



(12) 发明专利

(10) 授权公告号 CN 1573784 B

(45) 授权公告日 2012.11.07

(21) 申请号 200410063953.9

(56) 对比文件

(22) 申请日 2004.06.04

US 6615242 B1, 2003.09.02,

(30) 优先权数据

WO 02093428 A1, 2001.12.21,

10/454,168 2003.06.04 US

审查员 刘欢

(73) 专利权人 微软公司

地址 美国华盛顿州

(72) 发明人 J·T·古德曼 R·L·罗斯怀特

D·格沃兹 J·D·梅尔

N·D·豪威尔 M·C·鲁普斯伯格

B·T·斯塔白克

(74) 专利代理机构 上海专利商标事务所有限公司 31100

代理人 谢喜堂

(51) Int. Cl.

G06F 17/60 (2006.01)

权利要求书 3 页 说明书 21 页 附图 15 页

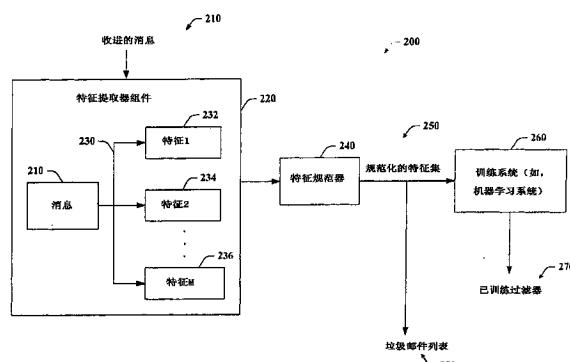
(54) 发明名称

用于阻止垃圾邮件的源 / 目的地的特征和列

表

(57) 摘要

本发明包括一种易于从消息中提取用于垃圾邮件过滤的数据的系统和方法。所提取的数据可以是特征的形式，其能够与机器学习系统一同使用，以建立改进的过滤器。嵌入在消息体中的与源信息以及其它信息相关联的数据能够作为特征被提取，该消息允许消息的收件人联系和 / 或者响应消息的发件人。在被用作机器学习系统的特征之前，该特征或者其子集能够被规范化和 / 或者被摆脱困惑。该（已摆脱困惑的）特征能被用于填充多个易于检测和阻止垃圾邮件的特征列表。示范性的特征包括一个 email 地址，IP 地址，URL，指向 URL 的一个嵌入式图像，以及 / 或者其中的一部分。



1. 一种便于提取关于垃圾邮件处理的数据以阻止垃圾邮件的系统，包括：

一个部件，其接收一个项目并且提取一组特征，所述一组特征与使预定的收件人能够就消息进行联系、响应或者接收的所述消息的源和所述消息的部分以及信息中的一个或多个相关联，其中，所述一组特征包括主机名和域名；

一个规范化部件，其使一个特征子集摆脱困惑；

一个部件，其利用一个被提取的特征的子集建立过滤器，所述过滤器是垃圾邮件过滤器；以及

机器学习系统，利用所述垃圾邮件过滤器来识别和阻止垃圾邮件。

2. 权利要求 1 的系统，该过滤器是一个父控制过滤器，所述父控制过滤器通知用户所述消息是不适宜的，而且也能够表明这种不适宜的原因。

3. 权利要求 1 的系统，进一步包括一个机器学习系统部件，其利用已摆脱困惑的特征来学习至少垃圾邮件和非垃圾邮件其中之一。

4. 权利要求 1 的系统，所述特征的子集包括至少 IP 地址，该至少一个 IP 地址是回复地址、抄送地址、收件人地址、发件人地址、和定位在消息中的 URL 中的任何一个的至少一部分。

5. 权利要求 4 的系统，该 IP 地址包括一个数据块 ID，其中该数据块 ID 能被提取作为至少一个特征。

6. 权利要求 5 的系统，其中该数据块 ID 至少部分地通过查阅一个数据块目录来确定。

7. 权利要求 6 的系统，其中该数据块目录是 arin.net。

8. 权利要求 5 的系统，其中该数据块 ID 至少部分地通过推测来确定，从而提取 IP 地址的至少前 1 个比特，至少前 2 个比特，至少前 3 个比特，直到至少前 31 个比特中的任何一个作为特征。

9. 权利要求 1 的系统，其中特征的子集包括 IP 地址的前 1 个到前 31 个比特中的每一个。

10. 权利要求 1 的系统，特征的子集包括一个 URL。

11. 权利要求 10 的系统，其中 URL 地址被定位在至少其中一个消息体，作为文本嵌入在消息中，以及嵌入在消息中的图像中。

12. 权利要求 1 的系统，进一步包括一个部件，其使用至少一个已提取特征的子集来填充至少一个特征列表。

13. 权利要求 12 的系统，该至少一个特征列表是好用户的列表、垃圾邮件制作者列表、表示合法的发件人的肯定特征的列表、以及表示垃圾邮件的特征的列表中的任何一个。

14. 权利要求 1 的系统，其中该特征子集包括至少一个 URL。

15. 权利要求 14 的系统，其中 URL 作为文本被嵌入到消息体中。

16. 权利要求 14 的系统，其中 URL 是消息体中链接的至少一部分。

17. 权利要求 14 的系统，其中 URL 是作为图像嵌入到消息中链接的至少一部分。

18. 权利要求 1 的系统，特征的子集包括从 email 地址中提取的主机名称和域名中的至少一个。

19. 权利要求 1 的系统，特征的子集包括从 email 地址和 URL 任何一个中提取的至少一部分 FQDN。

20. 权利要求 1 的系统, 特征的子集包括从 email 地址和 URL 任何一个中提取的至少一部分域名。

21. 权利要求 1 的系统, 至少一部分被提取特征的子集在同机器学习系统一同使用之前被规范化。

22. 权利要求 1 的系统, 至少一部分被提取特征的子集在被用于填充至少一个特征列表之前被规范化。

23. 权利要求 1 的系统, 进一步包括一个分类部件, 其分类 URL、email 地址和 IP 地址中的至少一个的至少一部分, 作为成人、成人内容、不适合的、不适合某个年龄段的、适合于所有年龄的、不合宜的以及合宜的中的任何一个。

24. 权利要求 23 的系统, 其中该分类部件是一个父控制系统, 所述父控制系统通知用户所述消息是不适宜的, 而且也能够表明这种不适宜的原因。

25. 权利要求 23 的系统, 其中该分类部件分配一个或多个特征类型给 URL、网站地址和 IP 地址中的至少一个的已分类部分。

26. 权利要求 1 的系统, 其中该特征组包括至少一个非免费电话号码, 所述电话号码包含一个电话地区号, 以便于映射发件人或者与消息相关的联系者的地理位置。

27. 一种使用权利要求 1 的系统的计算机。

28. 一种提取与垃圾邮件处理有关的数据以阻止垃圾邮件的方法, 包括 :

接收一个消息 ;

提取一组特征, 所述一组特征与使预定的收件人就所述消息进行联系, 响应或者接收的所述消息的源和所述消息的部分以及信息中的一个或多个相关联, 其中, 该组特征包括一个 IP 地址的至少一部分, 并且提取所述 IP 地址的至少一部分包括执行下述至少一个行动 : 查找一个数据块 ID 的目录以确定至少一个对应于 IP 地址的数据块 ID, 以便该数据块 ID 被提取作为一个附加的特征或从 IP 地址中提取至少前 1 个比特直到前 31 个比特中的每一个 ;

通过规范化所提取的特征来使一个特征子集摆脱困惑 ;

利用一个被提取的特征的子集以建立过滤器, 所述过滤器是垃圾邮件过滤器 ; 以及利用所述垃圾邮件过滤器来识别和阻止垃圾邮件。

29. 权利要求 28 的方法, 其中已提取的 IP 地址对应于服务器。

30. 权利要求 29 的方法, 进一步包括提取该服务器作为一个附加的特征。

31. 权利要求 28 的方法, 进一步包括使至少一个从消息中提取的特征的子集摆脱困惑。

32. 权利要求 28 的方法, 进一步包括使从消息中提取的至少一个特征的至少一部分摆脱困惑。

33. 权利要求 32 的方法, 其中使从消息中提取的接收的发件人的 IP 地址摆脱困惑包括 : 追溯通过多个“添加到”IP 地址的搜索路径, 来核对“添加到”IP 地址的搜索路径的身份。

34. 权利要求 32 的方法, 进一步包括从网站地址提取附加的特征, 包括执行至少下列动作的其中一个 :

每次删除至少一个后缀, 从而产生相应的附加特征 ; 以及

每次删除至少一个前缀，从而产生相应的附加特征。

35. 权利要求 32 的方法，其中该组特征包括回复地址，抄送地址，收件人地址，URL，链接，和发件人地址中任何一个的至少一部分。

36. 权利要求 28 的方法，其中至少一个被提取特征的子集作为文本和图像其中之一被嵌入消息体中。

37. 权利要求 28 的方法，其中该组特征包括一个主机名和一个域名。

38. 权利要求 28 的方法，进一步包括分类一个或者多个已提取的特征以表明与该消息相关联的是适宜的和不适宜的内容中的任何一种，并且将这种分类用作一个附加的特征。

39. 权利要求 28 的方法，进一步包括分配一种特征类型给相应的被提取的特征以便至少部分地基于各个已提取的特征来通知用户消息内容，并且利用这种特征类型作为一个附加的特征。

40. 权利要求 39 的方法，进一步包括确定特征类型和特征中的至少一种是稀有的和通用的其中之一，并且利用特征的稀有性和通用性作为一个附加的特征。

41. 权利要求 28 的方法，其中特征的子集经由一个机器学习系统被采用来建立一个过滤器。

42. 权利要求 28 的方法，其中该过滤器是一个垃圾邮件过滤器。

43. 权利要求 28 的方法，其中该过滤器是一个父控制过滤器，所述父控制过滤器通知用户所述消息是不适宜的，而且也能够表明这种不适宜的原因。

44. 权利要求 28 的方法，进一步包括使用至少一个从消息中提取的特征的子集来填充一个或者多个特征列表。

45. 权利要求 44 的方法，其中特征列表包括表示非垃圾邮件制作者的肯定的特征列表，和表示垃圾邮件制作者的恶意的特征列表中的至少之一。

46. 权利要求 28 的方法，其中在被用作机器学习系统的特征之前，已提取的特征至少部分地被摆脱困惑。

47. 权利要求 28 的方法，其中在被用作填充特征列表的特征之前，已提取的特征至少部分地被摆脱困惑。

48. 一种提取与垃圾邮件处理过程有关的数据以阻止垃圾邮件的系统，包括：

用于接收消息的装置；

一种装置，用于提取一组特征，所述一组特征与能够使预定的收件人就所述消息进行联系、响应或者接收的所述消息的源和所述消息的部分以及信息中的一个或多个相关联，其中，所述一组特征包括主机名和域名；

用于通过规范化所提取的特征来使一个特征子集摆脱困惑的装置；以及

一种装置，用于利用一个被提取的特征的子集以建立过滤器，所述过滤器是垃圾邮件过滤器；

用于利用所述垃圾邮件过滤器来识别和阻止垃圾邮件。

用于阻止垃圾邮件的源 / 目的地的特征和列表

技术领域

[0001] 本发明涉及用于识别合法的（例如，好的邮件）和不希望得到的邮件，尤其涉及用于处理电子消息来提取数据以方便阻止垃圾邮件的系统和方法。

[0002] 发明背景

[0003] 诸如因特网这样的全球通信网络的出现已经为达到大量的潜在客户带来了商机。电子消息，尤其是电子邮件（“电子邮件”）作为一种向网络用户散布不需要的广告和宣传（也表示为“垃圾邮件”）的方式正在日益蔓延。

[0004] Radicati 集团有限公司，其是一个咨询销售研究公司，估计到 2002 年 8 月份为止，每天将发送二十亿条垃圾电子邮件消息。这个数量预计每两年翻三倍。个人和单位（例如，商业，政府机构）变得越来越不方便，而且时常被垃圾邮件搞得不愉快。同样，对于可靠的数据处理来说，垃圾邮件正在或者很快会变为一种主要的威胁。

[0005] 用于阻止垃圾邮件的关键的技术是使用过滤系统 / 方法。一种被证实的过滤技术是基于一种机器学习方法——机器学习过滤器分配给输入消息一个该消息是垃圾邮件的概率。在这种方法中，典型地从两种示例性消息（例如，垃圾邮件和非垃圾邮件消息）中提取特征，而且学习过滤器被应用于在两种类型之间进行概率区分。由于多种消息特征与内容（例如，在题目和 / 或者消息体中的单词和短语）相关，所以这种类型的过滤器通常称之为“基于内容的过滤器”。

[0006] 随着这种垃圾邮件过滤技术的冲击，许多垃圾邮件制作者已经想出了伪装它们身份以避免和 / 或者绕过垃圾邮件过滤器的方法。因此，在识别和阻止伪装了的垃圾邮件消息中，传统的基于内容的和自适应过滤器可能变得无效。

[0007] 发明概述

[0008] 为了提供对本发明某些方面的一个基本的理解，下面给出了本发明的一个简单的概述。这种概述不是本发明大范围的综述。不是为了识别本发明关键的 / 重要的要素，或者描绘本发明的范围。其唯一的目的是以简单的方式给出本发明的一些概念作为后面给出的更详细描述的开头。

[0009] 垃圾邮件制作者在它们的消息中几乎能够伪装所有的信息。例如，它们能够嵌入图像，所以没有作为用于机器学习系统的特征的字。图像甚至可能以失真的方式使得使用 ORC 软件变得困难，或者至少是耗时的。尽管如此，不管他们消除了多少特征，仍然存在有用的信息。首先，垃圾邮件制作者必须从某处发送该消息。我们能够检测消息是从哪个 IP 地址接收的。其次，垃圾邮件制作者几乎总是试图销售某物，因此必须包括联系他们的一种方式。这可能是免费号码，但是垃圾邮件制作者可能不愿使用该号码，因为抱怨高成本。其可能是非免费号码，但是因为较低的响应率，垃圾邮件制作者可能不愿这样做。作为选择，其可能是一个 URL（例如，<http://www.spamcorp.com/buyenlarger.htm>）。该 URL 可能被嵌入到一个图像中，使得过滤器和 / 或者软件更难检测到它。然而，垃圾邮件制作者可能不愿这样做，因为用户需要在他们的浏览器中键入该 URL，其可能使响应率较低。

[0010] 对垃圾邮件制作者来说，最可能的联系方式是嵌入链接，或者通过一个某种嵌入

电子邮件地址。例如，“点击这里可以了解更多”，其中“点击这里”包括一个具体网页的链接，根据本发明的一个方面，机器学习系统能够检测并使用该网页。同样，将要回复的地址（例如，典型地是“来自地址”，但有时是“回复”地址，如果存在一个的话），或者任何嵌入邮件到：链接（允许通过点击链接发送邮件消息的链接），或者任何其它电子邮件地址。另外，垃圾邮件制作者通常在消息中包括图像。因为反复邮寄大量的图像花费很高，所以垃圾邮件制作者通常仅将一个特殊的链接嵌入到图像中，这就会引起图像被下载。这些链接点的位置也能够作为特征被使用。

[0011] 关于从来自地址的邮件，邮件回复地址，嵌入邮件到的地址，外部链接，以及外部图像的链接中提取的信息，至少这种信息的一部分能被用作机器学习系统的一个特征，一个加权或者概率与之相关联；或者该信息可能被加入到一个列表中。例如，我们能够保存列表，这些列表是 IP 地址，或者来自只发送垃圾邮件的地址，或者仅发送好邮件的地址，或者发送 90% 以上是好邮件的地址等。事实是，在这样的列表上的一个特殊的链接或者地址既能够被用作机器学习系统的一个特征，又能够被用作任何其它垃圾邮件过滤系统的一部分，或者两者。

[0012] 本发明提供一种通过检查消息的特定部分来易于识别伪装的垃圾邮件消息的系统和方法。尤其是，本发明涉及处理一种诸如像电子邮件（电子邮件）这样的消息以提取源和 / 或者目的地数据，来区分垃圾邮件消息和合法的消息。该处理方法包括识别和分析 IP 地址信息，电子邮件地址信息，和 / 或者通用资源定位器（URL）信息的各种技术，以及将所提取的数据与垃圾邮件的属性（例如，好的用户对恶意的用户，或者好的发件人对恶意的发件人）相关联的各种技术。例如，一个恶意的用户或者恶意的发件人将被认为是一个垃圾邮件制作者（例如，发送垃圾邮件的那个人）。

