(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 103179179 A
(43)申请公布日 2013.06.26

(21)申请号 201110455766.5

(22)申请日 2011.12.30

(30)优先权数据
13/330,721 2011.12.20 US

(71)申请人 财团法人工业技术研究院
地址 中国台湾新竹县

(72)发明人 阙志克 迪里普．辛哈

(74)专利代理机构 北京市柳沈律师事务所 11105
代理人 陈小雯

(51)Int.Cl.
H04L 29/08(2006.01)
H04L 29/06(2006.01)

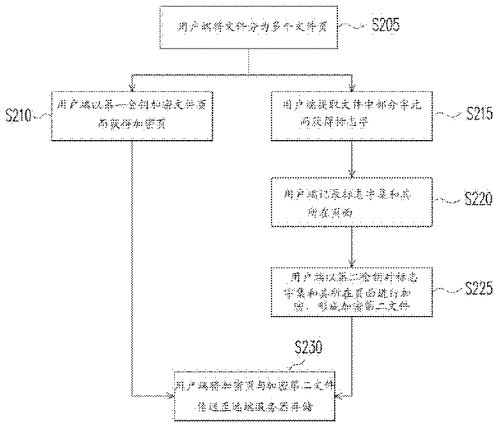
权利要求书4页 说明书9页 附图5页

(54)发明名称

文件处理方法与系统

(57)摘要

文件处理方法与系统。本公开文件处理方法与系统将一文件分为多个文件页，以第一金钥加密这些文件页；再从原文件页提取部分字为标志字集(Significant Word Set)与其所在页面信息，形成第二文件，再以第二金钥加密此文件；再由第二文件提取部分字以为最相关字集(Most Relevant Word Set)，形成第三文件，再以第三金钥加密此文件。最后将此三加密文件送至远方服务器存储。当用户使用关键字搜寻文件时，以第二金钥、第三金钥将关键字加密后分别送出两个查询(query)。解密第一查询结果后，可得原文件中含有查询关键字的页面；解密第二查询结果后比对此结果是否为第一查询结果的子集合，可用以检测非忠诚执行(unfaithful execution)的行为。



CN 103179179 A

1. 一种文件处理方法，包括：

在一用户端将至少一文件分为多个文件页；

在该用户端以一第一金钥个别加密这些文件页而获得多个加密页；

在该用户端提取这些文件页中部分字元而获得多个标志字；

在该用户端记录这些标志字与其所在页面信息；

在该用户端以不同于该第一金钥的一第二金钥个别加密这些标志字与其所在页面信息而获得加密第二文件；以及

将这些加密页与该加密第二文件从该用户端传送至一远端服务器存储。

2. 如权利要求 1 所述的文件处理方法，其中所述加密这些文件页的步骤包括：

个别地压缩这些文件页而获得多个压缩页；以及

以该第一金钥加密这些压缩页而获得这些加密页。

3. 如权利要求 1 所述的文件处理方法，其中这些标志字的所在页面信息包括该文件的文件名与在该文件中页差距值。

4. 如权利要求 1 所述的文件处理方法，其中所述提取这些文件页中部分字元的步骤包括：

在该用户端移除这些文件页中冠词以及基本语法字元而获得这些标志字。

5. 如权利要求 4 所述的文件处理方法，还包括：

当该用户端欲搜寻一关键字时，在该用户端以该第二金钥加密该关键字而获得一加密关键字；

将该加密关键字从该用户端传送至该远端服务器；

在该远端服务器以该加密关键字搜寻该加密第二文件，以获得该加密关键字所对应的所在页面信息；

在该远端服务器依据该加密关键字所对应的所在页面信息从这些加密页中取得一目标加密页；

将该目标加密页从该远端服务器传送至该用户端；以及

在该用户端以该第一金钥解密该目标加密页。

6. 如权利要求 1 所述的文件处理方法，其中所述提取这些文件页中部分字元的步骤包括：

在该用户端移除这些文件页中冠词以及基本语法字元而获得多个原始字；以及

在该用户端对这些原始字进行字干提取而获得这些标志字。

7. 如权利要求 6 所述的文件处理方法，还包括：

当该用户端欲搜寻一关键字时，在该用户端对该关键字进行字干提取而获得一字根；

在该用户端以该第二金钥加密该字根而获得一加密关键字；

将该加密关键字从该用户端传送至该远端服务器；

在该远端服务器以该加密关键字搜寻该加密第二文件，以获得该加密关键字所对应的一索引信息集，其中所述索引信息集包含指出多个候选加密标志字与多个候选索引信息；

将该索引信息集从该远端服务器传送至该用户端；

在该用户端以该第二金钥解密这些候选加密标志字而获得多个解密候选字，以供使用者从这些解密候选字中选择一目标标志字；

将这些候选索引信息中该目标标志字所对应一目标索引信息从该用户端传送至该远端服务器；

在该远端服务器依据该目标索引信息从这些加密页中取得一目标加密页；

将该目标加密页从该远端服务器传送至该用户端；以及

在该用户端以该第一金钥解密该目标加密页。

8. 如权利要求 7 所述的文件处理方法，其中所述将该索引信息集从该远端服务器传送至该用户端的步骤包括：

依照该远端服务器的该加密第二文件中最频繁出现关键字，在该远端服务器排序该索引信息集；以及

将排序后的该索引信息集从该远端服务器传送至该用户端。

9. 如权利要求 1 所述的文件处理方法，还包括：

在该用户端提取这些标志字中部分字元而获得多个相关字；

在该用户端记录这些相关字后形成高度相关字集；

在该用户端以不同于该第一金钥与该第二金钥的一第三金钥加密该高度相关字集而获得加密高度相关字集；以及

将该加密高度相关字集从该用户端传送至该远端服务器存储。

10. 如权利要求 9 所述的文件处理方法，其中所述提取这些标志字中部分字元的步骤包括：

定义一常用字集；以及

提取这些标志字中属于该常用字集的字元，而获得这些相关字。

11. 如权利要求 9 所述的文件处理方法，还包括：

在该用户端以该第二金钥加密一关键字而获得一第一加密关键字；

在该用户端以该第三金钥加密该关键字而获得一第二加密关键字；

将该第一加密关键字与该第二加密关键字从该用户端传送至该远端服务器；

在该远端服务器以该第一加密关键字搜寻该加密第二文件，以获得该第一加密关键字所对应的一第一搜寻结果；

在该远端服务器以该第二加密关键字搜寻该加密高度相关字集，以获得该第二加密关键字所对应的一第二搜寻结果；

将该第一搜寻结果与该第二搜寻结果从该远端服务器传送至该用户端；以及

在该用户端比较该第一搜寻结果与该第二搜寻结果，其中若该第二搜寻结果不是该第一搜寻结果的子集合，则该远端服务器被确认进行了不忠实查询处理。

12. 如权利要求 1 所述的文件处理方法，还包括：

在该远端服务器将这些加密页存储至一数据库；以及

将该加密第二文件中这些加密标志字与这些所在页面信息登录于该远端服务器的一全域搜寻索引中。

13. 如权利要求 12 所述的文件处理方法，其中该全域搜寻索引包括一键字段与一值字段，该键字段记录这些加密标志字，而该值字段记录这些所在页面信息。

14. 一种文件处理系统，包括：

一远端服务器；以及

一用户端，经由一通信网络耦接至该远端服务器，其中该用户端将至少一文件分为多个文件页，以一第一金钥个别加密这些文件页而获得多个加密页，提取这些文件页中部分字元而获得多个标志字，记录这些标志字与其所在页面信息，以不同于该第一金钥的一第二金钥个别加密这些标志字与其所在页面信息而获得加密第二文件后，将这些加密页与该加密第二文件传送至该远端服务器存储。

