(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 104982011 A
(43)申请公布日 2015.10.14

(21)申请号 201480007764.1
(22)申请日 2014.02.04
(30)优先权数据
13/790,636 2013.03.08 US
(85)PCT国际申请进入国家阶段日
2015.08.06
(86)PCT国际申请的申请数据
PCT/RO2014/000007 2014.02.04
(87)PCT国际申请的公布数据
WO2014/137233 EN 2014.09.12
(71)申请人 比特梵德知识产权管理有限公司
地址 塞浦路斯尼科西亚
(72)发明人 阿德里安·托马
马里厄斯·尼古拉·蒂贝卡
(74)专利代理机构 北京律盟知识产权代理有限责任公司 11287
代理人 张世俊
(51)Int.Cl.
*H04L 12/58*(2006.01)

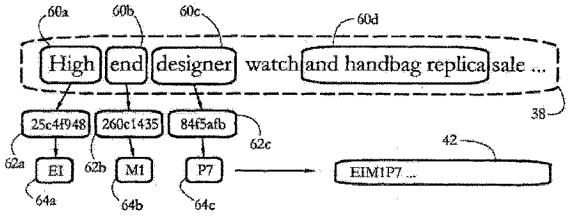
权利要求书4页 说明书14页 附图10页

(54)发明名称

使用多尺度文本指纹的文档分类

(57)摘要

所描述的系统及方法允许根据文档特定的文本指纹进行例如电子邮件消息及HTML文档的电子文档的分类。所述文本指纹是针对每一目标文档的文本块予以计算，且包括根据所述相应文本块的多个文本标记而确定的字符序列。在一些实施例中，通过针对短文本块进行放大且针对长文本块进行缩小，将所述文本指纹的长度强制为在预定长度范围（例如，在129与256个字符之间）内，而不管所述文本块的长度如何。例如，分类可包含确定电子文档表示未经请求的通信（垃圾邮件）还是例如网络钓鱼的网上诈骗。




CN 104982011 A

1. 一种客户端计算机系统，其包括至少一个处理器，所述至少一个处理器经配置以确定目标电子文档的文本指纹，使得所述文本指纹的长度约束在下限与上限之间，其中所述下限及上限为预定的，且其中确定所述文本指纹包括：

选择所述目标电子文档的多个文本标记；

响应于选择所述多个文本标记，根据所述上限及下限且根据所述所选择的多个文本标记的计数而确定指纹片段大小；

确定多个指纹片段，所述多个指纹片段中的每一指纹片段是根据所述所选择的多个文本标记中的相异文本标记的散列而确定，每一指纹片段由字符序列组成，所述序列的长度经选择为等于所述指纹片段大小；及

级联所述多个指纹片段以形成所述文本指纹。

2. 根据权利要求 1 所述的客户端计算机系统，其中所述至少一个处理器经进一步配置以：

将所述文本指纹发送到服务器计算机系统；及

从所述服务器计算机系统接收针对所述目标电子文档所确定的目标标签，所述目标标签指示所述目标电子文档所属的文档类别，其中确定所述目标标签包括：

从参考指纹的数据库检索参考指纹，所述参考指纹是针对属于所述类别的参考电子文档而确定，所述参考指纹是根据所述参考指纹的长度而选择，使得所述参考指纹的所述长度在所述上限与下限之间；及

根据比较所述文本指纹与所述参考指纹的结果而确定所述目标电子文档是否属于所述类别。

3. 根据权利要求 2 所述的客户端计算机系统，其中所述文档类别为垃圾邮件类别。

4. 根据权利要求 2 所述的客户端计算机系统，其中所述文档类别为诈骗性文档类别。

5. 根据权利要求 1 所述的客户端计算机系统，其中确定所述文本指纹进一步包括：根据所述相异文本标记的所述散列的位的相异群组而确定所述字符序列中的每一字符。

6. 根据权利要求 1 所述的客户端计算机系统，其中选择所述多个文本标记包括：

选择所述目标电子文档的初步多个文本标记；

确定所述初步多个文本标记的计数；及

作为响应，当所述初步多个文本标记的所述计数超过预定阈值时，修剪所述初步多个文本标记以形成所述所选择的多个文本标记，使得所述所选择的多个标记的所述计数不超过所述预定阈值。

7. 根据权利要求 6 所述的客户端计算机系统，其中修剪所述初步多个文本标记包括：根据所述初步多个文本标记中的目标文本标记的散列而将所述目标文本标记选择为所述所选择的多个文本标记。

8. 根据权利要求 6 所述的客户端计算机系统，其中修剪所述初步多个文本标记进一步包括：

确定所述目标文本标记的所述散列是否能被缩小因数整除；及

作为响应，当所述目标文本标记能被所述缩小因数整除时，将所述目标文本标记选择为所述所选择的多个文本标记。

9. 根据权利要求 1 所述的客户端计算机系统，其中选择所述多个文本标记包括：

选择所述目标电子文档的初步多个文本标记；

确定所述初步多个文本标记的计数；及

作为响应，当所述初步多个文本标记的所述计数超过预定阈值时，

确定多个聚合文本标记，所述多个聚合文本标记中的每一聚合文本标记包括所述初步多个文本标记的文本标记集合的级联；及

根据所述聚合文本标记的散列而将所述多个聚合标记中的聚合标记选择为所述所选择的多个文本标记。

10. 根据权利要求 1 所述的客户端计算机系统，其中所述目标电子文档是选自由电子邮件消息及超文本标记语言 HTML 文档组成的群组。

11. 根据权利要求 1 所述的客户端计算机系统，其中所述相异文本标记包括选自由目标电子通信的字、电子邮件地址及统一资源定位符 URL 组成的群组的项目。

12. 一种服务器计算机系统，其包括至少一个处理器，所述至少一个处理器经配置以执行与多个客户端系统进行的事务，其中事务包括：

从所述多个客户端系统中的客户端系统接收文本指纹，所述文本指纹是针对目标电子文档而确定，使得所述文本指纹的长度约束在下限与上限之间，其中所述下限及上限为预定的；及

向所述客户端系统发送指示所述目标电子文档所属的文档类别的目标标签，

其中确定所述文本指纹包括：

选择所述目标电子文档的多个文本标记；

响应于选择所述多个文本标记，根据所述上限及下限且根据所述所选择的多个文本标记的计数而确定指纹片段大小；

确定多个指纹片段，所述多个指纹片段中的每一指纹片段是根据所述所选择的多个文本标记中的相异文本标记的散列而确定，每一指纹片段由字符序列组成，所述序列的长度经选择为等于所述指纹片段大小；及

级联所述多个指纹片段以形成所述文本指纹，

且其中确定所述目标标签包括：

从参考指纹的数据库检索参考指纹，所述参考指纹是针对属于所述类别的参考电子文档而确定，所述参考指纹是根据所述参考指纹的长度而选择，使得所述参考指纹的所述长度在所述上限与下限之间；及

根据比较所述文本指纹与所述参考指纹的结果而确定所述目标电子文档是否属于所述类别。

13. 根据权利要求 12 所述的服务器计算机系统，其中所述文档类别为垃圾邮件类别。

14. 根据权利要求 12 所述的服务器计算机系统，其中所述文档类别为诈骗性文档类别。

15. 根据权利要求 12 所述的服务器计算机系统，其中确定所述文本指纹进一步包括：根据所述相异文本标记的所述散列的位的相异群组而确定所述字符序列中的每一字符。

16. 根据权利要求 12 所述的服务器计算机系统，其中选择所述多个文本标记包括：

选择所述目标电子文档的初步多个文本标记；

确定所述初步多个文本标记的计数；及

作为响应，当所述初步多个文本标记的所述计数超过预定阈值时，修剪所述初步多个文本标记以形成所述所选择的多个文本标记，使得所述所选择的多个标记的所述计数不超过所述预定阈值。

17. 根据权利要求 16 所述的服务器计算机系统，其中修剪所述初步多个文本标记包括：根据所述初步多个文本标记中的目标文本标记的散列而将所述目标文本标记选择为所述所选择的多个文本标记。

18. 根据权利要求 17 所述的服务器计算机系统，其中修剪所述初步多个文本标记进一步包括：

确定所述目标文本标记的所述散列是否能被缩小因数整除；及

作为响应，当所述目标文本标记能被所述缩小因数整除时，将所述目标文本标记选择为所述所选择的多个文本标记。

19. 根据权利要求 12 所述的服务器计算机系统，其中选择所述多个文本标记包括：

选择所述目标电子文档的初步多个文本标记；

确定所述初步多个文本标记的计数；及

作为响应，当所述初步多个文本标记的所述计数超过预定阈值时，

确定多个聚合文本标记，所述多个聚合文本标记中的每一聚合文本标记包括所述初步多个文本标记的文本标记集合的级联；及

根据所述聚合文本标记的散列而将所述多个聚合标记中的聚合标记选择为所述所选择的多个文本标记。

20. 根据权利要求 12 所述的服务器计算机系统，其中所述目标电子文档是选自由电子邮件消息及超文本标记语言 HTML 文档组成的群组。

21. 根据权利要求 12 所述的服务器计算机系统，其中所述相异文本标记包括选自由目标电子通信的字、电子邮件地址及统一资源定位符 URL 组成的群组的项目。

22. 一种方法，其包括使用客户端计算机系统的至少一个处理器以确定目标电子文档的文本指纹，使得所述文本指纹的长度约束在下限与上限之间，其中所述下限及上限为预定的，且其中确定所述文本指纹包括：

选择所述目标电子文档的多个文本标记；

响应于选择所述多个文本标记，根据所述上限及下限且根据所述所选择的多个文本标记的计数而确定指纹片段大小；

确定多个指纹片段，所述多个指纹片段中的每一指纹片段是根据所述所选择的多个文本标记中的相异文本标记的散列而确定，每一指纹片段由字符序列组成，所述序列的长度经选择为等于所述指纹片段大小；及

级联所述多个指纹片段以形成所述文本指纹。

23. 根据权利要求 22 所述的方法，其进一步包括使用所述至少一个处理器以根据所述文本指纹而确定所述目标电子文档所属的文档类别。

24. 一种方法，其包括使用经配置以执行与多个客户端系统进行的事务的服务器计算机系统的至少一个处理器以：

从所述多个客户端系统中的客户端系统接收文本指纹，所述文本指纹是针对目标电子文档而确定，使得所述文本指纹的长度约束在下限与上限之间，其中所述下限及上限为预

定的；及

向所述客户端系统发送针对所述目标电子文档所确定的目标标签，所述目标标签指示所述目标电子文档所属的文档类别，

其中确定所述文本指纹包括：

选择所述目标电子文档的多个文本标记；

响应于选择所述多个文本标记，根据所述上限及下限且根据所述所选择的多个文本标记的计数而确定指纹片段大小；

确定多个指纹片段，所述多个指纹片段中的每一指纹片段是根据所述所选择的多个文本标记中的相异文本标记的散列而确定，每一指纹片段由字符序列组成，所述序列的长度经选择为等于所述指纹片段大小；及

级联所述多个指纹片段以形成所述文本指纹，

且其中确定所述目标标签包括：

从参考指纹的数据库检索参考指纹，所述参考指纹是针对属于所述类别的参考电子文档而确定，所述参考指纹是根据所述参考指纹的长度而选择，使得所述参考指纹的所述长度在所述上限与下限之间；及

根据比较所述文本指纹与所述参考指纹的结果而确定所述目标电子文档是否属于所述类别。

# 使用多尺度文本指纹的文档分类

## 背景技术

[0001] 本发明涉及用于分类电子文档的方法及系统，且尤其涉及用于筛选未经请求的电子通信（垃圾邮件）且检测诈骗性网上文档的系统及方法。

[0002] 未经请求的电子通信（也称为垃圾邮件）形成全球通信业务的显著部分，从而影响计算机消息传递服务及电话消息传递服务两者。垃圾邮件可呈许多形式，从未经请求的电子邮件通信到伪装成各种互联网站点（例如，网志及社交网络站点）上的用户评论的垃圾邮件消息。垃圾邮件占用宝贵的硬件资源、影响生产率，且被通信服务及/或互联网的许多用户视为讨厌的及打扰的。

[0003] 网上诈骗（尤其是呈网络钓鱼及身份盗用的形式）已正对全球互联网用户造成日益增加的威胁。由在互联网上操作的国际犯罪网络诈骗性地获得的敏感身份信息（例如用户姓名、ID、密码、身份证号码及医疗记录、银行及信用卡明细）用于提取私人资金及/或进一步卖给第三方。除了给个人造成直接的金融损失以外，网上诈骗也造成一系列有害的副作用，例如公司日益增加的安全成本、较高的零售价格及银行收费、下跌的股票价值、较低的工资及下降的税收收入。

