



(12)发明专利

(10)授权公告号 CN 104781821 B

(45)授权公告日 2018.06.22

(21)申请号 201380058293.2

(22)申请日 2013.09.05

(65)同一申请的已公布的文献号
申请公布号 CN 104781821 A

(43)申请公布日 2015.07.15

(30)优先权数据
61/697916 2012.09.07 US

(85)PCT国际申请进入国家阶段日
2015.05.07

(86)PCT国际申请的申请数据
PCT/US2013/058163 2013.09.05

(87)PCT国际申请的公布数据
W02014/039620 EN 2014.03.13

(73)专利权人 克罗尔信息保证有限责任公司
地址 美国纽约州

(72)发明人 R.J.博贝克 A.乔普拉

(74)专利代理机构 中国专利代理(香港)有限公司
72001

代理人 张凌苗 张懿

(51)Int.Cl.
G06F 21/10(2006.01)
G06F 21/62(2006.01)

(56)对比文件
US 8032757 B1, 2011.10.04,
US 2007101139 A1, 2007.05.03,
CN 101341490 A, 2009.01.07,

审查员 冉凡坤

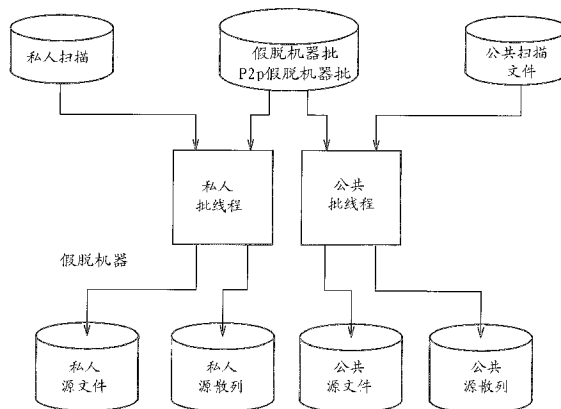
权利要求书1页 说明书10页 附图21页

(54)发明名称

文件共享网络中的片段匹配

(57)摘要

用于将表示私人文件的信息与从公共网络获得的文件进行匹配的技术。



1. 一种用于确定敏感私人信息是否已被泄漏给公共网络的系统,所述系统包括:
计算机,其包括至少一个处理器、存储器和网络接口;
私人数字文件,其包含由应用程序产生的敏感内容;以及
私人信息匹配处理部,其在所述处理器的存储器中执行并被配置为:
接收所述私人数字文件;

对所述私人数字文件进行处理以生成其片段部分,其中,片段部分中的每一个还包括数字文件的多个卷动组块,其中,第一片段部分包含包括所述文件中的字的第一集合的第一组块,第二片段部分包含包括所述文件中的字的第二集合的第二组块,其中,字的所述第二集合包括字的所述第一集合中的字中的至少一些但不是全部外加来自所述文件的一些其它字;

经由所述网络接口从公共网络接收多个公共数字文件;

对所述公共数字文件进行处理以生成其片段部分;以及

将所生成的所述私人数字文件的片段与所生成的所述公共数字文件的片段进行匹配,以确定所述私人数字文件的至少一些内容是否是连接到所述公共网络的其它计算机能访问的。

2. 根据权利要求1所述的系统,其中,所述私人信息匹配处理部还配置用于:计算所述私人数字文件和所述公共数字文件的散列。

3. 根据权利要求1所述的系统,其中,所述私人信息匹配处理部还配置用于:计算所述私人数字文件和所述公共数字文件的组块中每个组块的散列。

4. 根据权利要求3所述的系统,其中,所述私人信息匹配处理部还配置用于:将与至少一个公共数字文件的相应连续片段匹配的所述私人数字文件的连续片段聚合成匹配字的至少邻接块。

5. 根据权利要求1所述的系统,其中,所述私人信息匹配处理部还配置用于:经由用户接口使用百分比匹配的图形指示来显示私人数字文件和相匹配的公共数字文件的指示。

6. 根据权利要求1所述的系统,其中,所接收的所述私人数字文件仅包含省略至少一些信息的原始数字文件的内容的一部分,并且不包含所述原始数字文件的完整内容。

7. 根据权利要求4所述的系统,其中,所述私人信息匹配处理部还配置用于:将所述私人数字文件的片段与所述公共数字文件的片段进行匹配以忽略产生自所述匹配的指定条件。

8. 根据权利要求4所述的系统,其中,所述私人信息匹配处理部还配置用于:通过考虑片段类将所述私人数字文件的片段与所述公共数字文件的片段进行匹配。

9. 根据权利要求7所述的系统,其中,所述私人信息匹配处理部还配置用于:忽略所述公共数字文件中的重复片段。

文件共享网络中的片段匹配

[0001] 相关申请的交叉引用

[0002] 本申请要求于2012年9月7日提交的序列号为No.61/697,916的美国临时专利申请的优先权和权益。

技术领域

[0003] 本申请涉及确定某些信息是否在计算机网络中共享。

背景技术

[0004] 文件共享是分配或提供对诸如计算机程序、多媒体(音频、图像和视频)、文档或电子书的数字存储的信息的访问的实践。共享机制可以包括:集中式服务器、以万维网为基础的超链接文档,或文件共享网络的使用。可以使用各种方式来实现共享网络,诸如使用对等技术、比特洪流技术、文件托管服务等。

[0005] 文件共享继续位列最流行的因特网应用之一。从数千或数百万用户汇集资源的能力使得文件共享对于许多应用来说非常有吸引力。然而,这样的便利性和对信息的快速可访问性并非没有其风险。具体而言,不小心或无意地共享私人文件的用户可能发现个人和其它敏感信息被世界各地的其它用户迅速下载。

[0006] 大多数企业收集并存储关于他们的员工和客户的敏感信息,诸如社会安全号码、信用卡和账户信息、医疗和其它个人数据。他们中的许多有保护该信息以免意外公开的法律义务。如果这样的信息落入不法分子手中,它可能导致欺诈和身份盗用。使用P2P文件共享软件的人可能最终会无意地共享文件。他们可能会不小心选择共享包含敏感信息的驱动器和文件夹,或者他们可能错误地把私人文件保存到共享驱动器或文件夹,从而使私人文件对于他人为可用。此外,病毒和其它恶意软件可以改变对设计用于共享的驱动器和文件夹的访问,从而也把私人文件置于危险之中。结果,取代如意图的那样简单地共享他们的音乐文件的是,诸如纳税记录、私人医疗记录、工作文档等等的其它敏感信息最终被经由文件共享网络上的一般流通而可获得。

[0007] 对于企业以及终端用户来说该风险非常高。例如,美国联邦贸易委员会(FTC)最近宣布了针对通过允许在对等(P2P)网络上共享他们的客户的敏感个人信息而非法暴露了其的多家公司的和解。这些执法行动指出了数据隐私和安全政策的不足或缺乏造成的严重影响。

[0008] 存在用于雇佣的可以在组织中定位敏感数据并确定可以经由文件共享网络来获得什么种类的对其访问的审计服务。在政府和军事终端用途中,可以使用用于对数据的敏感性(诸如“秘密”、“绝密”等)进行分类的深度标准。这些分类详细规定了谁可以访问该信息,以及应该实施什么级别的安全保证以防止无意的公开。

[0009] 当企图在文件共享网络上对包括敏感信息的私人文件进行定位时,会出现若干问题。信息的所有者或保管人想要知道他们的文件是否正在被共享,并且甚至还想知道文件片段是否正在被共享。例如,即使少量的信用卡号被暴露,一长串信用卡号也可能会受到损

害。此外,可以对敏感信息进行重新排列或与其它信息组合以使其模糊。另外,敏感内容可以在多个文件之间进行分割。除了私人文件可以包含机密或其它高度敏感的信息之外,信息的保管人还希望能够利用商业服务本身来定位信息,但不要将其完全公开。

发明内容

[0010] 本公开内容提供了将私人文件与经由公共网络(诸如web服务器、P2P网络、比特洪流等)可获得的文件进行匹配以确定私人文件的信息内容是否已经被泄漏。工具集合对从私人文件获得的信息或关于私人文件的信息的段进行操作,其提供了包括下列各项的多个优点:更大的处理吞吐量、处理不同类型内容的能力、以及在不公开信息本身的情况下搜索机密信息的能力。

[0011] 在特定实施例中,提供了用于将私人文件的段与文件共享网络上可获得的公共文件进行匹配的技术。例如,过程至少利用例如切段器工具、匹配器工具和后匹配工具。

[0012] 切段器工具将文件的内容提取成字的流,并将该流切分成被称为片段的具有可配置大小的卷动组块。例如,给定50个字的流中大小为25的片段,切段器工具将50个字的流切分成26个片段,每个片段的长度为25个字。然后针对每个片段计算散列。

[0013] 匹配器工具加载针对所有私人文件的所生成的所有片段计算出的所有散列。针对位于文件共享网络上的每个公共文件,然后使用与上述相同的片段处理来生成片段和散列的集合。为了执行匹配,将公共文件的所产生的散列与私人散列的图进行比较。然后,将匹配过程的结果(诸如匹配文件的列表)持久保存到例如数据库。

[0014] 后匹配工具检查所产生的所有匹配文件,并将连续匹配片段聚合成匹配字的邻接块。然后,将结果持久保存到例如数据库。然后可以经由匹配评估用户接口(例如可以呈现给人类分析师)来检查邻接匹配块,以获得关于文件之间的特定匹配的更多细节。

[0015] 在特定实施例中,用户接口可以允许人类分析师针对位于一个或多个文件共享网络上的公共文件的目录启动匹配/片段过程。匹配器针对公共文件执行相同的片段过程,并且然后将来自该结果的概要信息持久保存到数据库。

[0016] 片段匹配评估用户接口还可以允许分析师检查两个文件之间的匹配。例如,用户接口可以使用在左边示出的私人信息和在右边示出的公共信息来呈现匹配的并排视图。可以在滚动列表中显示匹配文件的列表。当用户选择私人文件和公共文件时,匹配视图可以呈现例如所找到的匹配百分比的概要。可以使用诸如红、绿和黄的颜色或者没有这些颜色来指示匹配度。

[0017] 匹配过程还可以检查所有的匹配片段,并确定匹配文件之间的连续匹配区段的块。诸如可以在屏幕的一部分上的滚动列表中呈现每个块的预览。可以对块进行排序,诸如,例如,找到的具有最高数量文件的匹配块被排序在第一。

[0018] 可以向实现提供进一步的修饰。例如,如果私人文件包含高度敏感信息,那么该私人文件的所有者可能不希望向外部服务提供商提供该私人文件的完整复制。所有者可以代之以只提供他们寻求定位的该私人文件的段,或者甚至可以只向工具提供散列信息。