[0013] 所提取的数据，或者至少其中的一部分能够被用于为机器学习系统产生特征设置。机器学习技术检查消息的内容以确定该消息是否是垃圾邮件。垃圾邮件制作者能够使消息的大部分内容变得混乱，诸如通过将它们的大部分信息放入到难以处理的图像中。然而，消息的起源不能被完全地伪装，由于垃圾邮件制造者需要为收件人提供某种易于联系它们的方式。这样的实例包括使用一个链接（例如，URL）和 / 或者一个电子邮件地址（例如，IP 地址）。这些类型的信息或变体，或其中的部分能被用作垃圾邮件检测器的特征。尤其是，例如，该信息借助于机器学习系统能被用于训练一个垃圾邮件检测器和 / 或者垃圾邮件过滤器。

[0014] 本发明也能够与父控制系统合作。父控制系统可能通知用户该消息是不适宜的，而且也能够表明这种不适宜的原因，诸如包括色情资料。根据本发明的一个方面，一个或者多个被提取的并且被规范化的特征（例如，URL）能够通过一个父控制系统或者过滤器来获得父控制系统的分类。这种分类可能被用作该机器学习系统的一个附加的特征以方便建立和 / 或者改善垃圾邮件过滤器。

[0015] 此外，所提取的特征能够通过类型来分类，能够根据垃圾邮件的程度来加权，而且能够指明要么是肯定的（例如，很可能不是垃圾邮件），要么是否定的（很可能是垃圾邮件）特征。该特征也能够被用于创建诸如非垃圾邮件制造者列表和垃圾邮件制造者列表这样的列表。

[0016] 为了完成上述和相关的目的，这里结合下面的描述和附图描述了本发明的某些示

例性的方面。然而,这些方面表明了可以使用本发明的原理的几种不同的方式,而且本发明试图包括所有的这些方面及其它们的等价物。当结合附图考虑时,本发明的其它优点和新的特征从下面本发明的详细描述中将变得显而易见。

[0017] 附图简述

[0018] 图 1 是根据本发明的一个方面的易于阻止垃圾邮件的一个系统的高级框图;

[0019] 图 2 是根据本发明的一个方面,通过从输入的消息中提取一种或者多种特征以易于阻止垃圾邮件的系统的框图。

[0020] 图 3 是根据本发明的一个方面,能够从一个 IP 地址中提取的多个特征的示意性框图。

[0021] 图 4 是根据本发明的一个方面,能够从一个 FQDN 中提取的多个特征的示意性框图。

[0022] 图 5 是根据本发明的一个方面,能够从一个电子邮件地址中提取的多个特征的示意性框图。

[0023] 图 6 是根据本发明的一个方面,能够从一个 URL 或者网址中提取的多个特征的示意性框图。

[0024] 图 7 是根据本发明的一个方面与训练过滤器有关的示例性方法的流程图。

[0025] 图 8 是根据本发明的一个方面与使用一个训练过滤器有关的示例性方法的流程图。

[0026] 图 9 是根据本发明的一个方面与创建一个列表有关的示例性方法的流程图。

[0027] 图 10 是根据本发明的一个方面与使用一个列表来训练过滤器有关的示例性方法的流程图。

[0028] 图 11 是根据本发明的一个方面,至少参考图 7 和 8 的方法的处理过程的流程图。

[0029] 图 12 是根据本发明的一个方面,易于在合法的和伪造的发件人的 IP 地址中作出区分的处理过程的流程图。

[0030] 图 13 是根据本发明的一个方面,在来自输入消息的特征的生成和提取中结合父控制系统的方法的流程图。

[0031] 图 14 是根据本发明的一个方面,易于创建将在机器学习系统中使用的特征集的方法的流程图。

[0032] 图 15 是用于实施本发明各个方面的一种示例性的环境。

[0033] 发明详述

[0034] 现在将参考附图描述本发明,其中相似的参考数字完全被用于参照相似的元件。在下面的描述中,为了提供对本发明总体上的理解,出于解释的目的,阐明了多个具体细节。然而,很显然没有这些具体的细节也可以实施本发明。在另外的例子中,为了便于描述本发明,以方框图的形式示出了熟知的结构和设备。

[0035] 正如在该申请中所使用的,术语“组成部分”和“系统”是指与计算机相关的一个实体,要么是硬件,硬件和软件的组合,软件,要么是运行中的软件。例如,一个组成部分可能是,但不被限制为在处理器上运行的一个处理过程,一个处理器,一个对象,一个可执行的,一种执行的线程,一段程序,和 / 或者一台计算机。通过举例说明,在服务器上运行的应用程序和该服务器都可能是一个组成部分。一个或者多个组成部分可能驻留在一个处理器

中,和 / 或者执行的线程中,以及一个组成部分可以被定位在一台计算机上,和 / 或者在两台或者多台计算机之间分布。

[0036] 本发明可能包括各种推断方案和 / 或者技术,这些方案和 / 或者技术是关于为学习垃圾邮件过滤的机器产生训练数据。正如在这里所使用的,术语“推断”一般认为是与推断系统状态的过程,环境,和 / 或者来自一组经由事件和 / 或者数据而被捕获的观察的用户有关。例如,推断能够被用于识别一个具体的上下文或者动作,或者能够产生基于状态的概率分布。这种推断可能是概率性的,即基于数据和事件的考虑,基于感兴趣的状态来计算概率分布。推断也可能是指用于从一组事件和 / 或者数据中构成更高级别事件的技术。这种推断导致了从一组已观察的事件和 / 或者所存储的事件数据中构造新的事件或者动作,无论在密切临时接近中的事件是否相关联,以及是否该事件和数据来自一个或者多个事件或者数据源。

[0037] 应当理解尽管术语消息在整个说明书中被广泛的使用,但是这样的术语并没有从本质上限制电子邮件,但是可能更适合于包括能够在任何合适的通信结构上分布的任何形式的电子消息。例如,易于在两个或者多个人(例如,交互聊天程序,以及立即通知的程序)之间进行会议的会议应用和程序也能够利用这里公开的过滤的优点,由于不需要的文本在用户交换消息时,能够被电子地散布到通常的聊天消息中,和 / 或作为开始消息,结束消息被插入,或上述的全部。在这个特殊的应用中,为了捕获不希望的内容(例如商业广告节目,推销做广告,或广告)并且将其加标签为垃圾邮件,过滤器能被训练为自动过滤特殊的消息内容(文本和图像)。

[0038] 在本发明中,术语“收件人”指引入消息或邮件内容的地址。术语“用户”可能指收件人或发件人,这由上下文而定。例如,用户可能是指发送垃圾邮件的电子邮件用户,和 / 或用户可能是指接收垃圾邮件的电子邮件收件人,这由上下文和术语的应用而定。

[0039] 网际协议(IP)地址是一个32比特数字,典型地代表国际互联网上的一台机器。在当两台机器通信时使用这些数字。典型地以“XXX.XXX.XXX.XXX”的形式代表了它们,其中每个XXX在0和255之间。不幸地是,IP地址很难记忆。因为这个原因,就创造了“域名”和“主机名”协定。“域名”是指国际互联网上的一组机器的名字(可能是一台机器),并且典型的形式为“x.com”,或“y.edu”,或“courts.wa.gov”。

[0040] 一个正式域名(FQDN)是国际互联网上的一台特殊的机器,例如“b.x.com”或“c.y.edu”或“www.courts.wa.gov”,域名部分分别是“x.com”或“y.edu”或“courts.wa.gov”。“b”“c”和“www”部分分别被称为FQDN的主机名部分。通常,IP地址能被用在域名可能使用的任何情形中(例如“DN/IP”说明两种可能性存在)。而且通常,IP地址能被用在FQDN可能使用的任何情形中(例如“FQDN/IP”说明两种可能性存在)。一个电子邮件地址由用户名和域名或IP地址(DN/IP)组成,例如“a@x.com”或“a@1.2.3.4”。在两个例子中,用户名都是“a”。

[0041] 统一资源定位器(URL)典型的形式是“服务名称:FQDN/IP/url-path。”例如,“http://www.microsoft.com/windows/help.htm”。“http”部分是服务器名。“/www.microsoft.com”部分是FQDN,以及“windows/help.htm”是URL路径。这是某种URL的简化,但是对本发明来说已经是足够了。

[0042] 参考图1,示出了根据本发明的一个方面的特征提取和训练系统100的大体的框

图。特征提取和训练系统 100 包括处理输入消息 110 以便从消息中提取特征数据。这种特征能够从至少一部分源和 / 或者目的地信息中提取，这些信息是在消息和 / 或者其变型中提供。尤其是，一个或者多个输入消息 110 能够通过系统 100 经由消息接收部件 120 被接收。消息接收部件 120 能够被定位在一个电子邮件或者消息服务器上，例如，用来接收输入消息 110。尽管某些消息（例如，至少一个）对于现存的过滤器（例如，垃圾邮件，父控制过滤器）来说是易于攻击的，因此转向了一个垃圾箱或者垃圾邮件文件夹中，至少部分的源和 / 或者目的地数据能够被提取或者被理解，用于与机器学习系统或者填充一个特征列表有关的用途。

[0043] 消息接收部件 120 能够将输入消息，或者其中的一个子集传递到特征提取部件 130。该特征提取部件 130 能够从接收的消息 110 中提取数据，以便产生特征集以方便过滤器训练和最终的垃圾邮件检测。从消息中提取的数据或者特征与在其中被发现的和 / 或者嵌入的源和 / 或者目的地信息相关。数据或者特征的例子包括发件人的 IP 地址，回复的电子邮件地址，cc：（例如，副本）电子邮件地址，各种 URL（包括基于文本的链接，基于图像的链接，以及以文本形式的 URL 或者其中的一部分），非长途免费电话号码（例如，尤其是一个区号），长途免费的电话号码，邮寄到：电子邮件地址链接，文本形式的电子邮件地址，在 SMTPHELO 命令中的 FQDN，SMTP MAIL FROM 地址 / 返回路径地址，以及 / 或者至少任何上述中的一部分。

[0044] 特征提取部分 130 能够执行任何合适的数字处理，以便从消息 110 中提取各种特征集，随后在机器学习系统中使用。另外作为选择，特征集能被用于填充用于其它过滤器训练技术的列表。

[0045] 例如，诸如 a. x. com 这样的 FQDN 能够被翻译成一般被称作 IP 地址的号码。IP 地址典型地以有点的十进制的形式被观察，包括 4 个数字数据块。每个数据块分别由小数点或者点分开，而且每个数字数据块的范围是从 0 到 255，其中每个号码的变化对应于不同的英特网名称。例如，a. x. com 可能被翻译为 123. 124. 125. 126，而 121. 124. 125. 126 可能代表 qustuv. com。因为数字不如单词容易识别或者记忆。IP 地址通常通过它们各自的 FQDN 来被查阅。以有点的十进制格式表示的相同的 IP 地址也能够以可选择的下面将要描述的形式被表示。

[0046] 根据本发明的一个方面，特征提取部件 130 能够集中到包括在消息 110 中的发件人 IP 地址。发件人 IP 地址至少部分地基于发件人 IP 信息。一般来说，在英特网上的邮件发送是从服务器到服务器的传送，有时只包括两个服务器（例如，一个发件人和一个收件人）。更罕见的一种情况是，客户机能够直接发送到一个服务器。在某些情况下，能够包括更多的服务器，例如，由于防火墙的出现，邮件或者消息能够从一个服务器被发送到另一个服务器。尤其是，一些服务器能够被定位在防火墙之内，因此这些服务器就仅能够与防火墙另一侧的指定的服务器进行通信。这就引起了消息从发件人到收件人过程中，消息要经过的跳数的增加。接收线路包含 IP 地址，以方便跟踪消息的路径来确定消息从哪里发起。

[0047] 当消息 110 从服务器到服务器传播时，每个被联系的服务器将它从其接收消息的 IP 地址识别预先考虑到发送字段（即，接收的字段），也预先考虑服务器被断定的 FQDN 的名字，该服务器正在与它对话。该 FQDN 由发送服务器通过 SMTP 协议的 HELO 命令告诉接收服务器，因此如果发送服务器在该体系结构的外部时，那么收到的 FQDN 就不可信。例如，该

消息从具有 5 个预先考虑的 IP 地址和 FQDN 的线路中被接收五次,因此表明其已经通过六个不同的服务器(即已经通过 5 次),这些线路在相反的顺序中被预先考虑(即最近的开始)。然而,每个服务器都具有修改任何较低的(早期预先考虑的)线路的能力。当消息已经在多个服务器之间传播时,这可能尤其有问题。因为每个中间的服务器都能够改变任何早期所写的(较低的)发件人线路。垃圾邮件制作者能够预先考虑消息的发件人的伪 IP 地址,以伪装发件人的 IP 信息或者垃圾邮件消息的发件人。例如,垃圾邮件消息可能最初出现,好像其从 trusteddomain.com 被发送,因此错误地显示了到收件人的消息的真正的来源。

[0048] 对于垃圾邮件软件来说,容易地识别体系结构之外的 IP 地址是重要的,该 IP 地址被发送到体系结构内部的服务器上。由于该 IP 地址被接收服务器写入,所以在该体系结构内部,其可能被作为一个正确的 IP 地址来对待。所有其它的在该体系结构外部的 IP 地址都不被信任,由于它们被在体系结构之外的服务器写入,因此,很可能被修改。可能存在许多包括在到接收体系结构的路径中的发送服务器的 IP 地址,但是由于仅有一个地址能够被信任,所以我们仅参考可信赖的这一个作为“发件人”的 IP 地址。

[0049] 对于垃圾邮件过滤软件来说,一种用于找到发件人 IP 地址的方法是弄清楚在一个体系结构处的邮件服务器的配置。一般来说,如果一个服务器知道了哪一台机器通过其它的在状态中的机器,则其能够确定发件人的 IP 地址。然而,描述服务器的配置,尤其对于安装在 email 客户机上的垃圾邮件过滤软件来说,不是很方便的。一种可替换的方法包括利用 MX 记录来确定消息的真正来源。MX 记录列表,用于每个域名,用于该域名的邮件收件人的 FQDN。通过发件人的列表能够跟踪回一个 IP 地址,直到发现一个 IP 地址为止,该 IP 地址对应于一个 FQDN,该 FQDN 对应于在域名 MX 记录中的一个实体。机器接收的 IP 地址是发件人的 IP 地址。想像 1.2.3.101 是用于 x.com 的唯一的 MX 记录。然后,通过找到从 1.2.3.101 接收的线路,能够知道对应于 x.com 的输入邮件服务器的下一个线路,因此在该线路中的 IP 地址对应于发送到 x.com 的 IP 地址。

[0050] 下表描述了一种示例性的分析,正如讨论先前确定的消息的真正来源一样:

[0051]

行	注释
Received :from a.x.com({1.2.3.100})by b.x.com Tue, April122,200313:11:48-0700	在 x.com 的内部
Received:from mailserver.x..com({1.2.3.101})by b.x.com Tue April122,2003,12:11:48-0700	1.2.3.101 是用于 x.com 的一条 MX 记录, 所以我们知道下 一条线路是 x.com 的内部的第一条
Received :from outside.com({4.5.6.7})by mailserver.x.comTue April122,2003 11:11:48-0700	这是所接收的 x.com 的消息 : 这是最后一条可信的线路, 使用 4.5.6.7 作为发件人的 IP 地址
Received::from trustedsender.com({8.9.10.11})by outside.com Tue April122,2003,10:11:48-0700	通过服务器在 4.5.6.7 构造的线路可能是假的

[0052] 当前,没有用于列出输出邮件服务器的可接受的标准,例如,如果该启发式的邮件服务器可能失败的话,在一个体系结构之内的 IP 地址不同于在一个体系结构之外的那些 IP 地址,或者如果一个体系结构从 MX 记录中列出的机器直接发送邮件到 MX 记录中列出的另一个机器。此外,在特殊的情况下,即如上所述发件人的 IP 被发现是在体系结构的内部,如果在 MX 记录中的一个机器可能发送到 MX 记录中的另一个机器时,如上所述的过