[0004] 在示范性网络钓鱼尝试中，虚假网站（也称为克隆）可伪装成属于网上零售商或金融机构的正版网页，要求用户输入一些个人信息（例如，用户名或密码）或一些金融信息（例如，信用卡号、账号或安全代码）。一旦毫无戒心的用户提交所述信息，其就可由所述虚假网站搜集。另外，用户可被引导到另一网页，其能够在用户的计算机上安装恶意软件。所述恶意软件（例如，病毒、特洛伊木马）能够通过记录由用户在访问某些网页时键入的密钥而继续窃取个人信息，且能够将用户的计算机变换成用于发动其它网络钓鱼及垃圾邮件攻击的平台。

[0005] 在垃圾电子邮件或电子邮件诈骗的情况下，在用户或电子邮件服务提供商的计算机系统上运行的软件可用于将电子邮件消息分类为垃圾邮件/非垃圾邮件（或诈骗性/合法），且甚至区分各种种类的消息，例如，区分产品提供、成人内容及尼日利亚诈骗。垃圾邮件/诈骗性消息可随后被引导到特殊文件夹或被删除。类似地，在内容提供商的计算机系统上运行的软件能够用于拦截发布到由相应内容提供商托管的网站的垃圾邮件/诈骗性消息，且防止显示相应消息，或向所述网站的用户显示所述相应消息可为诈骗性或垃圾邮件的警告。

[0006] 已提出用于识别垃圾邮件及/或网上诈骗的若干方法，其包含使消息的发端地址与已知违法或受信任地址列表（分别称为黑名单及白名单的技术）匹配、搜索某些字或字形（例如，再融资、Viagra®、股票），及分析消息标头。有时结合自动化数据分类方法（例如，贝叶斯(Bayesian)筛选、神经网络）而使用特征提取/匹配方法。

[0007] 一些所提出的方法使用散列以产生电子文本消息的紧凑表示。此类表示允许有效的消息间比较，其用于垃圾邮件或诈骗检测目的。

[0008] 垃圾邮件发送者及网上诈骗者试图通过使用各种迷惑方法（例如，拼错某些字、将垃圾邮件及/或诈骗性内容嵌入到伪装成合法文档的较大文本块中，及将消息的形式及

/ 或内容从一个分布波更改到另一分布波）而避开检测。使用散列的反垃圾邮件及反诈骗方法通常易受此类迷惑的干扰，这是因为文本的小改变可产生实质上不同的散列。成功的检测可因此受益于能够识别多态垃圾邮件及诈骗的方法及系统。

## 发明内容

[0009] 根据一个方面，一种客户端计算机系统包括至少一个处理器，其经配置以确定目标电子文档的文本指纹，使得所述文本指纹的长度约束在下限与上限之间，其中所述下限及上限为预定的。确定所述文本指纹包括：选择所述目标电子文档的多个文本标记；及响应于选择所述多个文本标记，根据所述上限及下限且根据所述所选择的多个文本标记的计数而确定指纹片段大小。确定所述文本指纹进一步包括：确定多个指纹片段，所述多个指纹片段中的每一指纹片段是根据所述所选择的多个文本标记中的相异文本标记的散列而确定，每一指纹片段由字符序列组成，所述序列的长度经选择为等于所述指纹片段大小；及级联所述多个指纹片段以形成所述文本指纹。

[0010] 根据另一方面，一种服务器计算机系统包括至少一个处理器，其经配置以执行与多个客户端系统进行的事务，其中事务包括：从所述多个客户端系统中的客户端系统接收文本指纹，所述文本指纹是针对目标电子文档而确定，使得所述文本指纹的长度约束在下限与上限之间，其中所述下限及上限为预定的；及向所述客户端系统发送指示所述目标电子文档所属的文档类别的目标标签。确定所述文本指纹包括：选择所述目标电子文档的多个文本标记；及响应于选择所述多个文本标记，根据所述上限及下限且根据所述所选择的多个文本标记的计数而确定指纹片段大小。确定所述文本指纹进一步包括：确定多个指纹片段，所述多个指纹片段中的每一指纹片段是根据所述所选择的多个文本标记中的相异文本标记的散列而确定，每一指纹片段由字符序列组成，所述序列的长度经选择为等于所述指纹片段大小；及级联所述多个指纹片段以形成所述文本指纹。确定所述目标标签包括：从参考指纹的数据库检索参考指纹，所述参考指纹是针对属于所述类别的参考电子文档而确定，所述参考指纹是根据所述参考指纹的长度而选择，使得所述参考指纹的所述长度在所述上限与下限之间；及根据比较所述文本指纹与所述参考指纹的结果而确定所述目标电子文档是否属于所述类别。

[0011] 根据另一方面，一种方法包括使用客户端计算机系统的至少一个处理器以确定目标电子文档的文本指纹，使得所述文本指纹的长度约束在下限与上限之间，其中所述下限及上限为预定的。确定所述文本指纹包括：选择所述目标电子文档的多个文本标记；及响应于选择所述多个文本标记，根据所述上限及下限且根据所述所选择的多个文本标记的计数而确定指纹片段大小。确定所述文本指纹进一步包括：确定多个指纹片段，所述多个指纹片段中的每一指纹片段是根据所述所选择的多个文本标记中的相异文本标记的散列而确定，每一指纹片段由字符序列组成，所述序列的长度经选择为等于所述指纹片段大小；及级联所述多个指纹片段以形成所述文本指纹。

[0012] 根据另一方面，一种方法包括使用经配置以执行与多个客户端系统进行的事务的服务器计算机系统的至少一个处理器以：从所述多个客户端系统中的客户端系统接收文本指纹，所述文本指纹是针对目标电子文档而确定，使得所述文本指纹的长度约束在下限与上限之间，其中所述下限及上限为预定的；及向所述客户端系统发送针对所述目标电子文

档所确定的目标标签，所述目标标签指示所述目标电子文档所属的文档类别。确定所述文本指纹包括：选择所述目标电子文档的多个文本标记；及响应于选择所述多个文本标记，根据所述上限及下限且根据所述所选择的多个文本标记的计数而确定指纹片段大小。确定所述文本指纹进一步包括：确定多个指纹片段，所述多个指纹片段中的每一指纹片段是根据所述所选择的多个文本标记中的相异文本标记的散列而确定，每一指纹片段由字符序列组成，所述序列的长度经选择为等于所述指纹片段大小；及级联所述多个指纹片段以形成所述文本指纹。确定所述目标标签包括：从参考指纹的数据库检索参考指纹，所述参考指纹是针对属于所述类别的参考电子文档而确定，所述参考指纹是根据所述参考指纹的长度而选择，使得所述参考指纹的所述长度在所述上限与下限之间；及根据比较所述文本指纹与所述参考指纹的结果而确定所述目标电子文档是否属于所述类别。

## 附图说明

[0013] 在阅读以下详细描述后及在参考图式后就将更好地理解本发明的前述方面及优点，在图式中：

[0014] 图 1 展示根据本发明的一些实施例的包括保护多个客户端系统的安全服务器的示范性反垃圾邮件 / 反诈骗系统。

[0015] 图 2-A 展示根据本发明的一些实施例的客户端计算机系统的示范性硬件配置。

[0016] 图 2-B 展示根据本发明的一些实施例的安全服务器计算机系统的示范性硬件配置。

[0017] 图 2-C 展示根据本发明的一些实施例的内容服务器计算机系统的示范性硬件配置。

[0018] 图 3-A 展示根据本发明的一些实施例的包括文本块的示范性垃圾电子邮件消息。

[0019] 图 3-B 展示根据本发明的一些实施例的包括文本块的示范性垃圾邮件网志评论。

[0020] 图 3-C 说明根据本发明的一些实施例的包括多个文本块的示范性诈骗性网页。

[0021] 图 4-A 说明根据本发明的一些实施例的客户端计算机与安全服务器之间的示范性垃圾邮件 / 诈骗检测事务。

[0022] 图 4-B 说明根据本发明的一些实施例的内容服务器与安全服务器之间的示范性垃圾邮件 / 诈骗检测事务。

[0023] 图 5 展示根据本发明的一些实施例的目标电子文档的示范性目标指示符，所述指示符包括文本指纹及其它垃圾邮件 / 诈骗识别数据。

[0024] 图 6 展示根据本发明的一些实施例的在客户端系统上执行的示范性应用程序集合的图解。

[0025] 图 7 说明根据本发明的一些实施例的由图 6 的指纹计算器执行的示范性步骤序列。

[0026] 图 8 展示根据本发明的一些实施例的目标文本块的文本指纹的示范性确定。

[0027] 图 9 展示根据本发明的一些实施例的针对处于各种放大及缩小因数的目标文本块而确定的多个指纹。

[0028] 图 10 说明根据本发明的一些实施例的由指纹计算器执行以确定缩小指纹的示范性步骤序列。

[0029] 图 11 展示根据本发明的一些实施例的在安全服务器上执行的示范性应用程序。

[0030] 图 12 展示根据本发明的一些实施例的在安全服务器上执行的示范性文档分类器的图解。

[0031] 图 13 展示在包括分析实际垃圾邮件消息流的计算机实验中获得的垃圾邮件检测率，所述分析是根据本发明的一些实施例而执行；比较所述检测率与通过常规方法而获得的检测率。

## 具体实施方式

[0032] 在以下描述中，应理解，结构之间的所有列举的连接可为直接操作连接或通过中介结构的间接操作连接。元件集合包含一或多个元件。元件的任何列举应被理解为是指至少一个元件。多个元件包含至少两个元件。除非另有要求，否则任何所描述的方法步骤未必需要按所说明的特定次序执行。来源于第二元件的第一元件（例如，数据）涵盖等于第二元件的第一元件，以及通过处理第二元件而产生的第一元件及任选的其它数据。根据参数做出确定或决定涵盖根据参数且任选地根据其它数据做出确定或决定。除非另有指定，否则一些数量 / 数据的指示符可为所述数量 / 数据自身，或为与所述数量 / 数据自身不同的指示符。除非另有指定，否则散列为散列函数的输出。除非另有指定，否则散列函数为将符号（例如，字符、位）序列映射成数字或位串的数学变换。计算机可读媒体涵盖例如磁性、光学及半导体存储媒体（例如，硬盘驱动器、光盘、闪速存储器、DRAM）的非暂时性媒体，以及例如导电电缆及光纤链路的通信链路。根据一些实施例，本发明尤其提供包括硬件（例如，一或多个处理器）以及计算机可读媒体的计算机系统，所述硬件经编程以执行本文中所描述的方法，所述计算机可读媒体编码指令以执行本文中所描述的方法。

[0033] 以下描述作为实例而未必作为限制来说明本发明的实施例。

[0034] 图 1 展示根据本发明的一些实施例的示范性反垃圾邮件 / 反诈骗系统 10。系统 10 包含内容服务器 12、发送器系统 13、安全服务器 14 及多个客户端系统 16a 到 c，其全部是由通信网络 18 连接。网络 18 可为广域网（例如，互联网），而网络 18 的部分也可包含局域网 (LAN)。

[0035] 在一些实施例中，内容服务器 12 经配置以从多个用户接收用户贡献内容（例如，文章、网志条目、媒体上传、评论等等），且组织、格式化及分布此类内容到第三方（例如，客户端系统 16a 到 c）。内容服务器 12 的示范性实施例为将电子消息递送提供到客户端系统 16a 到 c 的电子邮件服务器。内容服务器 12 的另一实施例为托管网志或社交联网站点的计算机系统。在一些实施例中，用户贡献内容以电子文档（在以下描述中也称为目标文档）的形式在网络 18 上流传。电子文档包含网页（例如，HTML 文档）及电子消息（例如，电子邮件及短消息服务 (SMS) 消息等等）。在服务器 12 处接收的用户贡献数据的部分可包括未经请求的及 / 或诈骗性消息及文档。

[0036] 在一些实施例中，发送器系统 13 包括向客户端系统 16a 到 c 发送未经请求的通信（例如，垃圾电子邮件消息）的计算机系统。可在服务器 12 处接收此类消息，且随后发送到客户端系统 16a 到 c。替代地，可使在服务器 12 处接收的消息可用（例如，通过 web 界面）以供客户端系统 16a 到 c 检索。在其他实施例中，发送器系统 13 可向内容服务器 12 发送未经请求的通信（例如，垃圾网志评论，或发布到社交联网站点的垃圾邮件）。客户端系统