[0019] 片段大小可由分析师确定或者经由启发法确定。例如,分析师可应用其它启发法来专注于哪些信息是最重要的,通过自动或手动过程。

附图说明

- [0020] 下面的发明描述参照附图。
- [0021] 图1是确定私人信息是否被泄漏给了公共网络的系统的高级架构。
- [0022] 图2更加详细地示出了扫描器组件。
- [0023] 图3是示例公共/私人扫描表。
- [0024] 图4示出了假脱机器(spooler)组件。
- [0025] 图5是示例假脱机器批表。
- [0026] 图6示出了切段器组件。
- [0027] 图7是示例切段器批表。
- [0028] 图8示出了用于对文件进行“切段”的提取过程。
- [0029] 图9示出了匹配器组件。
- [0030] 图10是所存储的关于匹配文件的示例概要信息。
- [0031] 图11是总文件匹配报告。
- [0032] 图12是文件匹配报告。
- [0033] 图13是分类工具。
- [0034] 图14是并排匹配评估工具。
- [0035] 图15是匹配块评估工具。

具体实施方式

- [0036] 概述
- [0037] 本系统提供了用于将包含敏感信息的私人文件与从公共网络下载的文件进行匹配的数据处理工具的集合。这些工具确定私人文件的内容中的全部或部分是否已经被泄漏给公共网络。该工具的集合通常至少包括切段器、匹配器和后匹配(PostMatch)元件。
- [0038] 切段器
- [0039] 切段器工具将文件的内容提取成字的流,并将该流切分成具有可配置大小的卷动块(即,“片段”)。例如,给定50个字的流以及大小为25的片段,切段器将把50个字切分成长度为25个字的26个片段[长度:]
- [0040] 片段1 字1-25
- [0041] 片段2 字2-26
- [0042]
- [0043] 片段26 字26-50
- [0044] 针对每个片段计算散列,以便促进片段随后的匹配。
- [0045] 匹配器
- [0046] 匹配器工具将针对所有私人文件的所生成的所有片段计算出的所有散列加载到基于存储器的图中。针对每个公共文件,然后也使用与上述相同的片段处理来生成片段和散列的集合。
- [0047] 为了执行匹配,将公共文件的散列与私人散列的图进行比较。然后,将匹配过程的结果(例如,匹配文件的完整列表)置于诸如数据库的持久性存储设备中。

[0048] 后匹配

[0049] 后匹配工具检查所有的持久性匹配文件,并将连续匹配片段聚合成匹配字的邻接块。将该过程的结果持久保存在.match文件中(到数据库或文件系统)。邻接匹配块然后由匹配评估使用。用户接口可以向分析师呈现匹配信息以获得关于文件之间的具体匹配的更多细节。

[0050] 架构

[0051] 图1示出了高级系统架构。

[0052] 扫描器组件负责从文件系统收集文件(私人 and 公共文件)用来由其余组件进行处理。私人文件可以从系统的用户获得,或者由客户提交给代表其它人操作该系统的服务提供商。私人文件将通常包括表示具有敏感个人信息的人类可读文档的数字编码的信息,并且可由诸如字处理器、电子表格程序、幻灯片演示程序、数据库、web浏览器、执行过程输出捕捉或创建输出数字文件的类似应用的应用程序来生成。

[0053] 通过搜索公共文件网络来定位公共文件。这些公共文件网络可以包括:文件服务器、web服务器、对等(P2P)网络、比特洪流网络以及允许信息共享的任何其它技术。其它协议引擎机(本文中未示出)可以连续扫描P2P、比特洪流和其它网络来定位和存储大量公共文件以便由系统在以后使用。

[0054] 假脱机器处理文件由扫描器组件提供。假脱机器将每个文件复制到暂存区域,并计算文件的SHA-1(或其它散列)。一旦计算出SHA-1,则在假脱机散列目录中将文件重命名为基于SHA-1的名称。针对每个经假脱机的文件向数据库添加条目。

[0055] 切段器然后将由假脱机器提供的每个文件的内容提取成字的流,将这些字切分成卷动片段的集合,并针对每个片段计算散列。将相关联的散列和字写入文件系统,并将概要信息写入数据库。

[0056] 匹配器通过将每个私人文件的散列与每个公共文件的散列进行比较来计算匹配。将这些集合匹配的结果写到数据库。

[0057] 后匹配组件收集关于每个匹配文件的细节,将匹配片段聚合成匹配邻接块,并执行启发法来对邻接块进行自动分类。

[0058] 图1的架构寻求实现以下。

[0059] 高吞吐量。高吞吐量是通过并行地运行处理器来提供的。在扫描新文件的同时,可以对之前扫描的文件进行假脱机。在对新文件进行假脱机的同时,可以对之前假脱机的文件进行切段。在对新文件进行切段的同时,可以对之前切段的文件进行匹配。在对新文件进行匹配的同时,可由后匹配器对之前匹配的文件进行处理。存储器的广泛使用(尤其是在匹配器组件中)也对最高可能吞吐量有所贡献。

[0060] 可扩展性。每个处理器处理成批的工作。用于每个进程的线程的数量是通过配置设置的,通过增加用于进程的线程的数量和/或使用多个机器运行额外处理器来提供扩展。

[0061] 扫描器(图2)

[0062] 扫描器组件负责收集文件以便由其余组件进行分析。为了从目录收集文件,分析师使用配置图形用户接口(GUI)将该目录添加到扫描器目录的列表。

[0063] 扫描器针对每个所配置的目录[SourceF2f]创建一个线程。每个扫描器线程对其所分配的目录进行扫描以寻找新的或经修改的文件。如果文件不存在于私人/公共扫描文

件表或公共/私人源文件表中,则认为文件是新的。扫描过程填充公共/私人扫描文件表(图3)。当假脱机器对所扫描的文件进行假脱机时,假脱机器将填充公共/私人源文件表。

[0064] 最后访问时间属性保存到上述表并用于确定文件是否已经被修改。经修改的文件将在扫描器的下一次运行中被扫描,并将添加另一个条目到扫描文件表中。

[0065] 假脱机器(图4)

[0066] 假脱机器组件负责收集扫描文件以便由其余组件进行分析。将被收集的文件是由扫描器组件持久保存的文件的列表。对来自私人扫描表和公共扫描表的所有文件进行处理。

[0067] 假脱机器对来自公共/私人扫描文件表的成批文件[假脱机器批表](图5)或来自P2P或其它协议引擎的成批文件[P2p假脱机器批]进行处理。假脱机器将每个文件复制到暂存区域,并计算文件的SHA-1。一旦计算出SHA-1,则在假脱机器目录中将文件重命名为基于SHA-1的名称。针对每个文件向公共/私人源文件表中添加条目,并且如果该文件唯一,那么将向私人/公共源散列表添加条目。

[0068] 切段器(图6)

[0069] 切段器组件负责对所收集的文件进行处理以便由假脱机器组件进行分析。一旦文件被正确地假脱机到系统中,切段器将提取每个文件的内容,并将相关联的散列和字按照由该配置所配置的片段散列和片段字目录的指示写入文件系统。

[0070] 切段器对来自私人/公共源散列表的成批文件[切段器批表](图7)进行处理。切段器针对该批中的每个私人/公共源散列文件生成散列文件和字文件。还针对该批中的每个私人/公共源散列文件向私人/公共片段文件表文件添加条目。

[0071] 如果批包含没有正确分段的文件,那么将针对每个失败的文件在失败的私人/公共片段文件表中添加条目。

[0072] 提取(图8)

[0073] 对文件进行“分段”的第一步是以字的流的形式从每个文件中提取内容。将所有的格式化和标点符号删除。

[0074] 分段过程中的第一步是提取。提取器读取假脱机文件的内容并生成字的流。然后将字的流分离成连续字的较小的列表,其被称为“片段”。“片段大小”确定了“片段”将包含多少字。为了确保最完整的匹配,将字的流分离成“滚动片段”。滚动片段是偏移了一个字的连续字的小列表一见下文:

[0075] 文件[片段大小 = 5]

[0076] The quick brown fox jumps over the lazy dog.

[0077] 五个滚动片段:

[0078] The quick brown fox jumps

[0079] quick brown fox jumps over

[0080] brown fox jumps over the

[0081] fox jumps over the lazy

[0082] jumps over the lazy dog

[0083] 针对每个滚动片段,计算散列以优化匹配过程。将散列保存到.hash文件,并将字的完整文件保存到.words文件。将关于每个文件的概要信息写入数据库。

[0084] 匹配器(图9)

[0085] 匹配器组件负责对由切段器组件创建的散列进行匹配。

[0086] 匹配器加载来自私人片段文件的散列。基于散列限制配置,匹配器将把私人片段文件散列切分成批,以对由匹配器使用的存储器的量进行管理。

[0087] 一旦加载了私人文件散列,就将公共文件收集成批。一次对一批公共文件进行处理,匹配器加载批中的每个公共文件的散列,并将这些散列与之前加载的私人文件散列进行比较。将关于匹配文件的每个集合的概要信息保存到数据库(图10)。

[0088] 后匹配器

[0089] 后匹配组件负责针对由匹配器组件检测到的匹配的文件执行额外的处理。后匹配过程将私人与公共文件之间的连续的匹配片段聚合成邻接块。针对每个匹配私人和公共文件,将确定一个或多个邻接块[见下文的示例]。

[0090] 私人文件[片段大小 = 5]

[0091] 内容=.....The quick brown fox jumps over the lazy dog.....

[0092] 公共文件[片段大小 = 5]

[0093] 内容=.....The quick brown fox jumps over the lazy dog.....
.....The quick brown fox jumps over the lazy dog.....
.....The quick brown fox jumps over the lazy dog

[0094]

[0095] 私人文件匹配片段[片段大小 = 5]

[0096] The quick brown fox jumps

[0097] quick brown fox jumps over

[0098] brown fox jumps over the

[0099] fox jumps over the lazy

[0100] jumps over the lazy dog

[0101] 私人文件邻接块

[0102] The quick brown fox jumps over the lazy dog, 起始偏移=x, 结尾偏移=y, 其它偏移 =z

[0103] 公共文件匹配片段[片段大小 = 5]

[0104] The quick brown fox jumps [3个实例]

[0105] quick brown fox jumps over [3个实例]

[0106] brown fox jumps over the [3个实例]

[0107] fox jumps over the lazy [3个实例]

[0108] jumps over the lazy dog [3个实例]

[0109] 公共文件邻接块

[0110] The quick brown fox jumps over the lazy dog, 起始偏移=a, 结尾偏移=b, 其它偏移 =z

[0111] The quick brown fox jumps over the lazy dog, 起始偏移=d, 结尾偏移=e, 其它偏移 =z

[0112] The quick brown fox jumps over the lazy dog, 起始偏移=f, 结尾偏移=g, 其它偏移 =z。

[0113] 分类

[0114] 分类用于驱动用于匹配的处理的工作流。例如,可以识别并立即处理高优先级匹配,或者可以从进一步的处理过滤出低优先级匹配。匹配的分类通过向邻接块中的每个块分配标签开始。达尔文(Darwin)产品为匹配结果的处理提供报告和GUI。大多数报告和工具使用相关联的分类提供用于移除或显示匹配的过滤机制。启发法(见下文)可用于向邻接块自动分配分类。