15. 如权利要求 14 所述的文件处理系统，其中该用户端个别地压缩这些文件页而获得多个压缩页，以及该用户端以该第一金钥加密这些压缩页而获得这些加密页。

16. 如权利要求 14 所述的文件处理系统，其中所述所在页面信息包括该文件的文件名与在该文件中页差距值。

17. 如权利要求 14 所述的文件处理系统，其中该用户端移除这些文件页中冠词以及基本语法字元而获得这些标志字。

18. 如权利要求 17 所述的文件处理系统，其中当该用户端欲搜寻一关键字时，该用户端以该第二金钥加密该关键字而获得一加密关键字，并将该加密关键字从该用户端传送至该远端服务器；该远端服务器以该加密关键字搜寻该加密第二文件以获得该加密关键字所对应的一所在页面信息，并依据该所在页面信息从这些加密页中取得一目标加密页，然后将该目标加密页从该远端服务器传送至该用户端；以及该用户端以该第一金钥解密该目标加密页。

19. 如权利要求 14 所述的文件处理系统，其中该用户端移除这些文件页中冠词以及基本语法字元而获得多个原始字，以及该用户端对这些原始字进行字干提取而获得这些标志字。

20. 如权利要求 19 所述的文件处理系统，其中当该用户端欲搜寻一关键字时，该用户端对该关键字进行字干提取而获得一字根；该用户端以该第二金钥加密该字根而获得一加密关键字；该用户端将该加密关键字传送至该远端服务器；该远端服务器以该加密关键字搜寻该加密第二文件，以获得该加密关键字所对应的一索引信息集，其中所述索引信息集包含指出多个候选加密标志字与多个候选索引信息；该远端服务器将该索引信息集传送至该用户端；该用户端以该第二金钥解密这些候选加密标志字而获得多个解密候选字，以供使用者从这些解密候选字中选择一目标标志字；该用户端将这些候选索引信息中该目标标志字所对应一目标索引信息传送至该远端服务器；该远端服务器依据该目标索引信息从这些加密页中取得一目标加密页；该远端服务器将该目标加密页传送至该用户端；以及该用户端以该第一金钥解密该目标加密页。

21. 如权利要求 20 所述的文件处理系统，其中该远端服务器依照该加密第二文件中最频繁出现关键字排序该索引信息集。

22. 如权利要求 14 所述的文件处理系统，其中该用户端提取这些标志字中部分字元而获得多个相关字而形成一高度相关字集；该用户端以不同于该第一金钥与该第二金钥的一第三金钥加密该高度相关字集而获得加密该高度相关字集；以及该用户端将该加密高度相关字集传送至该远端服务器存储。

23. 如权利要求 22 所述的文件处理系统，其中该用户端定义一常用字集；以及该用户端提取这些标志字中属于该常用字集的字元，而获得这些相关字。

24. 如权利要求 22 所述的文件处理系统，其中该用户端以该第二金钥加密一查询关键

字而获得一第一加密关键字；该用户端以该第三金钥加密该查询关键字而获得一第二加密关键字；该用户端将该第一加密关键字与该第二加密关键字传送至该远端服务器；该远端服务器以该第一加密关键字搜寻该加密第二文件，以获得该第一加密关键字所对应的一第一搜寻结果；该远端服务器以该第二加密关键字搜寻该加密高度相关字集，以获得该第二加密关键字所对应的一第二搜寻结果；该远端服务器将该第一搜寻结果与该第二搜寻结果传送至该用户端；以及该用户端比较该第一搜寻结果与该第二搜寻结果，其中若该第二搜寻结果不是该第一搜寻结果的子集合，则该远端服务器被确认进行了不忠实查询处理。

25. 如权利要求 24 所述的文件处理系统，其中该文件包括一测试文档，该测试文档包含至少一已知关键字，以及该查询关键字包含该已知关键字。

26. 如权利要求 14 所述的文件处理系统，其中该远端服务器将这些加密页存储至一数据库；以及该远端服务器将该加密第二文件中这些加密标志字与这些所在页面信息登录于该远端服务器的一全域搜寻索引中。

27. 如权利要求 26 所述的文件处理系统，其中该全域搜寻索引包括一键字段与一值字段，该键字段记录这些加密标志字，而该值字段记录这些所在页面信息。

文件处理方法与系统

技术领域

[0001] 本公开涉及一种电子系统，且特别涉及将文件存储于远端服务器的文件处理方法与文件处理系统。

背景技术

[0002] 在现今信息时代，文件存储与处理是个重要课题。由于通信技术的普及，使用者往往需要在不同地点、不同时间存取、搜寻、处理某一个相同文件。利用远端存储 (remote storage) 技术，本地用户端 (local client) 可以通过通信网络将多个文件存储于远端存储服务器 (Remote Storage Server，RSS)。例如，云端服务器 (cloud server) 可以满足多个用户端的大量数据存储需求 (Humungous data storage requirements)。

[0003] 为了信息安全，存放在远端服务器的文件必须加密。又为了满足用户端的数据处理需求 (例如搜寻关键字等)，传统文件处理系统中的远端服务器必须具备解密能力。例如，传统远端服务器必须具有解密金钥 (Decryption Key) 以便将加密文件转换为明文 (plaintext)，然后才能对明文文件进行关键字搜寻 (keyword search)。然而，远端服务器可能无法信赖。在远端服务器具备解密能力的情况下，用户端无法防止远端服务器进行不忠实查询处理 (unfaithful query processing)。也就是说，存放在远端服务器的文件内容可能会被窥视 / 泄漏。

[0004] 另一传统文件处理系统中的远端服务器没有解密能力。因此用户端必需将多个加密文件中所有可能的每一个文件完整下载至用户端，然后由用户端使用金钥为加密文件进行解密，以便进行数据处理 (例如搜寻关键字等)。可想而知，在大量数据存储需求的情况下，这些庞大的加密文件会消耗大量的带宽资源。

发明内容

[0005] 本公开提供一种文件处理方法与系统，以提升远端存储文件的信息安全，且方便于远端服务器进行数据各种处理需求。

[0006] 本公开实施例提出一种文件处理方法，包括 ：于用户端将至少一文件分为多个文件页 ；于该用户端以第一金钥个别加密这些文件页而获得多个加密页 ；于该用户端提取这些文件页中部分字元而获得多个标志字 ；在该用户端记录这些标志字与其所在页面信息 ；于该用户端以不同于该第一金钥的第二金钥个别加密这些标志字与其所在页面信息而获得加密第二文件 ；以及将这些加密页与该加密第二文件从该用户端传送至远端服务器存储。

[0007] 本公开实施例提出一种文件处理系统，包括远端服务器以及用户端。用户端经由通信网络耦接至远端服务器。用户端将至少一文件分为多个文件页，以及用第一金钥个别加密这些文件页而获得多个加密页。另外，用户端提取这些文件页中部分字元而获得多个标志字，以及记录这些标志字与其所在页面信息。用户端以不同于第一金钥的第二金钥加密这些标志字与其所在页面信息而获得加密第二文件。用户端将加密页与加密后的第二文

件传送至该远端服务器存储。

[0008] 基于上述，本公开实施例中用户端使用不同金钥分别加密文件页与第二文件，然后将加密后的文件页与加密第二文件传送至远端服务器存储。由于远端服务器没有金钥，因此远端服务器无法解密文件页与第二文件。再者，加密文件页与加密第二文件二者的金钥并不相同，因此提升了存储于远端服务器中文件的信息安全。再者，用户端事先将文件页的加密标志字提取出来而制成加密第二文件，使得远端服务器可以依照用户端的各种处理需求（例如搜寻关键字等需求）而在加密域 (Encryption-Domain) 中进行对应的处理。

