6。

[0042] 从垃圾邮件图 5 中,可以明显地看到散列值为 33690 和 39900 的两个垃圾邮件发送者在发送邮件。发送者只发不收,接收者之间没有任何通信关系,节点之间联结疏松。而在合法邮件图 6 中,节点之间存在着双向互通关系,并且节点间联结很紧密,形成联结紧密的关系网。这正是由于合法邮件之间存在合法的社会关系,使得他们之间有对应的通信关系的结果。

[0043] 邮件发送者和接收者之间是否拥有社会关系,是区分垃圾邮件和合法邮件最本质的特征之一,这种特征在垃圾邮件和合法邮件拓扑图中表现出极大的不同。所以可以根据图形理论知识,选取合适的度量,构建邮件拓扑图,然后用图形理论分析的方法来区分垃圾邮件与合法邮件。

[0044] 2) 垃圾邮件判定模型

[0045] 以用户邮箱地址为结点,用户之间通信关系作为边建立邮件拓扑图。分析图的拓扑特性,得到多个反映用户社会关系的用户类。同一个类内的用户相互通信的邮件为正常邮件,不同类之间用户单向通信的邮件为垃圾邮件。

[0046] 用户类别划分是以用户间是否相互通信为依据,所有的用户划分为若干个正常类和一个奇异类。当且仅当两个用户相互发送过邮件,两者归为同一个正常类,们所有不能和其他用户归为同一类的用户形成奇异类。随着用户之间的相互通信,类可以自动进化以反映当前的用户关系。

[0047] 图 7 是邮件拓扑关系的示意图,其中灰色圆圈围着的部分有相互通信关系,表示正常类。两个黑点没有相互通信关系,说明这两个黑点是垃圾邮件发送者,它们代表垃圾类。

[0048] 在识别模型的建立过程中,本文采用 MNTA(mail net topology arithmetic) 算法,结合图 1 该方法具体描述如下:

[0049] 设网络中所有节点集合为 U, from 表里存放可直达该节点的节点, to 表里存放该节点可直达的节点。

[0050] (1) 在 U 中任取一个节点 a, 把 a 放进集合 T1 中;

[0051] (2) 在 from 表中查找出 a 可到达的所有节点 ak, 并加入到 T1 中;

[0052] (3) 在 from 表中查找 ak 可达的所有节点,并加入到 T1 中(已有的不再加入),重复这种查找直到 T1 不再发生变化;

[0053] (4) 同样的方法在 to 表中进行查找,得到另一个集合 T2, 取 T1 和 T2 的交集 T 为节点 a 的类(当然也是 T 中任意一个元素的类);

[0054] (5) 在 U 中去掉 T 中元素,再选择一个节点,重复 2、3、4 过程得出新的类,如此下去直至 U 为空;

[0055] (6) 对每一个分出来的类,若其内元素个数大于等于 2 则为正常类,给其分配一个奇素数类号,其它所有节点都归为一个奇异类,为其分配类号为 1;

[0056] 按照这个算法,我们就把一个大的网络拓扑图分成若干个类。正常类里面包含的都是互相连通的、可以互达的、具有一定社会关系的节点。这些节点之间相互通信,认为是合法的,它们之间通信的邮件即为合法邮件。而奇异类中只有单向通信关系,节点之间不可以互达。由于节点之间不存在着确定的社会关系,因此,这里面的节点是可疑的。

[0057] 最后,我们把网络中的每个节点,即邮箱地址,分配一个由上面算法获得的类号。这样,识别模型就建立了起来。

[0058] 3) 邮件属性判定

[0059] 对于每新来的一封邮件,

[0060] (1) 首先要提取出 from 邮箱地址和 to 邮箱地址,检查它们的类号。

[0061] 如果两个邮箱中至少有一个没有类号,说明是新邮箱之间通信或已有的类与新邮箱通信,这时暂判为正常邮件,把没有类号的邮箱的类号记为 1(奇异类)并记录下通信关系。然后根据其以后的通信情况,再做相应判断和处理。否则,向下继续进行。