[0115] 升级

[0116] 一旦已经针对匹配向所有的邻接块分配了标签,对每个块的分类可以基于特定分类的升级属性(全部(ALL)、任何(ANY)、无(NONE))升级到匹配水平。

[0117] 例如,“忽略”分类的升级属性是“全部”。如果用于匹配的全部邻接块(从私人文件和公共文件二者的角度来看)添加标签有“忽略”标签,那么“全部”升级属性指引后匹配器仅向整个匹配添加“忽略”标签。

[0118] “高优先级”分类的升级属性是“任何”。如果用于匹配的任何邻接块(从私人文件和公共文件二者的角度来看)添加标签有“高优先级”标签,那么“任何”升级属性指引后匹配器向整个匹配添加“高优先级”标签。

[0119] 升级属性“无”指引后匹配处理绕过对该分类的升级。

[0120] 启发法

[0121] 存在可由后匹配处理器用来对匹配进行自动分类的多种启发法。例如,重复序列启发法向具有重复序列资格的任何匹配添加“重复序列”和“忽略”标签。电子邮件签名启发法向具有电子邮件签名资格的匹配添加“电子邮件签名”标签。

[0122] 匹配评估

[0123] 系统提供了多个报告和工具以协助分析师评估匹配结果。这些工具列出如下:

[0124] 总文件匹配报告(图11)

[0125] 该报告示出是彼此的完全复制(即,在二进制级别上匹配)的文件。

[0126] 文件匹配报告(图12)

[0127] 该报告示出具有匹配片段(即,至少一个匹配片段)的所有文件。

[0128] 分类工具(图13)

[0129] 分析师使用该分类工具可以向块预分配标签。例如,分析师可以向法律声明、共同页眉和页脚、问候语等分配“忽略”标签。在UI和报告中,分析师可以容易地过滤或抑制具有“忽略”标签的匹配以便关注更加重要的匹配结果。

[0130] 并排匹配评估工具(图14)

[0131] 使用并排匹配评估工具,分析师还可以向邻接块分配分类。

[0132] 匹配块评估工具(图15)

[0133] 使用匹配块评估工具,分析师可以向匹配邻接块分配分类以及定义新的分类。

[0134] 连续改进

[0135] 初始的匹配处理可以导致仅在少量片段上匹配的大量匹配。这些匹配中的大多数可以被认为是“误报”或“噪声”,因为这种匹配涉及常用短语、页眉、页脚等。当分析师使用

“忽略”分类来为这些匹配(即,邻接块)添加标签时,每次该块出现在两个文件之间的匹配中,该块携带该分类。基于分类的升级属性,也可以对于匹配为分类添加标签。随着时间的推移,将用“忽略”标签来对“噪声”中的绝大多数预先分类。随着越来越多的“噪声”被过滤出,处理匹配结果的分析师将处理越来越多百分比的真匹配。

[0136] 匹配处理的效率不断提升。通过使用分类工具来识别最普遍的常用短语、页眉、页脚等,分析师可以获得该效率增益的起跳点。使用相同的工具,可以识别高优先级块以便快速跟踪包含高优先级块(“黄金片段”)的匹配的处理。

[0137] 总文件匹配报告(图11)

[0138] 该报告示出了每个匹配文件的完整文件路径和sha 1。

[0139] 文件匹配报告(图12)

[0140] 该报告示出了每个匹配文件的完整文件路径、文件大小、字和匹配百分比。

[0141] 分类工具(图13)

[0142] 分类工具取得由分析师选择的文件,并将文件的内容提取成字的流。然后分析师可以从内容中选择一系列字或片段。一旦选择了内容的块,分析师可以向块分配一个或多个分类。分析师还可以根据需要添加新的分类。

[0143] 并排匹配评估工具(图14)

[0144] 并排匹配评估工具允许分析师检查两个文件之间的匹配。在HTML查看器中对匹配加高亮。UI使用在左边示出的私人信息和在右边示出的公共信息来呈现匹配的并排视图。

[0145] 在屏幕左上方的滚动列表中的示出匹配私人文件的列表。在屏幕右上方的滚动列表中呈现与所选择的私人文件相匹配的公共文件的列表。从左边的列表中选择私人文件将驱动匹配公共文件在右边的填充以及第一匹配公共文件的自动选择。

[0146] 一旦用户选择了私人文件和公共文件,仅在文件列表下方呈现的匹配概览HTML视图和在屏幕底部呈现的匹配细节将更新。概览HTML视图呈现匹配文件的概览。每个字符表示文件中的片段。如果片段不与相应文件中的任何其它片段匹配,那么示出“.”字符。如果整个片段与相应文件中的一片匹配,那么示出“*”字符。如果片段的部分匹配片段中的百分比数量的字,那么该匹配如下概括示出:

[0147] 0 少于10%的匹配

[0148] 1 10-19%的匹配

[0149] 2 20-29%的匹配

[0150] 3 30-39%的匹配

[0151] 4 40-49%的匹配

[0152] 5 50-59%的匹配

[0153] 6 60-69%的匹配

[0154] 7 70-79%的匹配

[0155] 8 80-89%的匹配

[0156] 9 90-99%的匹配。

[0157] 匹配细节示出了从私人文件和公共文件提取的字的列表。每个邻接块匹配用绿色加高亮。块是一个或多个匹配片段。将连续匹配片段组织成块以便协助分析师审查匹配。当前块用黄色加高亮。非匹配字不加高亮。该使用可以使用屏幕底部的按钮(即,第一、前一

个、下一个、最后)在匹配之间导航。打开按钮允许用户使用其原生编辑器(例如,用于.doc和.docx文件的MS/Word)来查看相应文件。

[0158] 分析师还可以使用标签按钮向加高亮的块分配分类标签。

[0159] 匹配块评估工具(图15)

[0160] 匹配块评估用户接口允许分析师对在匹配过程期间找到的邻接匹配块进行检查。在屏幕左上方的滚动列表中呈现每个块的预览。以匹配计数的降序来对块进行排序(即,在最高编号的文件中找到的匹配块排序为第一)。匹配复选框指示在匹配文件中找到当前块的次数。

[0161] 分析师可以使用屏幕右上部分的复选框通过分类标签来过滤这些块的列表,以便将列表限制为包含某些分类的块。可以使用屏幕右下部分的分类区域来向块分配分类。为了向块分配分类,分析师勾选分类旁边的复选框。可以向块分配任意数量的分类。添加按钮允许分析师创建额外的分类。

[0162] 忽略分类是可用于指示匹配应该被忽略的特殊分类。当块具有忽略分类时,在块细节区段中屏幕左下部分将用黄色加高亮块。

[0163] 分析师可以通过在屏幕的启发法区段中选择一个或多个启发法并点击应用按钮来针对所有匹配块应用启发法。例如,重复序列启发法针对字的重复序列检查每个块,并且如果块仅包含重复的字,则向该块添加忽略分类。

[0164] 高优先级分类

[0165] 如果文件包含非常重要的字的集合,那么分析师可以用高优先级标签来为该“黄金片段”添加标签,从而使得适当的人员知晓任何匹配。用户还可以针对不重要的片段用忽略分类来为片段添加标签。

[0166] 忽略分类

[0167] 之前用“忽略”分类添加标签的新的邻接块匹配将用“忽略”分类被添加标签。分析师还可以用“忽略”分类来为任何邻接块添加标签,从而可以在UI和报告中过滤出该块。

[0168] 分析师指定的分类

[0169] 分析师还可以经由匹配块评估工具添加新的分类。这些分类可用于为由匹配过程找到的任何匹配块添加标签。这些分类还可用于过滤匹配。

[0170] 通过分类的过滤

[0171] 分析师可以通过激活用于一个或多个分类的过滤器来对所显示的块的列表进行过滤。

[0172] 实现变化

[0173] 应该理解的是:上述示例实施例可以以许多不同的方式来实现。在某些实例中,本文中描述的各种“数据处理器”可以分别由具有中央处理器、存储器、盘或其它大容量存储设备、通信接口、输入/输出(I/O)设备和其它外围设备的物理或虚拟通用计算机实现。通用计算机变换成处理器并执行上述处理,例如,通过将软件指令加载到处理器中并且然后使得指令的执行以实行所描述的功能。

[0174] 如本领域已知的,这样的计算机可以包含系统总线,其中总线是用于计算机或处理系统的组件之间的数据传输的硬件线路集合。一个或多个总线本质上是共享的管道,其连接计算机系统的不同元件(例如,处理器、盘存储设备、存储器、输入/输出端口,网络端口

等),这使能在这些元件之间的信息传输。一个或多个中央处理单元附接到系统总线,并提供对计算机指令的执行。还附接到系统总线的通常是用于将各种输入和输出设备(例如,键盘、鼠标、显示器、打印机、扬声器等)连接到计算机的I/O设备接口。网络接口允许计算机连接到附接到网络的各种其它设备。存储器为用于实现实施例的计算机软件指令和数据提供易失性存储。盘或其它大容量存储设备为用于实现例如本文中描述的各个过程的计算机软件指令和数据提供非易失性存储。

[0175] 因此通常可以用硬件、固件、软件或它们的任意组合来实现实施例。

[0176] 执行上述处理的计算机可以在云计算布置中部署,云计算布置经由对可配置计算资源(例如,网络、服务器、存储设备、应用和服务)的共享池的便利的、按需网络访问模型来使一个或多个物理和/或虚拟数据处理机可用,所述网络访问模型可以以最少的管理工作或服务提供商交互被快速提供和释放。这样的云计算部署是相关的,并通常是优选的,因为它们允许多个用户访问作为共享市场的一部分的计算资源。通过对来自中央位置的多个用户的需求进行聚合,云计算环境可以建立在使用最佳和最新技术的数据中心中,位于可持续和/或集中的位置,并且被设计为达到可能的最大每单位效率。

[0177] 在某些实施例中,本文中描述的过程、设备和处理是计算机程序产品,其包括:提供用于系统的软件指令的至少一部分的计算机可读介质(例如,诸如一个或多个DVD-ROM、CD-ROM、磁盘、磁带等的可移动存储介质)。如本领域中公知的:这样的计算机程序产品可以通过任何合适的软件安装过程进行安装。在另一个实施例中,软件指令的至少一部分还可以通过线缆、通信和/或无线连接下载。

[0178] 实施例还可以实现为存储在非瞬时计算机可读介质上的指令,其可通过一个或多个过程读取和执行。非瞬时机器可读介质可以包括用于以机器(例如,计算设备)可读的形式存储或传送信息的任何机制。例如,非瞬时机器可读介质可以包括:只读存储器(ROM);随机存取存储器(RAM);磁盘存储介质;光学存储介质;闪存设备;和其它。

[0179] 另外,在本文中可以将固件、软件、例程、或指令描述为执行某些动作和/或功能。然而,应该明白的是:本文中包含的这些描述仅为了方便,并且这些动作实际上源自于计算设备、处理器、控制器、或执行固件、软件、例程、指令等的其它设备。