程被继续。另外,某个 IP 地址可能作为内部的 IP 地址被检测(因为它们是通过 172.31.y.z,或者通过 192.168.0.z 的形式 10.x.y.z 或者 172.16.y.z,一种仅用于内部 IP 地址的形式):任何到达体系结构内部的 IP 地址都能够被信任。最后,如果接收线路的形式是“从 a..x.com[1.2.3.100]”并且 a..x.com 的 IP 地址的查找输出 1.2.3.100,或者反向的 1.2.3.100 的 IP 地址的查找输出 a..x.com,如果 x.com 是一个体系结构,那么下一个线路也可能是可信任的。

[0053] 通过使用这些观察,找到发件人的 IP 地址通常是可能的,示例性的伪代码如下:

```
[0054]     bool fFoundHost InMX ;
[0055]     if(extemal IP address of MX records matches internal IP address of
MX records)
[0056]     {
[0057]         fFoundHost InMX=FALSE ;#it's worth looking for
[0058]     }else{
[0059]         fFoundHost InMX=TRUE ;#it's not worth looking for
[0060]         pretend we already found it
[0061]     }
[0062]     for each received from line of the form Received from a.b.c
[0063]         [i.j.k.1]{
[0064]             if i.j.k.1 in MX records of receiver domain
[0065]             {
[0066]                 fFoundHost InMX=TRUE ;
[0067]                 continue ;
[0068]             }
[0069]             if not fFoundHost InMx
[0070]             {
[0071]                 #Has not yet gone through an MX record, must be internal
[0072]                 continue ; ;
[0073]             }
[0074]             if i.j.k.1 is of form
[0075]                 10.x.y.z or
[0076]                 172.16.y.z to172.31.y.z or
[0077]                 192.168.0.z to192.168.255.z
[0078]             {
[0079]                 #Must be internal
[0080]                 continue ;
[0081]             }
[0082]             ifDNS lookup ofa.b.c yields i.j.k.1 and b.c is
[0083]             receiver domain
[0084]             {
```

```
[0085]      #Must be internal
[0086]      continue ;
[0087]      }
[0088]      Output sender's alleged FQDN a.b.c and sender's actual IP address
i.j.k.k
[0089]      }
[0090]      If we reach here, then Error:unable to identify sender's alleged
FQDN and
[0091]      sender's actual IP address.
```

[0092] 利用发件人的 IP 地址,同时利用其它的源和目的地特征能够做很多事情。首先,它们能一律被加到恶意发件人的列表中,有时候称为 Black 列表。Black 列表实际上能够被用于过滤,阻止,或者重新定向一个不可信赖的消息到一个适当的文件夹或者它们能够被进一步调查的一个位置。

[0093] 其它类型的列表也可能被产生并且作为过滤器在基于结构的客户机或者服务器上被使用。在客户机结构中,用户能够通知客户机电子邮件软件,他将从哪里接收邮件(例如,邮件列表,个人等)。对应于可信的电子邮件地址的记录的一个列表要么自动要么手动通过用户产生。因此,想像具有电子邮件地址 ‘b@zyx.com’ 的发件人发送给用户一个电子邮件消息。该发件人的电子邮件地址 b@zyx.com 包括用户名 ‘b’,以及 FQDN/IP ‘zyx.com’。当客户机从发件人 (b@zyx.com) 接收输入消息 110 时,它能够检索一个用于用户电子邮件地址的可信的发件人列表,以确定是否用户已经表明 ‘b@zyx.com’ 是一个有效的而且可信的地址。对于服务器结构来说,该列表能够被直接定位在服务器上。因此,当消息到达消息服务器时,它们的各个特征(例如,发件人的 IP 地址,在 MAILFROM 或者 HELO 字段中的域名,以及其它的源和 / 或者目的地信息)能够与定位在消息服务器上的列表相比较。根据基于客户或者基于服务器的传送协议,确定是来自有效发件人的消息能被传送到所希望的收件人。然而,确定包括了在有问题的或者不好的特征列表中的源或者目的地特征的消息,能够被移到垃圾邮件文件夹中以便删除,或者相反被特别地处理。

[0094] 作为一种填充可信的或者有害的源特征列表的选择,发件人的源特征(例如,IP 地址,合法的 From 地址)能够被提取作为一个或者多个特征,而且日后与机器学习技术一同用于过滤器的建立和 / 或者训练。

[0095] IP 地址能够从一个消息首部的任何部分中的 email 地址(例如,在发送者的地址或者答复地址中有关 FQDN 的 IP 查询)导出,或者从嵌入到消息实体中的一个 URL 链路的域名部分的 IP 地址查询中导出,或者直接从 IP 地址中导出,如果其作为 URL 的 FQDN/IP 部分出现的话。此外,如后面将要描述的,IP 地址具有若干种属性,其中的每一种属性能够被用作机器学习系统的特征,或者用作用户填充列表的一个元素。因此,在第二种方法中,特征提取部件 130 能够采用 IP 地址的多个子部分来产生附加的特征。

[0096] 如上所述任何特征的组合都能够从各个输入消息 110 中提取。典型地,尽管所有的消息都能被使用,但是消息能够被随机地,自动地,和 / 或者手动地选择来参与到特征提取中。已提取的特征集实际上被应用于一个过滤训练部件 140,诸如机器学习系统或者任何其它的建立和 / 或者训练象垃圾邮件过滤器这样的训练过滤器 150 的系统。

[0097] 现在参考图 2,根据本发明的一个方面示出了一个特征提取系统 200,该系统易于摆脱一个或者多个输入消息 210 的特征的困惑,或者规范一个或者多个输入消息 210 的特征。最后,至少部分地基于标准化的一个或者多个特征来建立一个过滤器。系统 200 包括一个特征提取部件 220,例如,正如所示出的,其要么直接地要么间接地借助于一个消息收件人(图 1)来接收一个输入信息 210。根据用户的优先选择,选择用于特征提取的或者在特征提取中的输入消息能够受系统 200 支配。作为选择,对于特征提取来说,实际上所有的输入消息都可能是有效的。

[0098] 特征提取包括抽取一个或者多个与来自消息 210 的源和 / 或者目的地信息相关联的特征 230(也被称为 FEATURE₁232,FEATURE₂234,和 FEATURE_M236,其中 M 是大于或等于 1 的整数)。源消息可能与表明消息的发送者的元素,服务器域名,以及相关的规定了消息来源的标识信息相关。目的地信息可能与的一个消息的元素相关,该消息表明收件人发送其对消息的响应给谁或者到哪里。能够在消息的首部以及消息体中发现源和目的地信息,对于消息收件人来说要么是可见的要么是不可见的。

[0099] 由于垃圾邮件制作者注意去伪装并且 / 或者迷惑它们通过传统的垃圾邮件过滤器检测的能力,所以系统 200 包括一个特征标准化部件 240,其易于摆脱一个或者多个被提取的特征 230 的困惑,或者至少其中的一部分。该特征标准化部件 240 能够处理和 / 或者分解已提取的特征 230,诸如通过分析已提取的特征 230(例如,FQDN- 查阅数据块的目录和 MX 记录,和 / 或者根据其当前的格式来翻译 FQDN),而且将它们与现存的垃圾邮件制作者列表,非垃圾邮件制作者,以及 / 或者父控制列表的数据库作一个比较。在正如下文中图 4 所讨论的某些情况中,诸如当已提取的特征是一个 URL 时,前缀和 / 或者后缀可能也被删除以便于规范化该特征,并且识别 URL 是否指向垃圾邮件制作者的网站,或者指向一个合法的源。

[0100] 一旦特征被规范化,至少 250 的一个子集能够通过诸如机器学习系统这样的训练系统 260 来使用,以建立和 / 或者更新一个过滤器 270。例如,该过滤器能够被训练用于一个垃圾邮件过滤器。此外,能够以肯定的特征来建立并且 / 或者训练过滤器,诸如表明非垃圾邮件来源(例如,发件人的 From 电子邮件地址,发件人的 IP 地址,嵌入式的电话号码,以及 / 或者 URL)这样的特征,以及 / 或者非垃圾邮件发件人,以及以否定的特征,诸如识别并且与一个垃圾邮件制作者相关的特征。

[0101] 作为选择,特征集能够被用于填充一个新的或者加入到一个现存的垃圾邮件特征列表 280 中。其它的列表也能够被产生以对应于特定的被提取的特征,诸如好的地址的列表,有害地址的列表,好的 URL 的列表,有害的 URL 的列表,好的电话号码的列表,以及有害的电话号码的列表。好的特征列表能够识别非垃圾邮件制作者,过去的合法的发件人,和 / 或者具有较高可能性的非垃圾邮件(例如,90% 的机会不是垃圾邮件来源)的发件人。相反,有害的特征列表可能对应于垃圾邮件制作者,潜在的垃圾邮件制作者,以及 / 或者具有相对高的可能性的垃圾邮件(例如,大约 90% 的垃圾邮件来源)的发件人。

[0102] 现在参考图 3-6,其中根据本发明的若干个方面示出了能够分别从 IP 地址,FQDN,电子邮件地址和 URL 中导出并提取的示例性的特征,以方便检测和阻止垃圾邮件。

[0103] 图 3 描述了根据本发明一个方面的 IP 地址 300 的示例性细分类。在以虚线的十进制格式(每 4 个数据块等于 3 位数,其中每个数据块通过周期来分开,而且其中 3 位数的

每个数据块是在 0 到 255 之间可除尽的任何数字) 表示时, IP 地址 300 是 32 比特长, 并且定位在数据块(例如, 网络数据块)中。这些数据块被分配为诸如等级 A, 等级 B 和等级 C 这样的等级。每个数据块包括一组 IP 地址, 其中每个数据块的 IP 地址的数量根据种类而不同。也就是说, 可能存在或多或少的分配给每个数据块的地址, 这取决于种类(即, A, B 或者 C)。数据块的尺寸通常是 2 的幂次, 并且在同一个数据块中的 IP 地址的集合将分享最初的 k 个二进制数字, 而且不同于最后的 32-k(例如, 32 减去 k) 个二进制数字。因此, 根据每个数据块所分得的最初的 k 个比特, 每个数据块都能够被识别(数据块 ID302)。为了确定与特定 IP 地址 300 相关联的数据块 ID302, 用户能够查阅诸如 arin.net 这样的数据块的目录。此外, 数据块 ID302 能够被提取并且用作一个特征。

[0104] 然而, 在某些情况下, 数据块 ID302 不能被容易地确定, 甚至通过 arin.net, 因为在一个数据块中的 IP 地址的组合能够被分开出售, 并且重复出售多次。在某些情况下, 在数据块 ID302 处, 对于各个 IP 地址来说, 用户或者提取系统能够作出一种或者多种推测。例如, 用户能够提取至少最初的 1 个比特 304, 最初的 2 个比特 306, 最初的 3 个比特 308, 最初的 M 个比特 310(即, M 是大于或者等于 1 的整数) 和 / 或者等于至少最初的 31 个比特 312 作为分开的特征, 对于随后通过一个机器学习系统和 / 或者作为有关特征列表(例如, 好的特征列表, 垃圾邮件特征列表等等) 的元素使用来说。

[0105] 例如, 实际上, IP 地址最初的 1 比特能够被提取并用作一个特征, 来确定是否该 IP 地址指向一个垃圾邮件制作者或者非垃圾邮件制作者。来自其它 IP 地址的从其它消息中提取的最初的 1 比特能够被比较, 以方便确定至少一个数据块 ID。然后, 识别至少一个数据块 ID 能够帮助鉴别该消息是否来自一个垃圾邮件制作者。此外, 共享最初 M 个比特的 IP 地址能够与它们的其它被提取的特征相比较, 以确定该 IP 地址是否来自合法的发件人和 / 或者相应的消息是否是垃圾邮件。

[0106] IP 地址也能够按照体系(314)来排列。也就是说, 一组高位比特可以被定位到一个特定的国家。那个国家能够定位一个子集到 ISP(英特网服务提供商), 以及该 ISP 可以定位一个子集到一家特定的公司。相应地, 不同的级别对于同一个 IP 地址来说是有意义的。例如, 来自一个定位在韩国的 IP 地址能够在确定 IP 地址是否与垃圾邮件制作者相关中使用。如果该 IP 地址是定位到以严格地政策来反对垃圾邮件制作者的 ISP 的数据块的一部分, 则这也可能在确定 IP 地址与一个垃圾邮件制作者不相关的进程中是有用的。因此, 通过使用 IP 地址的最初的 1-31 个比特中的每一个, 结合 IP 地址的至少一个子集的排列体系 314, 一个用户能够自动的在不同的级别得到信息, 而实际上不知道 IP 地址被定位的方式(例如, 不知道数据块 ID)。

[0107] 除了上面讨论的特征之外, 一种稀有的特征 316(例如, 特征的出现不是很普遍的)能够通过运行适当的运算和 / 或者使用比较频率和计数的统计数据来确定, 其中例如在输入消息的抽样中出现的特征。实际上, 不常用的 IP 地址 300 可以是被用于发送电子邮件的拨号上网线路的一个例子, 其通常由垃圾邮件制作者使用。垃圾邮件制作者尝试经常修改它们的身份和 / 或者位置。因此, 一个特征可能经常或者不经常是有用的信息。因此, 稀有的特征 316 能够被用作机器学习系统的一个特征和 / 或者作为至少一个列表(例如, 稀有特征列表)的一部分。

[0108] 图 4 示出了 FQDN400 的示例性特征的细分类, 诸如用于 Example.b.x.com。例如,

FQDN400 能够从一个 HELO 域中提取（例如，发件人的合理的 FQDN），并且典型地包括一个主机名 402，和域名 404。主机名 402 是指一个特定的计算机，其是例子中的“b”。域名 404 是指至少在英特网上的一个机器或者一组机器的名字。在该实例中，“x. com”表示域名 404。FQDN400 体系的细分类由 406 表示。尤其是，B. X. com408（整个 FQDN400）能被部分地剥离到 X. com410（部分 FQDN），然后被剥离到 COM412（部分 FQDN），由此每个 FQDN 部分都能被用作一个特征。

[0109] 某些特征，诸如从信息中接收的特征，首先以 IP 地址的形式存在。因此，将 FQDN400 转换到 IP 地址 300 可能是有用的，该 IP 地址又能够细分为另外的特征（如图 3 所示），因为创建一个新的主机名和域名是相对容易的，但是获得一个新的 IP 地址是相当困难的。

[0110] 不幸的是，域的主人显然能够使不同的机器全部映象到同一个位置。例如，命名为“a. x. com”的机器的主人可能与“b. x. com”的主人是一样的，其可能是“x. com”的同一个主人。因此，垃圾邮件制作者能够容易地误导一个传统的过滤器以确信该消息是来自 FQDN400 “b. x. com”而不是来自域 404 “x. com”，因此实际中允许消息通过垃圾邮件过滤器，域 404 “x. com”已经表明了该消息是垃圾邮件或者很可能是垃圾邮件。因此，在提取消息的源和 / 或者目的地信息时，剥离该地址以简化域名 404 是有用的。作为选择，整个 FQDN400 能够作为一个特征被提取。

[0111] 在某些情况下，附加的来源是有效的，诸如父控制系统。这种资源通常能够为主机名字和 / 或者 URL 分配一种“类型”或者质量评估，诸如色情或者暴力。通过使用这样一种资源，该提取的信息能够进一步通过“类型”来分类。该特征的特征类型 414 连同建立和 / 或者学习与垃圾邮件相关的改进的过滤器一起，能够被用作附加的特征。作为选择，能够产生对应于不同的先前已经识别的特征类型的列表。特征类型 414 可能包括，但是不被限制为，性和色情相关的特征，种族和 / 或者憎恨的语言相关的特征，物理增加特征，收入或者财务解决方案特征，家庭购买力特征等，其一般识别消息的主题。

