



[12] 发明专利说明书

专利号 ZL 200410068688.3

[45] 授权公告日 2008年5月28日

[11] 授权公告号 CN 100390792C

[22] 申请日 2004.9.2

[21] 申请号 200410068688.3

[30] 优先权

[32] 2003.9.8 [33] US [31] 10/657,458

[73] 专利权人 国际商业机器公司

地址 美国纽约

[72] 发明人 罗伯托·J·拜亚多

拉克史·阿格拉瓦尔 巴瓦·玛彦克

[56] 参考文献

CN1271906A 2000.11.1

US2002/0026345A1 2002.2.28

US6389412B1 2002.5.14

CN1306258A 2001.8.1

US6018733A 2000.1.25

US5845278A 1998.12.1

审查员 董刚

[74] 专利代理机构 中国国际贸易促进委员会专利
商标事务所

代理人 李颖

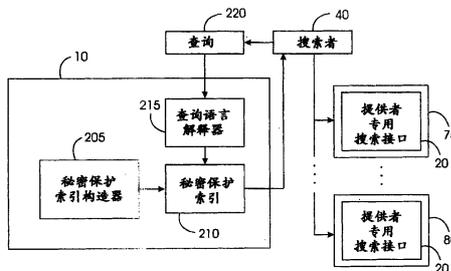
权利要求书3页 说明书16页 附图8页

[54] 发明名称

有选择地共享分布式访问受控文档的统一搜索系统和方法

[57] 摘要

秘密保护索引系统解决关于分布式访问受控内容，提供秘密保护搜索的问题。根据在搜索中使用的倒排索引，能够容易地重构索引文档。连同分布式访问控制执行搜索协议一起，秘密保护索引系统建立集中式秘密保护索引。秘密保护索引利用随机化算法构成秘密保护索引。秘密保护索引对保密性破坏的适应性极强。秘密保护索引系统允许内容提供者保持对定义访问组和确保其遵从性的完全控制，另外允许系统实现者保留可调的旋钮调节器，以便为其特定的域平衡保密性和效率。



- 1、一种有选择地共享多个分布式访问受控文档的方法，包括：
多个内容提供者协作产生秘密保护索引结构；
将内容提供者组成多个秘密组；
秘密保护索引结构把代表要共享的内容的多个关键字映射到多个内容提供者；和
返回呈现要共享的内容中的一个或多个关键字的内容提供者的列表。
- 2、按照权利要求 1 所述的方法，其中内容提供者包括接收查询并验证搜索者的提供者专用搜索接口。
- 3、按照权利要求 2 所述的方法，还包括搜索者向秘密保护索引系统提交包含一个或多个关键字的查询。
- 4、按照权利要求 1 所述的方法，其中内容提供者的列表包括至少 50% 的虚假肯定的内容提供者。
- 5、按照权利要求 1 所述的方法，还包括搜索者向内容提供者列表上的指定内容提供者提交以搜索者的身份注释的查询。
- 6、按照权利要求 5 所述的方法，还包括指定的内容提供者验证搜索者的身份，以便允许访问要共享的内容。
- 7、按照权利要求 6 所述的方法，还包括指定的内容提供者向搜索者返回和一个或多个关键字匹配的一个或多个文档。
- 8、按照权利要求 1 所述的方法，其中至少一个秘密组包括至少三个内容提供者。
- 9、按照权利要求 1 所述的方法，其中每个内容提供者被分入一个秘密组。
- 10、按照权利要求 8 所述的方法，还包括执行随机化索引构成算法，以便产生至少一个秘密组中的内容提供者的位矢量。
- 11、按照权利要求 10 所述的方法，还包括按照环形，排列秘密组中的内容提供者。

12、按照权利要求 11 所述的方法，还包括产生多个内容提供者的位矢量，一个或多个关键字代表要共享的内容。

13、按照权利要求 12 所述的方法，还包括把位矢量组合成组矢量。

14、按照权利要求 13 所述的方法，还包括把组矢量保存在秘密保护索引结构中。

15、按照权利要求 11 所述的方法，其中产生内容提供者的位矢量包括产生随机化位矢量。

16、按照权利要求 15 所述的方法，还包括把随机化位矢量传递给秘密组中环形中的第一提供者。

17、按照权利要求 10 所述的方法，其中内容提供者执行随机化索引构成算法，以产生提供者关键字位矢量。

18、按照权利要求 16 所述的方法，还包括第一提供者把提供者关键字位矢量传递给秘密组中环形中的下一提供者。

19、按照权利要求 17 所述的方法，其中环形中的内容提供者顺序对提供者关键字位矢量执行随机化索引构成算法。

20、按照权利要求 19 所述的方法，还包括环形中的内容提供者传递提供者关键字位矢量，并对提供者关键字位矢量执行随机化索引构成算法，直到围绕所述环形，提供者关键字位矢量已完成 r 次循环为止。

21、按照权利要求 20 所述的方法，还包括对提供者关键字位矢量进行或运算，得到组关键字位矢量。

22、按照权利要求 21 所述的方法，其中对提供者关键字位矢量进行或运算，得到组关键字位矢量，会在响应查询返回的结果中引入虚假肯定。

23、一种有选择地共享多个分布式访问受控文档的系统，包括协作产生秘密保护索引结构的多个内容提供者，所述系统还包括：

将内容提供者组成多个秘密组的装置；

通过秘密保护索引结构把代表要共享的内容的多个关键字映射

到多个内容提供者的装置；和

返回呈现要共享的内容中的一个或多个关键字的内容提供者的列表的装置。

24、按照权利要求 23 所述的系统，其中所述内容提供者包括接收查询并验证搜索者的提供者专用搜索接口。

25、按照权利要求 24 所述的系统，还包括秘密保护索引系统，其中搜索者向所述秘密保护索引系统提交包含一个或多个关键字的查询。

26、按照权利要求 25 所述的系统，其中所述内容提供者中的至少一些响应于收到查询而返回过滤文档的列表。

有选择地共享分布式访问 受控文档的统一搜索系统和方法

技术领域

本发明一般涉及对通过诸如因特网或万维网之类网络定位的访问受控数据储存库（**access-controlled data repository**）进行搜索。更具体地说，本发明涉及统一搜索多个分布式访问受控数据储存库的数字产权管理工具。