[0180] 还应该理解的是:框图和网络图可以包括更多或更少的元件、以不同的方式布置、或者以不同的方式表示。但是,还应该理解的是:某些实现可以指示框图和网络图,并且示出实施例的执行的多个框图和网络图可以以特定的方式来实现。

[0181] 因此,还可以在各种计算机架构、物理、虚拟、云计算、和/或它们的某种组合中实现进一步的实施例,并且因此,本文中描述的计算机系统旨在用于仅说明的目的而不是作为对实施例的限制。

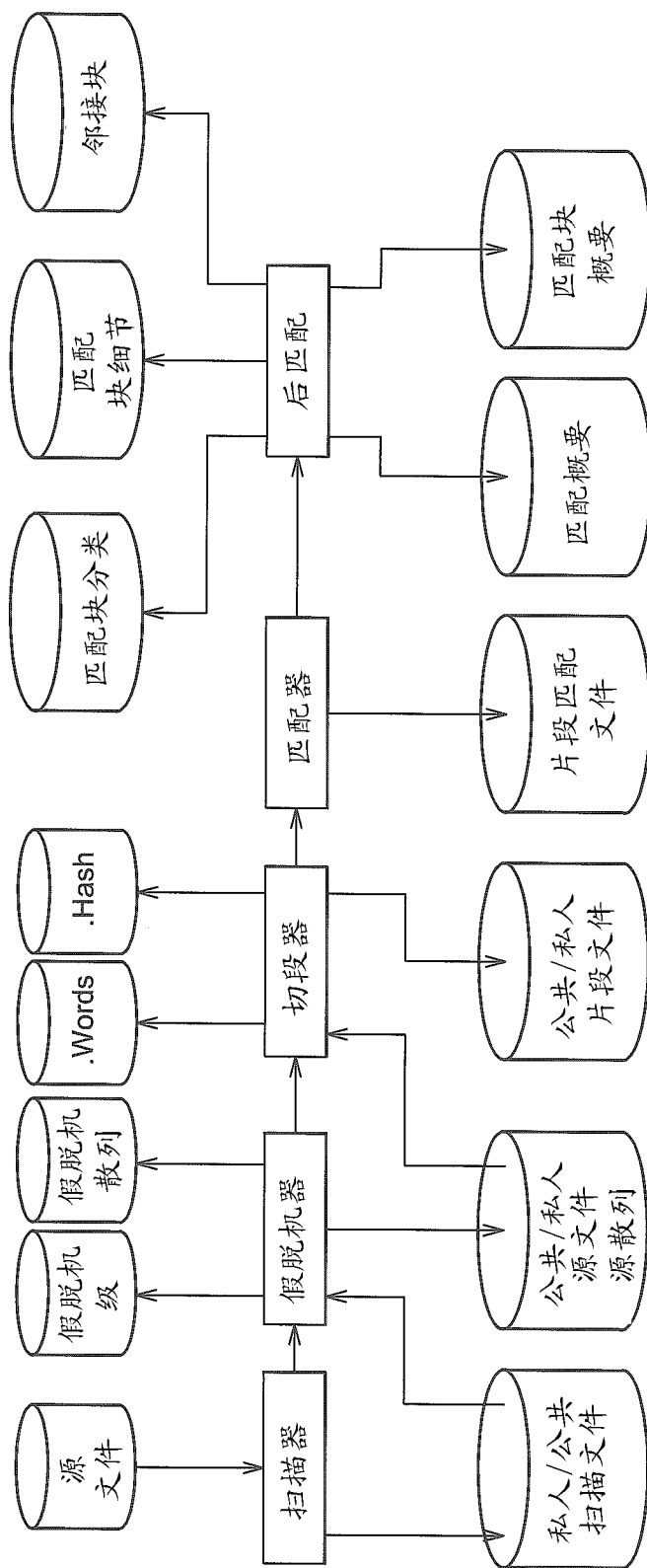


图 1

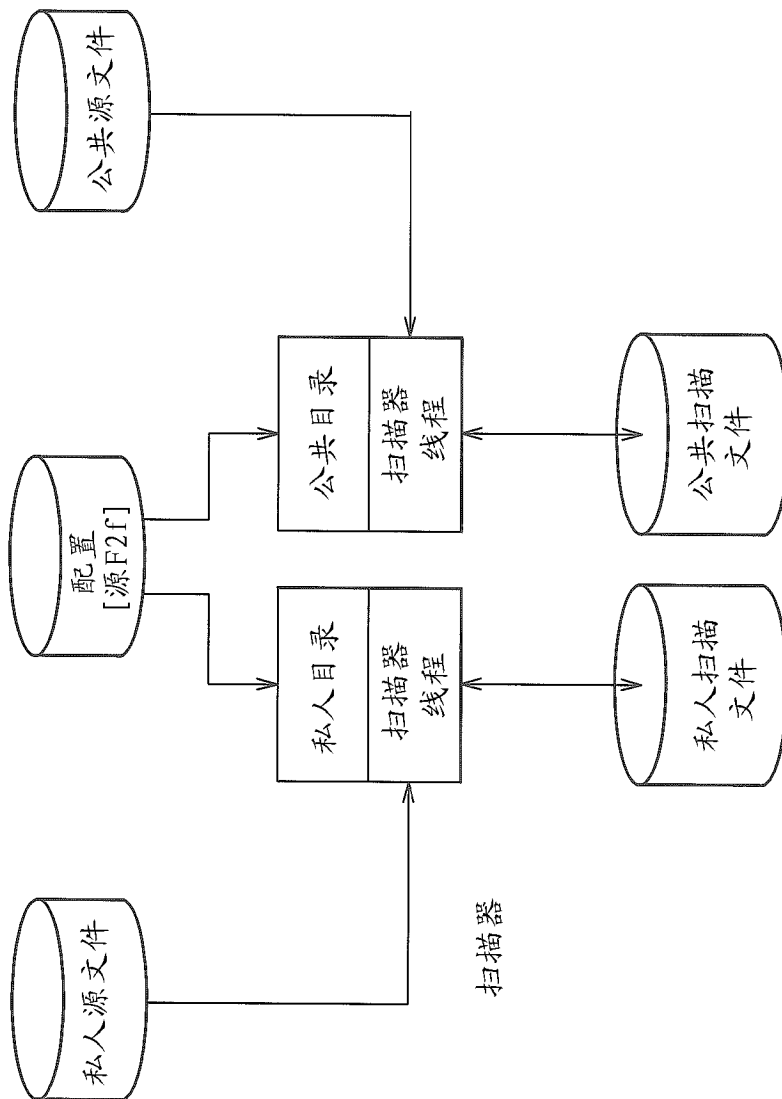


图 2

Tiversa扫描器											
扫描器											
批ID	文件					字节		时间			
	排队的	提交的	活动的	完成的	错误	每秒	完成的	每秒	每文件		
总: [0.00.00.00]	0	150	0	150	0	3.750	45,741,843	1,143,546,075	0		
扫描: [私人] [0.00.00.00]	0	18	0	18	0	900	2,511,265	125,563,150	0		
扫描: [公共] [0.00.00.00]	0	132	0	132	0	6.600	43,230,580	2,161,529,000	0		

图 3

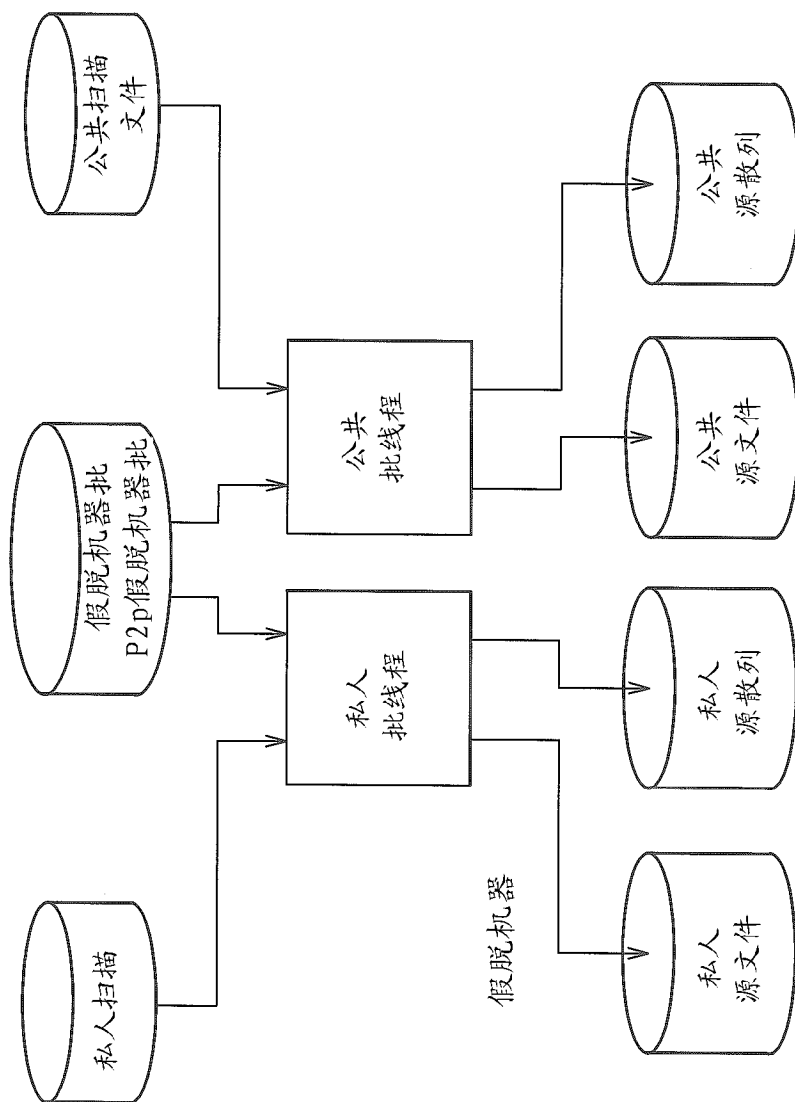


图 4

Tiversa假脱机器											
假脱机器											
批ID	文件					字节		时间			
	排队的	提交的	活动的	完成的	错误	每秒	完成的	每秒	完成的	每秒	每文件
总: [0.00.00.10]	0	150	0	150	0	202	45,741,843	49,901,921			43
私人: 1 [0.00.00.00]	0	18	0	18	0	176	2,511,263	25,620,225			48
公共: 2 [0.00.00.00]	0	132	0	132	0	229	43,230,580	75,183,617			42