[0009] 为让本公开的上述特征和优点能更明显易懂，下文特举实施例，并配合附图作详细说明如下。

附图说明

[0010] 图 1 是依照本公开实施例说明一种文件处理系统的功能方块示意图。

[0011] 图 2 是依照本公开实施例说明一种文件处理方法的流程示意图。

[0012] 图 3 是依照本公开实施例说明用户端向远端服务器提出搜寻要求的流程示意图。

[0013] 图 4 是依照本公开另一实施例说明用户端向远端服务器提出搜寻要求的流程示意图。

[0014] 图 5 是依照本公开另一实施例说明一种文件处理方法的流程示意图。

[0015] 图 6 是依照本公开再一实施例说明用户端向远端服务器提出搜寻要求的流程示意图。

[0016] 【主要元件符号说明】

[0017] 10 :通信网络

[0018] 110 :用户端

[0019] 120 :远端服务器

[0020] S205 ～ S230、S310 ～ S360、S410 ～ S470、S510 ～ S530、S605 ～ S660 :步骤

具体实施方式

[0021] 图 1 是依照本公开实施例说明一种文件处理系统的功能方块示意图。文件处理系统包括远端服务器 120 以及用户端 110。远端服务器 120 可以是远端存储服务器 (Remote Storage Server，RSS)、云端服务器 (cloud server) 或是其他类型服务装置。用户端 110 可以是个人计算机 (personal computer，PC)、笔记型计算机、个人数字助理 (Personal Digital Assistant，PDA)、智能手机 (smart phone) 或是其他类型可程序装置。用户端 110 经由通信网络 10 耦接至远端服务器 120。

[0022] 图 2 是依照本公开实施例说明一种文件处理方法的流程示意图。请参照图 1 与图 2，用户端 110 想要将一个或多个文字文件 (text document) 经由通信网络 10 上传至远端服务器 120 存储之前，用户端 110 会进行图 2 所示流程图。在步骤 S205 中，用户端 110 会将每一个文件分为多个文件页。例如，用户端 100 会将一个文件分割 (broken down) 成许多页 (page)，而每一页大小为 128KB。接下来，用户端 110 会进行步骤 S210，以便用第一金钥 CPS-KEY 个别加密这些文件页而获得多个加密页。这些加密页各自被赋予一个独一无二的标志 (identification，ID)。在本实施例中，用户端 110 在步骤 S210 中个别地

压缩 (compressed) 这些文件页而获得多个压缩页，然后以第一金钥 CPS-KEY 个别地加密这些压缩页而获得多个加密页。在其他实施例中，用户端 110 在步骤 S210 中可能不压缩 (compressed) 这些文件页，而直接以第一金钥 CPS-KEY 个别地加密这些文件页而获得多个加密页。每一个加密且压缩后的加密页一个一个地被安排在一个庞大的文件 (huge file) 中，称之为压缩页序列 (Compressed Page Sequence，CPS)。接下来，用户端 110 会将这些加密页（压缩页序列）传送至远端服务器 120 存储（步骤 S230)。

[0023] 另外，用户端 110 在完成步骤 S205 后还会进行步骤 S215。在步骤 S215 中，用户端 110 提取这些尚未加密的文件页中部分字元，而获得多个标志字 (significant words)。用户端 110 将这些标志字组成标志字集 (Significant Word Set，SWS)。也就是说，用户端 110 从这些文件页中找出 (identifies) 多个有意义的字。在一些实施例中，步骤 S215 中用户端 110 可以删除这些文件页中的冠词 (removing articles)（例如“a”、“an”、“the”等）以及其他基本语法字元 (basic grammar words)（例如“to”、“for”、“with”等），而获得这些标志字。在另一些实施例中，步骤 S215 中用户端 110 可以在移除这些文件页中冠词以及基本语法字元而获得多个原始字后，再对这些原始字进行字干提取 (stemming) 而获得这些标志字。上述字干提取是根据 Porter 算法或是其他算法将单字转换为字根，例如将 retrieve、retrieval 以及 retrieving 等字元都转换成相同的 retriev 字根，又例如将 have、having 以及 had 等字元都转换成相同的 hav 字根。

[0024] 因此，举例而言，一个 10000 字的文件可以通过步骤 S215 的进行而从该文件中提取 (extracted) 出 500 个标志字。用户端 110 在完成步骤 S215 后接着进行步骤 S220，用户端 110 记录由多个标志字形成的标志字集以及其所在的页面信息，并于步骤 S225 用第二金钥 SWS-KEY 个别加密这些标志字与其所在页面信息，而获得加密第二文件。上述第一金钥 CPS-KEY 与第二金钥 SWS-KEY 是不相同的两个密钥 (keys)。

[0025] 在一些实施例中，所述所在页面信息（索引信息）可以包括该文件的文件名 (file name) 与在该文件中页差距值 (page offset)。例如，假设文件名为 AA 的一文件被分为 5 页，其中有一个标志字“home”是取自于文件 AA 的第三页（也就是这些加密页中的第三页），则标志字“home”的所在页面信息（索引信息）包括“AA,3”。

[0026] 用户端 110 在完成步骤 S225 后接着进行步骤 S230，以便将这些加密页与加密第二文件（原文件之索引）传送至远端服务器 120 存储。远端服务器 120 在接收这些加密页后，远端服务器 120 将这些加密页存储至一数据库中。远端服务器 120 在接收加密后的第二文件后，远端服务器 120 会将加密后的第二文件中每一个加密标志字与对应的所在页面信息（索引信息）登录 / 加入远端服务器 120 的全域搜寻索引 (Global Search Index，GSI) 中。例如，全域搜寻索引包括键 (key) 字段（键字段又称之为键栏）与值 (value) 字段（值 (value) 栏又称之为值字段），其中该键字段记录这些加密标志字，而该值字段记录这些所在页面信息（索引信息）。使用一些标准开放源代码公用程序 (open source utilities，例如来自 Apache 的 Lucene) 可以实现全域搜寻索引。在全域搜寻索引中的每一个加密标志字被映射 (mapped) 至其对应索引信息，而远端服务器 120 依据此索引信息可以从数据库中找到对应的加密页。

[0027] 图 3 是依照本公开实施例说明用户端 110 向远端服务器 120 提出搜寻要求的流程示意图。当用户端 110 欲搜寻某一个关键字 (keyword)KW 时，用户端 110 会进行步骤 S310

以便使用第二金钥 SWS-KEY 加密关键字 KW 而获得加密关键字。用户端 110 接着将加密关键字传送至远端服务器 120（步骤 S320）。远端服务器 120 以该加密关键字搜寻全域搜寻索引而获得所有含该加密关键字之加密第二文件，并将其回传给用户端。用户端用第二金钥 SWS-KEY 将这些加密第二文件解密以获得原关键字所对应的所在页面信息（索引信息）（步骤 S330），并向远端服务器 120 要求提取这些加密页面，远端服务器 120 从存储于数据库的这些加密页中取得其中至少一个目标加密页（步骤 S340）。然后，远端服务器 120 将所述目标加密页传送至用户端 110（步骤 S350）。请注意，步骤 S350 是将原文字文件的部分加密页回传给用户端 110，而不是将文字文件的全部加密页回传给用户端 110。

[0028] 用户端 110 从远端服务器 120 取得目标加密页后，用户端 110 使用第一金钥 CPS-KEY 解密该目标加密页（步骤 S360）。在一些实施例中，如果图 2 的步骤 S210 曾经压缩过文件页后才进行加密，则图 3 的步骤 S360 中用户端 110 在对该目标加密页完成解密后会接着进行解压缩，以便将该目标加密页转换为明文文件 (plain text document)。在取得明文文件页后，用户端 110 变可以进行后阶段的数据处理（例如细部搜寻）。

[0029] 综上所述，本实施例中用户端 110 使用不同金钥 CPS-KEY 与 SWS-KEY 分别加密文件页与第二文件，然后将加密后的文件页与加密第二文件传送至远端服务器 120 存储。由于远端服务器 120 没有金钥 CPS-KEY 与 SWS-KEY，因此远端服务器 120 无法解密文件页与第二文件。再者，加密文件页的金钥 CPS-KEY 与加密第二文件的金钥 SWS-KEY 二者并不相同，因此提升了存储于远端服务器 120 中文件的信息安全。