16a 到 c 可随后经由协议（例如，超文本传输协议 (HTTP)）而检索此类通信。

[0037] 安全服务器 14 可包含一或多个计算机系统，其执行电子文档的分类（如下文详细地所展示）。执行此类分类可包含识别未经请求的消息（垃圾邮件）及 / 或诈骗性电子文档（例如，网络钓鱼消息及网页）。在一些实施例中，执行所述分类包含安全服务器 14 与内容服务器 12 之间及 / 或安全服务器 14 与客户端系统 16a 到 b 之间进行的协作式垃圾邮件 / 诈骗检测事务。

[0038] 客户端系统 16a 到 c 可包含终端用户计算机，其各自具有处理器、存储器及存储装置，且运行操作系统（例如，Windows®、MacOS®或 Linux）。一些客户端计算机系统 16a 到 c 可为移动计算及 / 或电信装置，例如，平板 PC、移动电话、个人数字助理 (PDA)，及家用装置（例如，电视机或音乐播放器等等）。在一些实施例中，客户端系统 16a 到 c 可表示个别客户，或若干客户端系统可属于同一客户。客户端系统 16a 到 c 可通过从发送器系统 13 接收电子文档（例如，电子邮件消息）且将其存储在本地收件箱中或通过在网络 18 上检索此类文档（例如，从由内容服务器 12 服务的网站）而存取此类文档。

[0039] 图 2-A 展示客户端系统 16（例如，图 1 的系统 16a 到 c）的示范性硬件配置。图 2-A 展示用于说明性目的的计算机系统；其它装置（例如，移动电话）的硬件配置可不同。在一些实施例中，客户端系统 16 包括处理器 20、存储器单元 22、输入装置 24 的集合、输出装置 26 的集合、存储装置 28 的集合及通信接口控制器 30，其全部是由总线 34 的集合连接。

[0040] 在一些实施例中，处理器 20 包括物理装置（例如，多核集成电路），其经配置以用信号及 / 或数据集合来执行计算及 / 或逻辑运算。在一些实施例中，此类逻辑运算是以处理器指令序列（例如，机器码或其它软件类型）的形式递送到处理器 20。存储器单元 22 可包括易失性计算机可读媒体（例如，RAM），其存储由处理器 20 在进行指令期间存取或产生的数据 / 信号。输入装置 24 可包含计算机键盘、鼠标及麦克风等等，其包含允许用户将数据及 / 或指令引入到系统 16 中的相应硬件接口及 / 或适配器。输出装置 26 可包含显示装置（例如，显示器及扬声器等等），以及硬件接口 / 适配器（例如，图形卡），其允许系统 16 向用户传达数据。在一些实施例中，输入装置 24 及输出装置 26 可共享硬件的公用部分，在触摸屏装置的情况下就是如此。存储装置 28 包含计算机可读媒体，其实现软件指令及 / 或数据的非易失性存储、读取及写入。示范性存储装置 28 包含磁盘与光盘及闪速存储器装置，以及可移动媒体（例如，CD 及 / 或 DVD 盘与驱动器）。通信接口控制器 30 使系统 16 能够连接到网络 18 及 / 或其它装置 / 计算机系统。总线 34 共同地表示多个系统、外围设备及芯片集总线，及 / 或实现客户端系统 16 的装置 20 到 30 的内部通信的所有其它电路。举例来说，总线 34 可包括将处理器 20 连接到存储器 22 的北桥，及 / 或将处理器 20 连接到装置 24 到 30 的南桥等等。

[0041] 图 2-B 展示根据本发明的一些实施例的安全服务器 14 的示范性硬件配置。安全服务器 14 包含处理器 120 及存储器单元 122，且可进一步包括存储装置 128 的集合及至少一个通信接口控制器 130，其全部是经由总线 134 的集合而互连。在一些实施例中，处理器 120、存储器 122 及存储装置 128 的操作可分别类似于项目 20、22 及 28 的操作，如上文关于图 2-A 所描述。存储器单元 122 存储由处理器 120 在进行指令期间存取或产生的数据 / 信号。控制器 130 使安全服务器 14 能够连接到网络 18，以向连接到网络 18 的其它系统发射数据及 / 或从连接到网络 18 的其它系统接收数据。

[0042] 图 2-C 展示根据本发明的一些实施例的内容服务器 12 的示范性硬件配置。内容服务器 12 包含处理器 220 及存储器单元 222，且可进一步包括存储装置 228 的集合及至少一个通信接口控制器 230，其全部是由总线 234 的集合互连。在一些实施例中，处理器 220、存储器 222 及存储装置 228 的操作可分别类似于项目 20、22 及 28 的操作，如上文所描述。存储器单元 222 存储由处理器 220 在进行指令期间存取或产生的数据 / 信号。在一些实施例中，接口控制器 230 使内容服务器 12 能够连接到网络 18，且向连接到网络 18 的其它系统发射数据及 / 或从连接到网络 18 的其它系统接收数据。

[0043] 图 3-A 展示根据本发明的一些实施例的包括垃圾电子邮件的示范性目标文档 36a。目标文档 36a 可包括标头及有效负载，所述标头包含消息路由数据（例如，发件人的指示符及 / 或收件人的指示符），及 / 或其它数据（例如，时间戳及内容类型（例如，多用途互联网邮件扩展 (MIME) 类型）的指示符）。所述有效负载可包含作为文本及 / 或图像显示给用户的数据。在内容服务器 12 及 / 或客户端系统 16a 到 c 上执行的软件可处理所述有效负载以产生目标文档 36a 的目标文本块 38a。在一些实施例中，目标文本块 38a 包括意在被解释为文本的标志及 / 或符号序列。文本块 38a 可包含比如标点符号的特殊字符，以及表示网络地址、统一资源定位符 (URL)、电子邮件地址、假名及别名的字符序列等等。目标文本块 38a 可直接嵌入到目标文档 36a 中（例如，作为纯文本 MIME 部分），或可包括处理嵌入在文档 36a 中的计算机指令集合的结果。举例来说，目标文本块 38a 可包含呈现超文本标记语言 (HTML) 指令集合的结果，或执行嵌入在目标文档 36a 中的客户端脚本指令或服务器端脚本指令集合（例如，PHP、Javascript）的结果。在另一实施例中，目标文本块 38a 可嵌入到图像中，在图像垃圾邮件的情况下就是如此。

[0044] 图 3-B 展示另一示范性目标文档 36b，其包括发布在网页（例如，网志、网上新闻页面或社交联网页面）上的评论。在一些实施例中，文档 36b 包括数据字段（例如，嵌入在 HTML 文档中的表单的字段）集合的内容。例如，填写此类表单字段可由人类操作者远程执行，及 / 或由在发送器系统 13 上执行的软件部分自动地执行。在一些实施例中，文档 36b 的显示包括文本块 38b，其由字符及 / 或符号序列组成，所述字符及 / 或符号序列意在由存取相应网站的用户解释为文本。文本块 38b 可包含超链接、特殊字符、表情符及图像等等。

[0045] 图 3-C 说明另一示范性目标文档 36c，其包括网络钓鱼网页。文档 36c 可作为 HTML 及 / 或服务器端或客户端脚本指令集合而递送，其在执行时确定文档查看器（例如，web 浏览器）以产生图像集合及 / 或文本块集合。图 3-C 中说明两个此类示范性文本块 38c 到 d。文本块 36c 到 d 可包含超链接及电子邮件地址。

[0046] 图 4-A 展示根据本发明的一些实施例的示范性客户端系统 16（例如，图 1 的客户端系统 16a 到 c）与安全服务器 14 之间的示范性垃圾邮件 / 诈骗检测事务。图 4-A 中说明的交换发生（例如）在系统 10 的实施例中，系统 10 经配置以检测电子邮件垃圾邮件。在从内容服务器 12 接收到目标文档 36（例如，电子邮件消息）之后，客户端系统 16 可确定目标文档 36 的目标指示符 40，且可将目标指示符 40 发送到安全服务器 14。目标指示符 40 包括允许安全服务器 14 执行目标文档 36 的分类的数据以确定（例如）文档 36 是否为垃圾邮件。响应于接收到目标指示符 40，安全服务器 14 可向相应客户端系统 16 发送指示文档 36 是否为垃圾邮件的目标标签 50。

[0047] 图 4-B 中说明垃圾邮件检测事务的另一实施例，且其发生在内容服务器 12 与安全

服务器 14 之间。此类交换可发生（例如）以检测发布到网志及 / 或社交网络网站的未经请求的通信，或检测网络钓鱼网页。托管及 / 或显示相应网站的内容服务器 12 可接收目标文档 36（例如，网志评论）。内容服务器 12 可处理相应通信以产生相应文档的目标指示符 40，且可将目标指示符发送到安全服务器 14。作为回报，服务器 14 可确定指示相应文档是否为垃圾邮件或诈骗性的目标标签 50，且将标签 50 发送到内容服务器 12。

[0048] 图 5 展示针对示范性目标文档 36（例如，图 3-A 中的电子邮件消息 36a）所确定的示范性目标指示符 40。在一些实施例中，目标指示符 40 为数据结构，其包含与目标文档 36 唯一地相关联的消息标识符 41（例如，散列索引），及针对文档 36 的文本块（例如，图 3-A 中的文本块 38a）所确定的文本指纹 42。目标指示符 40 可进一步包括指示文档 36 的发件人的发件人指示符 44、指示发出文档 36 的网络地址（例如，IP 地址）的路由指示符 46，及指示文档 36 被发送及 / 或接收的时刻的时间戳 48。在一些实施例中，目标指示符 40 可包括文档 36 的其它垃圾邮件指示及 / 或诈骗指示特征，例如，指示文档 36 是否包含图像的旗标、指示文档 36 是否包含超链接的旗标，及针对文档 36 所确定的文档布局指示符等等。

[0049] 图 6 展示根据本发明的一些实施例的在客户端系统 16 上执行的示范性组件集合。图 6 中说明的配置适合于（例如）检测在客户端系统 16 处接收的垃圾电子邮件消息。系统 16 包括文档消解仪 52 及连接到文档消解仪 52 的文档显示管理器 54。文档消解仪 52 可进一步包括指纹计算器 56。在一些实施例中，文档消解仪 52 接收目标文档 36（例如，电子邮件消息），且处理文档 36 以产生目标指示符 40。处理文档 36 可包含剖析文档 36 以识别相异数据字段及 / 或类型，且区分标头数据与有效负载数据等等。当文档 36 为电子邮件消息时，示范性剖析可产生针对相应消息的发件人、IP 地址、主题、时间戳及内容等等的相异数据对象。当文档 36 的内容包含多个 MIME 类型的数据时，剖析可产生针对每一 MIME 类型（例如，纯文本、HTML 及图像等等）的相异数据对象。文档消解仪 52 可随后制定目标指示符 40，例如通过填写目标指示符 40 的相应字段（例如，发件人、路由地址及时间戳等等）。客户端系统 16 的软件组件可随后将目标指示符 40 发射到安全服务器 14 以供分析。

[0050] 在一些实施例中，文档显示管理器 54 接收目标文档 36，将其转化为视觉形式且将其显示在客户端系统 16 的输出装置上。显示管理器 54 的一些实施例也可允许客户端系统 16 的用户与所显示的内容交互。显示管理器 54 可与现成的文档显示软件（例如，web 浏览器、电子邮件阅读器、电子书阅读器及媒体播放器等等）集成。例如，此类集成可以软件插件的形式而实现。显示管理器 54 可经配置以将目标文档 36（例如，传入电子邮件）指派到文档类别（例如，文档的垃圾邮件、合法及 / 或各种其它类别与子类别）。此类分类可根据从安全服务器 14 接收的目标标签 50 而确定。显示管理器 54 可经进一步配置以将垃圾邮件 / 诈骗消息分组为单独的文件夹及 / 或仅向用户显示合法消息。管理器 54 也可根据此类分类而对文档 36 加标签。例如，文档显示管理器 54 可以相异的颜色显示垃圾邮件 / 诈骗消息，或紧接于每一垃圾邮件 / 诈骗消息而显示指示相应消息（例如，垃圾邮件、网络钓鱼等等）的分类的旗标。类似地，当文档 36 为诈骗性网页时，显示管理器 54 可阻止用户存取相应页面及 / 或向用户显示警告。

[0051] 在经配置以检测作为评论发布在网志及社交网络站点上的垃圾邮件 / 诈骗的实施例中，文档消解仪 52 及显示管理器 54 可在内容服务器 12（代替图 6 中展示的客户端系统 16a 到 c）上执行。此类软件可以服务器端脚本的形式在内容服务器 12 上实施，其可进