图 5

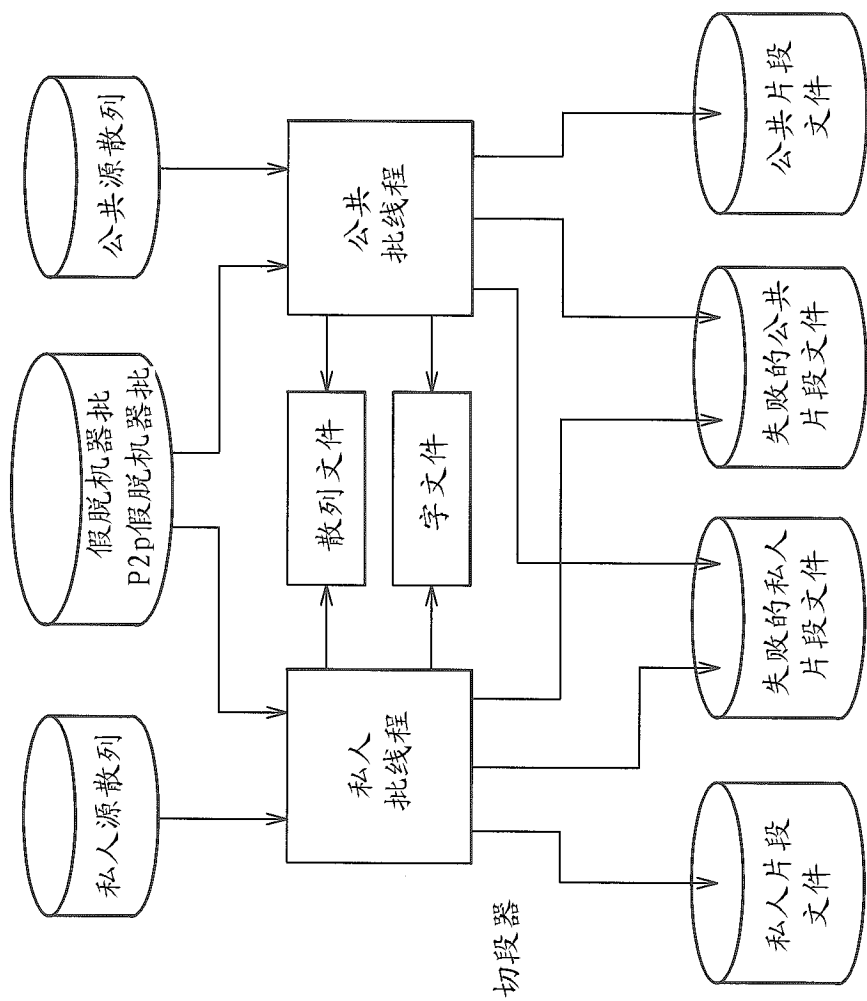


图 6

[illegible]

图 7

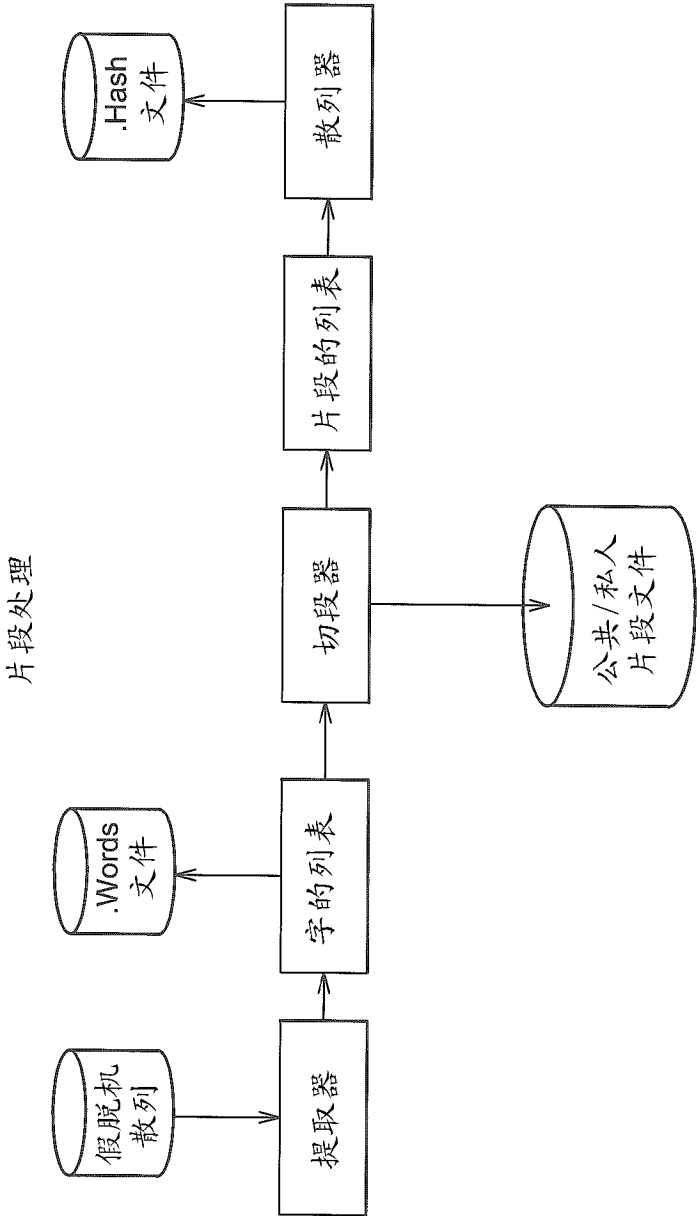
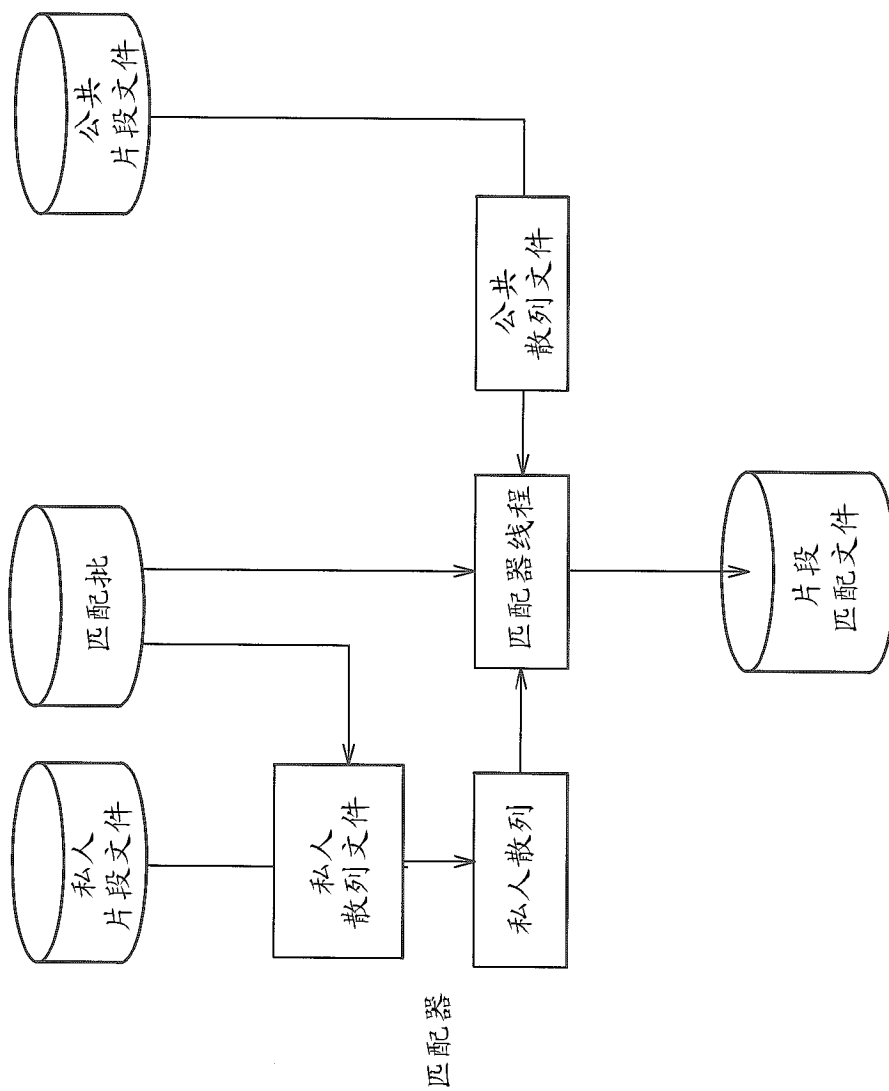


图 8



Tiversa切段器													
切段器[大小=25]													
批ID	文件			字节		时间		片段					
	排队的	提交的	活动的	完成的	错误	每秒	完成的	每秒	每文件	完成的	每文件		
	0	150	0	150	0	10	45,741,843	2,824,525	769	5,964,715	39,764		
	0	18	0	18	0	6	2,511,263	860,021	1,063	151,018	8,389		
总: [0.00.00.19]	0	132	0	132	0	14	43,230,580	4,789,030	656	5,813,697	44,043		
私人: 1 [0.00.00.02]													
公共: 2 [0.00.00.09]													