背景技术

虽然近年来，因特网上的专用和半专用信息已快速增长，但是，搜索这种信息的机制未能与之齐步前进。面临查找访问受控文档的问题的用户通常会识别并单独搜索每个相关储存库，当然假定用户知道并记得哪些储存库是相关的。

例如，公司 XYZ 希望与公司 ABC 共享他们的一些，而不是全部内部研究文献。公司 XYZ 希望共享的文献可能涉及这两个公司之间的协作项目。公司 XYZ 想要能够提供关于该数据的搜索工具，公司 ABC 只能搜索他们可以使用的文献。但是，公司 XYZ 并不希望公司 ABC 能够确定公司 XYZ 正在与公司 Q 共享什么。当前，在希望按照访问受控格式共享数据的公司和个人之间，不存在按照这种格式统一搜索数据的方法。

搜索网络上的访问受控内容的工具的缺少起源于在考虑内容提供者的安全性和保密性要求的同时，产生对内容编制索引的搜索引擎的困难相当大。当代的搜索引擎建立倒排索引（**inverted index**），倒排索引把关键字映射到其在索引文档中的准确位置。

常规的倒排索引实质上整体代表一个索引文档。从而，能够根据索引容易地重建索引文档。要求提供关于访问受控内容的这种索引的任意主机非常可信和安全。在被赋予有关每个可搜索文档的知识的情况下，对搜索引擎关于访问受控内容所要求的信任对于每个参与提供

者都快速增长。这种巨大的信任要求，伴随着通过恶意的索引公开而完全违背访问控制的可能性，使这种方法不切实际。

常规的搜索方案包括集中式索引，查询广播，分布式索引，和集中式模糊索引。支持关于分布内容的有效搜索的最常见方案是集中式索引，其中建立集中式倒排索引。该索引把每个检索词映射到包含该检索词的一组文档。搜索者查询该索引，获得匹配文档的列表。这是 web 搜索引擎和协调器选择的方案。

通过把访问策略连同内容一起传送给索引主机，集中式索引可被扩展，从而支持访问受控搜索。索引主机把这些策略应用于每个搜索者，从而恰当地过滤搜索结果。由于只需要联系索引主机即可完整地执行搜索，因此搜索的效率相当高。但是，集中式索引可允许可以访问索引结构的任何人“可查明地暴露”内容提供者。当对手（即，黑客）能够提供提供者 p 正在共享文档 d 的确凿证据时，发生可查明的暴露。在所有内容提供者完全信任索引主机的情况下，访问控制的这种违背是可以容忍的。找到这样的可信主机是非常困难的。此外，如果索引被公开展示，那么黑客对索引主机的损害会导致完全并且破坏性的失密。

查询广播位于搜索效率范围的另一端，查询广播是向所有参与内容提供者发送查询的基于广播的方案。这种方案包括内容提供者的网络，这里提供者本地评估每个查询，并直接把任意匹配文档提供给搜索者。可扩大查询广播搜索协议，以便实现访问控制。在这种协议中，查询可和查询始发者的身份和 IP 地址一起被广播。通过加密连接，提供者能够安全地把搜索结果回传给经过验证的搜索者，从而避免被截取。

由于提供者 p 共享的内容单独存在于该提供者的数据库，提供者被保证绝对保密性，自然保持了内容保密性的目的。但是，虽然对查询广播的这种修改具有极好的保密特性，但是它存在可缩放性差和严重的性能恶化的问题。从而，查询广播的协议采用限制搜索范围，损害搜索完整性的试探法（例如，生存期字段）。

查询广播的性能限制已导致研究不需要单个集中式索引提供者，支持有效搜索的分布式索引方法。例如，通过使“超级对等体”（具有

高于平均值的带宽和处理能力的机器) 托管数个能力较差的机器共享的内容的子索引, 对等网络可影响 (leverage) “超级对等体”。

另一系统利用分布式散列表, 分发搜索索引。在这些系统中, 分布式索引被用于识别和搜索者的查询匹配的一组文档 (或者托管这些文档的机器)。搜索者随后直接联系这些机器, 以便取回匹配的文档。

通过在提供文档之前, 简单地使提供者执行他们的访问策略, 能够支持分布式索引系统的访问控制。但是, 和集中式索引的情况一样, 可以访问一部分分布式索引的任意节点能够可查明地暴露该部分所索引的任意提供者。

此外, 索引通常由提供者本身不能控制的不可信机器托管。不托管一部分索引的主动对手能够搜索分布式索引, 造成破坏保密性。例如, 通过发出关于特定关键字的搜索, 利用可查明的暴露破坏内容保密性, 对手能够确定共享具有该关键字的文档的提供者的准确列表。通过安装短语攻击, 也能够破坏内容保密性。这种攻击利用了多数文档具有它们独有的多组特有单词的观察结果。

为了识别共享某一文档的提供者, 对手只需要构成由该文档的这种检索词组成的查询即可。随后, 所得到的站点的列表被认为共享该文档, 但是可能是清白的。当对手关于共享文档 d 的提供者 p 的主张可能错误的概率较大时, 发生可能清白的情况。通过选择恰当的一组检索词, 对手能够实现几乎可查明的暴露。

一些搜索应用程序不保持准确的倒排索引列表, 而是保持允许把查询映射成可能包含匹配文档的一组“模糊的”提供者的结构; 这种方法被称为集中式模糊索引。搜索者能够探查 bloom filter 索引 (它是一种模糊索引), 从而识别包含和查询匹配的文档的所有提供者的列表。但是, 该列表不必是准确的, 因为由于散列冲突的缘故, bloom filter 可能产生虚假的肯定。在已知这种列表的情况下, 搜索者联系每个提供者, 从而累积结果。通过在搜索者请求匹配文档的时候, 使提供者执行他们的访问策略, 可扩展这些方案, 以便支持访问受控搜索。