一步并入（例如，作为插件）到较大的脚本包中（例如，作为针对 Wordpress® 或 Drupal® 网上出版平台的反垃圾邮件 / 反诈骗插件）。一旦确定目标文档 36 为垃圾邮件或诈骗性。显示管理器 54 就可经配置以阻止相应消息，从而防止其在相应网站内显示。

[0052] 指纹计算器 56（图 6）经配置以确定目标文档 36 的文本指纹，其构成目标指示符 40 的部分（例如，图 5 中的项目 42）。在一些实施例中，针对目标电子文档所确定的指纹包括字符序列，所述序列的长度约束在预定上限与下限之间（例如，在 129 与 256 个字符之间，129 及 256 包含在内）。使此类指纹在预定长度范围内可为期望的，其允许与参考指纹集合的有效比较，以识别包括垃圾邮件及 / 或诈骗的文本块，如下文更详细地所展示。在一些实施例中，形成指纹的字符可包括字母数字字符、特殊字符及符号（例如，*、/、$ 等等）等等。用于形成文本指纹的其它示范性字符包含用于在各种编码中表示数目的数字或其它符号，例如，二进制、十六进制及 Base64 等等。

[0053] 图 7 说明由指纹计算器 56 执行以确定文本指纹的示范性步骤序列。在步骤 402 中，指纹计算器可选择用于指纹计算的目标文档 36 的目标文本块。在一些实施例中，目标文本块可实质上由目标文档 36 的全部文本内容（例如，文档 36 的纯文本 MIME 部分）组成。在一些实施例中，目标文本块可由文档 36 的文本部分的单一段落组成。在经配置以筛选基于 web 的垃圾邮件的实施例中，目标文本块可由网志评论的内容组成，或由用户发送且意在发布在相应网站上的另一种类的消息（例如，Facebook® wall post、Twitter® tweet 等等）组成。在一些实施例中，目标文本块包括 HTML 文档的章节的内容（例如，由 DIV 或 SPAN 标签指示的章节）。

[0054] 在步骤 404 中，指纹计算器 56 可将目标文本块分成文本标记。图 8 展示将文本块 38 分成多个文本标记 60a 到 c 的示范性分段。在一些实施例中，文本标记为由任何定界符字符 / 符号集合而与其它文本标记分离的字符 / 符号序列。用于西方语言脚本的示范性定界符包含空格、断行、制表符、‘\r’、‘\0’、句号、逗号、冒号、分号、圆括号及 / 或方括号、反向及 / 或正向斜线、双斜线、数学符号（例如，‘+’、‘-’、‘*’、‘^’）、标点符号（例如，‘!’及‘?’）及特殊字符（例如，‘$’及‘|’等等）。图 8 中的示范性标记为个别字；文本标记的其它实例可包含多字序列、电子邮件地址及 URL 等等。为了识别文本块 38 的个别标记，指纹计算器可使用所属领域中所知的任何串标记化算法。指纹计算器 56 的一些实施例可考虑某些标记（例如，英语中的常见字（例如，‘a’及‘the’）），此针对指纹计算为不合格的。在一些实施例中，超过预定最大长度的标记被进一步分成较短标记。

[0055] 在一些实施例中，由计算器 56 确定的文本指纹的长度约束在预定范围（例如，在 129 与 256 个字符之间，129 及 256 包含在内）内，而不管相应目标文本块的长度或标记计数如何。为了计算此类指纹，在步骤 406 中，指纹计算器 56 可首先确定目标文本块的文本标记的计数，且比较所述计数与预定上限阈值，所述预定上限阈值是根据指纹长度的上限而确定。当标记计数超过上限阈值（例如，256）时，在步骤 408 中，计算器 56 可确定缩小指纹，如下文详细地所展示。

[0056] 当标记计数降到低于上限阈值时，在步骤 410 中，指纹计算器可计算每一文本标记的散列。图 8 展示示范性散列 62a 到 c，其分别是针对文本标记 60a 到 c 而确定。散列 62a 到 c 是以十六进制记数法而展示。在一些实施例中，此类散列为将散列函数应用于每一标记 60a 到 c 的结果。许多此类散列函数及算法在所属领域中为已知的。Naïve 散列算法

快速，但通常产生大量冲突（相异标记具有相同散列的情况）。更复杂的散列（例如，通过比如 MD5 的消息摘要算法而计算的散列）据称为无冲突，但其计算费用巨大。本发明的一些实施例使用在计算速度与冲突避免之间提供权衡的散列算法来计算散列 62a 到 c。此类算法的实例归因于罗伯特．赛奇威克 (Robert Sedgewick)，且在所属领域中称为 RSHash。下文展示 RSHash 的伪代码：

[0057]

```
foreach ( byte x ; byte ) {
    value = value * a + x ;
    a * = b ;
}
return value ;
```

[0058] 其中 a 及 b 表示整数，例如，a ＝ 63,689，且 b ＝ 378,551。

[0059] 散列 62a 到 c 的大小（位的数目）可影响冲突的可能性，且因此影响垃圾邮件检测率。一般来说，使用小散列会增加冲突的可能性。较大散列通常不易产生冲突，但在计算速度及存储器方面更加昂贵。指纹计算器 56 的一些实施例计算项目 62a 到 c 作为 30 位散列。

[0060] 指纹计算器 56 可现在确定目标文本块的实际文本指纹。图 8 进一步说明针对目标文本块 38 所确定的示范性指纹 42。文本指纹 42 包括根据在步骤 410 中确定的散列 62a 到 c 所确定的字符序列。在一些实施例中，针对每一标记 60a 到 c，指纹计算器 56 确定指纹片段，其在图 8 中被说明作为项目 64a 到 c。在一些实施例中，随后级联此类片段以产生指纹 42。

[0061] 每一指纹片段 64a 到 c 可包括根据相应标记 60a 到 c 的散列 62a 到 c 所确定的字符序列。在一些实施例中，所有指纹片段 64a 到 c 具有相同的长度：在图 8 的实例中，每一片段 64a 到 c 由两个字符组成。确定指纹片段的所述长度，使得相应指纹具有在所要范围（例如，在 129 到 256 个字符）内的长度。在一些实施例中，指纹片段的长度称为放大因数 k。例如，长度 1 的片段为无缩放片段（放大因数 1），从而产生无缩放指纹；长度 2 的片段为 2 倍放大片段（放大因数 2），从而产生 2 倍放大片段，等等。图 9 展示针对处于各种放大因数 k 的文本块 38 所确定多个文本指纹 42a 到 c。

[0062] 在步骤 412（图 7）中，指纹计算器 56 确定放大因数 k 的值，其产生在所要的预定范围内的指纹长度。例如，当标记计数大于下限阈值（其是根据所要指纹长度的下限而确定）时，指纹计算器可决定计算无缩放指纹 (k ＝ 1)，这是因为无缩放指纹已在所要长度范围内。例如，当文本块 38 具有过少的标记时，指纹计算器可计算 2 倍或 3 倍放大指纹。

[0063] 然后，在步骤 414 中，根据每一标记的相应散列，指纹计算器 56 计算针对所述标记的指纹片段。为了确定片段 64a 到 c，指纹计算器 56 可使用所属领域中所知的任何编码方案（例如，散列 62a 到 c 的二进制或 Base64 表示）。此类编码方案建立数目与来自预定字母表的字符序列之间的一对一映射。例如，当使用 Base64 表示时，散列的六个连续位中的每一群组可被映射成字符。

[0064] 在一些实施例中，通过改变用于表示相应散列的字符的数目，可针对每一散列而

确定多个指纹片段。为了产生长度 1(例如,放大因数 1)的片段,一些实施例仅使用相应散列的六个最低有效位。可使用相应散列的额外六个位来产生长度 2(例如,放大因数 2)的片段,等等。在 Base64 表示中,30 位散列可因此得到高达 5 个字符长的指纹片段,其对应于五个放大因数。表 1 展示在各种放大因数下计算的示范性指纹片段(来自图 9 中的示范性文本块 38)。

[0065] 表 1

[0066]

| 标记 | 散列 | 处于放大因数 1 的片段 | 处于放大因数 2 的片段 | 处于放大因数 4 的片段 |
|---|---|---|---|---|
| high | 25c4f948 | I | EI | IE51 |
| end | 260c1435 | 1 | M1 | mMU1 |
| designer | 84f5afb | 7 | P7 | IPa7 |
| watch | 34f5dc75 | 1 | 11 | 01c1 |
| and | 2367c3d9 | Z | nZ | jnDZ |
| handbag | 1aa88b79 | 4 | o5 | aoL5 |
| replica | 33381eca | K | 4K | z4eK |
| sale | e96c2eb | r | Wr | OWCr |
| compare | 1c947587 | H | UH | cU1H |

[0067]

| | | | | |
|---|---|---|---|---|
| our | 24b80bd8 | Y | 4Y | k4LY |
| price | 3b54d80d | N | UN | 7UYN |
| on | 1777af4f | P | 3P | X3vP |
| a | 61 | h | Ah | AAAh |
| handful | 380be94e | O | LO | 4LpO |
| of | 1777af47 | H | 3H | X3vH |
| our | 24b80bd8 | Y | 4Y | k4LY |
| high | 3f155a68 | o | Vo | /Vao |
| end | 260c1435 | 1 | M1 | mMU1 |
| replicas | ad4c229 | p | Up | KUCp |

[0068] 在步骤 416(图 7)中,指纹计算器 56 组合文本指纹 42,例如,通过级联步骤 414 中计算的片段。

[0069] 回到步骤 406,当标记计数被发现为大于上限阈值时,指纹计算器 56 确定相应文本块的缩小指纹。在一些实施例中,缩小包括仅从文本块 38 的标记的子集计算指纹 42。选择子集可包括根据散列选择准则而修剪步骤 404 中确定的多个文本标记。图 10 中说明执行此类计算的示范性步骤序列。步骤 422 选择用于指纹计算的缩小因数。在一些实施例中,表示为 k 的缩小因数平均起来指示文本块 38 的标记的仅 1/k 用于指纹计算。指纹计算器 56 可因此根据步骤 406(图 7)中确定的标记计数而选择缩小因数。在一些实施例中,缩小因数的最初选择可能无法产生在所要长度范围内的指纹(参见下文);在此类情况下,可在一循环中以试误方式执行步骤 422 到 430,直到产生适当长度的指纹。例如,指纹计算器 56 可最初选择缩小因数 k = 2;当此值未能产生足够短的指纹时,计算器 56 可选择 k = 3,等等。

[0070] 然后,指纹计算器可根据散列选择准则而选择标记。当缩小时,指纹计算器 56 可使用已在步骤 404(图 7)中确定的标记,或可从文本块 38 计算新标记。在图 10 所说明的

实例中，在步骤 424 中，指纹计算器 56 确定文本块 38 的聚合标记集合。在一些实施例中，通过级联连续个别标记而确定聚合标记（在图 9 中被说明为项目 60d）。用于形成聚合标记的标记的计数可根据缩小因数而变化。

[0071] 在步骤 426 中，针对每一聚合标记而计算散列（例如，使用上文所描述的方法）。在步骤 428 中，计算器 56 根据散列选择准则而选择聚合标记子集。在一些实施例中，针对缩小因数 k，所述选择准则要求针对所选择的子集的成员所确定的全部散列等于模数 k。例如，为了确定 2 倍缩小指纹，计算器 56 可仅考虑聚合标记，其散列等于模数 2（即，仅奇散列，或仅偶散列）。在一些实施例中，散列选择准则包括仅选择其散列能被缩小因数 k 整除的标记。

[0072] 在步骤 430 中，指纹计算器 56 可检查步骤 428 中选择的标记的计数是否在所要指纹长度范围内。如果所述计数不在所要指纹长度范围内，那么计算器 56 可返回到步骤 422 且以另一缩小因数 k 重新开始。当所选择的标记的计数在范围内时，在步骤 432 中，计算器 56 根据所选择的标记的每一散列而确定指纹片段。在步骤 434 中，组合此类片段以产生指纹 42。图 9 说明针对文本块 38 所确定的一些缩小指纹 42d 到 h。表 2 展示针对图 9 中的同一文本块 38 所确定的示范性指纹片段（在各种缩小因数下）。