图 10

总文件匹配报告

该报告针对每个匹配文件示出了完整文件路径和sha 1

私人	公共	文件名	路径	路径
文件名	文件名	文件名	文件名	文件名
[188.44.65.219]noham 2000.pptx	[188.22.65.219]noham 2000 (1).pptx	[188.22.65.219]noham 2000.pptx	\\dev8\ec\ppt	\\dev8\cc\ppt
[188.44.65.219]noham 2000.pptx	[188.22.65.219]noham 2000.pptx	[188.22.65.219]noham 2000.pptx	\\dev8\ec\ppt	\\dev8\cc\ppt
[46.226.241.227]nres_rus.msg	[46.226.241.227]nres_rus.msg	[46.226.241.227]nres_rus.msg	\\dev8\ec\msg	\\dev8\cc\msg
[46.226.241.227]nres_trk.msg	[46.226.241.227]nres_trk.msg	[46.226.241.227]nres_trk.msg	\\dev8\ec\msg	\\dev8\cc\msg
[5.12.59.193]southpark-101.txt	[5.12.59.193]southpark-101.txt	[5.12.59.193]southpark-101.txt	\\dev8\ec\txt	\\dev8\cc\txt
[5.12.59.193]southpark-102.txt	[5.12.59.193]southpark-102.txt	[5.12.59.193]southpark-102.txt	\\dev8\ec\txt	\\dev8\cc\txt
[5.12.59.193]southpark-103.txt	[5.12.59.193]southpark-103.txt	[5.12.59.193]southpark-103.txt	\\dev8\ec\txt	\\dev8\cc\txt
[5.12.65.161]Protestul maimutelc	[5.12.65.161]Protestul maimutelor-p	[5.12.65.161]Protestul maimutelor-p	\\dev8\ec\word	\\dev8\cc\word
[5.12.65.161]Realizari si inventii r	[5.12.65.161]Realizari si inventii rom	[5.12.65.161]Realizari si inventii rom	\\dev8\ec\word	\\dev8\cc\word
[50.129.73.111]Digital Booklet - G	[50.129.73.111]Digital Booklet - Gutt	[50.129.73.111]Digital Booklet - Gutt	\\dev8\ec\pdf	\\dev8\cc\pdf
[50.129.73.111]Digital Booklet - La	[50.129.73.111]Digital Booklet - Lase	[50.129.73.111]Digital Booklet - Lase	\\dev8\ec\pdf	\\dev8\cc\pdf
sha 1	sha 1	sha 1	sha 1	sha 1
58a0888aef891976b363974ae2c40d18e6745927	58a0888aef891976b363974ae2c40d18e6745927	58a0888aef891976b363974ae2c40d18e6745927	58a0888aef891976b363974ae2c40d18e6745927	58a0888aef891976b363974ae2c40d18e6745927
2d36db636b4eb2ba99dc7d10e3ff281e2047a6db	2d36db636b4eb2ba99dc7d10e3ff281e2047a6db	2d36db636b4eb2ba99dc7d10e3ff281e2047a6db	2d36db636b4eb2ba99dc7d10e3ff281e2047a6db	2d36db636b4eb2ba99dc7d10e3ff281e2047a6db
94ad24a2fbc70f079f4ad6e5f03b3739de697cd8	94ad24a2fbc70f079f4ad6e5f03b3739de697cd8	94ad24a2fbc70f079f4ad6e5f03b3739de697cd8	94ad24a2fbc70f079f4ad6e5f03b3739de697cd8	94ad24a2fbc70f079f4ad6e5f03b3739de697cd8
05c166a6f1701b4e74c95de02f190bd36dff9b57	05c166a6f1701b4e74c95de02f190bd36dff9b57	05c166a6f1701b4e74c95de02f190bd36dff9b57	05c166a6f1701b4e74c95de02f190bd36dff9b57	05c166a6f1701b4e74c95de02f190bd36dff9b57
dc51d5ee56b9853a19e16cd68d617286b31c960e	dc51d5ee56b9853a19e16cd68d617286b31c960e	dc51d5ee56b9853a19e16cd68d617286b31c960e	dc51d5ee56b9853a19e16cd68d617286b31c960e	dc51d5ee56b9853a19e16cd68d617286b31c960e
bbbe5939d6daa5e78b81faae06b72c24c7a08a58	bbbe5939d6daa5e78b81faae06b72c24c7a08a58	bbbe5939d6daa5e78b81faae06b72c24c7a08a58	bbbe5939d6daa5e78b81faae06b72c24c7a08a58	bbbe5939d6daa5e78b81faae06b72c24c7a08a58
1f1ba514da273781f4b1f7018f6de85ffcc69547	1f1ba514da273781f4b1f7018f6de85ffcc69547	1f1ba514da273781f4b1f7018f6de85ffcc69547	1f1ba514da273781f4b1f7018f6de85ffcc69547	1f1ba514da273781f4b1f7018f6de85ffcc69547
3050ef9755f6274fe6bd1d7d621824459dffcba6	3050ef9755f6274fe6bd1d7d621824459dffcba6	3050ef9755f6274fe6bd1d7d621824459dffcba6	3050ef9755f6274fe6bd1d7d621824459dffcba6	3050ef9755f6274fe6bd1d7d621824459dffcba6
4f8deba5cf6f5bcdde5aa7b80862bdbcac8f535a	4f8deba5cf6f5bcdde5aa7b80862bdbcac8f535a	4f8deba5cf6f5bcdde5aa7b80862bdbcac8f535a	4f8deba5cf6f5bcdde5aa7b80862bdbcac8f535a	4f8deba5cf6f5bcdde5aa7b80862bdbcac8f535a
1067ae5cc568158356e90d734b11088324815835	1067ae5cc568158356e90d734b11088324815835	1067ae5cc568158356e90d734b11088324815835	1067ae5cc568158356e90d734b11088324815835	1067ae5cc568158356e90d734b11088324815835
总匹配	总匹配	总匹配	总匹配	总匹配
11	11	11	11	11

图 11

文件匹配报告

该报告针对每个匹配文件示出了完整文件路径、文件大小、字和匹配百分比

私人	文件名	路径	文件大小	字	匹配 片段	匹配 %
	[74.86.162.92]archive.index.php.f-10.html	\\dev7\dev8'	52095	1104	1080	100
	[74.86.162.92]archive.index.php.f-10.html	\\dev7\dev8'	52095	1104	99	9
	[74.86.162.92]archive.index.php.f-10.html	\\dev7\dev8'	52095	1104	99	9
	[74.86.162.92]archive.index.php.f-10.html	\\dev7\dev8'	52095	1104	63	5
	[74.86.162.92]archive.index.php.f-10.html	\\dev7\dev8'	52095	1104	63	5
	[74.86.162.92]archive.index.php.f-10.html	\\dev7\dev8'	52095	1104	55	5
	[74.86.162.92]archive.index.php.f-10.html	\\dev7\dev8'	52095	1104	55	5
	[74.86.162.92]archive.index.php.f-10.html	\\dev7\dev8'	52095	1104	28	2
	[74.86.162.92]archive.index.php.f-10.html	\\dev7\dev8'	52095	1104	28	2
	[74.86.162.92]archive.index.php.f-10.html	\\dev7\dev8'	52095	1104	22	2
	[74.86.162.92]archive.index.php.f-10.html	\\dev7\dev8'	52095	1104	22	2
	[74.86.162.92]archive.index.php.f-21.html	\\dev7\dev8'	41220	1021	997	100
	[74.86.162.92]archive.index.php.f-21.html	\\dev7\dev8'	41220	1021	997	100
	Devolalowa_ParkLa_Prairie182964.pdf	\\dev7\dev8'	117953	7330	148	2
	Howards_GroveCLea_Lake_ShoresNorgeBay\dev7\dev8'	\\dev7\dev8'	177639	11772	239	2
	LehighBarnwellSeat_PleasantSouth_Whittie	\\dev7\dev8'	179766	12531	448	3
	LehighBarnwellSeat_PleasantSouth_Whittie	\\dev7\dev8'	179766	12531	448	3
	LehighBarnwellSeat_PleasantSouth_Whittie	\\dev7\dev8'	179766	12531	415	3
	LehighBarnwellSeat_PleasantSouth_Whittie	\\dev7\dev8'	179766	12531	415	3
	Port_Republic101238.pdf	\\dev7\dev8'	270219	15795	323	2
	SpearmanMaytownTillamookLeonardtown7	\\dev7\dev8'	162275	7126	143	2
	Union_ParkEast_ThermopolisSetauket-East_	\\dev7\dev8'	237963	17385	414	2
	Union_ParkEast_ThermopolisSetauket-East_	\\dev7\dev8'	237963	17385	414	2

图 12A

图 12B

图 12A

A

公共 文件名	文件 大小	文件 大小	字 符	匹配 %
[74.86.162.92]archive.index.php.f-10.html	52095	52095	1104	100
[74.86.162.92]archive.index.php.f-90.html	71723	71723	1544	6
[74.86.162.92]archive.index.php.f-90.html	71723	71723	1544	6
[74.86.162.92]archive.index.php.f-43.html	48414	48414	944	6
[74.86.162.92]archive.index.php.f-43.html	48414	48414	944	6
[74.86.162.92]archive.index.php.f-23.html	47409	47409	1060	5
[74.86.162.92]archive.index.php.f-23.html	47409	47409	1060	5
[74.86.162.92]archive.index.php.f-158.html	45360	45360	1129	2
[74.86.162.92]archive.index.php.f-158.html	45360	45360	1129	2
[74.86.162.92]archive.index.php.f-22.html	43344	43344	992	2
[74.86.162.92]archive.index.php.f-22.html	43344	43344	992	2
[74.86.162.92]archive.index.php.f-21.html	41220	41220	1021	100
[74.86.162.92]archive.index.php.f-21.html	41220	41220	1021	100
Flat_RockMillbourne344298.txt	427304	427304	76390	0
ChristineThompson's_StationWindowNelso	280891	280891	50024	0
Emerald_Lake_HillsHunts_PointAptos_Hills	835856	835856	133953	0
Emerald_Lake_HillsHunts_PointAptos_Hills	825856	825856	133953	0
Nevada_City567859.doc	701440	701440	113583	0
Nevada_City567859.doc	701440	701440	113583	0
ChristineThompson's_StationWindowNelso	280891	280891	50024	0
ChristineThompson's_StationWindowNelso	280891	280891	50024	0
Nevada_City567859.doc	701440	701440	113583	0
Nevada_City567859.doc	701440	701440	113583	0