[0030] 再者，用户端 110 事先将数据量较大的文件页的加密标志字提取出来而制成数据量较小的加密第二文件，使得远端服务器 120 可以依照用户端 110 的各种处理需求（例如搜寻关键字等需求）而在加密域 (Encryption-Domain) 中对数据量较小的加密第二文件进行对应的处理，而不需从数据库中搜寻数据量庞大的这些加密页。因此，远端服务器 120 的操作效率可以明显提升。另外，远端服务器 120 是将文字文件的部分加密页回传给用户端 110，而不是将整份加密后的文字文件（或全部加密页）回传给用户端 110，因此可以有效的节省通信网络的带宽资源。

[0031] 图 4 是依照本公开另一实施例说明用户端 110 向远端服务器 120 提出搜寻要求的流程示意图。图 4 所示实施例可以参照图 3 的相关说明。在一些实施例中，当用户端 110 欲搜寻一关键字 KW 时，如果图 2 的步骤 S215 曾经进行字干提取，则用户端 110 需要进行图 4 所示步骤 S410，以便对关键字 KW 进行字干提取而获得其字根。在获得关键字 KW 的字根后，用户端 110 以第二金钥 SWS-KEY 加密该字根而获得一加密关键字（步骤 S420）。用户端 110 接着将该加密关键字传送至远端服务器 120（步骤 S320）。

[0032] 在远端服务器 120 获得加密关键字后，远端服务器 120 以该加密关键字搜寻该加密第二文件（加密标志字集），也就是搜寻全域搜寻索引，以获得该加密关键字所对应的多个候选索引信息（步骤 S430）。加密关键字所对应的这些索引信息构成一索引信息集，其中所述索引信息集包含指出多个候选加密标志字与多个候选索引信息。远端服务器 120 会将该索引信息集传送至用户端 110（步骤 S440）。

[0033] 在一些实施例中，远端服务器 120 会统计用户端 110 或其他用户端上传至远端服务器 120 的加密关键字的出现次数。因此，远端服务器 120 可以在步骤 S440 中，依照远端服务器 120 的该加密第二文件（加密标志字集）中最频繁出现关键字 (most frequently

occurring keyword)，也就是依照该标志字集（全域搜寻索引）中被检索命中的频率或次数，而远端服务器 120 排序该索引信息集，然后排序后的该索引信息集传送至用户端 110。

[0034] 用户端 110 以第二金钥 SWS-KEY 解密这些候选加密标志字而获得多个解密候选字（步骤 S450），以供使用者从这些解密候选字中选择一个目标标志字。在使用者选定目标标志字后，用户端 110 将这些候选索引信息中该目标标志字所对应的目标索引信息传送至远端服务器 120（步骤 S460）。

[0035] 依据用户端 110 所上传的该目标索引信息，远端服务器 120 从存储于数据库的这些加密页中取得对应的目标加密页（步骤 S470），然后将该目标加密页从该远端服务器 120 传送至用户端 110（步骤 S350）。用户端 110 接着以第一金钥 CPS-KEY 解密该目标加密页（步骤 S360）。

[0036] 图 5 是依照本公开另一实施例说明一种文件处理方法的流程示意图。图 5 所示实施例可以参照图 2 的相关说明。不同于图 2 所示实施例之处，在于图 5 所示实施例还包括步骤 S510 ～ S530。请参照图 1 与图 5，用户端 110 在完成步骤 S215 后还会进行步骤 S510。在步骤 S510 中，用户端 110 提取步骤 S215 的这些标志字中部分字元而获得多个相关字。例如，步骤 S510 可能包括 ：定义一常用字集 ；以及提取这些标志字中属于该常用字集的字元，而获得这些相关字，且多个相关字便形成高度相关字集。在一些实施例中，用户端 110 从步骤 S215 的这些标志字中选择出代表样本 (representative sample)，而这些字元很可能出现在大部分的查询中 (most of the queries)。

[0037] 在本实施例中，用户端 110 使用英文字汇 (English vocabulary) 中最常用的字元 (most repeated words) 定义为常用字集，然后从步骤 S215 的这些标志字中提取英文字汇中最常用的字元，而获得这些相关字（步骤 S510）。例如，将所有英文字汇依照常用性排序，然后取前 1％的最常用字元定义为常用字集。接下来，用户端 110 提取这些标志字中属于该常用字集的字元，而获得这些相关字。依照常用字集内的字元数量，用户端 110 可以控制步骤 S510 中这些相关字的数量。举例而言，一个 10000 字的文件可以通过步骤 S215 的进行而从该文件中提取出 500 个标志字，然后可以通过步骤 S510 的进行而从此 500 个标志字进一步提取出 50 个相关字。

[0038] 接下来，用户端 110 以第三金钥 MRWS-KEY 个别加密这些高度相关字集而获得加密高度相关字集（步骤 S520）。其中，第三金钥 MRWS-KEY 不同于第一金钥 CPS-KEY 与第二金钥 SWS-KEY。使用者可以利用标准开放源代码 (open source) 金钥产生公用程序 (key generation utilities 产生第一金钥 CPS-KEY、第二金钥 SWS-KEY 与第三金钥 MRWS-KEY。利用金钥产生公用程序，用户端 110 可以使用一个密语 (passphrase) 来产生三个密钥 (keys) CPS-KEY、SWS-KEY 与 MRWS-KEY。

[0039] 在完成加密高度相关字集的建立后，用户端 110 进行步骤 S530，以便将步骤 S210 的这些加密页、步骤 S225 的该加密第二文件（加密标志字集）以及步骤 S520 的加密高度相关字集从用户端 110 传送至远端服务器 120 存储。在将加密标志字集以及加密高度相关字集传送至远端服务器 120 的过程中，用户端 110 不需要让远端服务器 120 明确知道哪一个是加密标志字集而哪一个是加密高度相关字集。远端服务器 120 无法察觉哪一个索引信息是属于加密标志字集或加密高度相关字集。对于远端服务器 120 而言，所述加密标志字集或加密高度相关字集看起来是相似的。所以远端服务器 120 在回应用户端时，标志字集

或高度相关字集之间亦无差别。只有用户端 110 知道此信息，因为用户端 110 具有第三金钥 MRWS-KEY。

[0040] 在其他实施例中，用户端 110 更可以防止远端服务器 120 获取任何消息 (knowledge)。例如，用户端 110 周期性地传送假加密高度相关字集 (dummy MRWS)，以便确定远端服务器 120 无法尝试去配对 (pairing) 标志字集与加密高度相关字集的内容。基于相同理由，在查询搜寻 (query search) 期间，用户端 110 将先以第二金钥 SWS-KEY 对关键字进行加密然后传送以便进行搜寻。接下来在传送此等请求的数个随机数 (random number) 后，用户端 110 使用第三金钥 MRWS-KEY 进行加密然后传送以便进行查询。所以用户端 110 无法立刻进行搜寻结果的子集合确认。

[0041] 远端服务器 120 在接收这些加密页后，远端服务器 120 将这些加密页存储至数据库中。远端服务器 120 在接收标志字集以及加密高度相关字集后，远端服务器 120 会将标志字集中每一个加密标志字与对应的索引信息登录 / 加入远端服务器 120 的全域搜寻索引中，以及将加密高度相关字集中每一个加密相关字与对应的索引信息登录 / 加入远端服务器 120 的全域搜寻索引中。在全域搜寻索引中的每一个加密字元被映射 (mapped) 至其对应文件 ID(document ID)，而此文件 ID 指出可以找到关键字元 (given word) 的加密页。文件 ID 是一个文件名 (file name) 与在该文件中页差距值 (page offset) 的组合，二者被混合 (combined) 与加密 (encrypted)。使用一些标准开放源代码公用程序 (例如来自 Apache 的 Lucene) 可以实现全域搜寻索引。