[0073] 表 2

[0074]

| 聚合标记 | 散列 | 缩小因数 2 | 缩小因数 3 | 缩小因数 4 | 缩小因数 5 | 缩小因数 6 |
|---|---|---|---|---|---|---|
| high end designer | 54206878 | 4 | | 4 | 4 | |
| end designer watch | 63514ba5 | | 1 | | 1 | |
| designer watch and | 60acfb49 | | | | | |
| watch and handbag | 73062bc7 | | H | | | |
| and handbag replica | 71486e1c | c | | c | | |
| handbag replica sale | 5c776d2e | u | u | | | u |
| replica sale compare | 5e63573c | 8 | | 8 | 8 | |
| sale compare our | 4fe3444a | K | | | | |
| compare our price | 7ca1596c | s | | s | | |
| our price on | 77849334 | 0 | 0 | 0 | 0 | 0 |
| price on a | 52cc87bd | | | | 9 | |
| on a handful | 4f8398fe | + | | | | |
| a handful of | 4f8398f6 | 2 | | | | |
| handful of our | 743ba46d | | | | | |
| of our high | 7b451587 | | H | | | |
| our high end | 89d97a75 | | | | | |
| high end replicas | 6ff630c6 | G | G | | | G |

[0075] 图 11 展示根据本发明的一些实施例的在安全服务器（也参见图 1）上执行的示范性组件。安全服务器 14 包括文档分类器 72，其连接到通信管理器 74 及指纹数据库 70。通信管理器 74 管理与客户端系统 16a 到 c 进行的垃圾邮件 / 诈骗检测事务，如上文关于图 4-A 到 B 所展示。在一些实施例中，文档分类器 72 经配置以经由通信管理器 74 而接收目标指示符 40，且确定指示目标文档 36 的分类的目标标签 50。

[0076] 在一些实施例中，分类目标文档 36 包括根据针对文档 36 所确定的文本指纹与参考指纹集合之间的比较而将文档 36 指派到文档类别，每一参考指纹指示文档类别。例如，

分类文档 36 可包含确定文档 36 是否为垃圾邮件及 / 或诈骗性，及确定文档 36 属于垃圾邮件 / 诈骗的子类别（例如，产品提供、网络钓鱼或尼日利亚诈骗）。为了分类文档 36，文档分类器 72 可结合指纹比较而使用所属领域中所知的任何方法。此类方法包含黑及白名单、图案匹配算法等等。例如，文档分类器 72 可计算多个个别得分，其中每一得分指示到特定文档类别（例如，垃圾邮件）的文档 36 的成员，每一得分是通过相异分类方法（例如，指纹比较、黑名单等等）而确定。分类器 72 可随后根据被确定为个别得分的复合得分而确定文档 36 的分类。

[0077] 文档分类器 72 可进一步包括指纹比较器 78（如图 12 中所展示），其经配置以通过比较目标文档的指纹与存储在数据库 70 中的参考指纹集合而分类目标文档 36。在一些实施例中，指纹数据库 70 包括针对参考文档集合所确定的文本指纹的存储库（例如，电子邮件消息、网页及网站评论等等）。数据库 70 可包括垃圾邮件 / 诈骗的指纹，但也包括合法文档的指纹。针对每一参考指纹，数据库 70 可存储相应指纹与文档类别（例如，垃圾邮件）之间的关联的指示符。

[0078] 在一些实施例中，数据库 70 中的参考指纹子集中的所有指纹具有在预定范围（例如，在 129 与 256 个字符之间）内的长度。此外，所述范围与由指纹计算器 56（图 6）针对目标文档所确定的目标指纹的长度范围相一致。此类配置（其中所有参考指纹具有大致相同的大小，且其中参考指纹所具有的长度大致等于目标指纹的长度）可促进用于文档分类目的的目标指纹与参考指纹之间的比较。

[0079] 针对每一参考指纹，数据库 70 的一些实施例可存储文本块的长度的指示符，针对所述文本块的长度而确定相应指纹。此类指示符的实例包含相应文本块的串长度、确定相应指纹时使用的片段长度，及放大 / 缩小因数等等。存储具有每一指纹的文本块长度的指示符可促进文档比较，这是通过使指纹比较器 78 能够选择性地检索表示在长度上与产生目标指纹 42 的文本块类似的文本块的参考指纹而实现。

[0080] 为了分类目标文档 36，分类器 72 可接收目标指示符 40，从指示符 40 提取目标指纹 42 且将指纹 42 转送到指纹比较器 78。比较器 78 可与数据库 70 进行接口连接，以选择性地检索用于与目标指纹 42 比较的参考指纹 82。在一些实施例中，指纹比较器 78 可优选地检索针对具有与目标文本块的长度类似的长度的文本块所计算的参考指纹。

[0081] 文档分类器 72 根据目标指纹 42 与从数据库 70 检索的参考指纹的比较而进一步确定目标文档 42 的分类。一些实施例中，所述比较包含计算指示指纹 42 与 82 的类似度的类似性得分。例如，此类似性得分可被确定为：

[0082] $$S = 1 - \frac{2d(f_T, f_R)}{|f_T| + |f_R|} \qquad [1]$$

[0083] 其中 $f_T$ 及 $f_R$ 分别表示目标指纹及参考指纹，$d(f_T, f_R)$ 表示两个指纹之间的编辑距离（例如，莱文斯坦 (Levenshtein) 距离），且其中 $|f_T|$ 及 $|f_R|$ 分别表示目标指纹及参考指纹的长度。得分 S 可取 0 与 1 之间的任何值，接近 1 的值指示两个指纹之间的高类似度。在示范性实施例中，当得分 S 超过预定阈值 T（例如，0.9）时，目标指纹 42 据称匹配于参考指纹 82。当目标指纹 42 匹配于来自数据库 70 的至少一个参考指纹时，文档分类器 72 可根据相应参考指纹的文档类别指示符而分类目标文档，且可制定目标标签 50 以反映所述分类。例如，当目标指纹 42 匹配于针对垃圾邮件消息所确定的参考指纹时，目标文档 36 可被分类

为垃圾邮件，且目标标签 50 可指示垃圾邮件分类。

[0084] 上文所描述的示范性系统及方法允许电子消息传递系统（例如，电子邮件及用户贡献网站）中的未经请求的通信（垃圾邮件）的检测，以及诈骗性电子文档（例如，网络钓鱼网站）的检测。在一些实施例中，针对每一目标文档而计算文本指纹，所述指纹包括根据相应文档的多个文本标记而确定的字符序列。所述指纹随后与针对文档集合所确定的参考指纹（包含垃圾邮件 / 诈骗性及合法文档）比较。当目标指纹与针对垃圾邮件 / 诈骗性消息所确定的参考指纹相匹配时，目标通信可被加标签为垃圾邮件 / 诈骗。

[0085] 当将目标通信肯定地识别为垃圾邮件 / 诈骗时，反垃圾邮件 / 反诈骗系统的组件可修改相应文档的显示。例如，一些实施例可阻止相应文档的显示（例如，不允许在网站上显示垃圾邮件评论），可在单独的位置（例如，垃圾电子邮件文件夹、单独的浏览器窗口）中显示相应文档，及 / 或可显示警报。

[0086] 在一些实施例中，文本标记可包含目标文本的个别字或字序列，以及电子邮件地址及 / 或网络地址（例如，包含于目标文档的文本部分中的统一资源定位符 (URL)）。本发明的一些实施例识别在目标文档内的多个此类文本标记。针对每一标记而计算散列，且根据相应散列而确定指纹片段。在一些实施例中，指纹片段随后通过（例如）级联而组合以产生相应文档的文本指纹。

[0087] 一些电子文档（例如，电子邮件消息）可在长度上有很大变化。在一些常规反垃圾邮件 / 反诈骗系统中，针对此类文档所确定的指纹的长度相应地变化。相比之下，在本发明的一些实施例中，文本指纹的长度约束在预定长度范围（例如，在 129 与 256 个字符之间）内，而不管目标文本块或文档的长度如何。使所有文本指纹在预定长度界限内可实质上改善消息间比较的效率。

[0088] 为了确定预定长度范围内的指纹，本发明的一些实施例使用放大及缩小方法。当文本块相对短时，通过调整指纹片段的长度而获得放大以产生所要长度的指纹。在示范性实施例中，30 位散列的每 6 个位可转换成一字符（使用（例如）Base64 表示），因此，相应散列可产生长度在 1 与 5 个字符之间的指纹片段。

[0089] 针对相对长的文本块，本发明的一些实施例通过从标记子集计算指纹而实现缩小，所述子集是根据散列选择准则而选择。示范性散列选择准则包括仅选择其散列能被整数 k（例如，2、3 或 6）整除的标记。针对给定实例，此类选择引起分别从可用标记的约 1/2、1/3 或 1/6 计算指纹。在一些实施例中，缩小可进一步包括将此类标记选择应用于多个聚合标记，其中每一聚合标记包括若干标记的级联（例如，相应电子文档的字序列）。

[0090] 各种散列函数可用于指纹片段的确定。在计算机实验中，将所属领域中所知的各种散列函数应用于从呈各种语言的电子邮件消息提取的 122,000 个字的集合，其目的是确定散列冲突（相异字产生相同散列）的数目，所述散列冲突为每一散列函数产生实际垃圾邮件。表 3 中说明的结果展示所属领域中称为 RSHash 的散列函数产生所有所测试的散列函数的最少冲突。

[0091] 表 3

[0092]

| 散列函数 | 32 位散列冲突 | 30 位散列冲突 |
|---|---|---|

| | | |
|---|---|---|
| RSHash | 0 | 4 |
| BKDRHash | 1 | 6 |
| SDBMHash | 2 | 7 |
| OneAtATimeHash | 2 | 6 |
| APHash | 4 | 6 |
| FNVHash | 7 | 10 |
| FNV1aHash | 7 | 10 |
| JSHash | 266 | 277 |
| DJBHash | 266 | 268 |
| DEKHash | 435 | 720 |
| PJWHash | 1687 | 1687 |
| ELFHash | 1687 | 1687 |
| BPHash | 61907 | 70909 |

[0093] 在另一计算机实验中，使用本发明的一些实施例来分析电子邮件消息集合（由企业服务器在一周期间接收的电子邮件的总量组成，且包括垃圾邮件及合法消息两者）。为了确定长度在129与256个字符之间的文本指纹，20.8％的消息要求无缩放，18.5％的消息要求2倍缩小，8.1％的消息要求3倍缩小，且8.7％的消息要求6倍缩小。在相同消息集合之中，14.8％的消息要求2倍放大，9.7％的消息要求4倍放大，且11.7％的消息要求8倍放大。以上结果表明在129到256个字符之间的指纹长度对于检测电子邮件垃圾邮件可为最佳的，这是因为根据放大及／或缩小因数而将实际电子邮件流分成群组的上述分割产生相对均匀填入的群组；此类情况针对指纹比较是有利的，这是因为可在大致相同的时间搜索所有群组。

[0094] 在另一计算机实验中，由遍及15小时而收集的大约865,000个消息组成的连续垃圾邮件流被分成消息集合，每一集合由在相异的10分钟间隔期间接收的消息组成。使用根据本发明的一些实施例而构造的文档分类器来分析每一消息集合（例如，参见图11到12）。针对每一消息集合，指纹数据库70由针对属于较早时间间隔的垃圾邮件消息所确定的指纹组成。图13中展示使用方程式[1]及阈值T＝0.75而获得的垃圾邮件检测率（实线），其与使用常规垃圾邮件检测方法（在所属领域中称为模糊散列）对相同消息集合而获得的垃圾邮件检测率（虚线）相比较。