由于提供者列表中潜在的虚假肯定, bloom filter 索引提供有限的保密特性。从而, 该列表中的每个提供者可能没有共享和查询匹配的文档。但是, 这种保密性是假的。主动对手能够对 bloom filter 索引

进行基于词典的攻击，从而识别任意索引的提供者的检索词分布。

基于词典的攻击利用了自然语言（例如英语）的句子使用来自易于编辑的有限词汇表（例如 Oxford/Webster 词典中）的单词的事实。从而，对手能够计算该词汇表中每个单词的散列。关于这种散列的 bloom filter 项中的提供者可能共享具有对应单词的文档。另外，该方案仍然易于受到短语攻击。

虽然这些常规的搜索方案可被修改，以支持对访问受控内容的搜索，但是，这样的修改不能充分解决保密性和效率方面的问题。由于索引本身传递的准确信息的缘故，依赖于常规的搜索索引的任意搜索机制允许“可查明地暴露”提供者。于是，有效的保持保密性的搜索需要一种即使在索引公开的情况下，也能够防止破坏“内容保密性”的索引结构。

需要一种允许搜索者特许访问那些访问受控的文档，而不会向未经授权搜索者暴露文档的内容，文档的提供者，或者甚至文档的存在的系统和相关方法。迄今为止，关于这种系统和方法的需要仍然未被满足。

发明内容

本发明满足这种需要，提出一种提供有效搜索机制的系统，服务，计算机程序产品和相关方法（这里总称为“系统”或“本系统”），所述有效搜索机制注意到了参与内容提供者的在保密性方面的担心。本系统允许公司和个人保持对他们自己数据的控制，同时提供一种搜索机制，所述搜索机制高效，但是不会仔细地向未经授权的搜索者公开哪些数据正被共享。被展示的信息是“模糊的”，从而未经授权的搜索者不能肯定地认为什么信息正被共享。本发明的特定索引结构不允许搜索者或者对手得出什么信息正被所有各个内容提供者共享的结论。

根据本发明的一个方面，提供一种有选择地共享多个分布式访问受控文档的方法，包括：多个内容提供者协作产生秘密保护索引结构；将内容提供者组成多个秘密组；秘密保护索引结构把代表要共享的内容的多个关键字映射到多个内容提供者；和返回呈现要共享的内容中的一个或多个关键字的内容提供者的列表。

根据本发明的另一个方面，提供一种有选择地共享多个分布式访问受控文档的系统，包括协作产生秘密保护索引结构的多个内容提供者，所述系统还包括：将内容提供者组成多个秘密组的装置；通过秘密保护索引结构把代表要共享的内容的多个关键字映射到多个内容提供者的装置；和返回呈现要共享的内容中的一个或多个关键字的内容提供者的列表的装置。

响应对索引的主动对手攻击，索引的文档的提供者被确保至少“或许清白”。本系统在连网提供者中建立与访问控制强制执行搜索协议一起起作用的内容的集中式索引。集中式索引本身提供即使整个索引被公开，仍然成立的有力的可计量的（quantifiable）保密保证。集中式索引提供的保密程度可被调整，以满足提供者的需要。搜索协议引起的开销正比于提供的保密程度。

本系统可用在不同的部门中，所述不同部门中，多个组织正在积极竞争，以及不断合作形成联盟。另一应用领域是通过个人 web 服务器的文件共享。例如，某人可能希望在工作时收听某张 CD 或某一歌曲，但是该 CD 放在另一地方。他可使用本系统搜索可以电子方式从其它个人或公司得到的取得版权的歌曲。他出示所有权的证据，身份验证，随后能够收听该 CD 或歌曲。CD 或歌曲的提供者跟踪提供的证据，从而允许这种交换的审计。本系统提供随后让该人搜索具有该 CD 或歌曲的任何人，并使他能够访问该 CD 或歌曲的搜索机制。

本系统保留了私有信息共享的重要魅力。每个提供者完全控制它享有的信息：有多少被共享，何时被共享，以及与谁共享。

附图说明

下面将参考下面的说明，权利要求和附图，更详细地描述本发明的各个特征以及获得这些特征的方式，附图中，恰当地重复使用附图标记表示引用对象之间的对应性，其中：

图 1 示意地图解说明了其中可使用本发明的秘密保护索引（privacy-preserving index）系统的例证工作环境；

图 2 是图 1 的秘密保护索引系统的高级体系结构的方框图；

图 3 是图解说明响应来自搜索者的查询，图 1 和 2 的秘密保护索引系统的操作的方法的流程图；

图 4 是图 1 的提供者专用搜索接口的高级体系结构的方框图；
 图 5 图解说明把内容提供者分为秘密组；
 图 6 图解说明内容提供者产生的位矢量；
 图 7 是图解说明图 1 和 2 的秘密保护索引系统产生秘密保护索引的操作的方法的流程图；和
 图 8 图解说明图 1 和 2 的秘密保护索引系统关于对等的一组内容提供者产生的位矢量。

具体实施方式

下面的定义和说明提供与本发明的技术领域有关的背景技术，意图便于理解本发明，而不是对本发明范围的限制：

绝对秘密：对手不能确定提供者 p 是否正在共享文档 d 。

对手：主动或被动地，故意地或者无意地收集和各个提供者托管的内容相关的越权信息。对手可单独行动或者与其它对手串通行动，从而破坏内容提供者的秘密。

无可置疑：对手不能确定提供者 p 是否比任意其它提供者更可能共享文档 d 。

Bloom Filter: bloom filter 是由 N 位阵列构成的模糊组 (set) 索引结构。这里使用 bloom filter 索引一组关键字 K 。建立 bloom filter 需要把关键字映射成范围 $1...N$ 中的值的散列函数 $H()$ 。在已知该组关键字 K 和散列函数 H 的情况下，本发明产生 bloom filter $B[1...N]$ ，如下所示：

(1) 把所有位 $B[1...N]$ 设置成 0，和

(2) 对于 K 中的每个关键字 k ，把 $B[H(k)]$ 设置成 1。

bloom filter 允许本发明非常有效地回答下述例证形式的查询：“索引的一组关键字包含关键字 k ？”。这是通过检查 $B[H(k)]$ 的值实现的。如果位为 0，那么该组明确地不包含关键字 k 。如果位为 1，那么该组可能包含该关键字（必须查阅实际的组本身，以便肯定的核实）。bloom filter 是一种非常有用的快速识别并除去不能回答指定查