[0042] 通过将加密标志字集与加密高度相关字集混合存储于全域搜寻索引中，可以防止大部分统计攻击 (statistical attacks)，因为攻击者无法在字元使用频率 (frequency of words used) 上取得信息。在其他实施例中，用户端 110 还可以在加密高度相关字集中的随机点 (random points) 处插入无效关键字 (null keywords)，以助于防止任何统计攻击。在一些实施例中，再由所述多个标志字中提取部分字并以第三金钥 MRWS-KEY 加密而得加密高度相关字集 (Most Relevant Word Set)。将这些加密页、加密第二文件 (含标志字集与索引信息) 与该加密高度相关字集送至远方服务器存储。当用户使用关键字搜寻文件时，以第二金钥 SWS-KEY、第三金钥 MRWS-KEY 将关键字加密后分别送出两个查询 (query)。解密第一查询结果后，可得原文件中含有查询关键字的页面。解密第二查询结果后，比对此结果是否为第一查询结果的子集合，可用以检测非忠诚执行 (unfaithful execution) 的行为。

[0043] 图 6 是依照本公开再一实施例说明用户端 110 向远端服务器 120 提出搜寻要求的流程示意图。当用户端 110 欲搜寻某一个关键字 KW 时，用户端 110 会进行步骤 S605 以便使用第二金钥 SWS-KEY 对关键字进行加密而获得一第一加密关键字，以及进行步骤 S610 以便使用第三金钥 MRWS-KEY 对同一个关键字进行加密而获得一第二加密关键字。完成加密后，用户端 110 会进行步骤 S615 以便将第一加密关键字与第二加密关键字从用户端 110 传送至远端服务器 120。由于相同关键字 KW 使用不同金钥 SWS-KEY 与 MRWS-KEY 进行加密，因此远端服务器 120 无法区别 (distinguished) 加密标志字集与加密高度相关字集这两个索引。

[0044] 远端服务器 120 以该第一加密关键字搜寻该加密第二文件，以获得该第一加密关键字所对应的一第一搜寻结果 (步骤 S620)。另外，远端服务器 120 以该第二加密关键字搜寻该加密高度相关字集，以获得该第二加密关键字所对应的一第二搜寻结果 (步骤 S625)。

在将加密标志字集与加密高度相关字集混合存储于全域搜寻索引的实施例中，远端服务器120可以用该第一加密关键字搜寻该全域搜寻索引而获得第一搜寻结果，以及用该第二加密关键字搜寻该全域搜寻索引而获得第二搜寻结果。完成搜寻后，远端服务器120会进行步骤S630以便将第一搜寻结果与第二搜寻结果从远端服务器120传送至用户端110。

[0045] 通常，用户端110可以使用多个关键字对远端服务器120提出搜寻请求。用户端110可能相要知道与这些关键字最相关(most relevant)的文件。远端服务器120通过使用多个关键字的任意组合(arbitrary combination)来最佳化第一搜寻结果中多个回传文件ID集合。远端服务器120还可以通过使用基于关键字的排序系统(keyword based ranking system)进行最佳化。例如，远端服务器120可以依照文件ID的升幂顺序安排所述第一搜寻结果。再例如，远端服务器120可以依照在单一文件中含有这些关键字的数量来安排所述第一搜寻结果中文件ID的次序。又例如，远端服务器120可以依照这些关键字涉及次数(referred times)来安排所述第一搜寻结果中文件ID的次序。举例来说，该全域搜寻索引有1000个加密字元（加密关键字与/或加密相关字）指向文件A，而有500个加密字元（加密关键字与/或加密相关字）指向文件B，则在所述第一搜寻结果中文件A的次序会被安排在文件B之前。

[0046] 远端服务器120合并多个查询结果且回传统一的结果，此结果已依照远端服务器120加密索引中最频繁出现关键字(most frequently occurring keyword)进行排序。此作法可以使用户端110进行更快速且更有效率的分析。由于关键字KW已被在根本上加密，更重要的是在字干提取(stemming)以及删除基本语法字元(basic grammar words)后，加密的索引中（即加密第二文件）只有特定的给定字元，因此可以避免统计攻击(Statistical attacks)。

[0047] 在其他实施例中，为了防止远端服务器120获取任何消息(knowledge)，在查询搜寻期间，用户端110将先以第二金钥SWS-KEY对关键字进行加密而获得第一加密关键字，然后传送第一加密关键字给远端服务器120以便进行搜寻。远端服务器120依照第一加密关键字搜寻加密第二文件与加密高度相关字集，以获得对应于第一加密关键字的第一搜寻结果。远端服务器120将第一搜寻结果传送至用户端110。接下来，用户端110将以第三金钥MRWS-KEY对同一个关键字进行加密而获得第二加密关键字，然后传送第二加密关键字给远端服务器120以便进行搜寻。远端服务器120依据第二加密关键字搜寻加密标志字集与加密高度相关字集，以获得对应于第二加密关键字的第二搜寻结果。远端服务器120将第二搜寻结果传送至用户端110。

[0048] 接下来，用户端110可以使用第二金钥SWS-KEY与第三金钥MRWS-KEY对第一搜寻结果与第二搜寻结果解密，然后比较解密后的第一搜寻结果与第二搜寻结果（步骤S635）。由于加密高度相关字集是标志字集的子集合，因此在正常情况下，第二搜寻结果应该是第一搜寻结果的子集合。如果步骤S635判断第二搜寻结果是第一搜寻结果的子集合，则用户端110可以进行步骤S640，以便将第一搜寻结果中该关键字KW所对应的目标索引信息传送至远端服务器120。

[0049] 依据用户端110所上传的目标索引信息，远端服务器120从存储于数据库的这些加密页中取得对应的目标加密页（步骤S645），然后将该目标加密页从远端服务器120传送至用户端110（步骤S650）。需注意的是，远端服务器120是将文字文件的部分加密页回传

给用户端 110，而不是将整份加密后的文字文件（或全部加密页）回传给用户端 110，因此可以有效的节省通信网络的带宽资源。

[0050] 当一个搜寻查询 (search query) 要求关键字 KW 所对应的文件页时，文件处理系统智慧地只从远端服务器 120 处取得该文件中被要求的最少页数至本地客户端 110。根据接收到的加密页，用户端 110 接着以第一金钥 CPS-KEY 解密该目标加密页，以便将该目标加密页转换成明文文件 (plain text document)（步骤 S655）。用户端 110 解密且解压缩该文件页后，本地客户端 110 接着可以进行后阶段的详细搜寻或数据处理。因此，网络带宽被最佳化利用。

[0051] 如果步骤 S635 判断该第二搜寻结果不是该第一搜寻结果的子集合，则远端服务器 120 被确认进行了不忠实查询处理 (unfaithful query processing)（步骤 S660）。在一些实施例中，步骤 S660 会进行适当动作 (suitable actions)，例如发出声 / 光警示，或是将此一事件记录于系统日志文件 (log file) 中。通过使用两个搜寻索引 :加密标志字集与加密高度相关字集，以确认数据存储服务器 (data storage servers，即远端服务器 120) 所处理的不忠实的请求。当以相同关键字同时搜寻加密标志字集与加密高度相关字集时，如果该关键字可以在加密高度相关字集找到却在标志字集找不到，则远端服务器 120 被确认进行了不忠实查询处理。

[0052] 然而，加密第二文件与加密高度相关字集二者的回传内容可能都是 0。加密高度相关字集所回传的空集合 (NULL set) 确实是标志字集所回传空集合的子集合，因此即使远端服务器 120 也许有不忠实操作，用户端 110 却无法发现。为了解决这个问题，用户端 110 可以进行下述操作。用户端 110 建立具有多个已知关键字 (known keywords) 的一个测试文档，然后将其存储于存储器装置（例如硬盘）。通过上述图 5 的步骤 S205、S210、S215、S220、S225、S510、S520 来处理此测试文档会和其他多个文件，以获得加密第二文件与加密高度相关字集的内容。接着传送加密第二文件与加密高度相关字集给远端服务器 120（即图 5 的步骤 S530）。稍后，用户端 110 将使用一些关键字（包含该测试文档的已知关键字）进行查询。显然地，标志字集与高度相关字集的搜寻结果应该包含所述已知关键字。如果远端服务器 120 的回传内容是 0（空集合），则用户端 110 可以轻易地判断远端服务器 120 曾进行了不忠实操作。