[0095] 所属领域的技术人员将清楚，在不脱离本发明的范围的情况下，可以多种方式更改以上实施例。因此，本发明的范围应由所附权利要求书及其合法等效物确定。

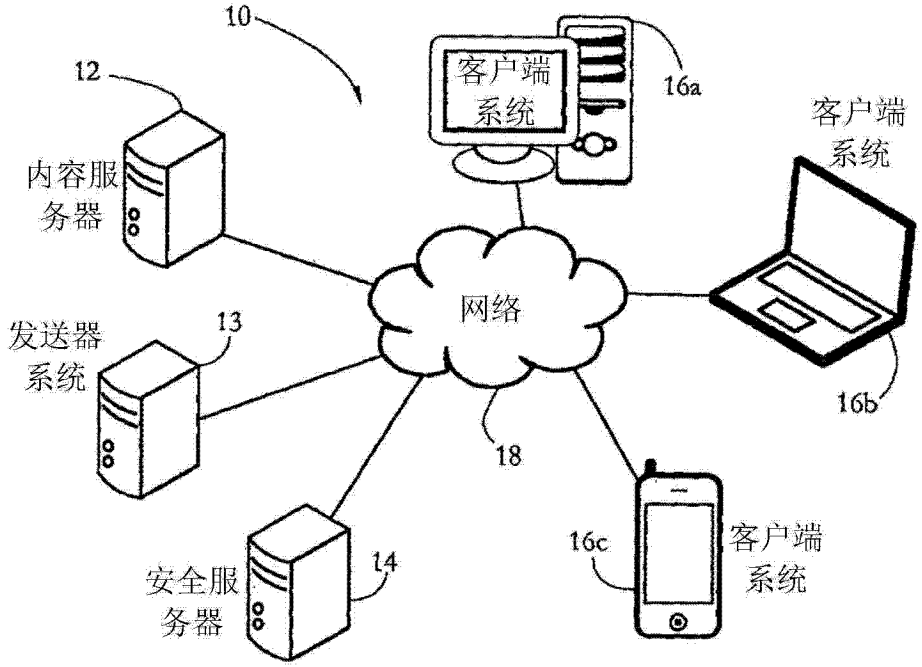

10
12
内容服务器
客户端系统
16a
客户端系统
网络
发送器系统
13
16b
18
安全服务器
14
16c
客户端系统

图 1

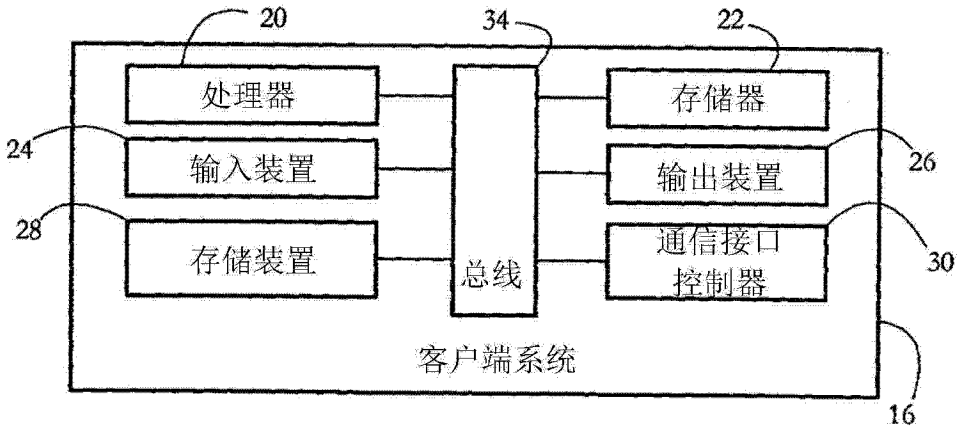

20
34
22
处理器
存储器
24
输入装置
输出装置
26
28
存储装置
总线
通信接口控制器
30
客户端系统
16

图 2-A

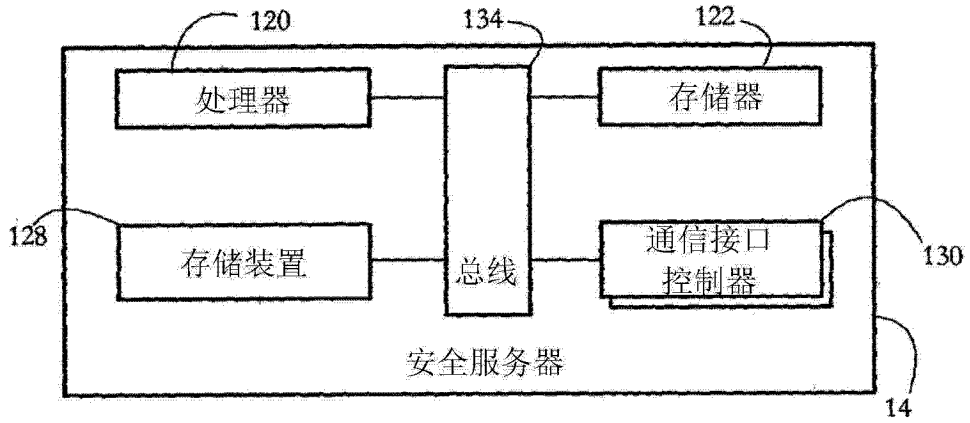

120
134
122
处理器
存储器
128
存储装置
总线
通信接口
控制器
130
安全服务器
14

图 2-B

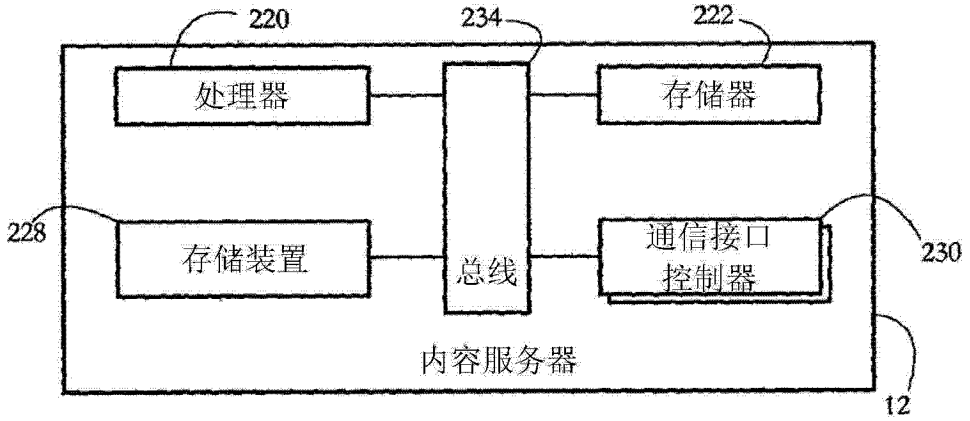

220
234
222
处理器
存储器
228
存储装置
总线
通信接口
控制器
230
内容服务器
12

图 2-C

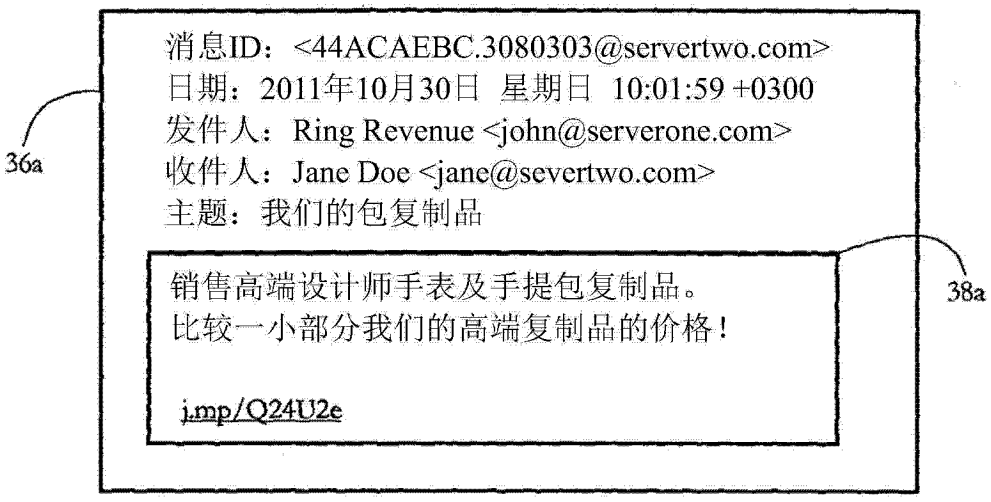

消息ID：<44ACAEBC.3080303@servertwo.com>
日期：2011年10月30日 星期日 10:01:59 +0300
发件人：Ring Revenue <john@serverone.com>
收件人：Jane Doe <jane@severtwo.com>
主题：我们的包复制品
销售高端设计师手表及手提包复制品。
比较一小部分我们的高端复制品的价格！
j.mp/Q24U2e
36a
38a

图 3-A

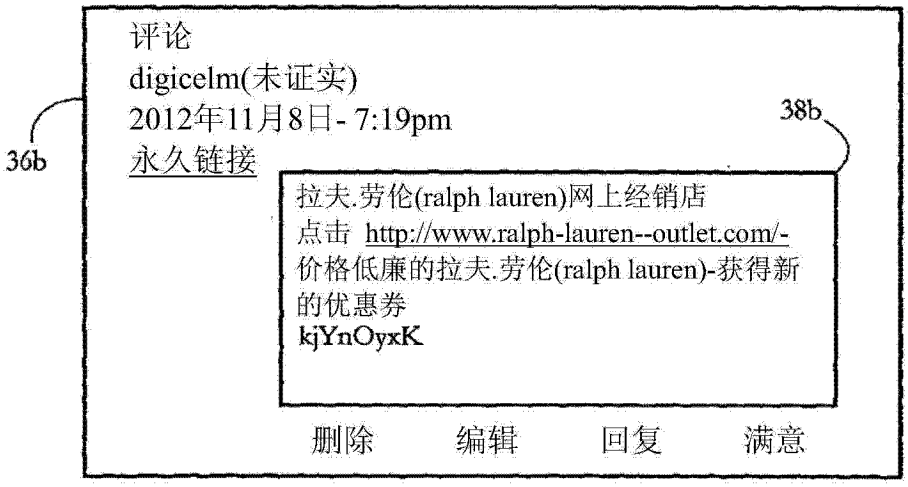

评论
digicelm(未证实)
2012年11月8日- 7:19pm
永久链接
36b
38b
拉夫.劳伦(ralph lauren)网上经销店
点击 http://www.ralph-lauren--outlet.com/-
价格低廉的拉夫.劳伦(ralph lauren)-获得新的优惠券
kjYnOyxK
删除
编辑
回复
满意

图 3-B

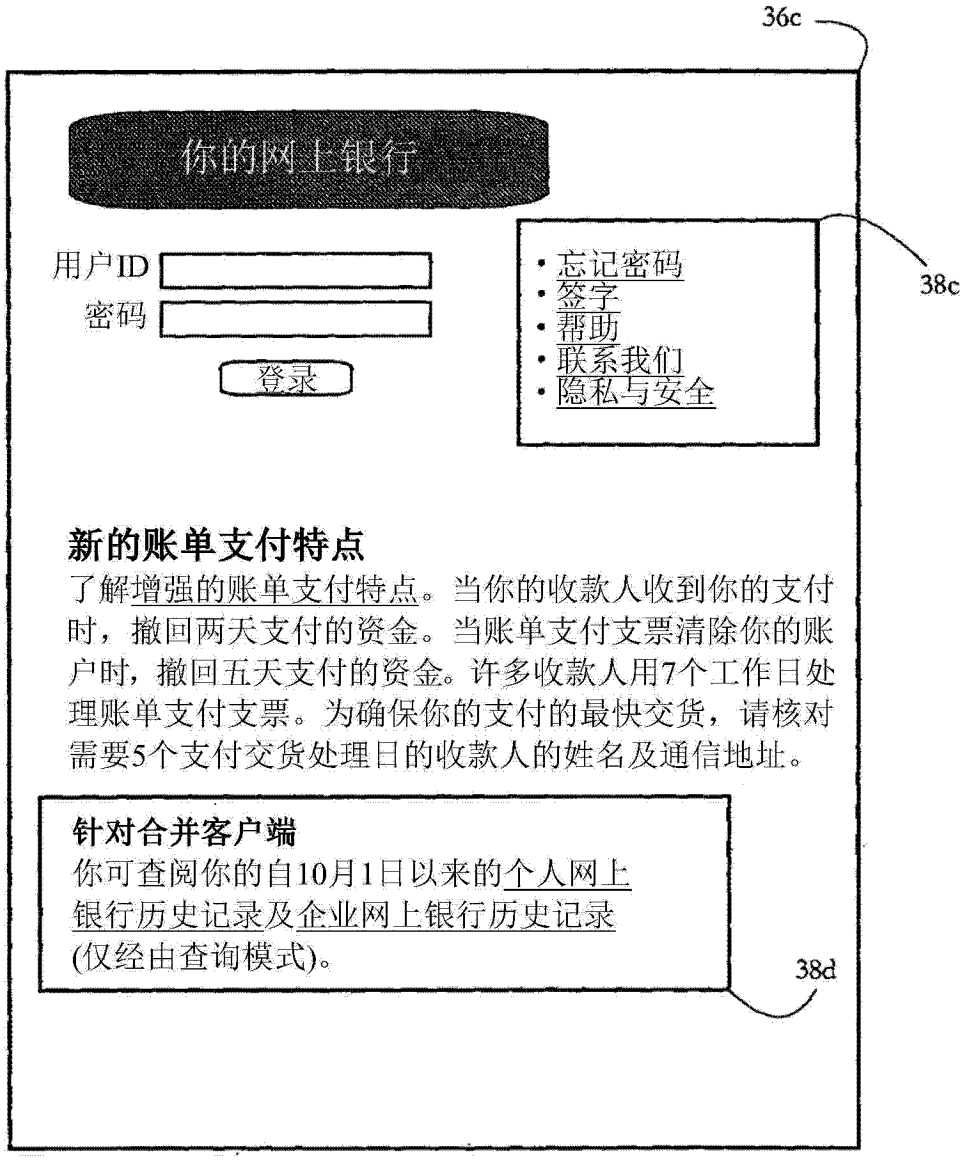

36c
你的网上银行
用户ID
密码
登录
• 忘记密码
• 签字
• 帮助
• 联系我们
• 隐私与安全
38c
新的账单支付特点
了解增强的账单支付特点。当你的收款人收到你的支付时，撤回两天支付的资金。当账单支付支票清除你的账户时，撤回五天支付的资金。许多收款人用7个工作日处理账单支付支票。为确保你的支付的最快交货，请核对需要5个支付交货处理日的收款人的姓名及通信地址。
针对合并客户端
你可查阅你的自10月1日以来的个人网上银行历史记录及企业网上银行历史记录(仅经由查询模式)。
38d