[0062] (2) 检查两个邮箱所属类号的最大公约数。

[0063] 如果最大公约数大于 1,则这封邮件被判为正常邮件。如果最大公约数为 1,向下继续进行。

[0064] (3) 看发送者是否在接受者已发送但并未回复的地址中,既判断这封邮件是否是一封回复邮件。如果是,则说明发送者和接收者在互相通信,则这封邮件被判为正常邮件。同时,还要更新类的信息。否则向下继续进行。

[0065] (4) 统计这个发送者向这个接收者已发送但并没有得到回复的邮件数目,并将其与我们设定的阈值相比较。如果小于阈值,就判为正常邮件。如果大于等于阈值,则判为垃圾邮件。

[0066] 4) 判定模型更新

[0067] 随着时间的推移,用户节点之间的关系会发生变化,而此时类也应能进化以表示新的用户关系,具体包括以下几种情形:

[0068] (1) 新节点和所有类节点进行单向通信,把新节点加入到奇异类中。

[0069] (2) 新节点与奇异类中节点进行双向通信,则它们生成新的正常类。

[0070] (3) 新节点与正常类中节点进行双向通信,则把新节点加入该正常类。

[0071] (4) 奇异类中的节点之间进行双向通信,则它们生成新的正常类。

[0072] (5) 奇异类中的节点和某正常类中节点进行双向通信后,把奇异类的节点也归为与其进行通信的节点的类中。

[0073] (6) 若两个不同的正常类节点进行双向通信,则这两个节点生成一个新的包含这两个类的聚类(原先的两个类称为该类的子类),它们的类号均新设为原先两个类号的积,但这两个节点都可正常与以前所属类别中的节点通信。

[0074] (7) 如果某个正常类里面已没有节点,则撤销该类,并把该类的聚类并入聚类的另一个子类中,撤销聚类。

[0075] 4) 实验及分析

[0076] 收集某大学校园邮件服务器日志信息,共 10586 条。其中 2000 条用来建立邮件拓扑图,形成垃圾邮件识别模型,剩余 8586 条用来测试。

[0077] 硬件环境 :曙光服务器一台

[0078] 软件环境 :Red hat 9.0 以上的 linux 操作系统

[0079] 在不同阈值的情况下,本文提出的基于拓扑的行为识别技术的召回率、准确率、精确率如图 8 所示。从图中可以看出,召回率和精确率会随着阈值的变大而降低;准确率随着阈值的增大而升高。准确率达到 100% 时,以后一直保持这个水平,准确率很高,召回率稍低。

[0080] 当阈值为 1 时,表示收发件人只有一次单向通信关系时,即判为垃圾邮件。这样,就会将没有来得及回信的邮件误判为垃圾邮件。所以图中当阈值为 1 时,准确率只有 70%,召回率 90%,有 30% 正常邮件被误判为垃圾邮件。阈值 2 时的情况,虽然准确率有所提高,但仍有 20% 的误判。然后准确率逐渐上升,当阈值为 6 时,准确率达到 100%,召回率 70%,精确率 72%,这时总的性能达到最好。

[0081] 另外,该技术平均处理每封邮件的时间仅为微秒级,而内容识别技术为毫秒级的,故其速度快、执行效率比较高。与内容识别技术在处理时间上的对比,如表 1 所示。