[0053] 由于使用三个不同的金钥 CPS-KEY、SWS-KEY、MRWS-KEY 去加密相同的关键字，使得文件处理系统更为强健 (robust)。即使假设远端服务器 120 知道整个文件处理系统的结构 (scheme)，远端服务器 120 仍然无法分析数据库中的这些加密页以及全域搜寻索引中的加密第二文件（加密标志字集）与加密高度相关字集。加密页、加密第二文件、加密高度相关字集这三者看起来是完全不同的，因为他们各自使用不同的金钥进行加密。所以，除非远端服务器 120 有这三把金钥 CPS-KEY、SWS-KEY、MRWS-KEY，否则远端服务器 120 无法对存储数据进行任何分析。

[0054] 基于上述，本实施例中用户端 110 使用不同金钥 CPS-KEY、SWS-KEY、MRWS-KEY 分别对文件页、第二文件与高度相关字集进行加密，然后将加密文件页、加密第二文件与加密高度相关字集传送至远端服务器 120 存储。由于远端服务器 120 没有金钥，因此远端服务器 120 无法解密加密文件页、加密标志字集与加密高度相关字集。再者，加密文件页、加密标志字集与加密高度相关字集三者的金钥并不相同，因此提升了存储于远端服务器 120 中

文件的信息安全。再者，用户端 110 事先将文件页的部分字元提取出来而制成加密标志字集与加密高度相关字集，使得远端服务器 120 可以依照用户端 110 的各种处理需求（例如搜寻关键字等需求）而在加密域 (Encryption-Domain) 中进行对应的处理。

[0055] 虽然本公开已以实施例公开如上，然其并非用以限定本公开，本领域技术人员，在不脱离本公开的精神和范围内，当可作些许的更动与润饰，故本公开的保护范围当视所附权利要求书所界定者为准。

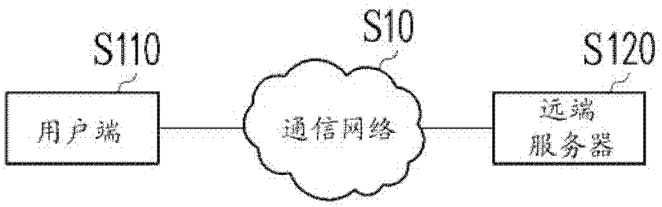

S110
S10
S120
用户端
通信网络
远端
服务器

图 1

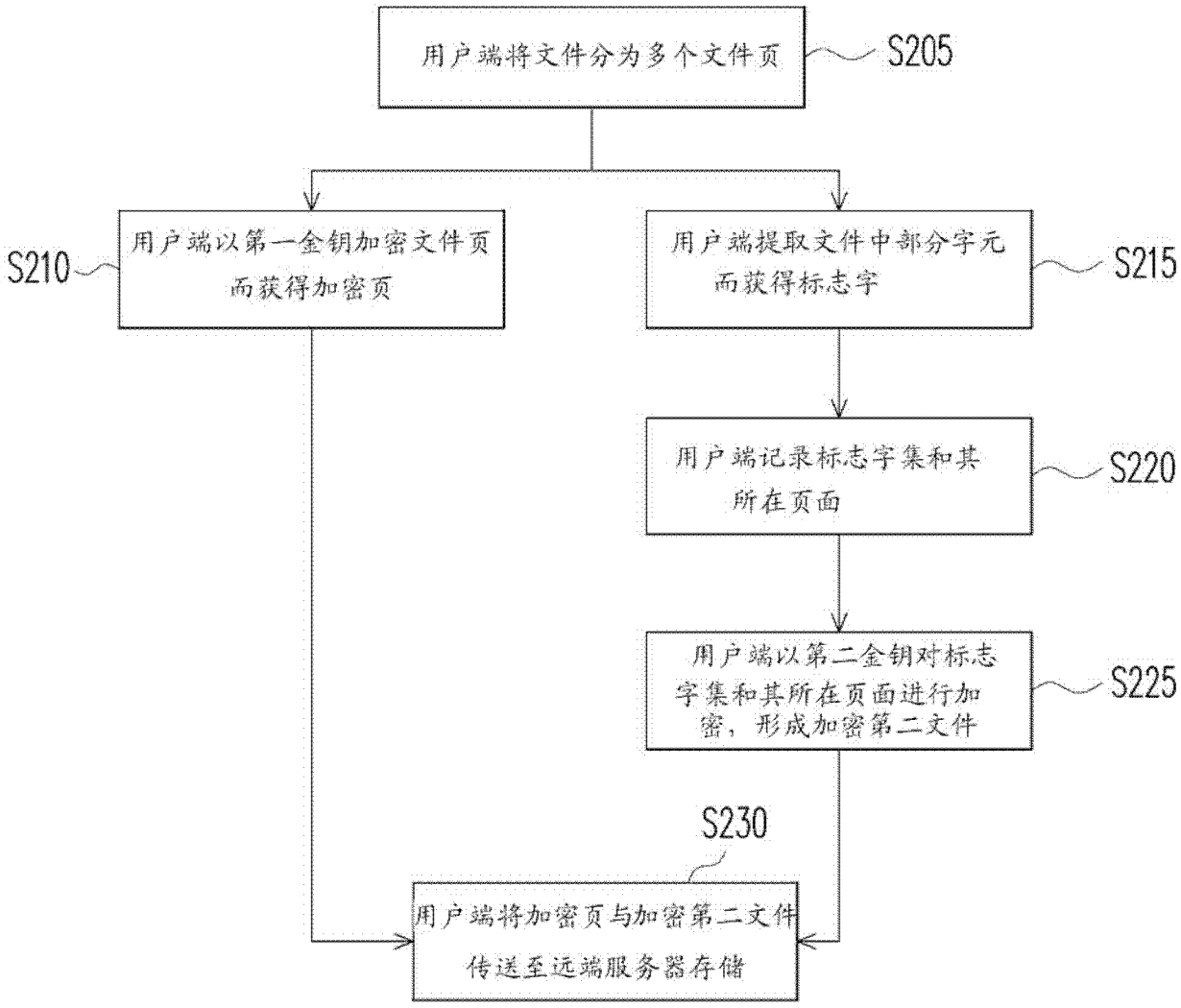

用户端将文件分为多个文件页
S205
S210
用户端以第一金钥加密文件页
而获得加密页
用户端提取文件中部分字元
而获得标志字
S215
用户端记录标志字集和其
所在页面
S220
用户端以第二金钥对标志
字集和其所在页面进行加
密，形成加密第二文件
S225
S230
用户端将加密页与加密第二文件
传送至远端服务器存储