图 3-C

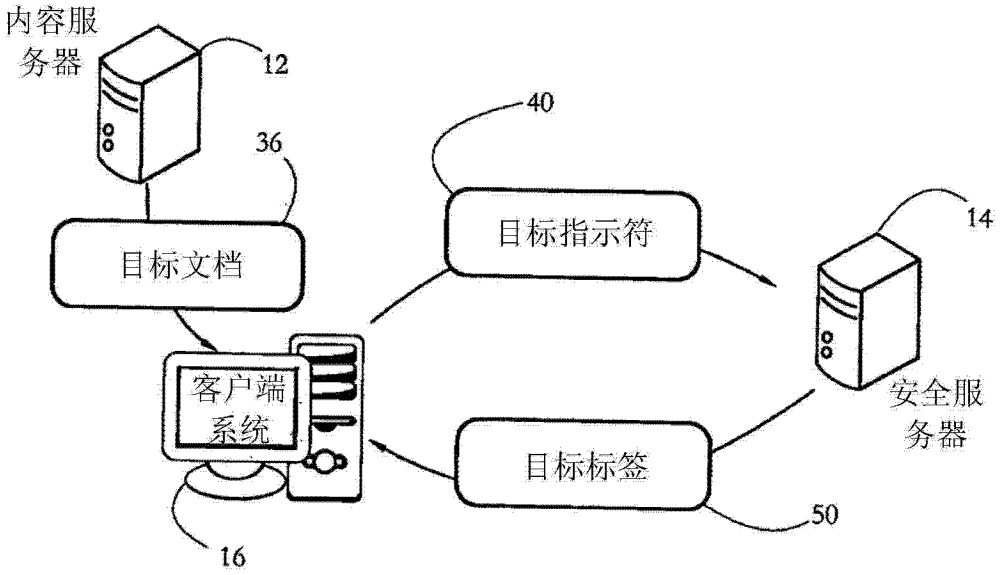

内容服务器
12
36
40
目标文档
目标指示符
14
客户端系统
安全服务器
目标标签
50
16

图 4-A

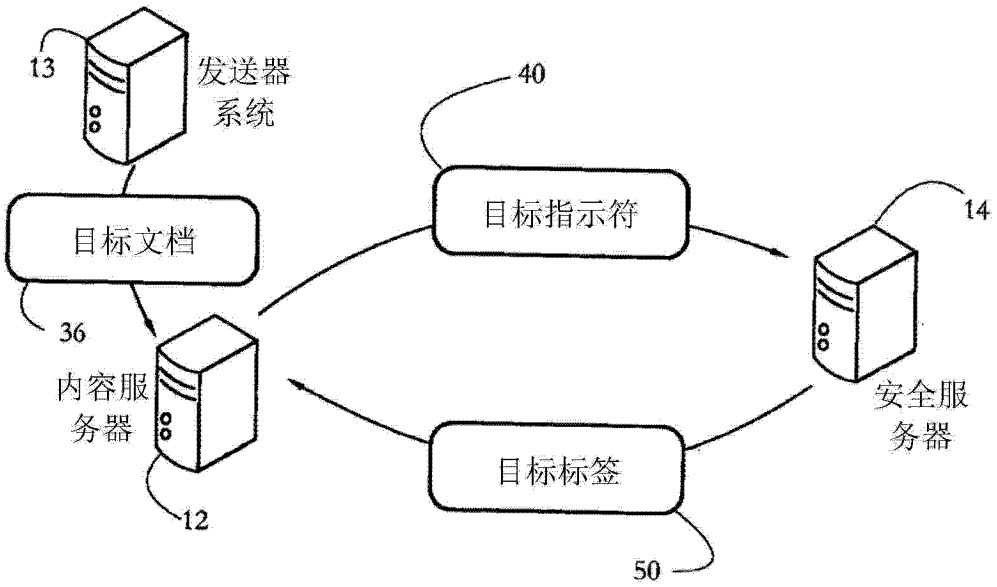

13
发送器系统
40
目标文档
目标指示符
14
36
内容服务器
安全服务器
目标标签
12
50

图 4-B

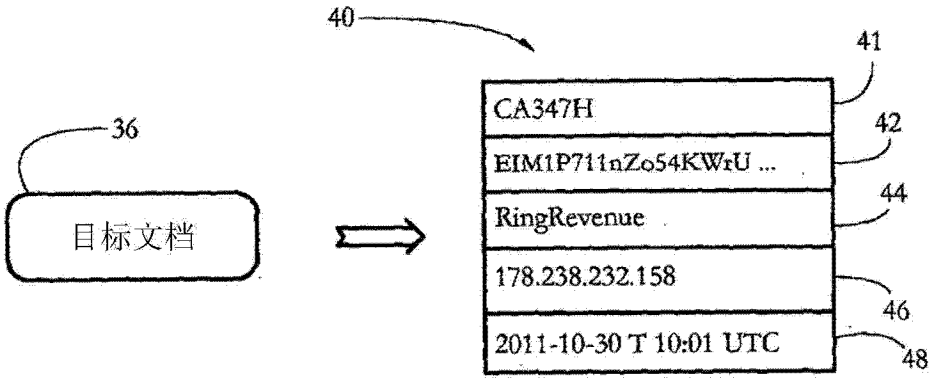

40
41
CA347H
42
EIM1P711nZo54KWrU ...
44
RingRevenue
178.238.232.158
46
2011-10-30 T 10:01 UTC
48
36
目标文档

图 5

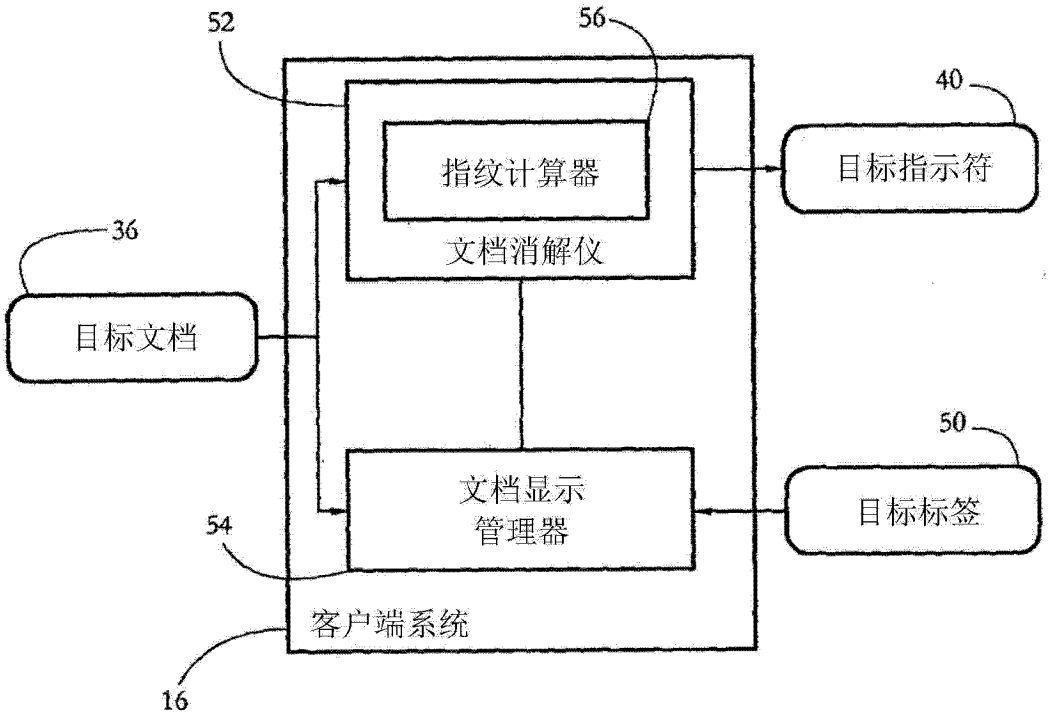

52
56
40
指纹计算器
目标指示符
文档消解仪
36
目标文档
50
文档显示
管理器
目标标签
54
客户端系统
16

图 6

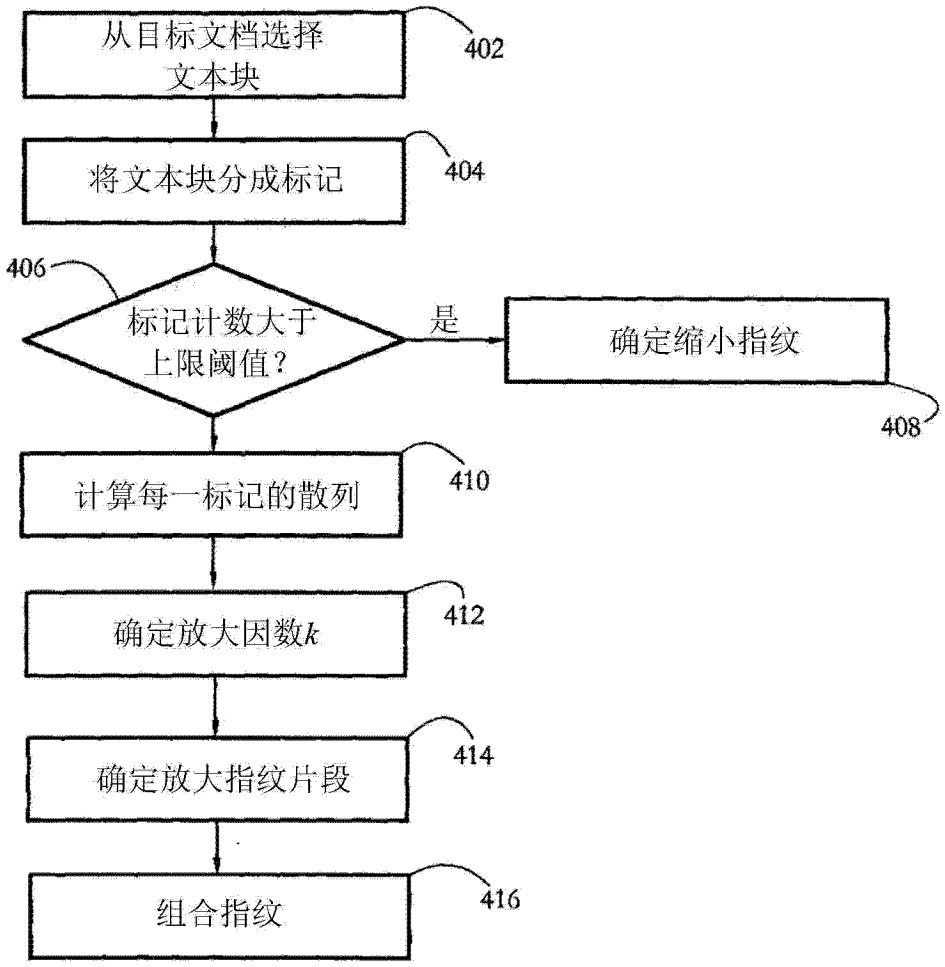

从目标文档选择文本块
402
将文本块分成标记
404
406
标记计数大于上限阈值?
是
确定缩小指纹
408
计算每一标记的散列
410
确定放大因数k
412
确定放大指纹片段
414
组合指纹
416

图 7

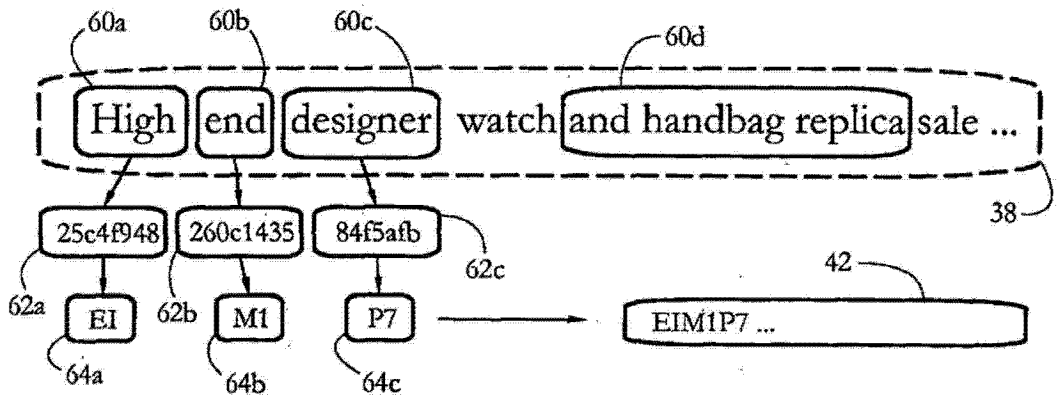

60a
60b
60c
60d
High
end
designer
watch
and handbag replica
sale ...
38
25c4f948
260c1435
84f5afb
62c
62a
62b
EI
M1
P7
64a
64b
64c
42
EIM1P7 ...