[0082] 综上,实验数据进一步证明了基于拓扑行为的垃圾邮件识别技术不仅快,而且准的特点。

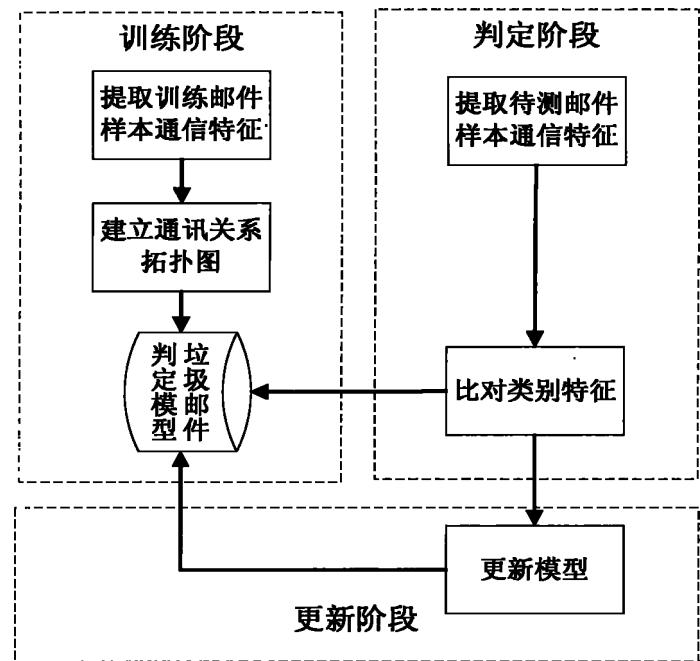


图 1

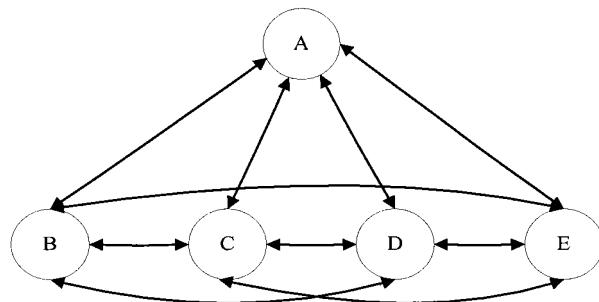


图 2

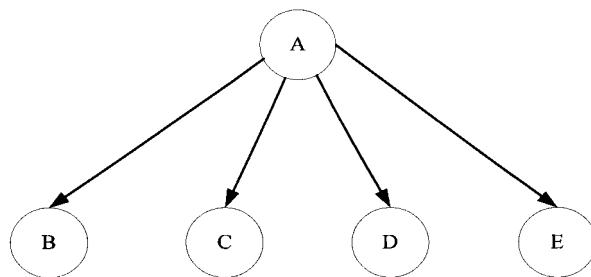