[0112] 最后，稀有的特征 316 或者特征类型（见上文中的图 3）可能是如上在图 3 中所讨论的另一个特征。例如，从一个消息中提取的诸如来自 FQDN400 “b. x. com”的主机名“B”402 这样的特征，可能是特征类型的一般的例子：色情资料。因此，当该特征从消息中提取并且发现了关于色情资料特征的列表时，可能得出结论即该消息很可能是垃圾邮件，或者对于所有的年龄是不合适 / 不恰当的，或者构成了成人内容（例如，成人电视节目），等等。因此，每个列表可能包括最普通的特定类型的特征。作为选择，对应的 IP 地址通常可能在垃圾邮件消息中被发现，因此指定作为垃圾邮件的公共的特征。此外，特征的通用性和 / 或者稀有性能够被用作一个用于机器学习或者其它基于系统的规则的单独的特征。

[0113] 图 5 示出了电子邮件地址 500 的示例性特征的细分类：a. @b. x. com，其包括 FQDN400 以及少量附加的特征，诸如用户名 502。该 email 地址 500 能够从 From 字段中提取，cc（副本）字段和消息的响应字段，以及来自任何的邮寄到：在消息（例如，邮寄到：链接是一种特定种类的链接，产生到一个特定地址的邮件）体中的链接，以及，如果有效，则来自在 SMTP 协议中使用的 MAIL FROM 命令。电子邮件地址 500 也能被嵌入到消息的文本中。在某些情况下，在响应该消息时，该消息的内容可能指导收件人使用“答复所有人”的功能。在这种情况下，在 cc 字段中的地址和 / 或者至少这些地址中的一部分包括在也将被

答复的“to”字段中（如果超过一个收件人被列出）。因此，这些地址中的每一个能够被提取作为一个或者多个特征，以便于识别和阻止垃圾邮件制作者。

[0114] Email 地址 500 ‘a. @b. x. com’能够被分解为各个要素或者子部分，而且这些要素能够被提取并用作特征。另外，电子邮件地址包括一个用户名 502 和 FQDN504（例如，见图 4 中的 FQDN400），其甚至能够被进一步分解到另外的特征中。出于几种实际的原因，诸如使用，识别和承认，电子邮件地址通常使用 FQDN 而不是 IP 地址被标记。

[0115] 在当前的实例中，‘a. @b. x. com’包括用户名 502 ‘a.’。因此，‘a.’能够被提取作为一个特征。同样，FQDN504 ‘b. x. com’能够从电子邮件地址中提取作为至少一个另外的特征。电子邮件地址 500 的 FQDN504 部分能够通过一个父控制系统，以方便确定特征类型 414，其在上面的图 4 中作了详细的描述。因此，与电子邮件地址 500 的 FQDN 部分相关的特征类型能够被用作另外的特征。

[0116] 另外的电子邮件地址，垃圾邮件制作者通常通过 URL 来联系。图 6 根据本发明的一个方面，描述了一种示例性的 URL600（例如，x. y. com/a. /b/c）连同多个被提取的特征。URL600 能够被嵌入到消息的文本中，和 / 或者作为消息文本的一个图像。例如，垃圾邮件消息可能包括到网站的指针，因此将收件人引到垃圾邮件制作者的网页或者相关的站点。

[0117] URL 可能以与 IP 地址同样的方式来摆脱困惑。最初，诸如 http://, http s://, ftp://, telnet:// 这样的任何的前缀（服务名称）能够在 URL600 摆脱困惑之前被删除。另外，如果“@”符号（例如 % 40 是十六进制的符号）出现在 URL 之中，则在前缀（例如 http://）和“@”符号之间的任何东西可能在规范化该 URL400 之前被删除。在前缀和“@”符号之间插入文本可能是另外一种形式的欺骗，这种欺骗是由垃圾邮件制作者用来迷惑消息收件人被引入的真实的网页位置。

[0118] 例如，http://www. amazon. com@121. 122. 123. 124/info. htm 被送至消息收件人，好像该网页被定位在 www. amazon. com。因此，收件人可以更加倾向于信任该链接，尤其重要的是消息的发送者。相反，真正的网页定位在 121. 122. 123. 124，实际上对应于与垃圾邮件相关的网页。然而，在某些情况下，合法的发件人可以结合鉴权信息，诸如在 URL400 部分的登录名和密码，以方便自动登录。

[0119] 一旦规范化并且摆脱了困惑，URL600 实际上就能够表达为 x. y. com/a/b/c，其中 x. y. com630 是机器（FQDN）的名字，而 a/b/c（例如后缀）是在机器上文件的位置。如果 x. y. com/a/b/c600 识别一个垃圾邮件制作者，则 x. y. com/a/b610 和 x. y. com/a620 很可能识别相同的或者相关的垃圾邮件制作者。因此，URL600 的结束部分或者路径每次都能被剥离一部分，以获得用于机器学习系统或者列表的附加的特征。这就使得对于垃圾邮件制作者来说，建立多种不同的位置就变得更加的困难，这些位置实际上都是以某种没有注意到的模式的方式指向它们。

[0120] 当后缀被剥离时，FQDN630 也能够进一步分析以获得附加的特征，正如先前在图 4 中所讨论的。此外，FQDN630 也能被转换为一个 IP 地址，正如在图 3 中所描述的。因此，各种与 IP 地址相关的特征也能被用作特征。

[0121] 以 IP 地址而不是 FQDN（例如，打点的十进制形式）来编写某些 URL，诸如 nnn. nnn. nnn. nnn/a. /b/c。这些后缀能够从“c”开始逐级逐次被删除，最终的（部分的）URL 能够被用作一个特征（例如，nnn. nnn. nnn. nnn/a/b, nnn. nnn. nnn. nnn/a, nnn. nnn. nnn. nnn 都是

可能的从 URL 中以打点的十进制形式来提取的特征)。接下来,IP 地址(例如,没有后缀和前缀)能被用作一个特征。然后,其能被映射到网络数据块。如果该网络数据块不是可确定的,则可能作出多种推测,使用前 1,2…中的每一个,直到 IP 地址的前 31 个比特为止作为独立的特征(见图 3)。

[0122] 除了打点的十进制格式以外,该 IP 地址能够以双字的格式(例如,在基数 10 中的两个每个 16 比特的二进制字),八进制的格式(例如,基数是 8)以及十六进制的格式(例如,基数是 16)来表达。实际上,垃圾邮件制作者能够混乱一个 IP 地址,一个 URL,一个 MAILTO 的链接,以及 / 或者例如,通过使用 % nn 符号(其中 nn 是一对十六进制数字)来编码域名部分的一个 FQDN。

[0123] 某些 URL 可能包括可以用于干扰或者欺骗用户的重定向器。重定向器是在 URL 的 IP 地址中跟随一个“?”的参数或者参数集,该 URL 指示一个浏览器重新将其定向到另一个网页。例如,该 URL 可以以“www.intendedpage.com? www.actualpage.com”出现,其中浏览器实际上指向“www.actualpage.com”,而且加载该页而不是预料中的“www.intendedpage.com”页。因此,包括在 URL 中的参数也可能考虑被提取作为特征。

[0124] 现在将通过一连串的动作来描述根据本发明的各种方法。应当理解本发明没有被动作的顺序所限制,从这里所描述和示出的可知,根据本发明的一些顺序可能以不同的顺序出现,或者与其它动作并行出现。例如,本领域的普通技术人员将理解一种方法可选择性地被表示为一系列相关联的状态或者事件,诸如在正式的图中。此外,不是所有的示例性的动作都可能需要执行根据本发明的方法。

[0125] 参考图 7,示出了示例性过程 700 的流程图,该过程便于根据本发明的一个方面来训练过滤器。过程 700 可能在 710 处接收一个消息(例如,至少一个消息)开始。该消息能够通过一个收件人来接收,例如,其中一个现存的过滤器(例如,一个垃圾邮件过滤器)能够分类该消息可能是垃圾邮件或者不可能是垃圾邮件,至少部分地基于一组提前通过过滤器学习的标准。该消息能够被分析以便在 720 处从中提取一个或者多个特征。在 725(在下文的图 11 中)处进一步详细地描述了特征的提取。特征的实例包括定位在接收字段,答复字段,cc 字段,邮寄到(mailto)字段,MAIL FROM SMTP 命令,HELO 字段,嵌入到文本中的或者作为一个图像的 URL 地址,和 / 或者非长途免费的电话号码(例如,映射到地理上的区域的电话区号),以及消息体内容部分中的信息(例如,发件人的 IP 地址)。

[0126] 所提取的特征(和 / 或者规范化)以及消息的分类(例如,垃圾邮件或者非垃圾邮件)能够在 730 处被加到一组训练数据中。对于所有其它的输入消息来说,在 740,上述所有的(例如,710,720 和 730)实际上都能被重复,直到它们能够被相应地处理为止。在 750,所出现的特征可能是有用的,或者最有用的特征能够从训练集中选择。这种选择的特征能够被用于训练一个过滤器,诸如机器学习过滤器,例如,在 760 处借助于机器学习算法。

[0127] 正如图 8 中通过一个示例性的方法 800 所描述的,一旦被训练,一个机器学习过滤器就能够被用于方便垃圾邮件的检测。该方法 800 以在 810 接收一个消息开始。在 820 处,一个或多个特征从该消息中被提取,正如在下文图 11 中所描述的。例如,在 830,被提取的特征通过一个过滤器,该过滤器通过一个机器学习系统来训练。接下来,从机器学习系统中获得一个诸如“垃圾邮件”、“非垃圾邮件”、或者消息可能是垃圾邮件的概率这样的判定。一旦获得有关消息内容的判定,就能够获得合适的动作。动作的类型包括,但不被限制为,检

测消息,将消息移动到一个特殊的文件夹中,隔离该消息,以及允许收件人访问该消息。

[0128] 作为选择,基于动作的列表能够以从消息中提取的特征来执行。参考图 9,示出了示例性过程 900 的一个流程图,用于建立并且填充列表,至少部分地基于所提取的特征和它们在所接收信息中的出现,这些信息被分类作为垃圾邮件或者非垃圾邮件(或者可能是垃圾邮件)。过程 900 通过接收一个消息开始。接下来,在 920 处提取一些感兴趣的特征,诸如发送 IP 地址的消息。例如,在接收了消息之后的某个时间,通过现存的过滤器,该消息能够被分为垃圾邮件或者非垃圾邮件。在 930 处,根据消息的分类(例如,垃圾邮件或者非垃圾邮件)能够增加特征的计数。这会在 940 处重复直到实际上所有的消息都被处理为止(例如,在 910,920 和 930 处)。此后,在 950 处,能够创建一个特征列表。例如,能够为 IP 地址创建一个特征列表,该 IP 地址 90% 是好的(例如,输入消息的 90% 是非垃圾邮件)。同样,另一个用于 90% 都是有害(垃圾邮件)的发件人 IP 地址的列表也能被创建。用于其它特征的列表也能够以同样的方式创建。

[0129] 应当理解这些列表可能是动态的。也就是说,当处理另外的新的消息组时,它们可能被更新。因此,对发件人的 IP 地址来说,首先发现好的列表是可能的;然后,在之后的某个时间,发现一个不好的列表,就象对于某些垃圾邮件制作者来说,实际上首先发送好的邮件(例如,获得“可信的”过滤器以及收件人),然后才开始发送垃圾邮件是很普遍的。

[0130] 可能以不同的方式来利用这些列表。例如,它们可能被用于产生通过机器学习系统使用的训练集,以便训练过滤器。这通过图 10 中描述的示例性的过程 1000 可以得到。根据图 10,过程 1000 能通过在 1010 上接收一个消息开始。该消息能被分类,例如,分为垃圾邮件或非垃圾邮件。在 1020 上,包括但不限于发件人的 IP 地址的特征能够从该消息中被提取。在 1030 上,被提取的特征和消息的分类被加到一个训练集上,其随后被用于训练机器学习系统。

[0131] 接下来,在 1040 上,与发件人的 IP 地址在其上的特殊列表相符合的一个具体特征被包括在训练集中。例如,如果发件人的 IP 地址在“90%好”列表上,则被加到训练集的特征将是“90%好列表”。在 1050 上,前述的步骤(例如 1010,1020,1030,和 1040)能被重复来随后处理所有的输入消息。对于过滤器训练的目的来说,因为这些特征可能比其它的特征更有用。最有用的特征部分地基于 1060 上的用户优先权被选择,并且被用于通过使用机器学习算法来训练诸如垃圾邮件过滤器这样的过滤器。

[0132] 此外,例如 IP 地址的动态列表能被构造以用于与测试消息,新消息,和 / 或可疑的消息相比较。然而,在这种情况下,IP 地址本身不是特征。而 IP 地址的属性是特征。作为选择,这些列表能以其它方式被利用。尤其是,例如,可疑 IP 地址的列表能被用来将发件人标记为有害的,并且相应地以可疑的方式来处理他们的消息。

[0133] 现在转到图 11,示出了与上述图 7-10 分别描述的过程 700,800,900 和 1000 相结合,从消息中提取特征的示范性方法 1100 的流程图。方法 1100 能够在接收的 IP 地址中开始,其中的一部分被提取并且在 1110 上被规范化。而且在 1110 上,为了从接收的 IP 地址中提取附加的特征,该 IP 地址可能经历比特方式处理(例如,如图 3 中讨论的,前 1 个比特,前 2 个比特,直到前 31 个比特为止)。此外,发件人的宣称的主机名也可能在 1110 上被提取。规范化的被接收的 IP 地址和发件人主机名特征现在能被用作计算机学习系统或相关的训练系统的特征。

[0134] 随意地,在 1120 上,“From”行的内容能被提取和 / 或规范化,并且随后被用作特征。在 1130 上,“MAIL FROM SMTP”命令的内容同样能被提取和 / 或被规范化用作特征。

[0135] 然后方法 1100 能继续寻找其它的可能被包括在消息中的特征。例如,它可以随意地提取和规范化(如果必要)1140 上的答复字段中的内容。在 1150 上,cc 字段的内容能随意地被提取或被规范化来用作至少一个特征。在 1160 上,非长途的免费电话号码从消息体中能被随意提取并且也被指定为特征。非电话的号码对于识别垃圾邮件制作者来说可能是有用,因为区号或电话号码的前三位数字能被用来映射出垃圾邮件制作者的位置。如果不止一个非长途的免费电话号码存在于消息中,那么每个号码都能被提取并且在 1160 上用作分离的特征。

[0136] 同样地,一个或多个 URL 和 / 或 MAILTO 链接或其中的部分,能分别在 1170 和 1180 上被随意地提取和 / 或规范化。尤其是,URL 可能经历路径剥离(例如 URL 的文件名部分),其中附加在 URL 的 FQDN 末端的一个或多个后缀可能被剥离。这就可能依赖于路径中的后缀的数字,导致一个或多个部分 URL。根据本发明,每个部分 URL 能被用作分离的特征。

[0137] 方法 1100 能继续扫描消息体来查找其它的电子邮件地址,也查找关键字和 / 或短语,其在垃圾邮件消息中比在合法消息更可能被找到,反之亦然。每个字或短语能被提取并且用作计算机学习系统的特征或列表单元的特征,或两者。

[0138] 如前面所讨论,在 Internet 上被发送的消息可能是从服务器到服务器发送,少到只包括两台服务器。与消息有联系的服务器的数量会由于防火墙和相关的网络结构的出现而增加。当消息从服务器到服务器被传送时,各个服务器预先考虑其从字段中接收的 IP 地址。每个服务器也具有修改任何容易考虑的接收地址的能力。不幸的是,垃圾邮件制作者能够利用这种能力的优点,而且能够进入在接收字段中的伪装的地址,以区分它们的位置和 / 或者身份,并且误导收件人有关消息的来源。

[0139] 图 12 示出了一个用于在输入消息的接收线路中区分合法的和伪装的(例如,垃圾邮件制作者)预先考虑的服务器 IP 地址的示范性过程 1200 的流程图。以它们被加入的顺序(例如,第一个是最近被加入的)能够检查该预先考虑的接收地址。因此,用户能够通过发送服务器 IP 地址的链接来追溯,以在 1210 确定最后确信的服务器 IP 地址。在 1220 处,最后确认的服务器 IP 地址(完全在体系结构之外的那个)能够被提取作为将被机器学习系统使用的特征。任何其它的在最后确信的 IP 地址之后的 IP 地址可能是有疑问的,不可靠的,而且可能被忽略,但是能够与好的(大部分)IP 地址的列表和(大部分)不好的 IP 地址的列表相互比较。