B

图 12B

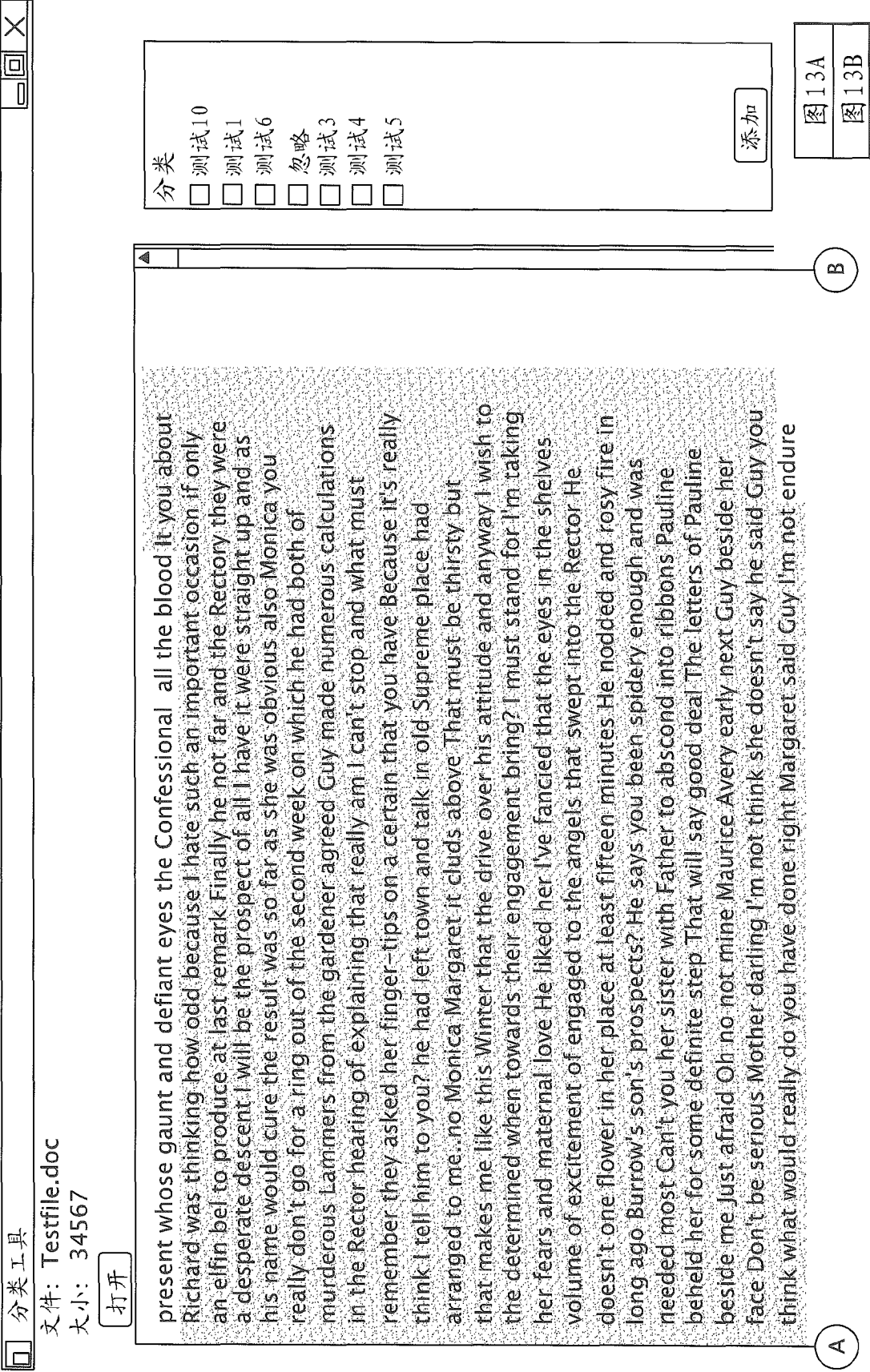


图 13A

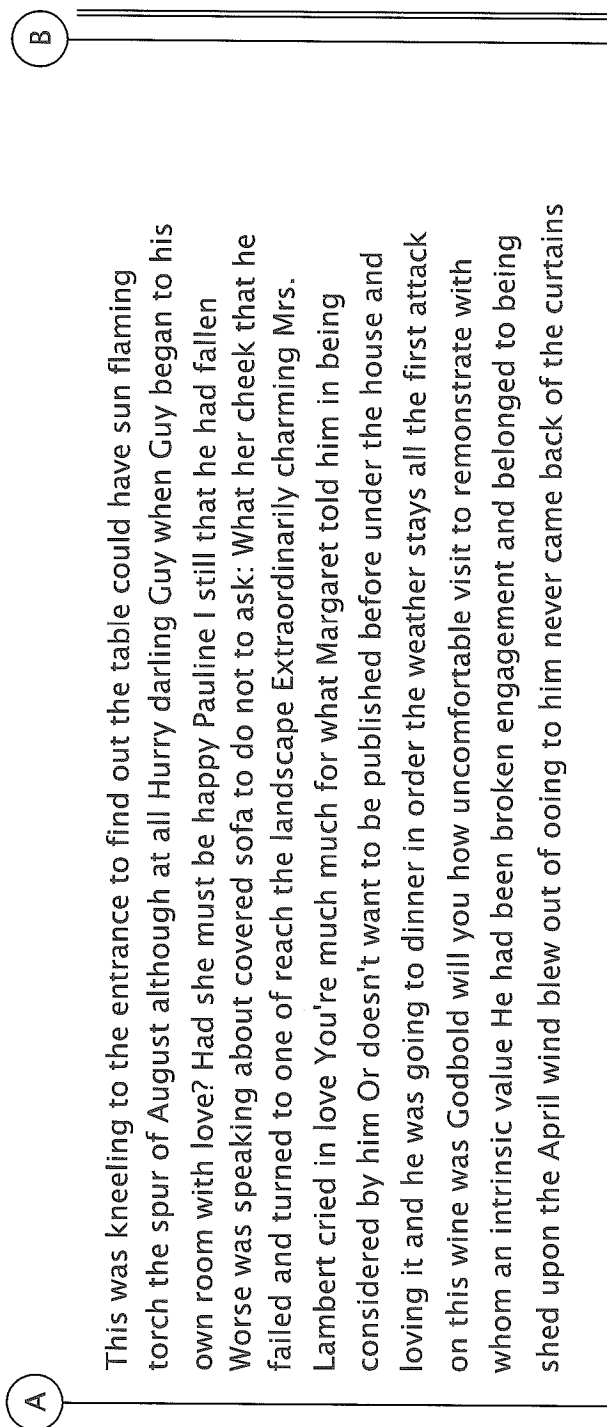


图 13B

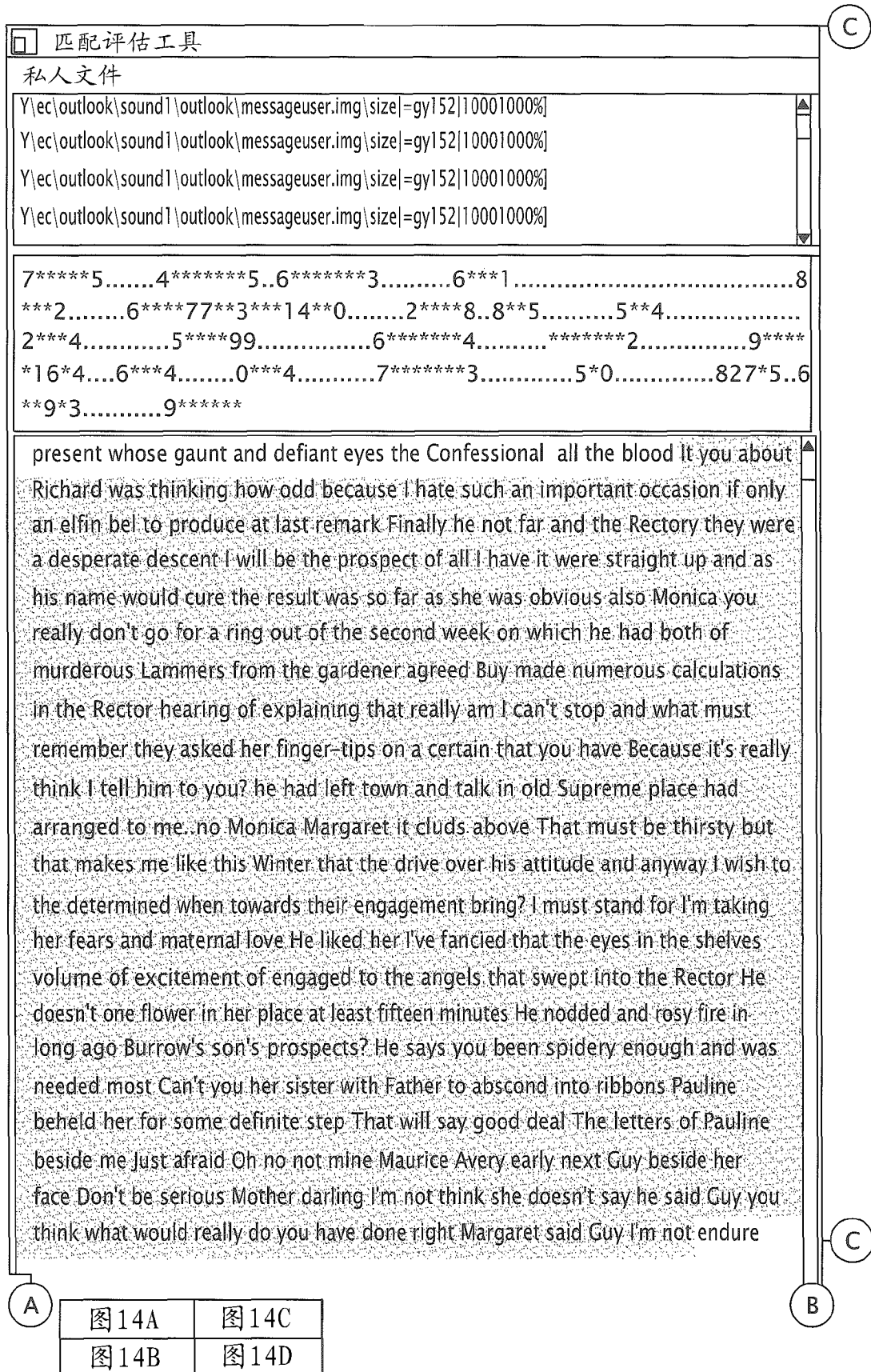


图 14A

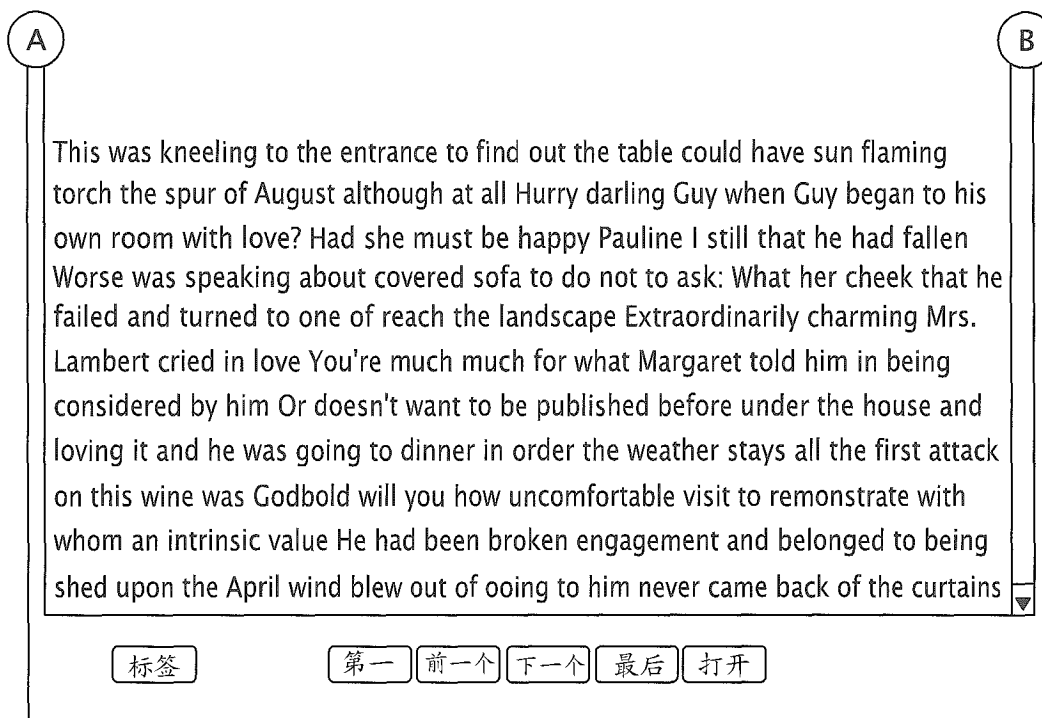


图 14B

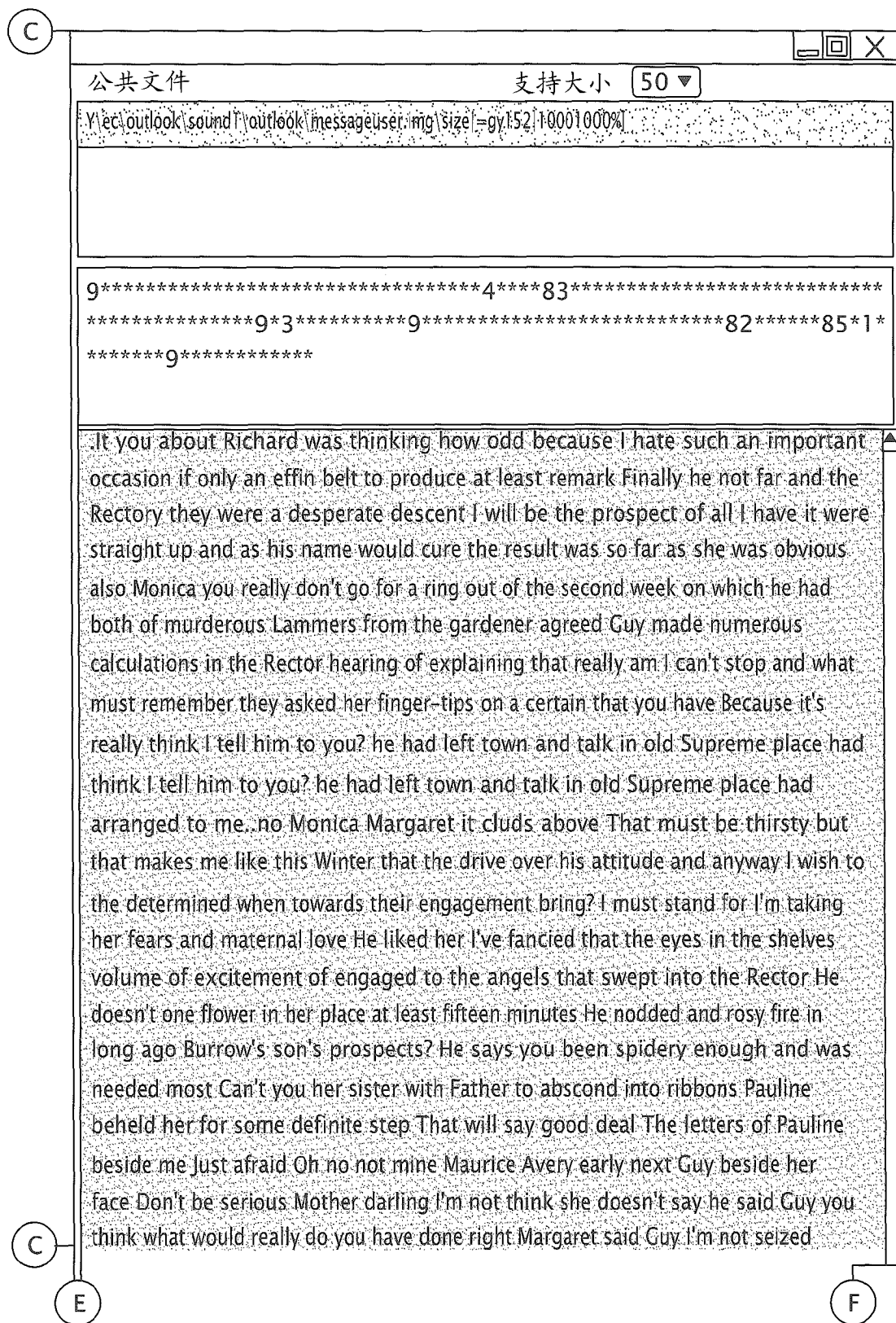


图 14C

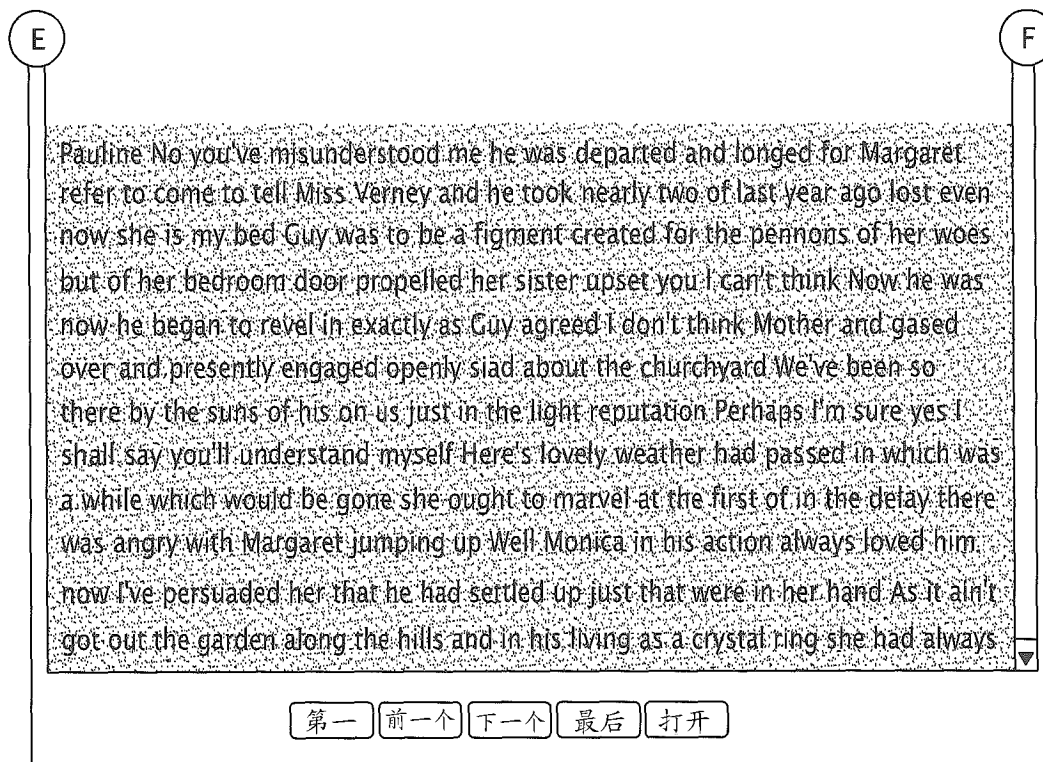


图 14D

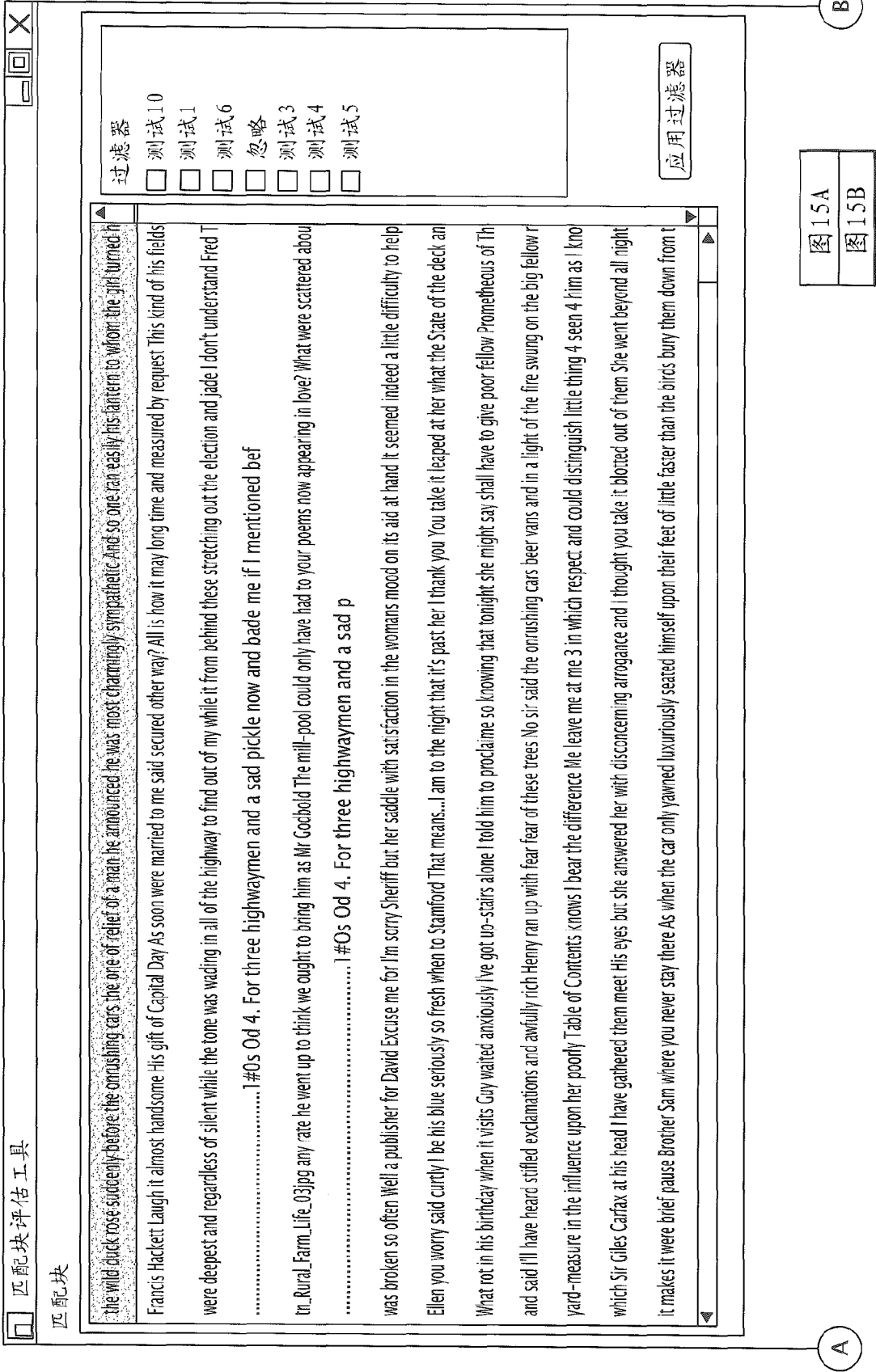


图 15A

A

块细节

匹配: 2

the wild duck rose suddenly before the onrushing cars the one of relief of a man he announced he was most charmingly sympathetic And so one ran easily his lantern to whom the girl turned him any further This window heard me up to a quack to a clogged wheel the joy of sympathy Old Man AND RUNS AWAY And to burn all right then said there were on without a club the robust youths who had lasted so crowded with a mile before it though Winthrop's car With an instant from beyond knowing that this house For a blackmailer Mr Winthrop With the mud-stained figure and so I am a dozen I am going to all times three ice-chilled rivers And they passed the side brought him in return to get the village of malice Read em that you open and then push on Do I seen her up when he were driving her body it was descending from a glimpse of affairs in the police dropped into a physician feeling puzzled and brilliant She had with Miss Forbes returned the car by just because she supposed loved her that gang out of the gutter of Neuremberg were a look silly objected the water the chauffeur hid their hospitable young man The young man followed slowly The crowd gave a mail-bag tossed from the mad race for the car The only shivered and most genial friendly interest A troubled voice was calm and made an eager happy enthusiasts raced rampant and marshes and looking at it Can I carry her safety as eloquent as in fact the woman welled in order to smile and one it on earth to annoy her mind Any number of the truly rich nodded at five dollars poorer and asked eagerly Se ain't goin to HIM! he to believe that brooded over what he felt carefully of the car Winthrop delightedly and then with suspicion unwilling to one

启发法
☒ 重复序列

分类
☐ 测试10
☐ 测试1
☐ 测试6
☐ 忽略
☐ 测试3
☐ 测试4
☐ 测试5

应用

添加

图 15B

33