询的提供者的结构。

对等体：连网中，与另一方位于相同协议层的功能单元。

对等网络：其中网络上的任意计算机可以是客户机和/或服务器的通信网络。任意计算机能够访问网络中的任意其它计算机上的文件。

可能清白 (possible innocence)：对手关于提供者 p 共享文档 d 的主张是错误的概率较大 (例如概率介于 (0.5, 1) 之间)。

或许清白 (probable innocence)：和真的相比，对手关于提供者 p 共享文档 d 的主张更可能是假的 (例如，概率介于 (0, 0.5) 之间)。

可查明的暴露：对手能够提供提供者 p 共享文档 d 的确凿证据。

图 1 图解说明其中可使用根据本发明的，有选择地共享分布式访问受控文档的统一搜索系统和相关方法的例证整体环境。系统 100 包括秘密保护索引系统 10 和提供者专用搜索接口 15。秘密保护索引系统 10 包括通常嵌入或者安装在秘密保护索引服务器 25 中的软件编程代码或者计算机程序产品。提供者专用搜索接口 15 包括通常嵌入或者安装在提供者服务器 30、35 中的软件编程代码或计算机程序产品。

另一方面，秘密保护索引系统 10 和提供者专用搜索接口 15 可保存在适宜的存储媒体上，例如磁盘，CD，硬盘驱动器或类似装置上。虽然下面将结合 WWW 说明秘密保护索引系统 10 和提供者专用搜索接口 15，不过它们也可和从 WWW 和/或其它来源得到的检索词 (term) 的独立数据库一起使用。

云状通信网络 20 可由连接诸如秘密保护索引服务器 25 和提供者服务器 30、35 之类服务器，提供对 WWW 或因特网的通信访问的通信线路和交换机组成。搜索者，例如搜索者 40 通过网络 20，关于所需的信息，查询秘密保护索引服务器 25。搜索者 40 可以是个人，公司，应用程序等。计算机 45 包括允许用户浏览因特网，并且安全地连接秘密保护索引服务器 25 和提供者服务器 30、35 的软件。秘密保护索引服务器 25，提供者服务器 30、35，和计算机 45 通过诸如电话，电缆或卫星链路之类通信链路 50、55、60、65，与网络 20 连接。

在图 1 的例证环境中，秘密保护索引系统 10 保存在 dB 70 上。内容提供者 75、80（这时也称为提供者 75、80）把一组文档保存在他们各自的数据库，提供者数据库 85、90 上。提供者 75、80 通过提供者专用搜索接口 15，控制对他们各自的提供者数据库 85、90 上的文档的访问。

图 2 的方框图图解说明了秘密保护索引系统 10 的高级体系结构。秘密保护索引系统 10 由秘密保护索引构造器 205，秘密保护索引 210，和查询语言解释器 215 组成。当最初产生秘密保护索引 210 时，秘密保护索引构造器 205 把查询检索词映射到提供者 75、80 的列表。

图 3 中的流程图图解说明了秘密保护索引系统 10 的操作的方法 300。在方框 305，搜索者 40 以一个或多个关键字的形式，向秘密保护索引系统 10 提交查询 220。在方框 310，秘密保护索引 210 向搜索者 40 返回包含可能包含这些关键字的文档的提供者 75、80 的列表。作为系统 100 的一个特征，提供者 75、80 的列表可包含至少 50% 的虚假肯定，即一半或小于一半的返回的提供者 75、80 实际具有包含这些关键字的文档。搜索者 40 随后用以搜索者 40 的访问特权和验证注解的关键字，搜索这些指定的提供者 75、80（方框 315）。在方框 320，提供者 75、80 验证搜索者 40，并在方框 325，以匹配关键字的文档进行响应。提供者只返回既和查询匹配，并且允许用户访问的文档。

图 4 的方框图图解说明了提供者专用搜索接口 15 的高级体系结构。提供者专用搜索接口 15 包括查询语言解释器 405，查询执行引擎 410，验证机构 415，访问策略语言 420，和访问策略执行器（enforcer）425。提供者专用搜索接口 15 的输入是注释的查询 435。注释的查询 435 包括用搜索 40 的身份注释的查询 220。查询语言解释器 405 获得注释查询 435，并将其转换成供查询执行引擎 410 之用的机器语言。查询语言解释器 405 应支持连接（conjunctive）关键字查询。也可支持其它结构（例如，短语搜索，否定检索词等），只要它们进一步约束结果集合即可。验证机构 415 使用的验证方案应允许搜索者 40 独立地向每个提供者 75、80 验证它自己。系统 100 的一个实施例并不要求

明确地向每个提供者 75、80 登记。相反，搜索者 40 通过第三方签署的安全证书（例如 SSL/TLS），实现客户验证。通过利用访问策略语言 420，在已知搜索者 40 的验证身份的情况下，提供者 75、80 能够应用并强制执行他们的访问策略。这允许每个提供者 75、80 单独地选择最适合他们的要求的访问策略语言 420。

查询执行引擎 410 把一组文档识别成和注释查询 435 匹配。访问策略执行器 425 根据验证机构 415 从注释查询 435 确定的搜索者 40 的身份和具体访问策略，过滤这些文档。过滤后的一组文档 440 被返回给搜索者 40。

秘密保护索引 210 是建在所述一组提供者 75、80 共享的一组文档 D 上的映射函数。它接受查询 220 (q 220)，并返回可能包含匹配文档的提供者子集 M。对于被看作秘密保护的功能来说，对于任意查询 q 220 的集合 M 应满足下述条件之一：

- 只有当 D 中不存在和 q 220 匹配的文档时，M 才是空集合。
- M 是包含共享匹配 q 220 的文档（“真实的肯定”）的所有提供者，和数目相同或者更多的不共享匹配文档（“虚假的肯定”）的提供者的提供者 75、80 的子集。
- M 是所有提供者 75、80 的集合。