[0140] 在 1230 处,发件人合理的 FQDN 也能够被提取以便于确定发件人是否是合法的或者是一个垃圾邮件制作者。尤其是,合法的 FQDN 能够通过域名剥离而被细分类,以产生一个或者多个部分 FQDN。例如,想像合法的 FQDN 是 a. b. c. x. com。这个合法的 FQDN 将以下面的方式被剥离以产生 :b. c. x. com → c. x. com → x. com → com。因此,每个 FQDN 字段部分以及整个 FQDN 能够被用作一个独立的特征,以帮助确定伪装的和合法的发件人。

[0141] 本发明也可以使用父控制系统。父控制系统能够至少部分地基于消息的内容,将一个消息分为不适合的,并且给出为什么不适合的原因。例如,一个 URL 可以被嵌入到一个消息中作为可点击的链接(要么基于文本要么基于图像),或者作为消息体中的文本。该父控制系统能够将嵌入的 URL 和一个或者多个其所存储的好和 / 或者有害的 URL 列表相比

较,以确定该消息的正确分类,或者利用其它的用于父控制分类的技术。然后,该分类能够被用作一个附加的特征,要么在机器学习系统中要么在一个特征列表中,或者在二者中。

[0142] 在图 13 中,示出了一个将至少父控制系统的一个方面结合到本发明中的示范性过程 1300 的流程图。在 1310 接收了一组消息之后,该消息能够被扫描用于 URL,邮件发送到的链接,或者类似于邮件发送到的链接的其它文本,一个 URL,或者在 1320 中的 URL 的一部分。如果该消息没有出现来获得 1330 处的任何的上述内容,则过程 1300 返回到 1310。然而,如果该消息没有表明这些,则至少被检测符号的一部分能够通过至少一个在 1340 处的父控制系统。

[0143] 在 1350 处,通过查阅一个或者多个 URL 数据库,该父控制系统能够分类该邮寄到的链接,URL 或者其一部分,URL 业务的名字,URL 路径,以及 FQDN(例如,诸如 URL 电子邮件地址等这样的 FQDN 部分)。例如,该消息可以被分为包括至少一个色情作品,逃避债务,赌博,以及其它类似的内容。这种分类能够被提取作为在 1360 中附加的特征。由于垃圾邮件消息的主题包括这些材料,所以合并的父控制系统在获得附加特征中可能是有用的,其中机器学习系统能够被用于训练并建立改进的过滤器。其它的分类也存在,包括但不被限制为这些,其中这种分类也可能被用作特征。垃圾邮件消息可能或者不可能包括涉及这种材料类型的主题,但是一个用户仍然可以想要这种类型的消息。

[0144] 实际上,不同的分类能够表明不同的垃圾邮件制作者的级别。例如,分类为憎恨语言的消息实际上可能表示没有垃圾邮件的等级(例如,因为其很可能不是垃圾邮件)。相反地,分类作为性内容 / 材料的消息可能反映一个相对高的垃圾邮件的级别(例如,大约 90% 的该消息是垃圾邮件的确认度)。机器学习系统能够建立一个说明垃圾邮件级别的过滤器。因此,过滤器能够被定制并且被个性化以满足用户的优先选择。

[0145] 正如已经讨论的,无数的特征能够从一个消息中被提取,并且用作由机器学习系统使用的训练数据,或者作为识别好坏特征列表的元素。特征的质量,除了特征本身之外,在检测和阻止垃圾邮件中可能是有用的。例如,想像一个特征是发件人的电子邮件地址。该电子邮件地址可能被用作一个特征,并且电子邮件地址在新的输入消息中出现的频率可能被用作另一个特征。

[0146] 图 14 描述了一个用于提取这种类型的特征(例如,与通用的或者稀有的提取特征相关的)的示范性过程 1400 的流程图。垃圾邮件制作者通常尽力快速去改变它们的位置,因此,很可能大多数用户从先前未知的地址发送邮件,或者以指示先前未知的机器的 URL 来发送邮件。因此,对于被提取的每一个特征类型来说(例如,接收的 IP 地址,URL,电子邮件地址,域名等等),假设用于每种类型的特征列表被保留,则可能跟踪特殊特征的出现率,频率或者数量。

[0147] 过程 1400 能够以一个或者多个特征从输入消息,和 / 或者在 1410 规范化一个特征开始。然后,该特征能够与一个或者多个特征列表相比较,这些特征先前已经被提取或者在 1420 中的多个先前的消息中已经被观察到。该过程 1400 能够确定当前的特征是否是通用的。一个特征的通用性能够通过已计算的近期出现的特征的频率,以及 / 或者先前的输入消息来确定。如果该消息在 1430 不是通用的或者不是足够通用的(例如,未能满足通用性的阈值),则在 1440,其稀有的特征能够被用作一个附加的特征。同样,该特征的通用性在 1450 也能被用作一个特征。

[0148] 根据上面所描述的本发明，下面的伪代码可以用于实施本发明的至少一个方面。所有的大写子母表明了不同的名称。应当注意，在伪代码的末端定义了两个函数，add-machine-features 和 add-ip-features。象“PREFIX-machine-MACHINE”这样的符号用于表示由 PREFIX 变量结合单词 machine 结合 MACHINE 变量组成的一个字符串。最后，函数 add-to-feature-list 将特征写入到与当前消息相关的特征列表中。

[0149] 示例性的伪代码如下所示：

```
[0150]      #for a given message, extract all the features
[0151]      IPADDRESS:=the last external IP address in the received from list ;
[0152]      Add-ipfeatures(received, IPADDRESS) ;
[0153]      SENDER-S-ALLEGED-FQDN:=FQDN in the last extemal IP
[0154]      Address in the recerved-from list ;
[0155]      Add-machine-features(senderfqdn, SENDER-S-ALLEGED-FQDN) ;
[0156]      For each 电子邮件 address type TYPE in (from, CC, to, reply-to,
[0157]      embedded-mailto-link, embedded-address, and SMTP MAIL FROM)
[0158]      {
[0159]          for each address ADDRESS of type TYPE in the message{
[0160]              deobfuscate ADDRESS if necessary ;
[0161]              add-to-feature-list TYPE-ADDRESS ;
[0162]              if ADDRESS is of the form NAME@MACHINE then
[0163]              {
[0164]                  add-machine-features(TYPE, MACHINE) ;
[0165]              }
[0166]              else
[0167]                  {#ADDRESS is of form NAME@IPADDRESS
[0168]                  add-ip-features(TYPE, IPADDRESS) ;
[0169]                  }
[0170]              }
[0171]          }
[0172]          for each url type TYPE in(clickable-links, text-based-links,
[0173]          embedded-image-links)
[0174]          {
[0175]              for each URL in the message of type TYPE
[0176]              {deobfuscate URL ;
[0177]              add-to-feature-fist TYPE-URL ;
[0178]              set PARENTALCLASS:=parental control system class of URL ;
[0179]              add-to-feature-fist TYPE-class-PARENTCLASS ;
[0180]              while URL has a location suffix
[0181]              {
[0182]                  remove location suffix from URL, i. e. x. y/a/b/c → ;x. y/a/b → x. y/a ;
```

```
x. y/a ;  
[0183]      #ALL suffixes have been removed ;URL is now either machine name or  
IP  
[0184]      address  
[0185]      if URL is machine name  
[0186]      {  
[0187]          add-machine-features(TYPE, URL) ;  
[0188]      }  
[0189]      else  
[0190]          {add-ip-features(TYPE, URL) ;  
[0191]      }  
[0192]      }  
[0193]      }  
[0194]      function add-machine-features(PREFIX, MACHINE)  
[0195]      {  
[0196]          add-ip-features(PREFIX-ip, nslookup(MACHINE) ;  
[0197]          while MACHINE not equal ""  
[0198]          {  
[0199]              add-to-feature-list PREFIX-machine-MACHINE ;  
[0200]              remove beginning from MACHINE#(i. e. a. x. com →  
[0201]                  x. com, or x. com → com) ;  
[0202]          }  
[0203]      }  
[0204]      fuction add-ip-features(PREFIX, IPADDRESS)  
[0205]      {  
[0206]          add-ip-feature-list PREFIX-ipaddress-IPADDRESS ;  
[0207]          find netblock NETBLOCK of IPADDRESS ;  
[0208]          Md-to-feature-list PREFIX-netblock-NETBLOCK ;  
[0209]          for N=1to31{  
[0210]              MASKED=first N bits of IPADDRESS ;  
[0211]              Add-to-feature-list PREFIX-masked-N-MASKED ;  
[0212]          }  
[0213]      }
```

[0214] 为了提供本发明各个方面的补充的背景,图 15 和下面的讨论想要为适宜的操作环境 1510 提供一个简短的全面的描述,其中可能实现了本发明的各个方面。尽管在计算机可执行的诸如程序模块这样的指令的通常的环境下描述了本发明,但是本领域的普通技术人员承认本发明也能够结合其它的程序模块,和 / 或者以软件和硬件的组合来执行。

[0215] 然而,一般来说,程序模块包括例行程序,程序,目标,部件,数据结构等,它们能够执行特定的任务或者执行特定的数据类型。操作环境 1510 仅仅是适宜的操作环境的一个

实例，并没有试图给出任何有关本发明的使用或者功能范围的限制。其它熟知的适合于与本发明一起使用的固定计算机系统，环境，和 / 或者配置包括但不限于个人计算机，手持或者膝上型设备，多处理器系统，基于系统的微处理器，可编程的用户电子，网络 PC，小型计算机，大型计算机，包括上述系统或者设备的分布式计算环境等。

[0216] 参考图 15，一个用于执行本发明各个方面的示范性的环境 1510 包括一个计算机 1512。该计算机 1512 包括一个处理单元 1514，系统存储器 1516，和系统总线 1518。该系统总线 1518 耦合一个系统部件，该系统部件包括但不限于用于处理单元 1514 的系统存储器 1516。处理单元 1514 可能是各种任何可用的处理器。双微处理器和其他的多处理器结构也可能用作处理单元 1514。

[0217] 系统总线 1518 可能是若干种总线结构类型中的一种，其包括存储器总线或者存储器控制器，外围总线或者外部总线，和 / 或者使用任何可用总线结构的本地总线，任何可用总线结构包括但不仅限制于 11 位总线，工业标准结构 (ISA)，微信道结构 (MSA)，扩展的 ISA (EISA)，智能设备电子 (IDE)，VESA 本地总线 (VLB)，外围部件互连 (PCI)，通用串行总线 (USB)，增强的图形端口 (AGP)，PC 机内存卡国际协会总线 (PCMCIA)，以及小型计算机系统接口 (SCSI)。

[0218] 该系统存储器 1516 包括易失的存储器 1520 和非易失的存储器 1522。基本输入 / 输出系统 (BIOS)，包括在计算机 1512 的范围内的组成部分之间传送信息的基本的例行程序，诸如在启动期间，被存储在非易失的存储器 1522 中。为了举例说明，同时不作为限制，非易失的存储器 1522 可能包括只读存储器 (ROM)，可编程 ROM (PROM)，可擦可编程只读存储器 (EPROM)，电可擦除可编程只读存储器 (EEPROM)，或者闪存。易失的存储器 1520 包括随机访问存储器 (RAM)，其作为一个外部缓存。为了举例说明，同时不作为限制，RAM 可以多种形式得到，诸如同步 RAM (SRAM)，动态 RAM (DRAM)，同步 DRAM (SDRAM)，双数据速率 SDRAM (DDR SDRAM)，增强型 SDRAM (ESDRAM)。同步链接 DRAM (SLDRAM)，以及直接随机存储器总线 RAM (DRRAM)。

[0219] 计算机 1512 也包括可移动的 / 不可移动的，易失的 / 非易失的计算机存储介质。图 15 说明了例如一个磁盘存储器 1524。该磁盘存储器 1524 包括但是不限制于像磁盘设备这样的设备，例如软盘，硬盘，磁带驱动器，Jaz 驱动器，邮政分区驱动器，LS-100 驱动器，闪存卡，或者存储棒。另外，磁盘存储器 1524 可能单独包括存储介质，或者与其它存储介质相结合，其它存储介质包括但不仅限制于诸如紧凑型磁盘 ROM 设备 (CD-ROM)，CD 可记录设备 (CD-R 驱动器)，CD 重写驱动器 (CD-RW 驱动器)，或者数字通用磁盘 ROM 驱动器 (DVD_ROM) 这样的光盘驱动器。为了方便磁盘存储设备 1524 和系统总线 1518 的连接，可移动的或者不可移动的接口典型地被用于诸如接口 1526 这样的接口。

[0220] 应当理解，图 15 描述了软件，该软件起到在用户和适当的操作环境 1510 中所描述的基本计算机设备之间的中间物的作用。这种软件包括一个操作系统 1528。操作系统 1528，其能被存储在磁盘存储器 1524 上，用于控制和定位计算机系统 1512 的资源。系统应用程序 1530 借助于操作系统 1528 通过程序模块 1532 和程序数据 1534 来利用管理和资源，程序数据 1534 被存储在系统存储器 1516 或者磁盘存储器 1524 上。应当理解，能够以各种操作系统或者操作系统的组合来实施本发明。

[0221] 一个用户通过输入设备 1536 将命令或者信息键入到计算机 1512 中。输入设备

1536 包括,但不仅限制于诸如鼠标,跟踪球,唱针,触模板,键盘,麦克风,操纵杆,游戏垫,圆盘式卫星电视天线,扫描仪,TV 调谐卡,数字相机,数字摄像机,网络摄像机等这样的点设备。这些或者其它的输入设备通过系统总线 1518 经由接口部分 1538 连接到处理单元 1514。接口部分 1538 包括例如串行端口,并行端口,游戏端口和通用串行总线 (USB)。输出设备 1540 使用某些同种类型的端口作为输入设备 1536。因此,例如 USB 端口可以被用于提供输入到计算机 1512,而且从计算机 1512 输出信息到一个输出设备 1540。输出适配器 1542 被提供以说明存在一些输出设备 1540 像监视器,扬声器,以及在要求特定适配器的其它输出设备 1540 中的打印机。该输出适配器 1542 包括,通过说明但不是限制,视频和声音卡,该卡提供一种在输出设备 1540 和系统总线 1518 之间连接的手段。应当注意,其它的设备和 / 或者系统提供诸如远程计算机 1544 这样的输入和输出性能。

[0222] 计算机 1512 能够在一个网络环境中通过使用与一个或多个诸如远程计算机 1544 这样的远程计算机的逻辑连接进行操作。远程计算机 1544 可能是个人计算机,服务器,网络,工作站,基于应用的微处理器,对等设备或其它通用网络节点等等,典型地包括多个或全部的所述的与计算机 1512 相关的组成部分。为了简洁,关于远程计算机 1544 只举例说明一个存储设备 1546。远程计算机通过网络接口 1548 被逻辑地连接到计算机 1512 上,然后经由通信连接 1550 物理地连接。网络接口 1548 包括诸如局域网 (LAN) 和广域网 (WAN) 这样的通信网络。LAN 技术包括光纤分布式数据接口 (FDDI),铜分布式数据接口 (CDDI),以太网 / IEEE1102.3,令牌环 / IEEE1102.5 等等。WAN 技术包括但是不被限于,点对点链接,电路交换网络,像 ISDN 以及在其上的变体,分组交换网络和用户数字线 (DSL)。

[0223] 通信连接 1550 指用于将网络接口 1548 连接到总线 1518 的硬件或软件。尽管为了在计算机 1512 的内部明确地说明而示出了通信连接 1510,但是它也可能是在计算机 1512 的外部。连接到网络接口 1548 的必要的硬件 / 软件包括,仅为示范目的,内部和外部的技术,诸如调制解调器,包括常规的电话类调制解调器,电缆调制解调器和 DSL 调制解调器,ISDN 适配器和以太网卡。

[0224] 上面的描述包括了本发明的实例。不可能描述每一种想得到的部件或者方法的组合,当然,为了描述本发明的目的,本领域的普通技术人员承认本发明的许多进一步的组合和置换是可能的。相应地,本发明意在包含所有落入到所附权利要求的精神和范围之内的改变,修改和变型。此外,为了扩展在详细的说明书或者权利要求中所使用的术语“包括”,当术语“包含”在权利要求中被用作一个过渡单词被解释时,该术语意在以类似于术语“包含”的方式被包含在内。