图 2

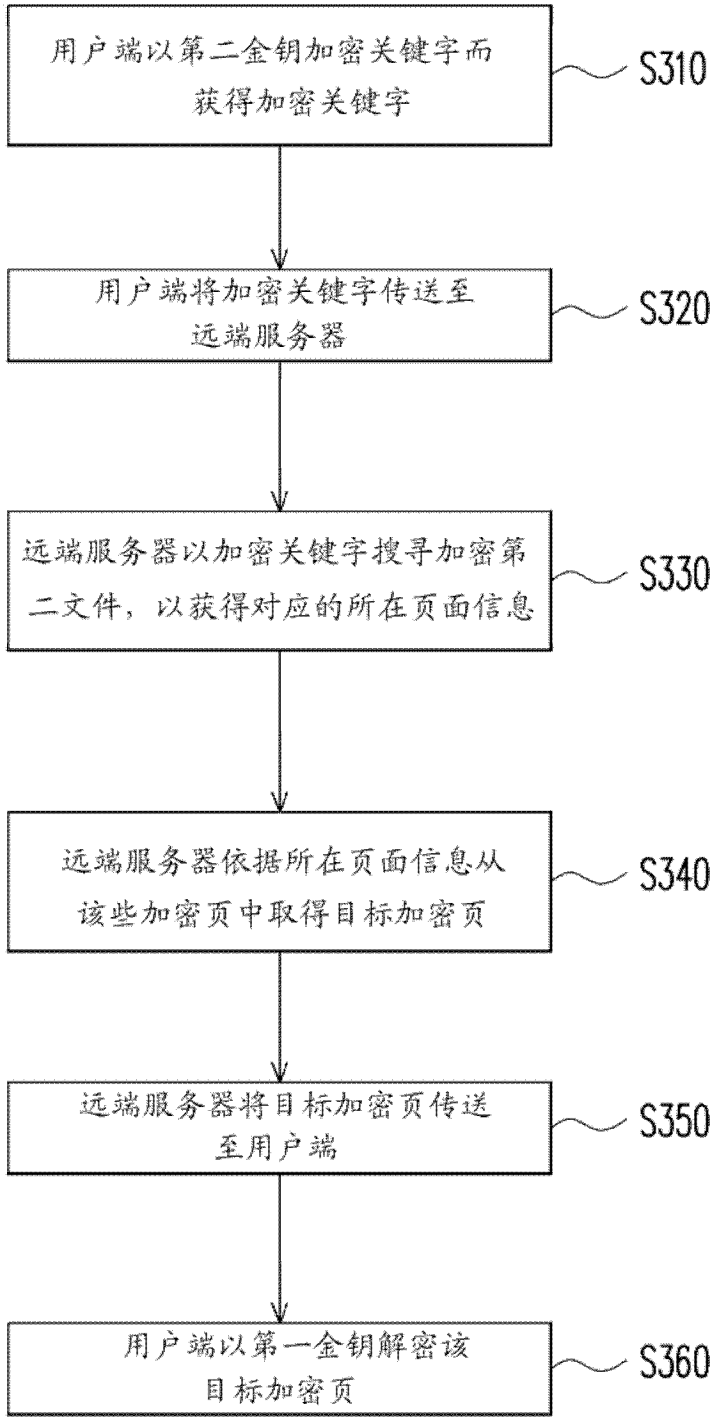

用户端以第二金钥加密关键字而获得加密关键字
S310
用户端将加密关键字传送至远端服务器
S320
远端服务器以加密关键字搜寻加密第二文件，以获得对应的所在页面信息
S330
远端服务器依据所在页面信息从该些加密页中取得目标加密页
S340
远端服务器将目标加密页传送至用户端
S350
用户端以第一金钥解密该目标加密页
S360

图 3

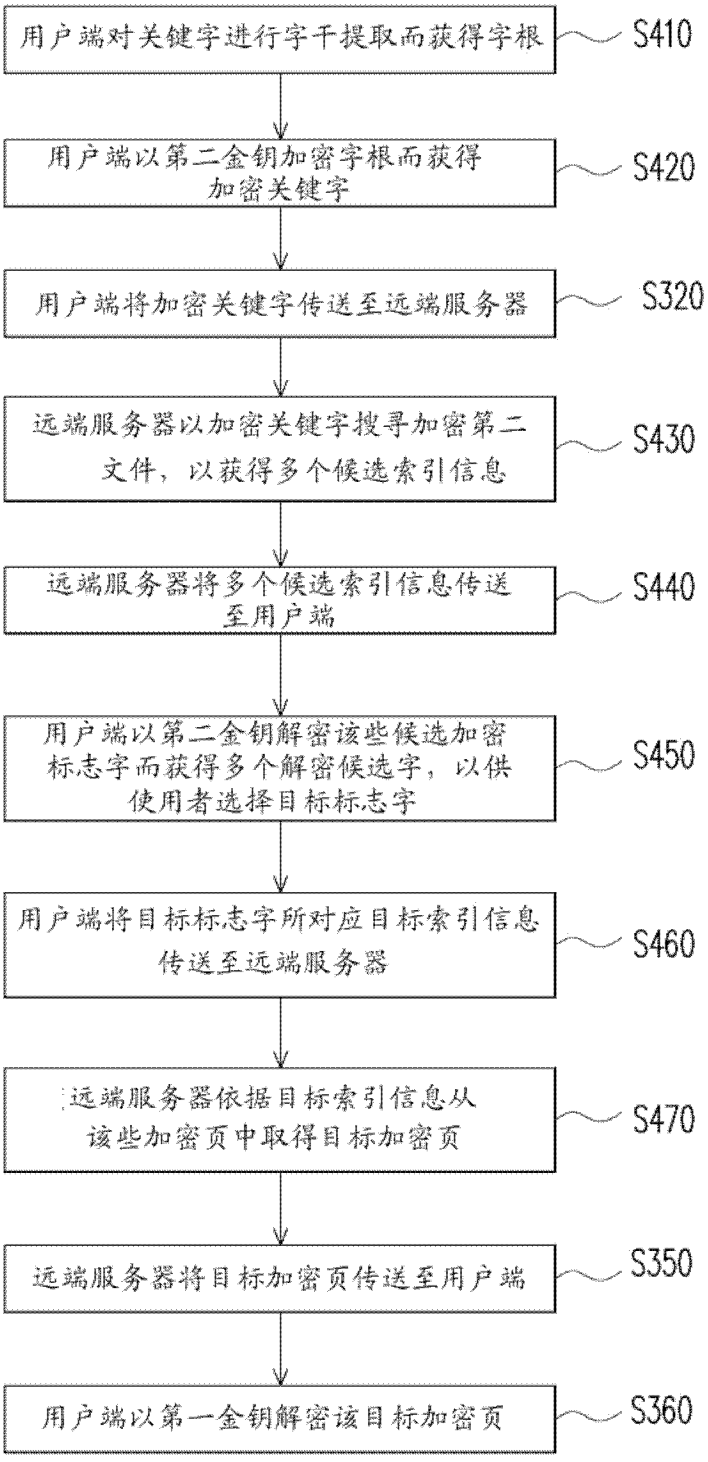

用户端对关键字进行字干提取而获得字根
S410
用户端以第二金钥加密字根而获得加密关键字
S420
用户端将加密关键字传送至远端服务器
S320
远端服务器以加密关键字搜寻加密第二文件，以获得多个候选索引信息
S430
远端服务器将多个候选索引信息传送至用户端
S440
用户端以第二金钥解密该些候选加密标志字而获得多个解密候选字，以供使用者选择目标标志字
S450
用户端将目标标志字所对应目标索引信息传送至远端服务器
S460
远端服务器依据目标索引信息从该些加密页中取得目标加密页
S470
远端服务器将目标加密页传送至用户端
S350
用户端以第一金钥解密该目标加密页
S360