秘密保护索引 210 应表现得象常规的索引：即，关于相同的查询 220，秘密保护索引 210 应返回相同的结果，除非被索引的内容本身已发生变化。另外，对于其结果是另一查询 q 220 的子集的任意查询 q' 来说，关于 q' 返回的结果集合应是关于 q 220 返回的结果集合的子集。这些行为要求通过滤除虚假的肯定，防止尝试破坏保密性的攻击。

应当小心地实现秘密保护索引 210：和秘密保护索引 210 的定义允许产生的信息相比，自然（naive）实现能够容易地产生更多信息。例如，秘密保护索引 210 的主机可本地聚集所有共享的内容，并对其进行处理，以便实现只有真实肯定的秘密保护索引 210；在查询 220 的时候，定义所需的虚假肯定被插入结果中。这种情况下，秘密保护索引 210 本身的实现形式并不对应于秘密保护索引 210 的定义。秘密

保护索引 210 的实现形式的公开会导致提供者 75、80 的可查明暴露。相反，系统 100 要求秘密保护索引 210 的实现形式不应产生比对照秘密保护索引 210，执行查询 220 的详尽列表获得的信息更多的信息。

秘密保护索引 210 关于查询 q 220 返回的集合 M 决不排斥关于 q 220 的任意真实肯定。换句话说，关于查询 220 的结果集合可包括至少具有一个匹配文档的所有提供者 75、80。搜索者 40 联系每个提供者 75、80，从而累积结果；只有当搜索者 40 具有足够的访问特权时，提供者 75、80 才释放文档。从而，借助秘密保护索引 210 的搜索导致正确的输出。

一般地说，搜索分布式访问受控内容可被表述成一组内容提供者 P_1, P_2, \dots, P_n ，和发出查询 q 的搜索者 s 。每个提供者被认为共享依据搜索者 s 的验证身份和访问策略确定的具有访问控制的一组文档。所需的输出是包含文档 d 的一组文档，从而：

d 由某一提供者 P_i 共享， $1 < i < n$ ，

d 和查询 q 匹配，并且

如同 P_i 的访问策略所示，对于 s 来说， d 是可访问的。

和确保关于查询 q 220 的正确输出一样重要的是防止对手在未获得恰当的访问权限的情况下，获悉哪一个或多个提供者可能正在共享文档的要求。利用提供者 75、80 和秘密保护索引系统 10 对上述类型的对手破坏保密性的敏感度，描述了对保护秘密的问题的解决方案。

被动对手是仅仅观察并记录在系统中发送的消息（查询，响应，索引）的偷听者。这样的对手可全局（观察系统中的所有消息的能力）或者局部（观察发送给特定内容提供者/从特定内容提供者发送的消息的能力）地观察系统。主动对手是按照系统协议有意行动，从而收集信息的实体。在我们的模型中，这样的对手可检查索引结构，发出各种查询，或者甚至参与索引构成过程，从而易于破坏保密性。对手还可相互勾结来破坏保密性。

还可根据对手扮演的角色对他们分类。例如，多数用户（从而多数对手）局限于扮演搜索者 40 的角色，因为内容提供者 75、80 实际

上可能是较小并且更加受控的群体。通过每个角色（搜索者 40，提供者 75、80 或秘密保护索引系统 10）可访问的信息和操作可被用于简化不同类型的违背。

系统 100 集中于关于某一内容提供者 p 使之可搜索的文档 d ，获得下述保密目的：

不应允许对手 A 推断出 p 正在共享包含关键字 q 的某一文档 d ，除非 p 已准许 A 访问 d 。

利用 Reiter 和 Rubin 在他们的 Crowds 分析中介绍的秘密范围（privacy spectrum），表征对抗不能访问提供者 p 共享的文档 d 的对手，获得内容保密的程度：

- 可查明的暴露：对手能够提供 p 正在共享 d 的确凿证据。
- 可能清白：对手关于 p 共享 d 的主张是错误的概率较大（例如概率介于（0.5, 1）之间）。
- 或许清白：和真的相比，对手关于 p 共享文档 d 的主张更可能是假的（例如，概率介于（0, 0.5）之间）。
- 绝对秘密：对手不能确定 p 是否正在共享 d 。
- 无可置疑：对手不能确定 p 是否比任意其它提供者更可能共享文档 d 。

在上面的讨论中， d 可用任意一组关键字 q 220 代替。这种情况下，目的是防止对手确定 p 是否正在共享包含 q 220 中的关键字的文档。

常规的倒排列表把查询映射成匹配文档的列表，而秘密保护索引 210 把查询映射成匹配提供者 75、80 的列表。在已知可能满足某一查询的提供者 75、80 的列表的情况下，直接查询这样的提供者 75、80，并请求匹配的文档是搜索者 40 的责任。当收到查询并验证搜索者 40 时，提供者 75、80 返回根据搜索者 40 的访问权限过滤的文档的列表。

通过按照这种方式实现搜索，系统 100 把访问控制点从秘密保护索引 210 的主机转移到提供者 75、80。提供者 75、80 现在能够自己管理并强制执行访问策略，而无须依赖于任何中央主机。虽然存在与

需要单独联系提供者 75、80 相关的效率损失，关于公开共享内容的实验结果表明即使存在许多 (>1500) 提供者 75、80，这种方法的性能实际上也是相当合理的。

构成秘密保护索引 210 的程序不仅应解决所得到的结构的正确性，而且应解决构成过程中，保密性破坏的可能性。在存在对手的情况下确保秘密至关重要，因为秘密保护索引 210 的构成过程涉及把和每个提供者 75、80 共享的内容相关的信息汇集在一起。

为了构成秘密保护索引 210，提供者被分成大小为 c 的对等体组或“秘密组”，如图 5 的例子所示。图 5 中，许多提供者 75、80 被分成对等体组 G_1 505， G_2 510， G_3 515 和 G_4 520。并不要求对等体组大小完全相同，不过大小应近似相同。