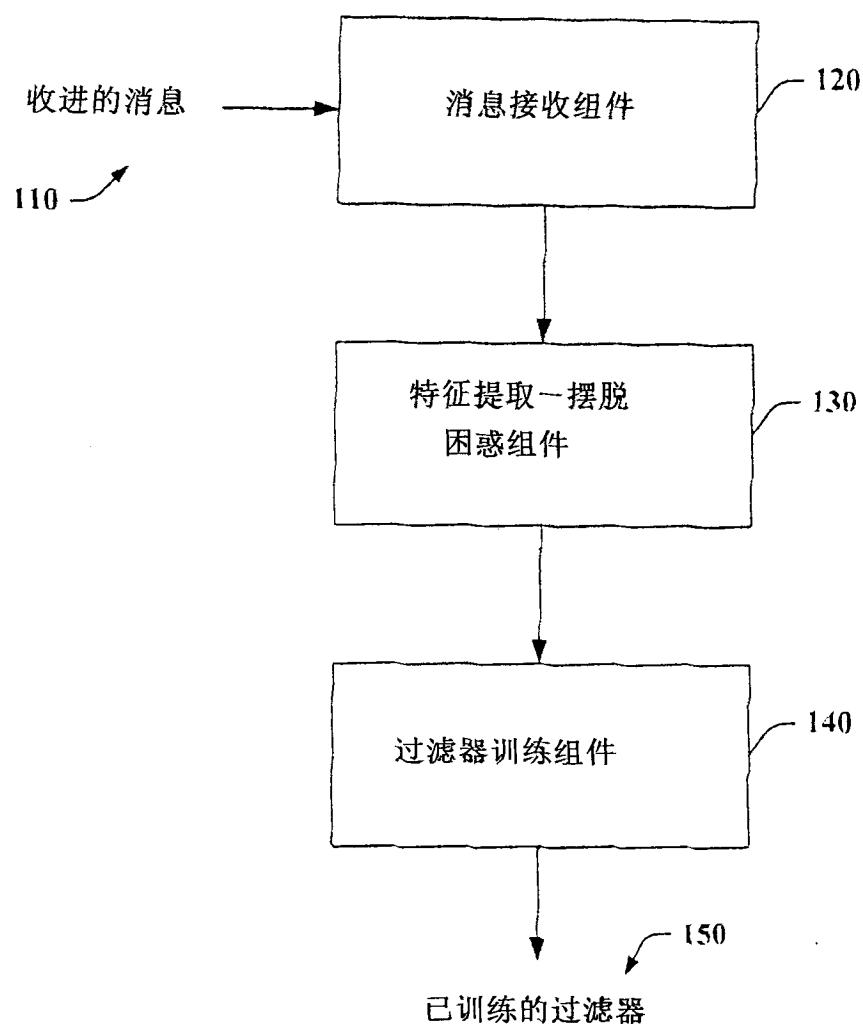


图 1

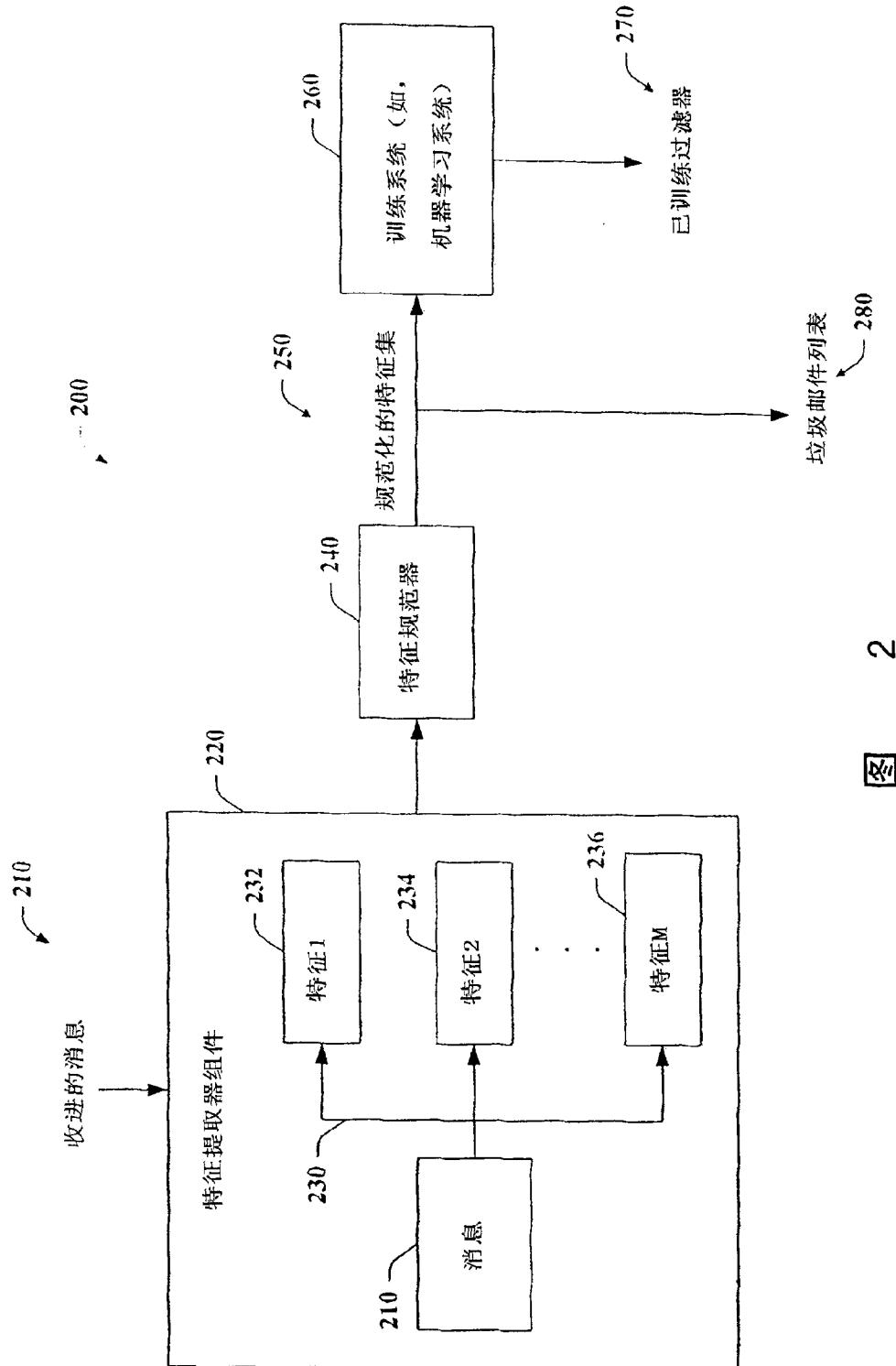


图 2

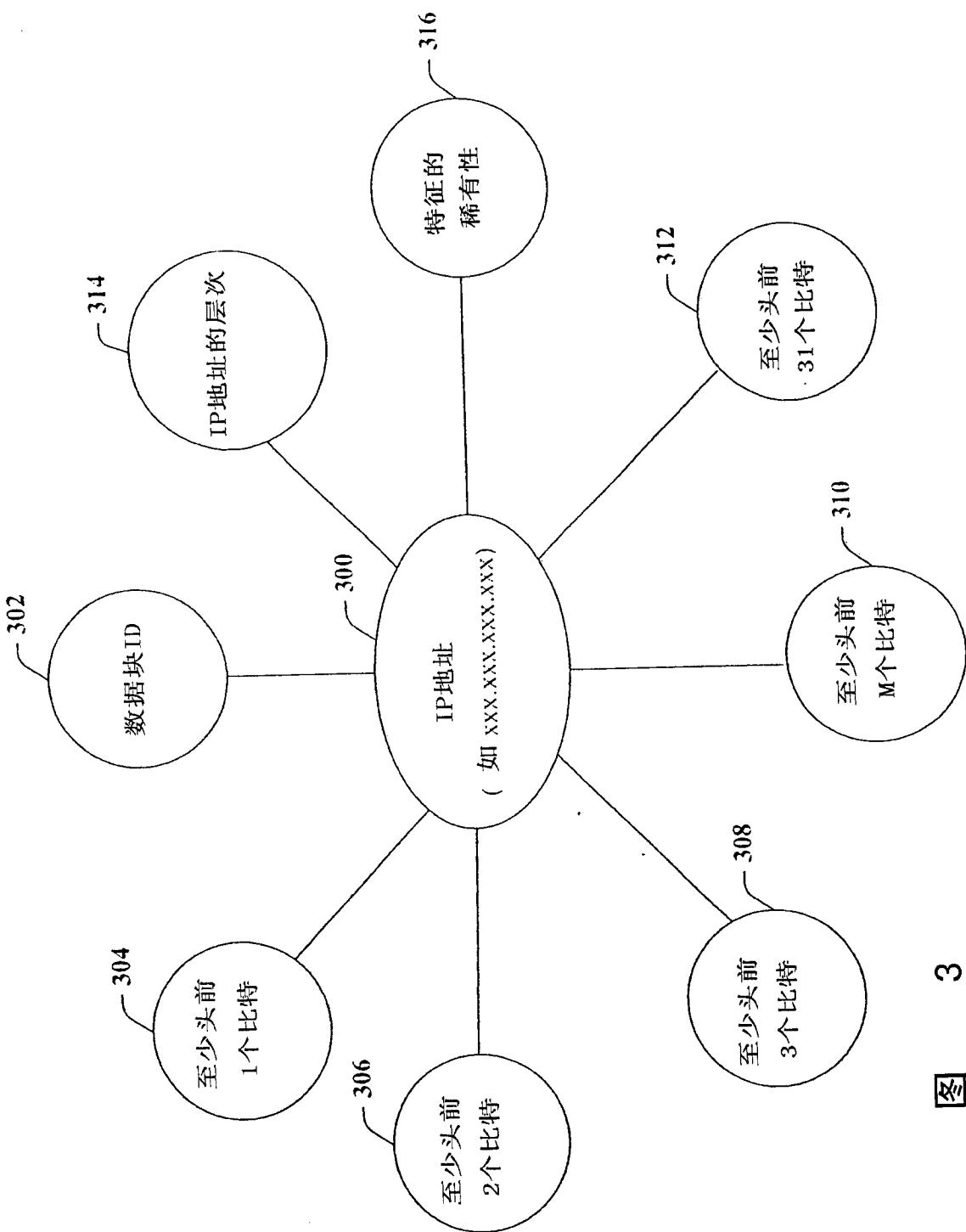
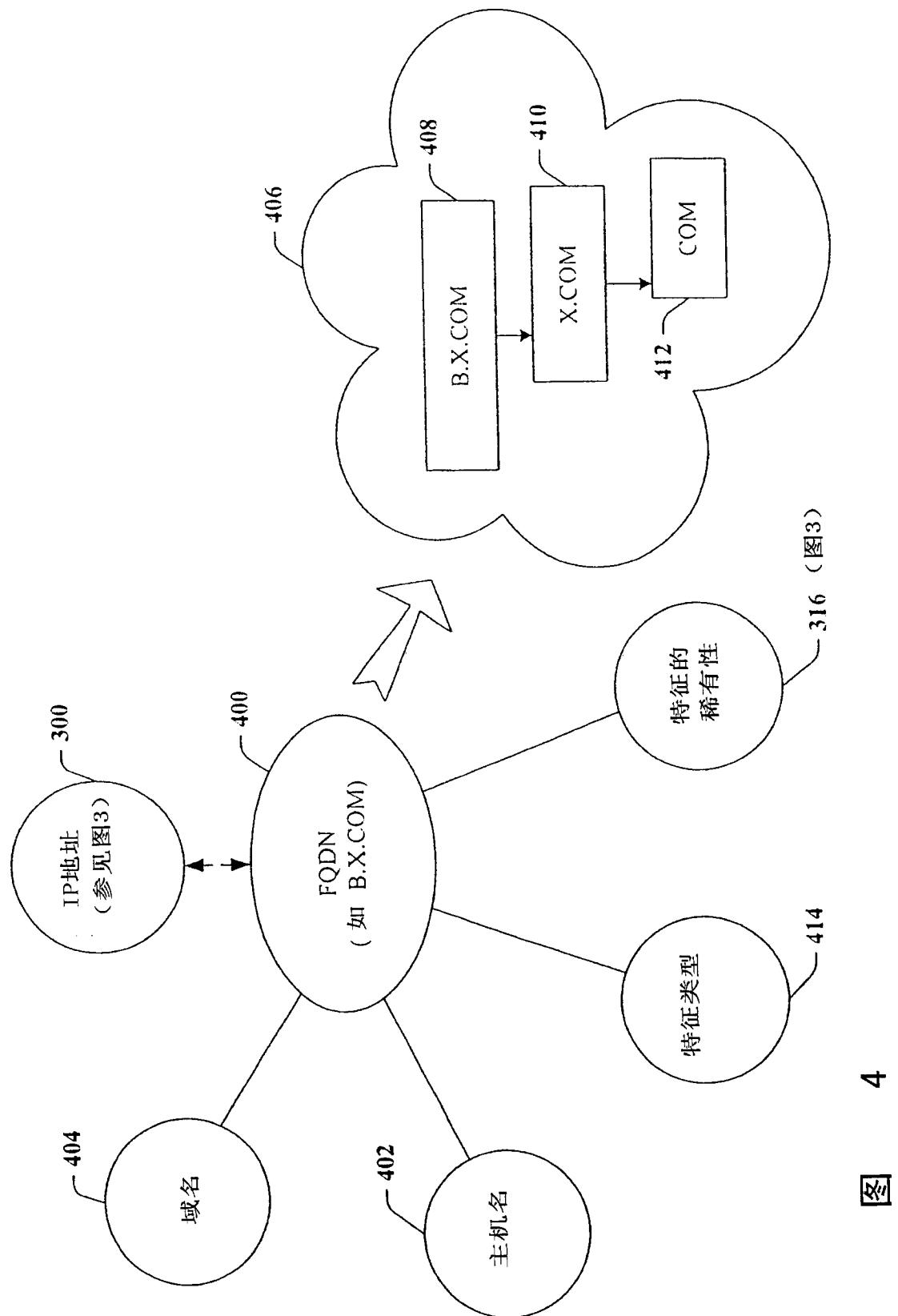
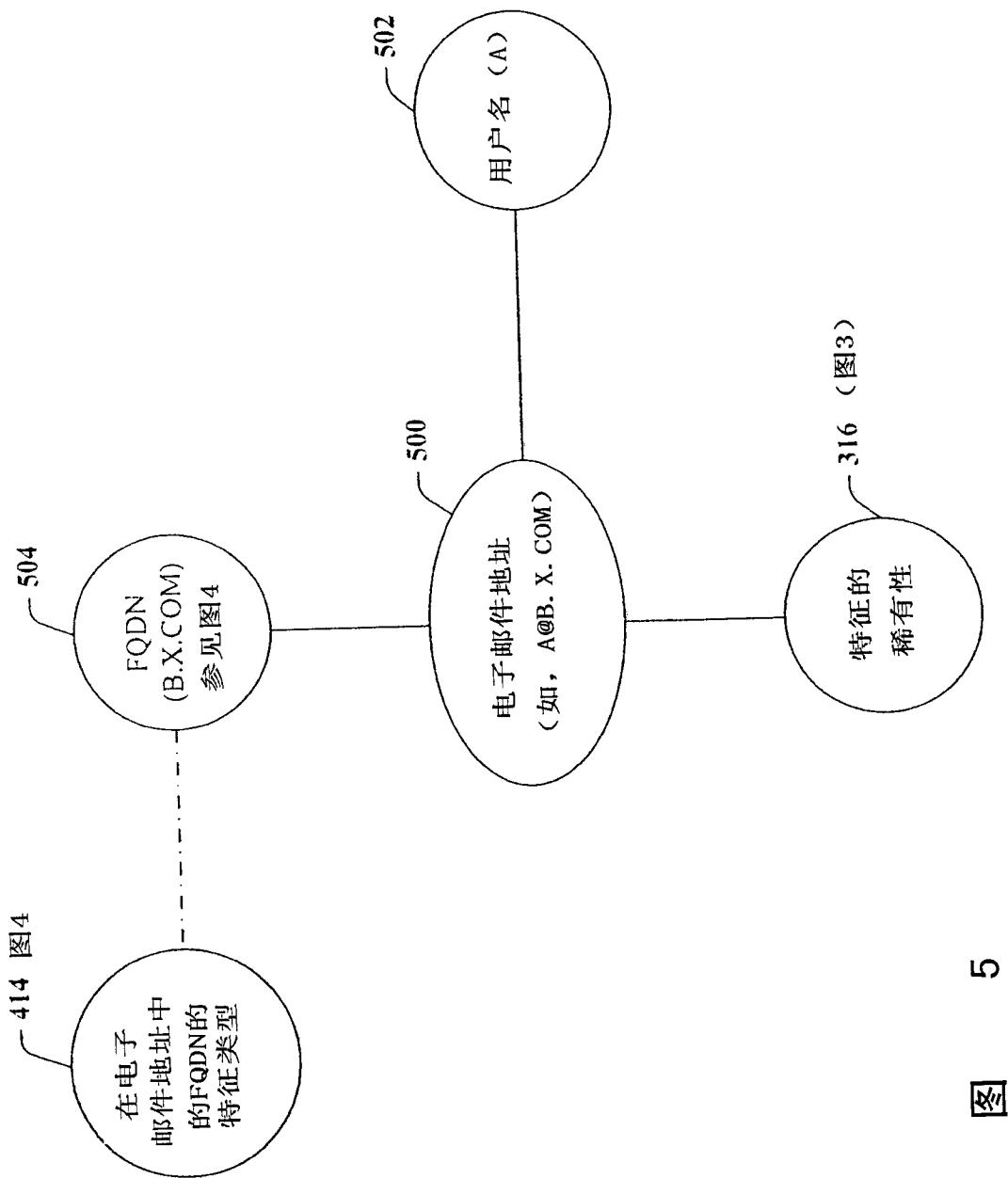