图 4

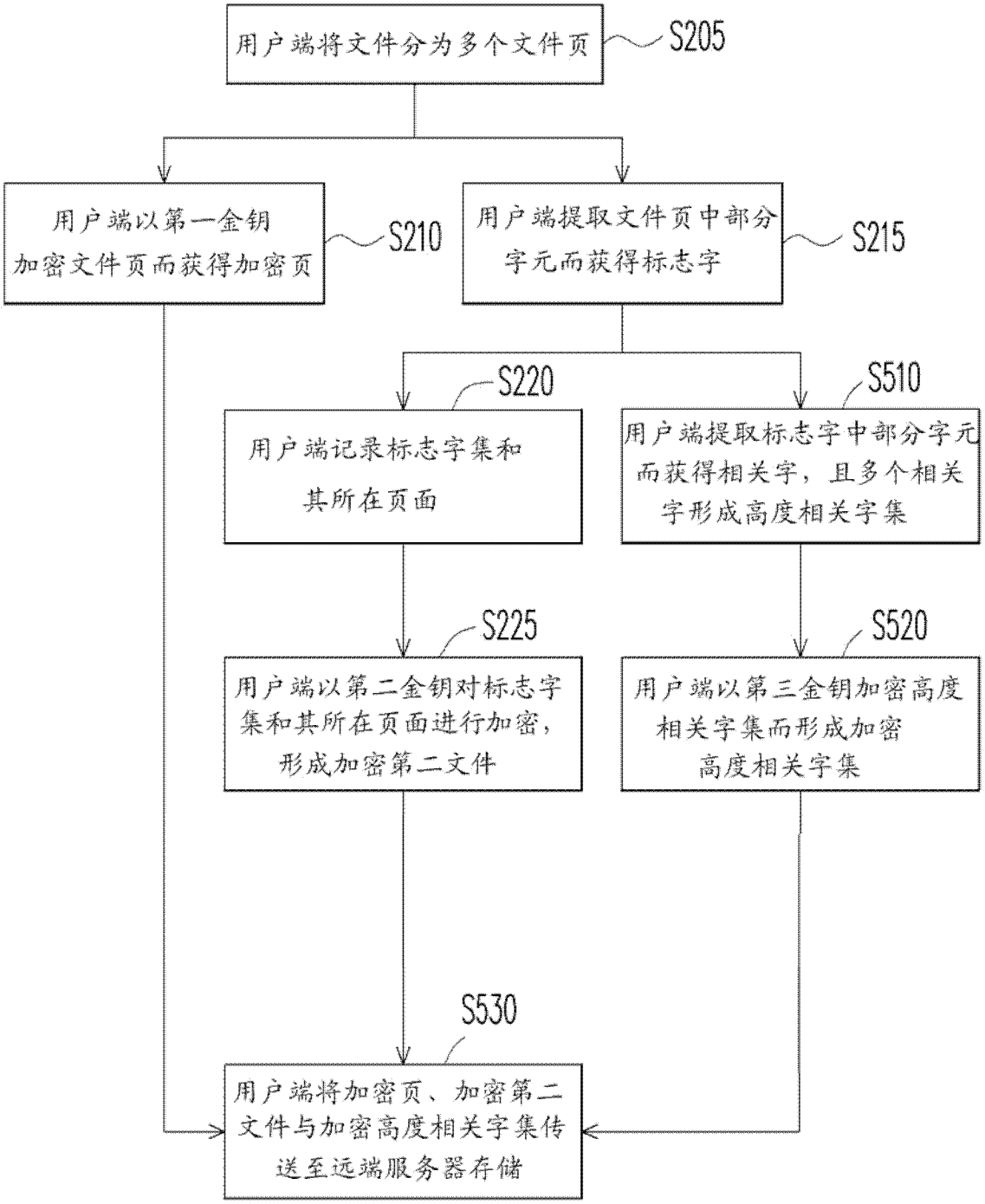

用户端将文件分为多个文件页
S205
用户端以第一金钥加密文件页而获得加密页
S210
用户端提取文件页中部分字元而获得标志字
S215
S220
用户端记录标志字集和其所在页面
S510
用户端提取标志字中部分字元而获得相关字，且多个相关字形成高度相关字集
S225
用户端以第二金钥对标志字集和其所在页面进行加密，形成加密第二文件
S520
用户端以第三金钥加密高度相关字集而形成加密高度相关字集
S530
用户端将加密页、加密第二文件与加密高度相关字集传送至远端服务器存储

图 5

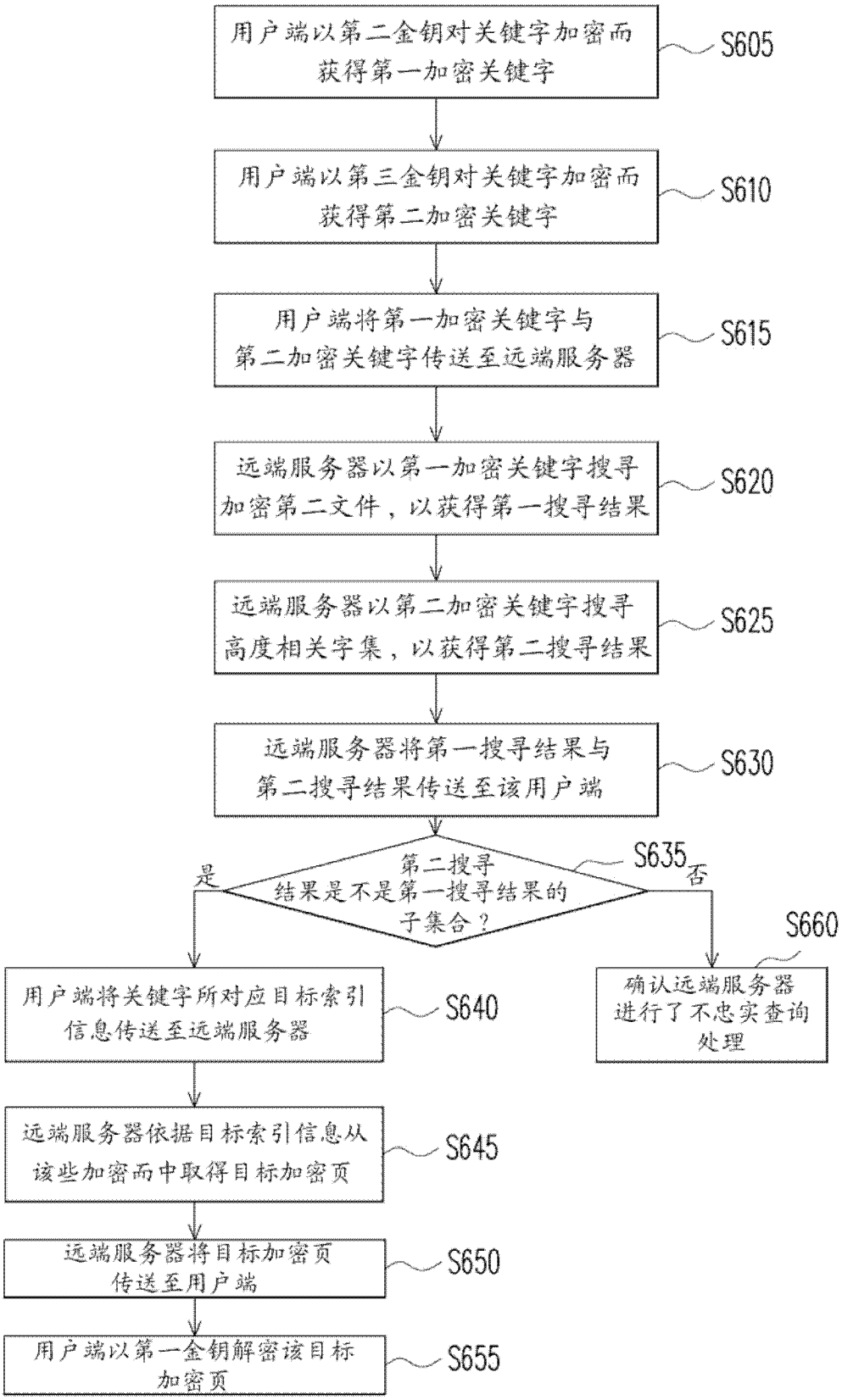

用户端以第二金钥对关键字加密而获得第一加密关键字
S605
用户端以第三金钥对关键字加密而获得第二加密关键字
S610
用户端将第一加密关键字与第二加密关键字传送至远端服务器
S615
远端服务器以第一加密关键字搜寻加密第二文件，以获得第一搜寻结果
S620
远端服务器以第二加密关键字搜寻高度相关字集，以获得第二搜寻结果
S625
远端服务器将第一搜寻结果与第二搜寻结果传送至该用户端
S630
第二搜寻结果是不是第一搜寻结果的子集合？
S635
是
否
用户端将关键字所对应目标索引信息传送至远端服务器
S640
确认远端服务器进行了不忠实查询处理
S660
远端服务器依据目标索引信息从该些加密而中取得目标加密页
S645
远端服务器将目标加密页传送至用户端
S650
用户端以第一金钥解密该目标加密页
S655

图 6