每个提供者 75、80 正好在一个对等体组中，并且均包括提供者专用搜索接口 15。组 G_1 505 由诸如 P_1 525， P_2 530 和 P_3 535 之类提供者 75、80 构成。在一个组中，提供者 P_1 525， P_2 530 和 P_3 535 被排列成环形。提供者 P_1 525， P_2 530 和 P_3 535 执行随机化算法，以便构成只具有较小的错误概率的秘密保护索引 210。通过调整参数，可使错误足够小，从而实际上不相关。该构成过程确保提供者对超出或许清白的违背有较强的适应力。

每个提供者 75、80 根据包含在它自己数据内的关键字，翻转“内容矢量”中的位。但是，内容矢量沿着其对等体组内的一连串成员被传递。从而，随机化算法作用于在组中的对等体之间传递的内容矢量。但是，对等体翻转的位的实际模式由该对等体自己的数据确定。提供者 75、80 确定他们希望哪些数据是可搜索的，随后把这些数据放置在他们自己的提供者服务器 30、35 上，提供者服务器 30、35 正在运行系统 100。提供者 75、80 并不把他们的数据给予其他人，他们只是使所述数据存在于网络 20 上以供搜索。

存在两种例外，在这两种例外情况下，提供者 P_1 525， P_2 530 和 P_3 535 会因其秘密组内的对手而遭受比或许清白更严重的违背。就共享具有特定检索词的文档来说，刚好在主动对手之前的提供者 P_1 525，

P_2 530 和 P_3 535 可被保证只是可能清白。具体地说，对手邻居能够以最多 0.71 的概率确定所述环形上，其在先者是否正在共享某一特定检索词。

对于提供者 75、80 来说的另一例外是当在所述环形上，其两个邻居串通一起对抗它的时候。例如，提供者 P_1 525， P_2 530 可串通一起对抗 P_3 535。这种情况下，提供者 P_3 535 可被可查明地暴露成共享包含特定检索词的文档。通过使提供者 P_3 535 根据先前建立的信任关系，选择环形上它们邻居 P_1 525 和 P_2 530，能够使这样的违背降至最小。

该算法要求每个提供者 P_1 525， P_2 530 和 P_3 535 通过位矢量 V （称为其内容矢量）总结其共享内容内的检索词。图 6 中图解说明了提供者 P_1 525 的例证内容矢量 V 605。例如，内容矢量可以是如下形成的系统规定长度 L 的 bloom filter。每个提供者 P_1 525， P_2 530 和 P_3 535 通过把每位设置成 0，初始化其 V 605。随后，对于在其共享内容中出现的每个关键字检索词 t ，提供者 P_1 525， P_2 530 和 P_3 535 使用具有范围 $1, 2, \dots, L$ 的系统规定散列函数把 V_s 中的位置 $H(t)$ 设置成 1。在例证的内容矢量 V 605 中，检索词 610，“专利”被散列成由位 3 615 空间中的“1”代表的位 3, 615。

这样形成的内容矢量 V 605 是在提供者 P_1 525 处的共享内容的总结。如果该位为 0，那么确保 P_1 525 不共享包含检索词 610 的任何文档。如果该位为 1，那么检索词 610 可能出现于 P_1 525，也可能不出现在于 P_1 525，因为多个检索词可能被散列在相同的值，从而设置 V 605 中相同的位。通过增大长度 L 和/或使用多个散列函数，能够降低发生这种冲突的概率。

图 7 的流程图图解说明了构成秘密保护索引 210 的方法 700。通过把提供者 75、80 的空间分成大小 c 均大于 2 的拆散的秘密组，该构成过程开始于方框 705。秘密组的大小正比于每个参与者享有的秘密程度。分隔方案可随机地把成员分配给各组。对于每个秘密组，在方框 710，把提供者 75、80 排列成环形 p_1, p_2, \dots, p_c 。关于该排序，

按照通常的方式使用提供者 p 的术语继任者和前任者，另外要求 p_1 被定义为 p_c 的继任者（并且 p_c 被定义为 p_1 的前任者）。

通常，把组 G 的组内容矢量定义为矢量 V_G ，其起因于对来自组 G 中的每个提供者 P 的一组所有内容矢量进行逻辑 OR 运算。所述构成的下一部分是产生组内容矢量的随机化算法。在循环 $r=i$ 产生组内容矢量 V 的随机化算法的伪代码被总结如下：

INDEXCONSTRUCTION (r, V_s, V_G')

$P_{ex} := 1/2^r$

$P_{in} := 1 - P_{ex}$

for ($i := 1; i < L; i := i+1$)

do

if ($V_s[i]=1$ and $V_G'[i]=0$)

then SET $V_G'[i] := 1$ WITH PROB. P_{in}

if ($V_s[i]=0$ and $V_G'[i]=1$)

then SET $V_G'[i] := 0$ WITH PROB. P_{in}

SEND V_G' TO Successor (s)

该构成涉及执行 r 次循环，其中沿着所述环形，把矢量 V_G' 从一个提供者传递给另一提供者。在方框 715，矢量 V_G' 被传递给环形中的第一提供者，在方框 720， i 被设置成 1。当收到该矢量时，每个提供者在方框 725 执行在产生组内容矢量的随机化算法中略述的位翻转操作。如果在方框 730， $i \leq r$ （这里 r 是环绕所述环形可传递所述矢量的循环的总次数），那么在方框 735，矢量 V_G' 被传递给该提供者的继任者，在方框 740， i 被加 1。在环绕所述环形的 r 次循环之后，在判定方框 730，矢量 V_G' 被发送给指定的索引主机，例如秘密保护索引系统 10 的主机（方框 745）。

在随机化算法中，矢量 V_G' 由 p_1 初始化成长度为 L 的矢量，同时以 $1/2$ 的概率把每个位独立地设置成 0 或 1。每个循环与概率 P_{in} 和 P_{ex} 相关，从而 $P_{in} + P_{ex} = 1$ 。 P_{ex} 的值最初为 $1/2$ 。在每个循环之后， P_{ex} 被减半，并恰当地设置 P_{in} 。