图 3



4

图



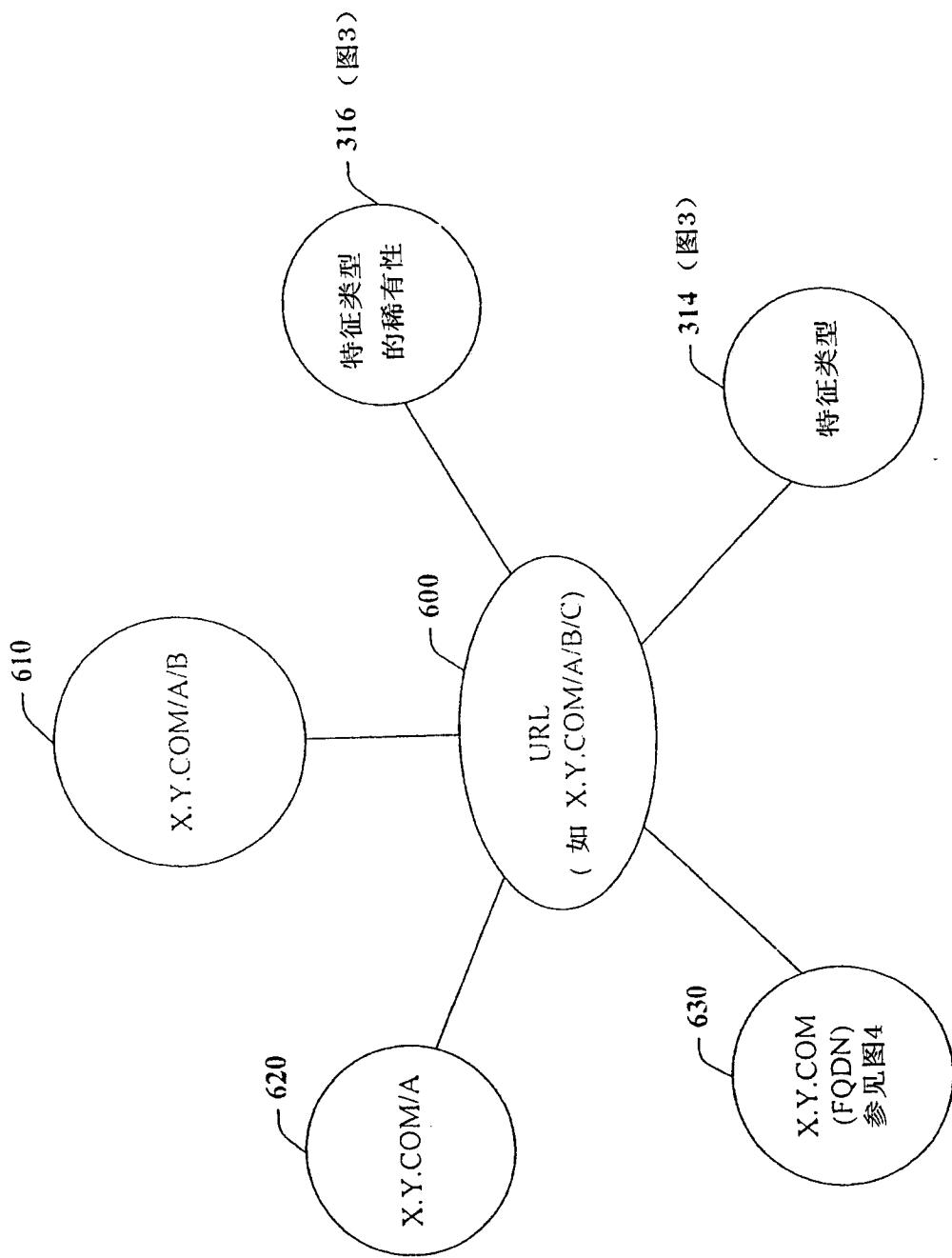


图 6

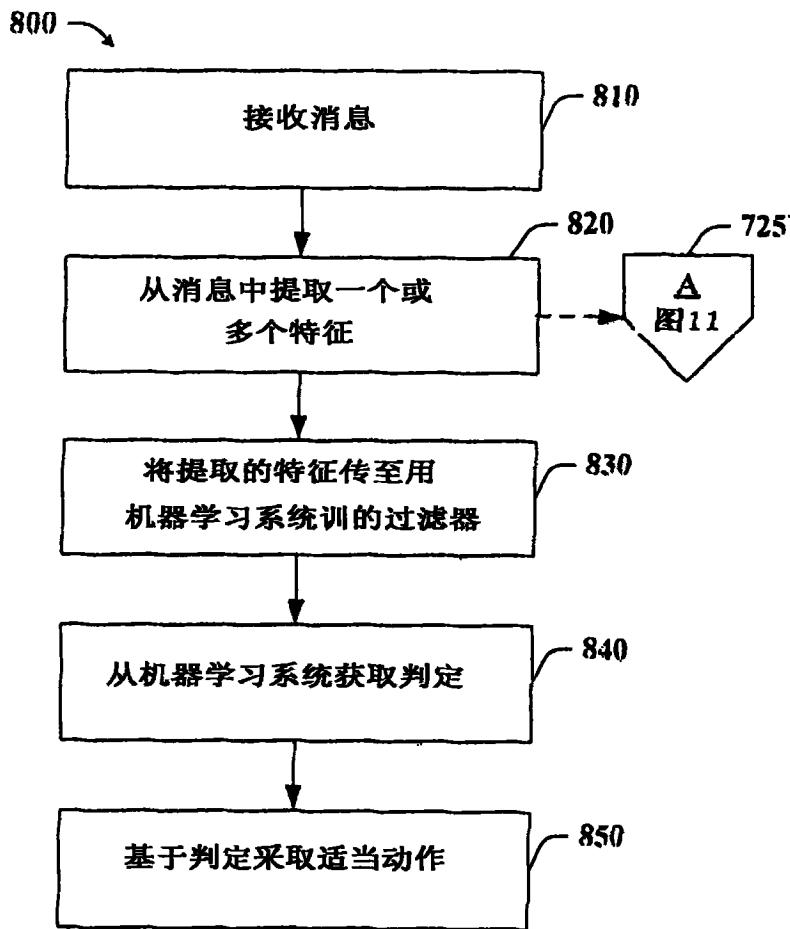


图 8

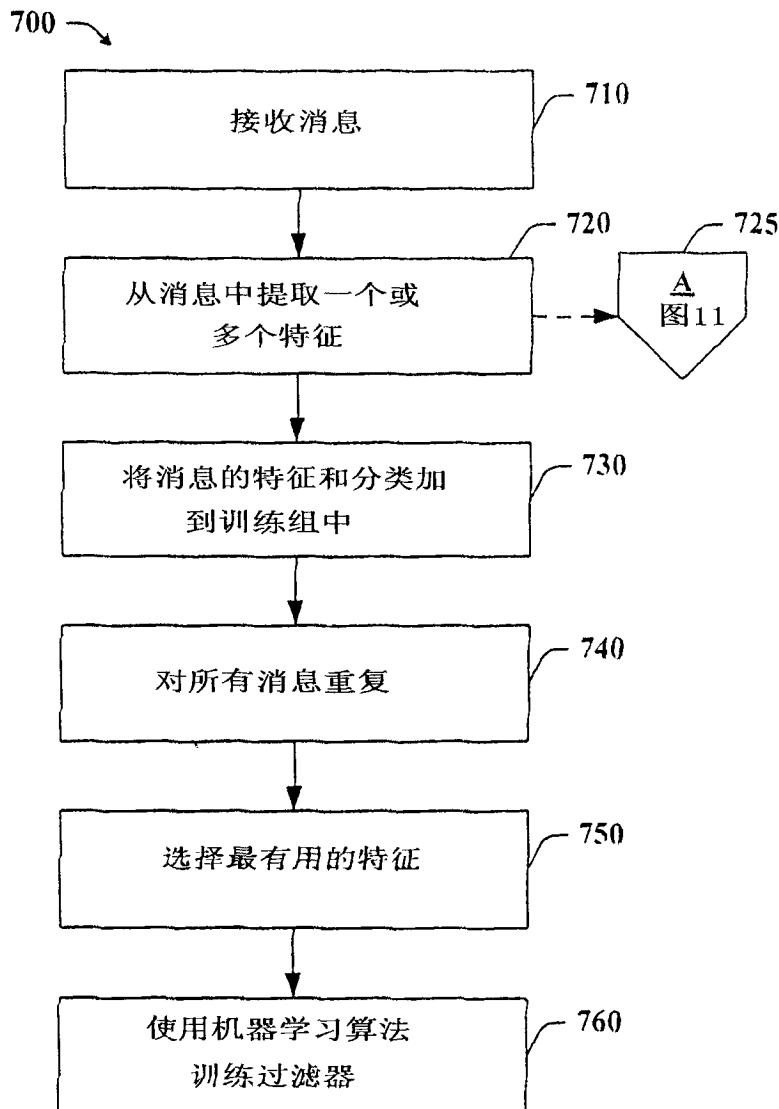


图 7

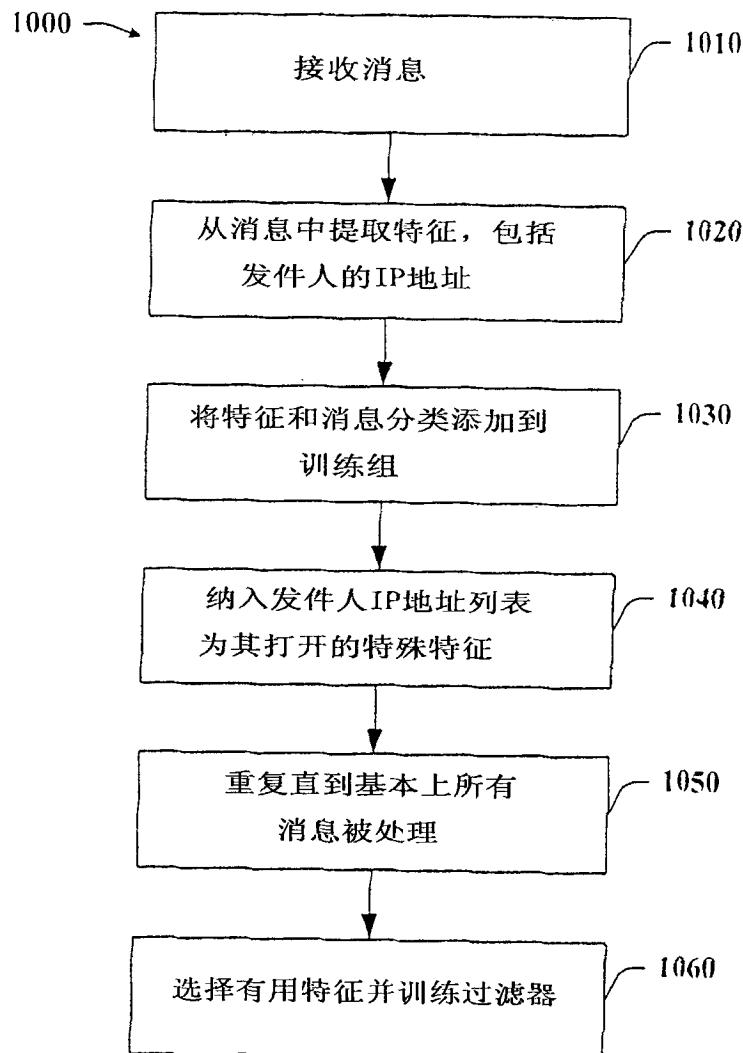


图 10

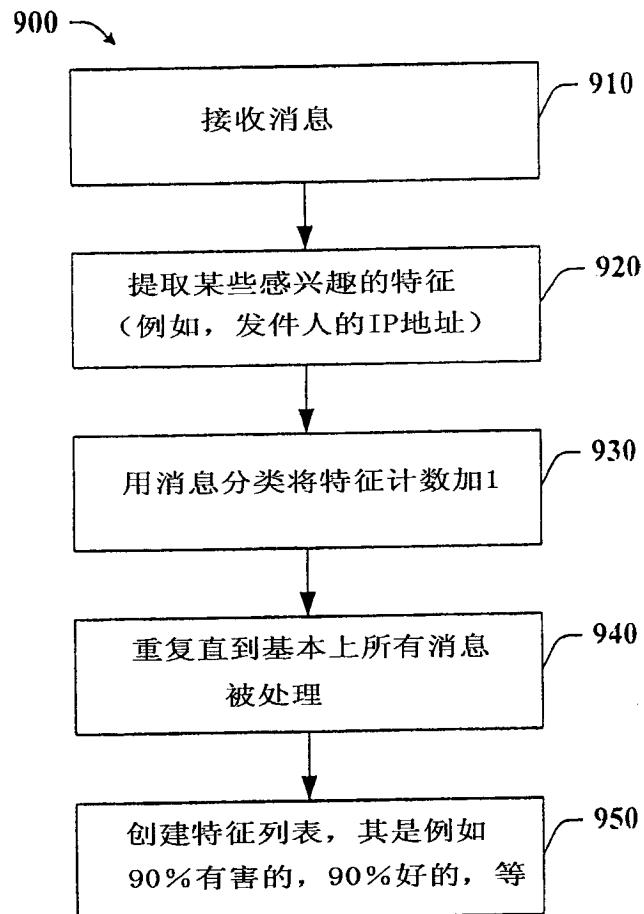