图 3

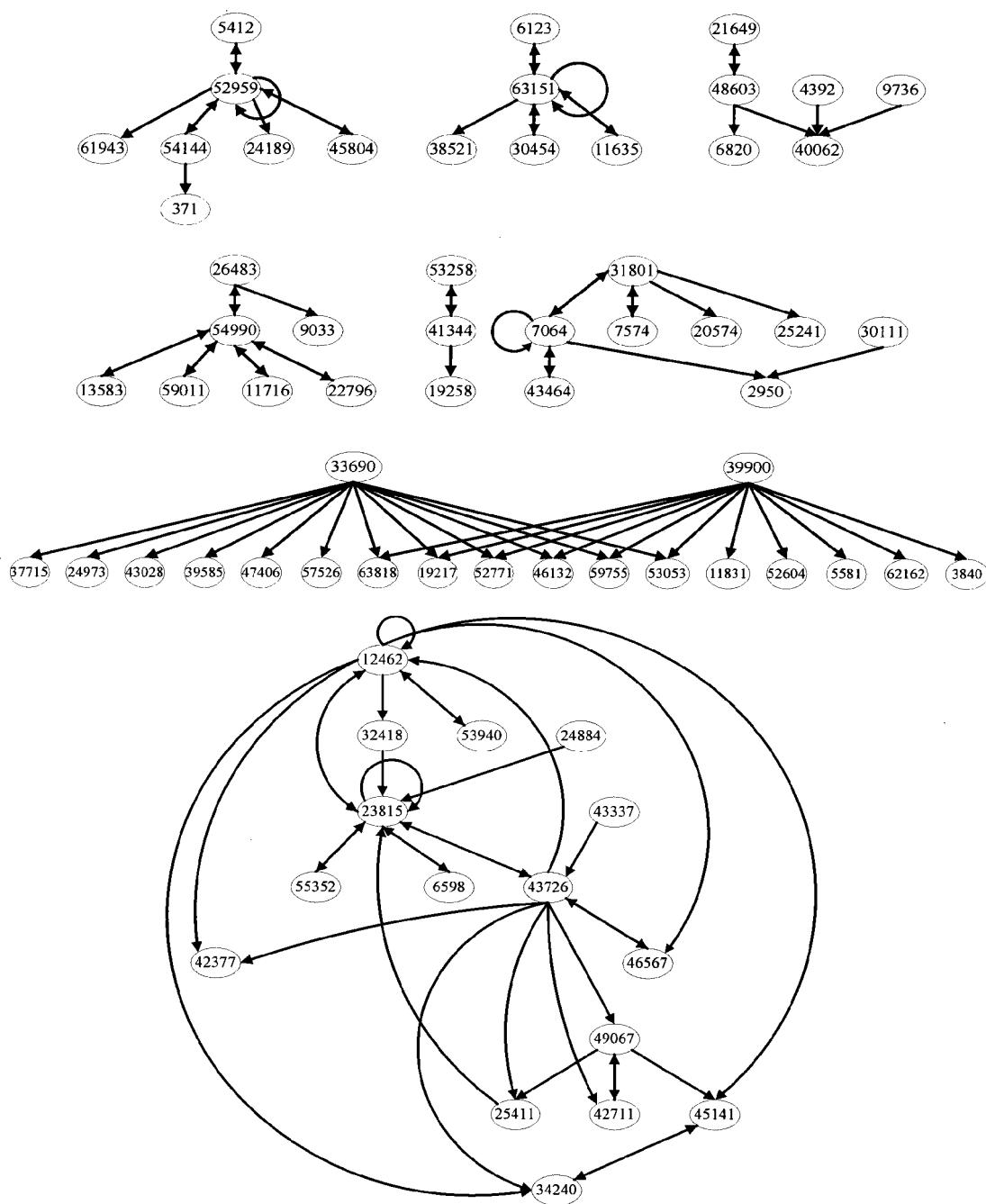


图 4

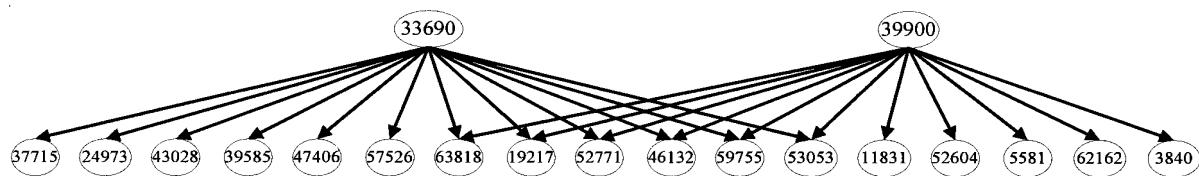


图 5

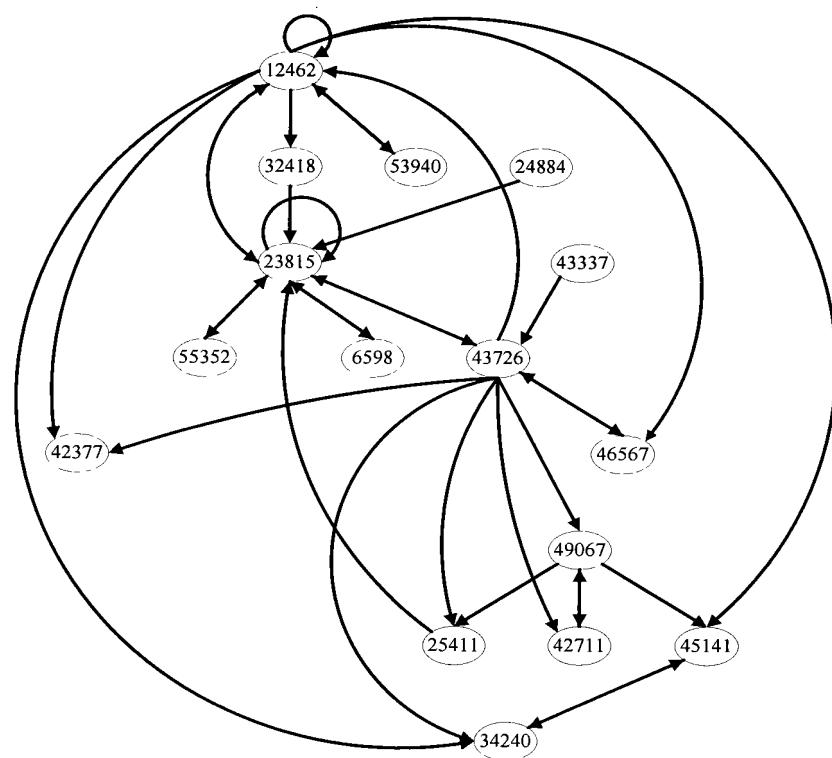


图 6

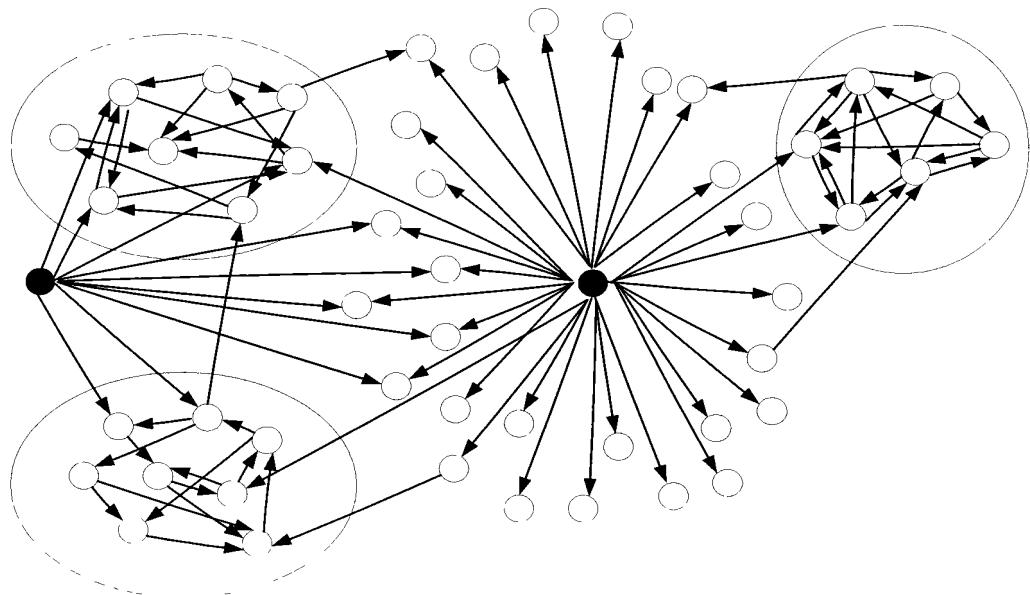