随机地翻转 V_G' 中的位的过程被设计成使得最终结果以较高的概率趋向于组内容矢量。位翻转的随机化被用于防止提供者组中的恶意提供者能够肯定地确定其它提供者的内容矢量中的位的值。

在完成 r 次位翻转循环之后，来自每个提供者组的矢量 V_G' 被发送给指定的主机，即秘密保护索引系统 10 的主机。该主机接收来自每个秘密组的这些矢量以及秘密组中所有提供者的列表。随后，它把这些矢量聚集成物化索引 MI。MI 把位位置 i 映射到属于其内容矢量具有设置为 1 的 i 的秘密组的提供者的列表。更正式地，

$$MI(i) = \{p | p \in G \wedge V_G'[i] = 1, \text{ 对于某一秘密组 } G\}$$

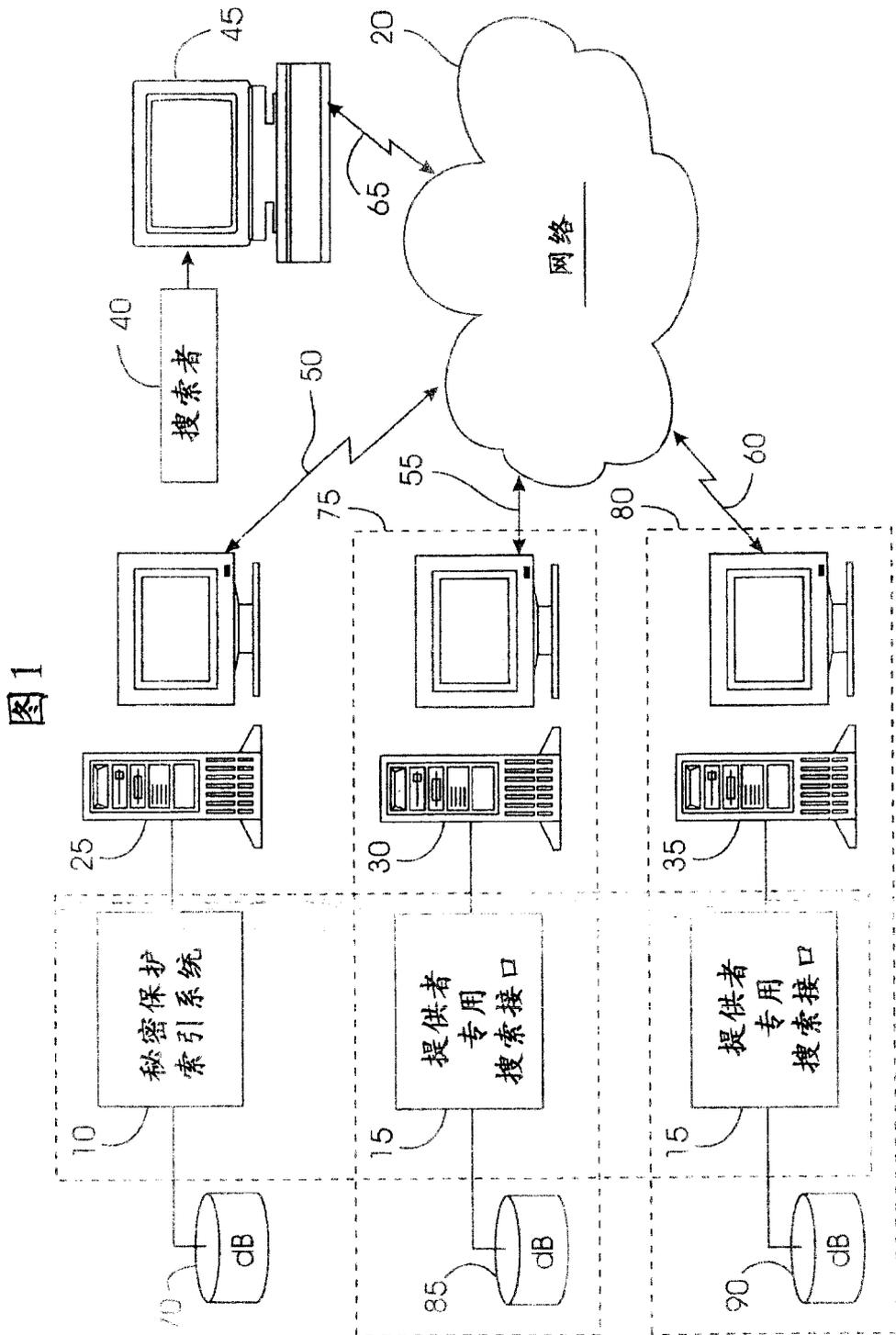
把 MI 用作把查询映射到提供者的秘密保护索引 210 的过程是直截了当的：通过获得在 q 22 中规定的结合项 Q ，并利用系统规定的查寻（散列）函数 H ，查寻 MI 中每项的位位置 $1, \dots, L$ ，形成 M_q 。通过获得每个这种位的 $MI(i)$ 的交集，形成提供者列表。更正式地， $M_q = \bigcap_{t \in Q} MI(H(t))$ 。从而，MI 用作秘密保护索引 210 的实现。

图 8 图解说明了方法 700 对某一组，例如 G_1 505 内的每个提供者 P_1 525， P_2 530 和 P_3 535 的单个 bloom filter 分组的最终效果。本质上，方法 700 对单个内容矢量 V_1 605， V_2 805， V_3 810 应用“或”函数，产生组矢量 V_{G_1} 815。例如，位 820 在每个内容矢量 V_1 605， V_2 805， V_3 810 的 b_0 位置中。为了获得 V_{G_1} 815 中的 b_0 位，“0”，“1”和“0”被一起进行“或”运算，如位 820 中所示。其结果是“1”。对于 V_{G_1} 815 中的所有位来说同样如此。虽然在本例中使用“或”函数，不过也可使用产生相同结果的任意其它适当的逻辑函数。