图 9

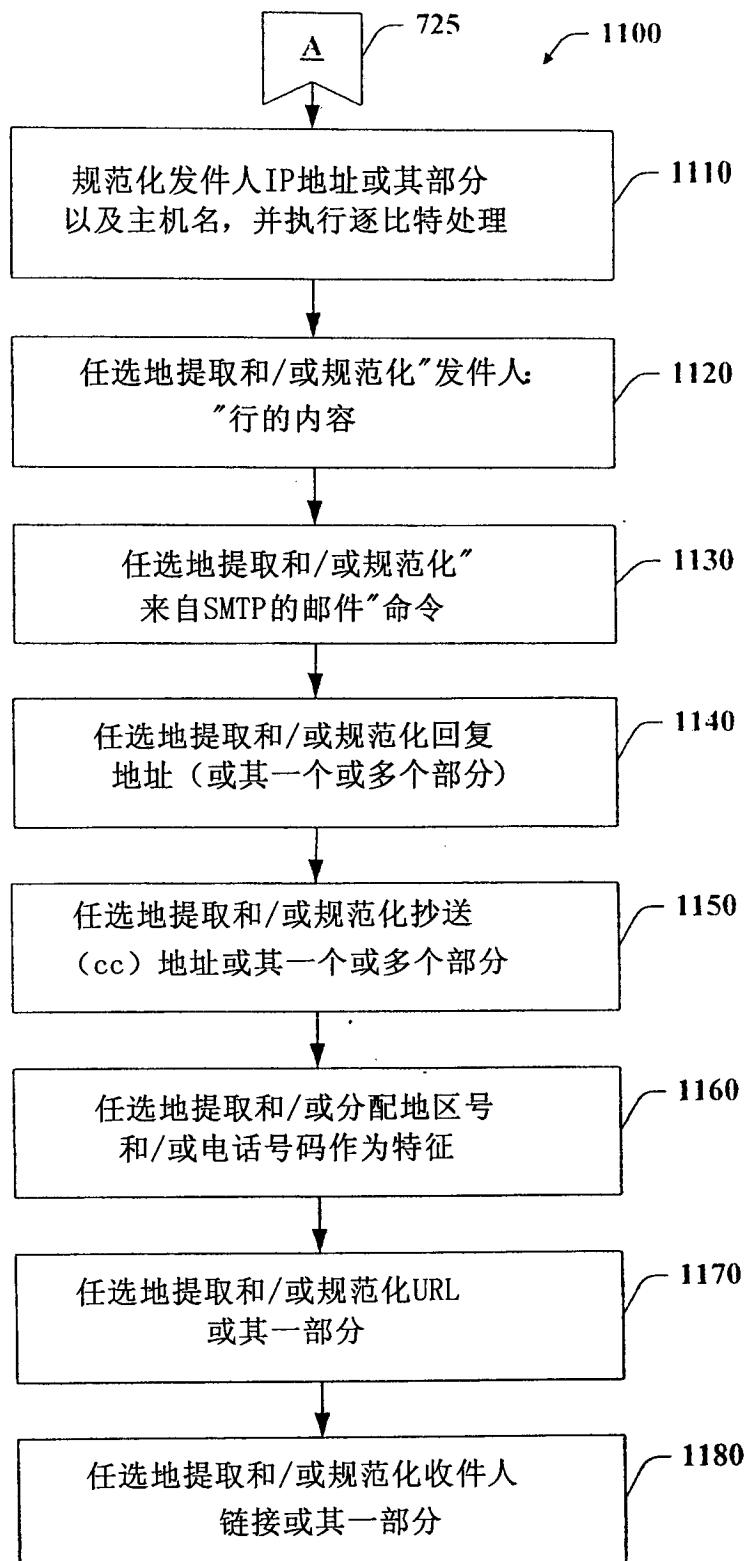


图 11

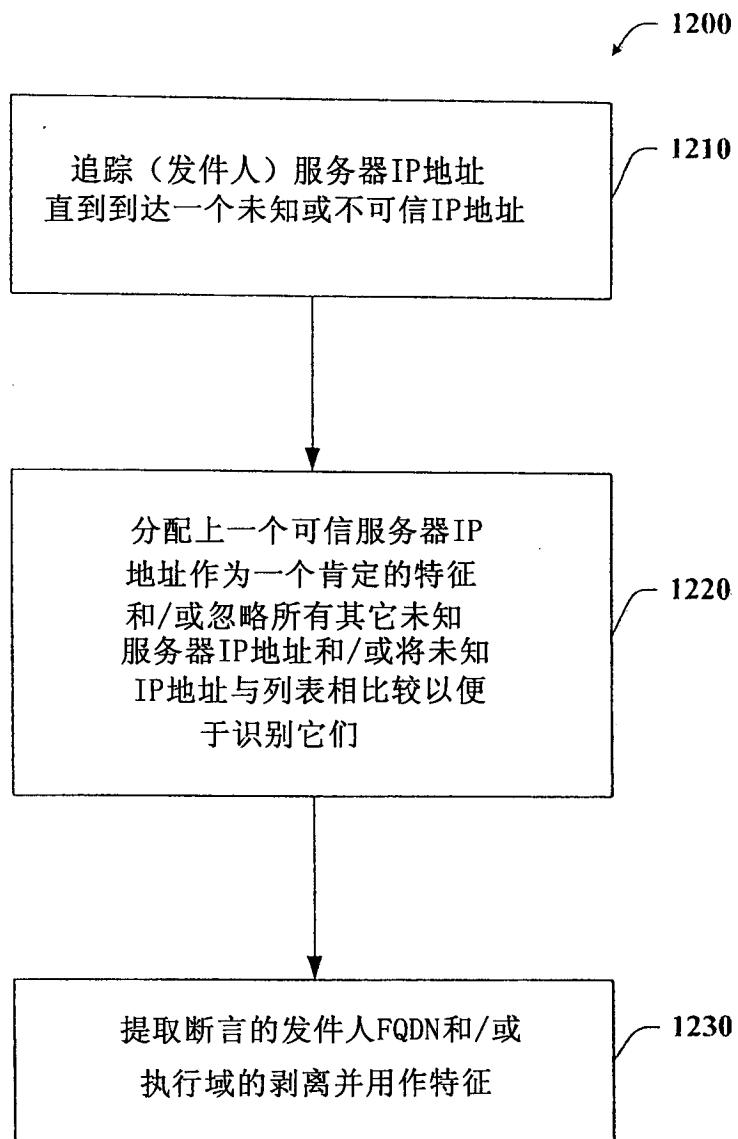


图 12

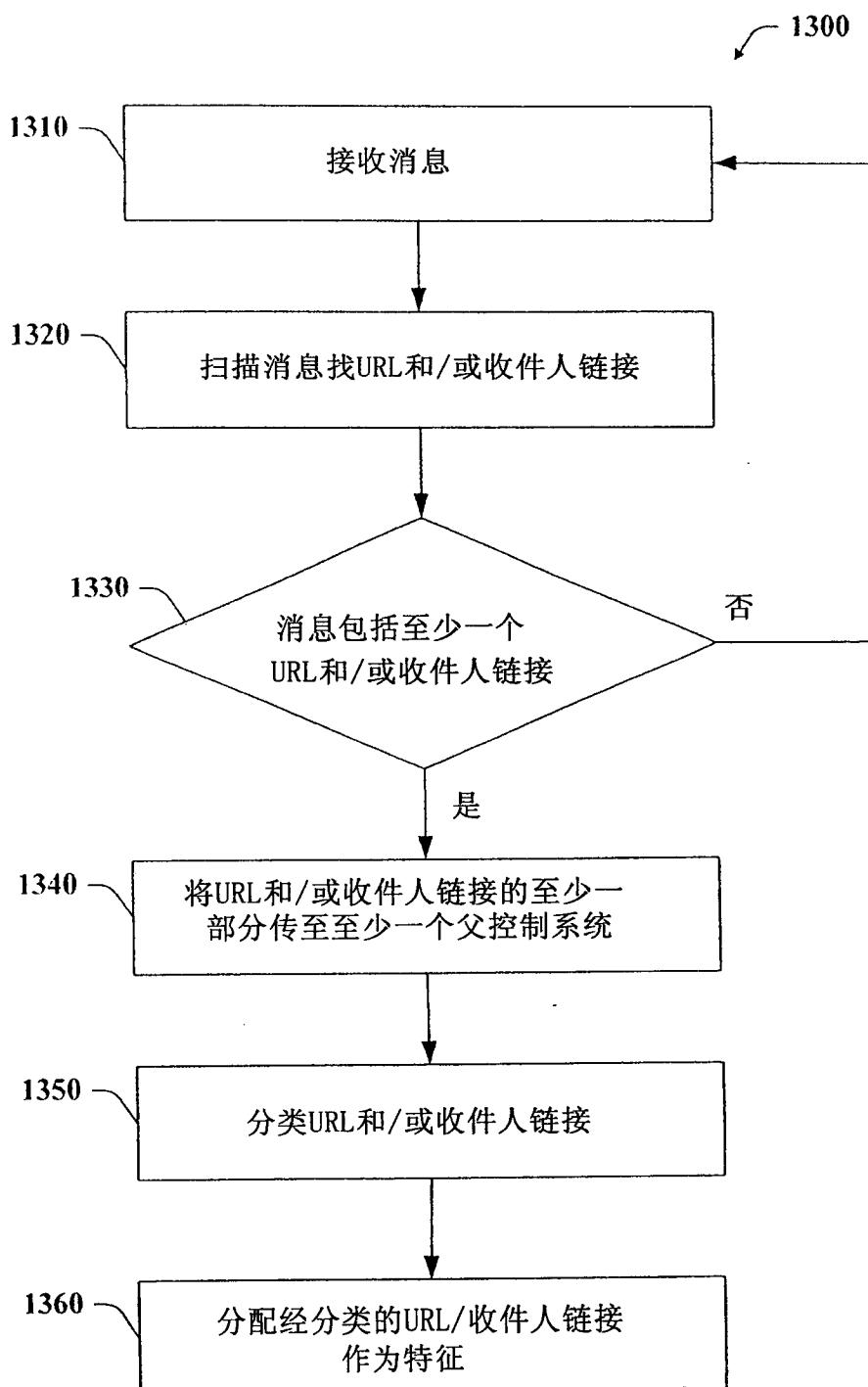


图 13

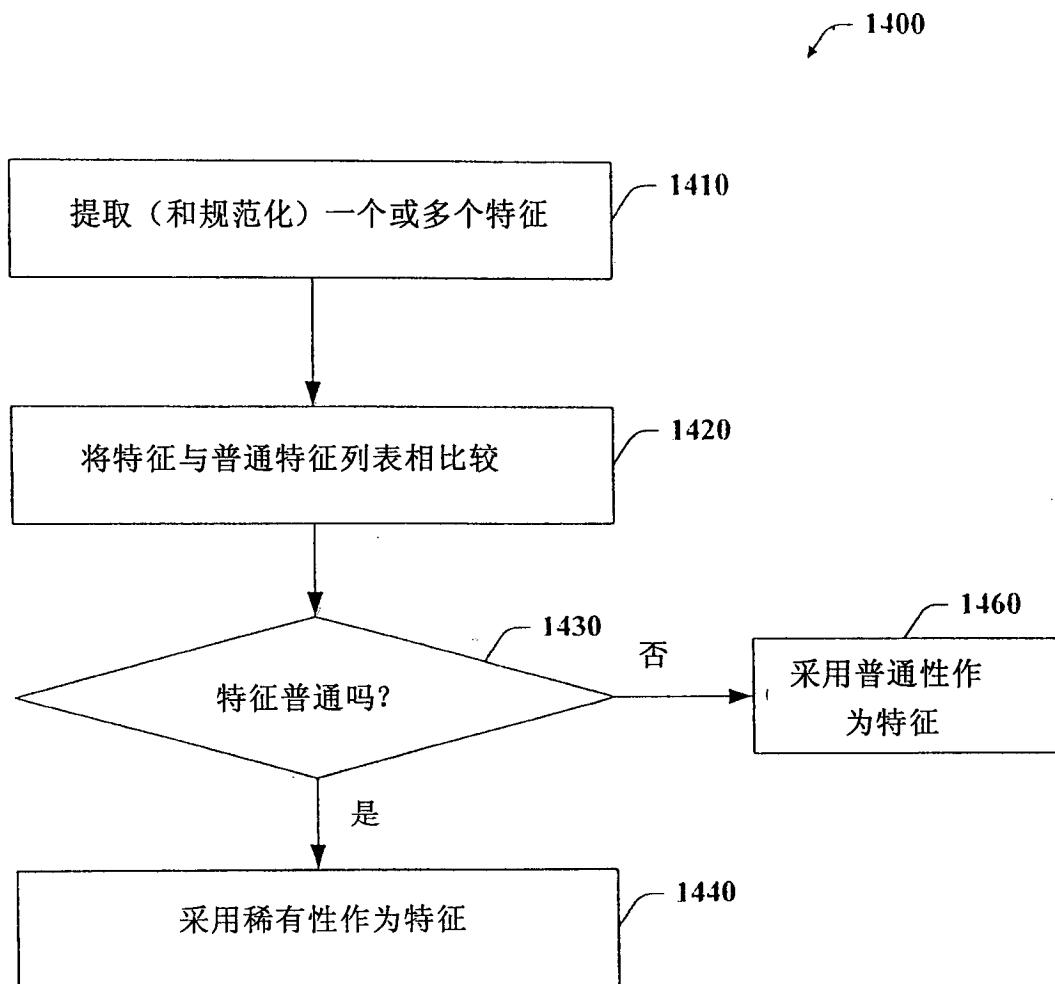


图 14

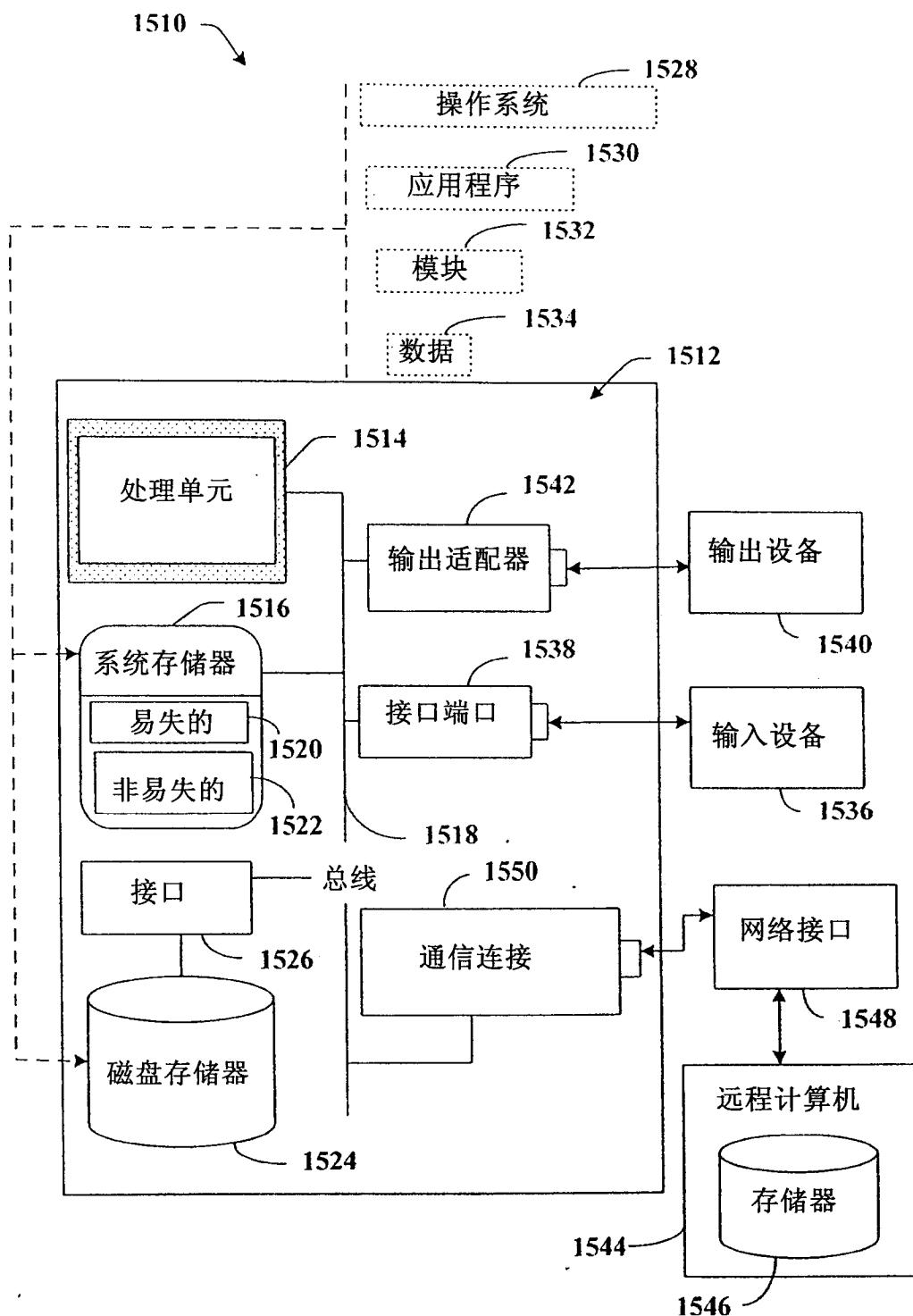


图 15