图 8

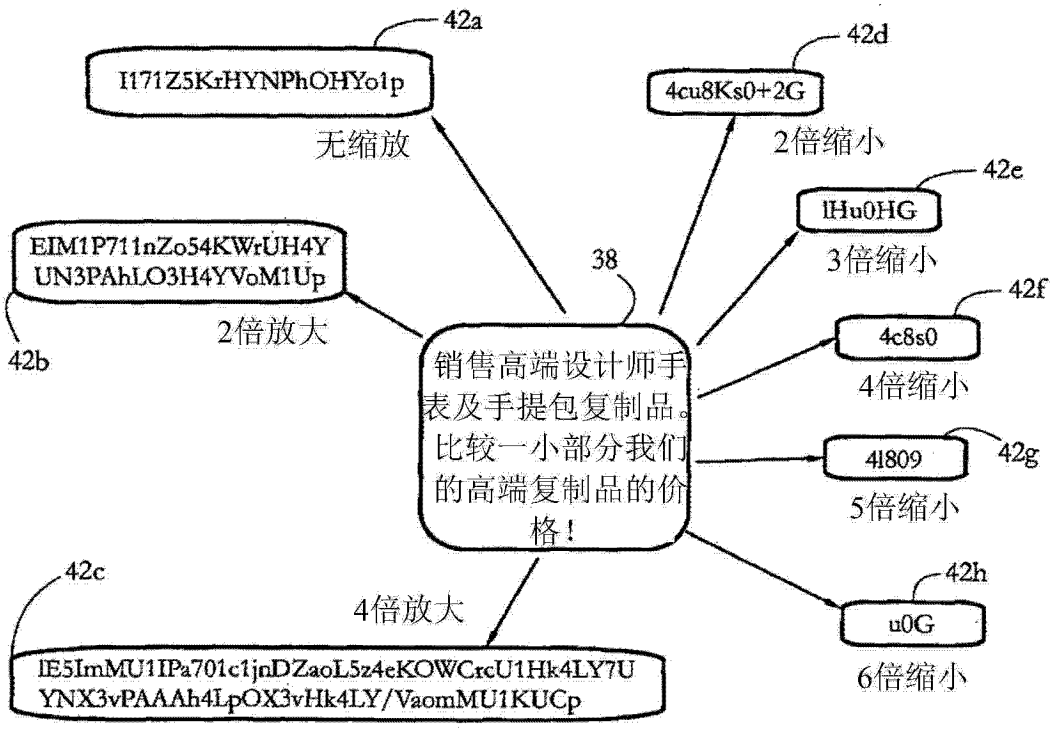

42a
I171Z5KrHYNPhOHYo1p
无缩放
42d
4cu8Ks0+2G
2倍缩小
42e
lHu0HG
3倍缩小
EIM1P711nZo54KWrUH4Y
UN3PAhLO3H4YVoM1Up
2倍放大
42b
38
销售高端设计师手表及手提包复制品。比较一小部分我们的高端复制品的价格！
42f
4c8s0
4倍缩小
41809
42g
5倍缩小
42c
4倍放大
lE5ImMU1IPa701c1jnDZaoL5z4eKOWCrcU1Hk4LY7U
YNX3vPAAAh4LpOX3vHk4LY/VaomMU1KUCp
42h
u0G
6倍缩小

图 9

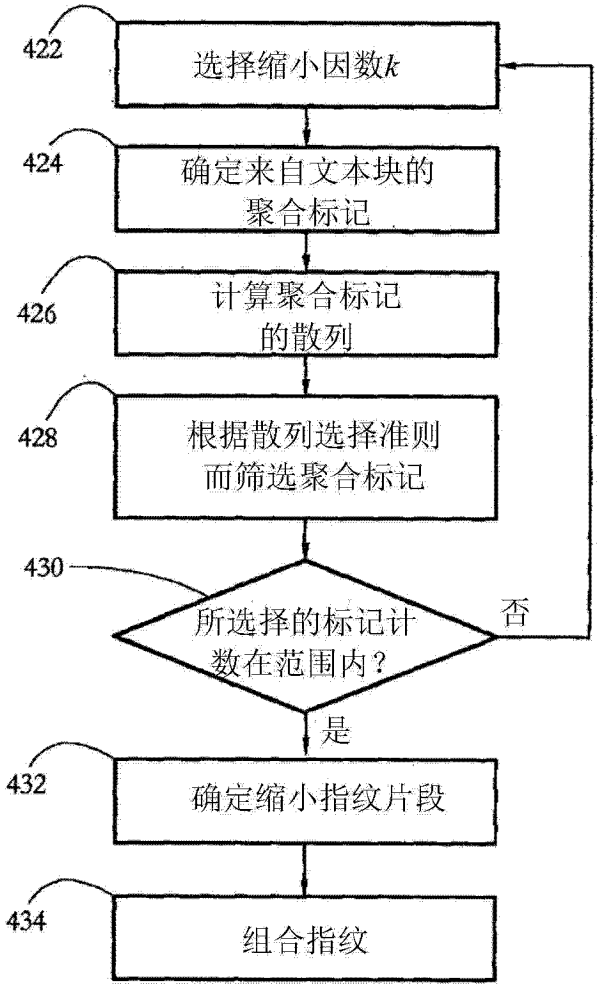

422
选择缩小因数k
424
确定来自文本块的聚合标记
426
计算聚合标记的散列
428
根据散列选择准则而筛选聚合标记
430
所选择的标记计数在范围内？
否
是
432
确定缩小指纹片段
434
组合指纹

图 10

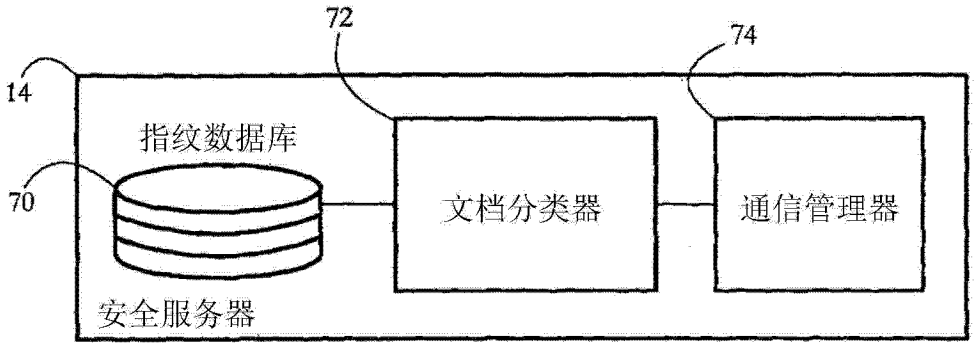

72
74
14
指纹数据库
70
文档分类器
通信管理器
安全服务器

图 11

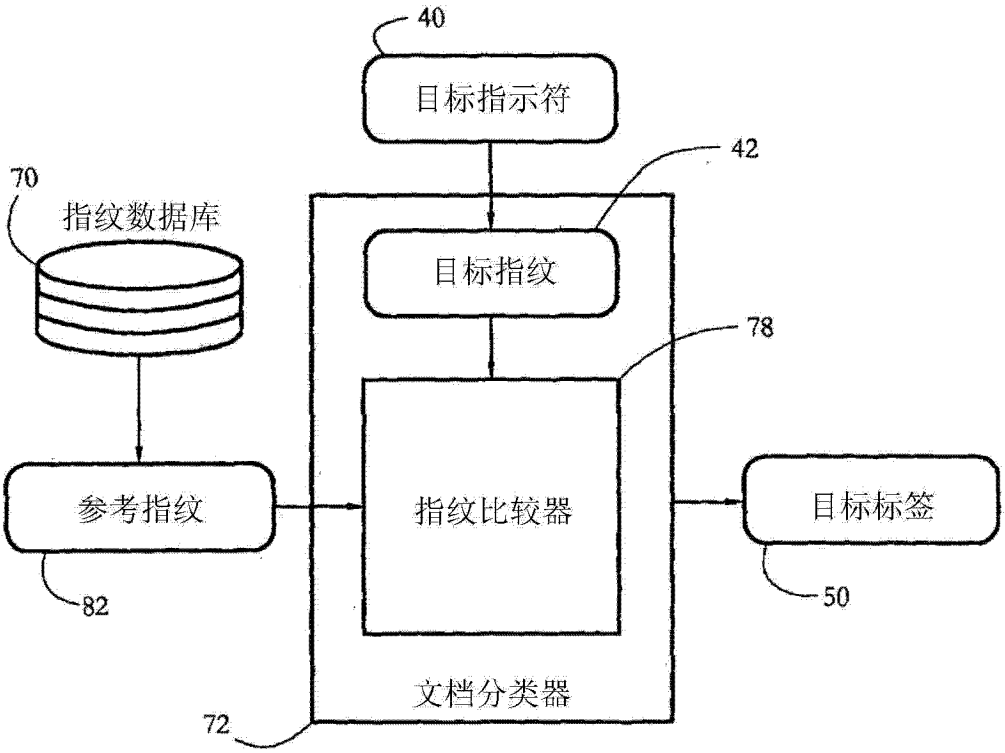

40
目标指示符
42
70
指纹数据库
目标指纹
78
参考指纹
指纹比较器
目标标签
82
50
文档分类器
72

图 12

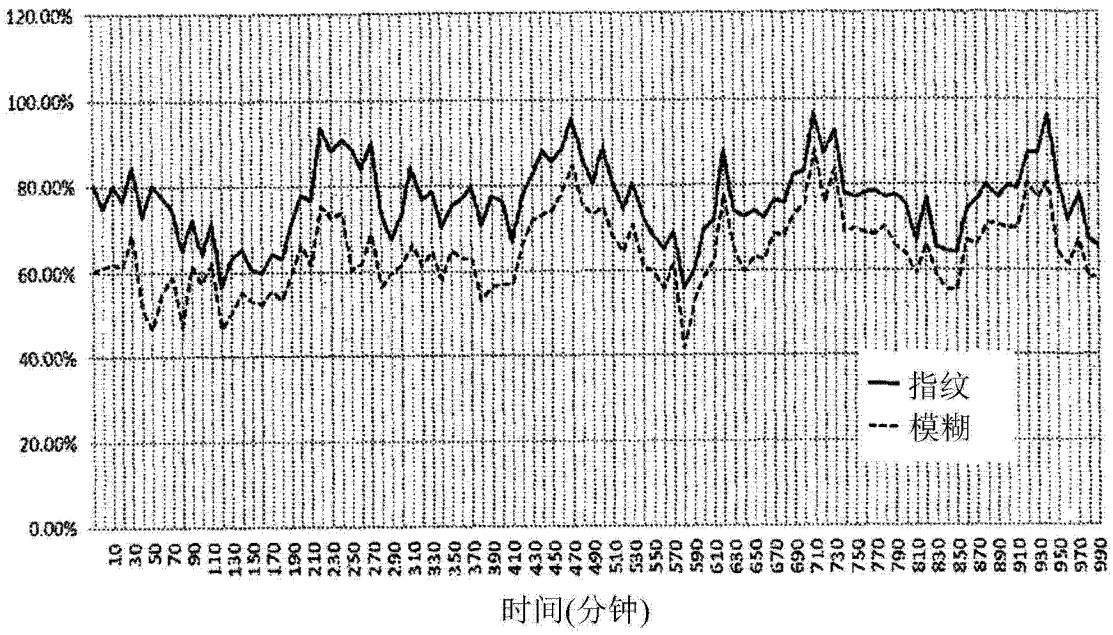

120.00%
100.00%
80.00%
60.00%
40.00%
20.00%
0.00%
指纹
模糊
时间(分钟)

图 13