图 7

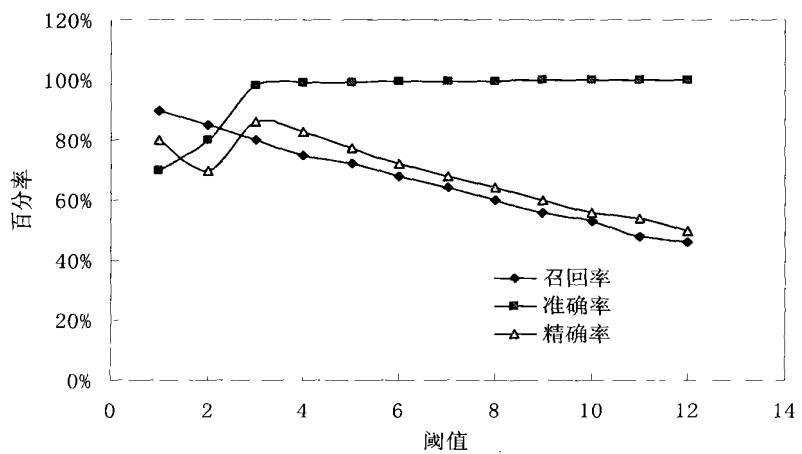


图 8

表 1

邮件数量	拓扑算法 (us/封)	Winnow 算法 (us/封)
1500	6.3	91761
3000	6.35	90862
4500	6.4	89755
6000	6.3	93256
7500	6.3	90023
8586	6.05	91253

图 9