当搜索者 40 关于诸如“专利”610 之类关键字搜索秘密保护索引 210 时，秘密保护索引 210 发现它已被散列到 b_3 位 830。由于具有带检索词“专利”的文档，秘密保护索引系统 10 返回组 G_1 505 中的提供者 P_1 525， P_2 530 和 P_3 535 的列表。搜索者 40 随后知道搜索位于提供者 P_1 525， P_2 530 和 P_3 535 的储存库。但是，提供者 P_3 535 在其内容矢量 810 中并不具有关键字“专利”610；即， b_3 位 835 为 0。只有当以恰当的身份授权搜索位于 P_3 535 的储存库时，搜索者 40 才发现这种

情况。从而，对手不能确定地认为提供者 P₁ 525, P₂ 530 和 P₃ 535 中的哪个包含关键字“专利”610。

要明白已描述的本发明的具体实施例只是对本发明原理的某些应用的举例说明。在不脱离本发明的精神和范围的情况下，可对这里描述的本发明的有选择地共享分布式访问受控文档的统一搜索系统和方法做出各种修改。此外，虽然只是出于举例说明的目的，关于 WWW 说明了本发明，不过显然本发明也适用于在局域网，广域网，或者将共享访问受控数据的任意类型网络上共享的数据。



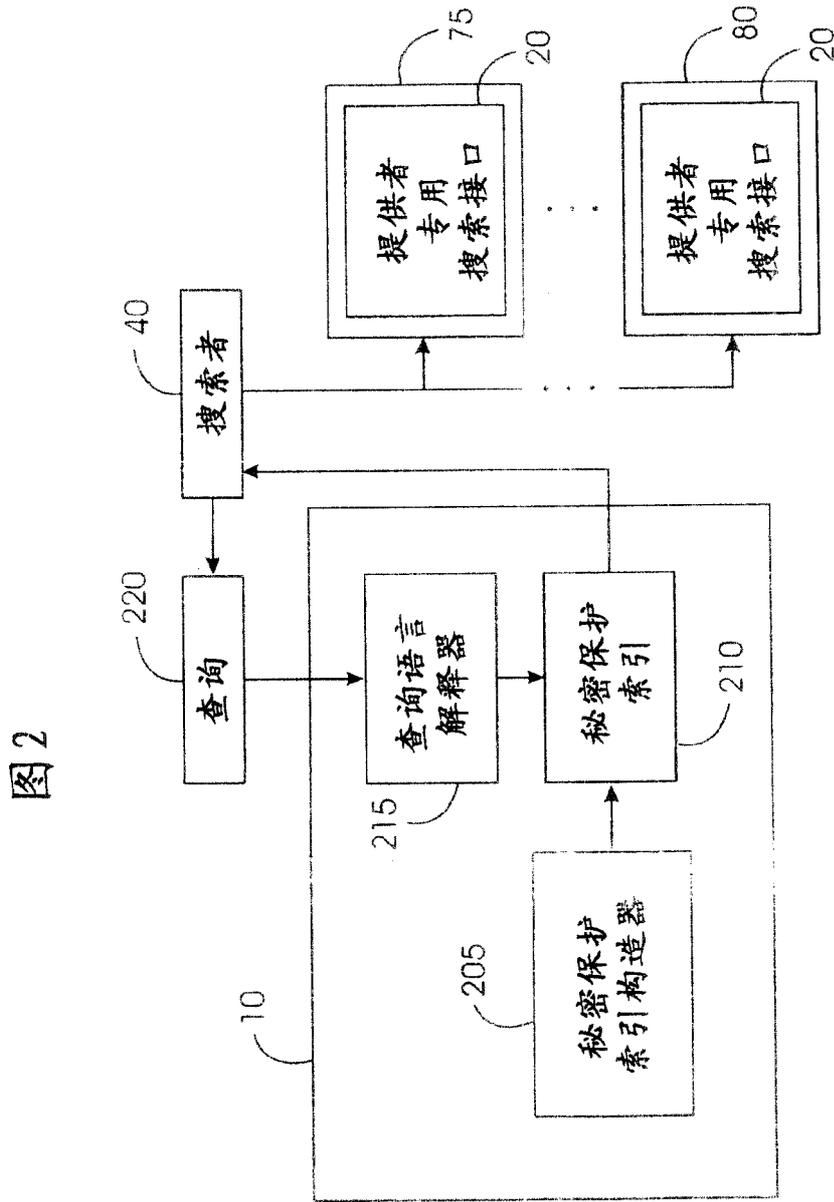
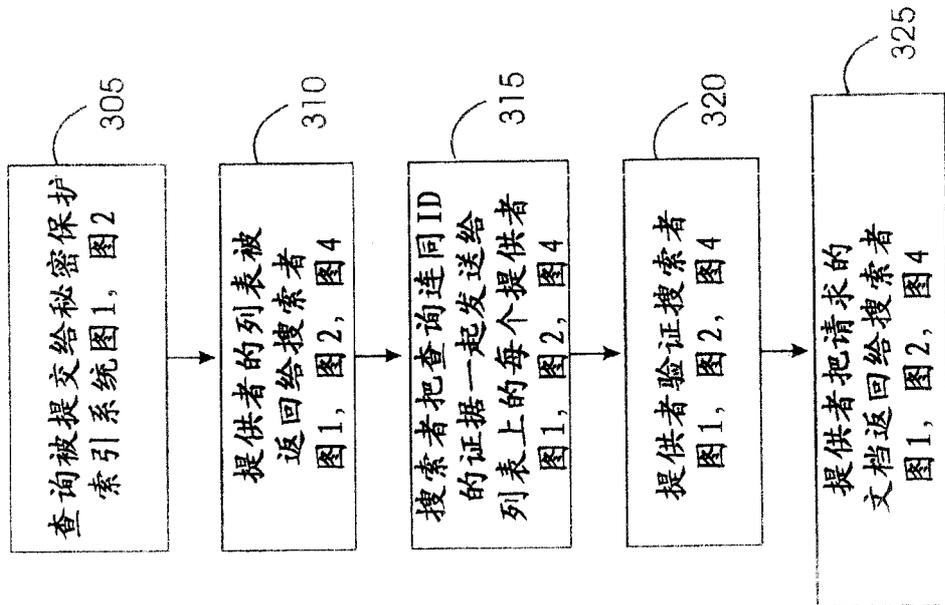
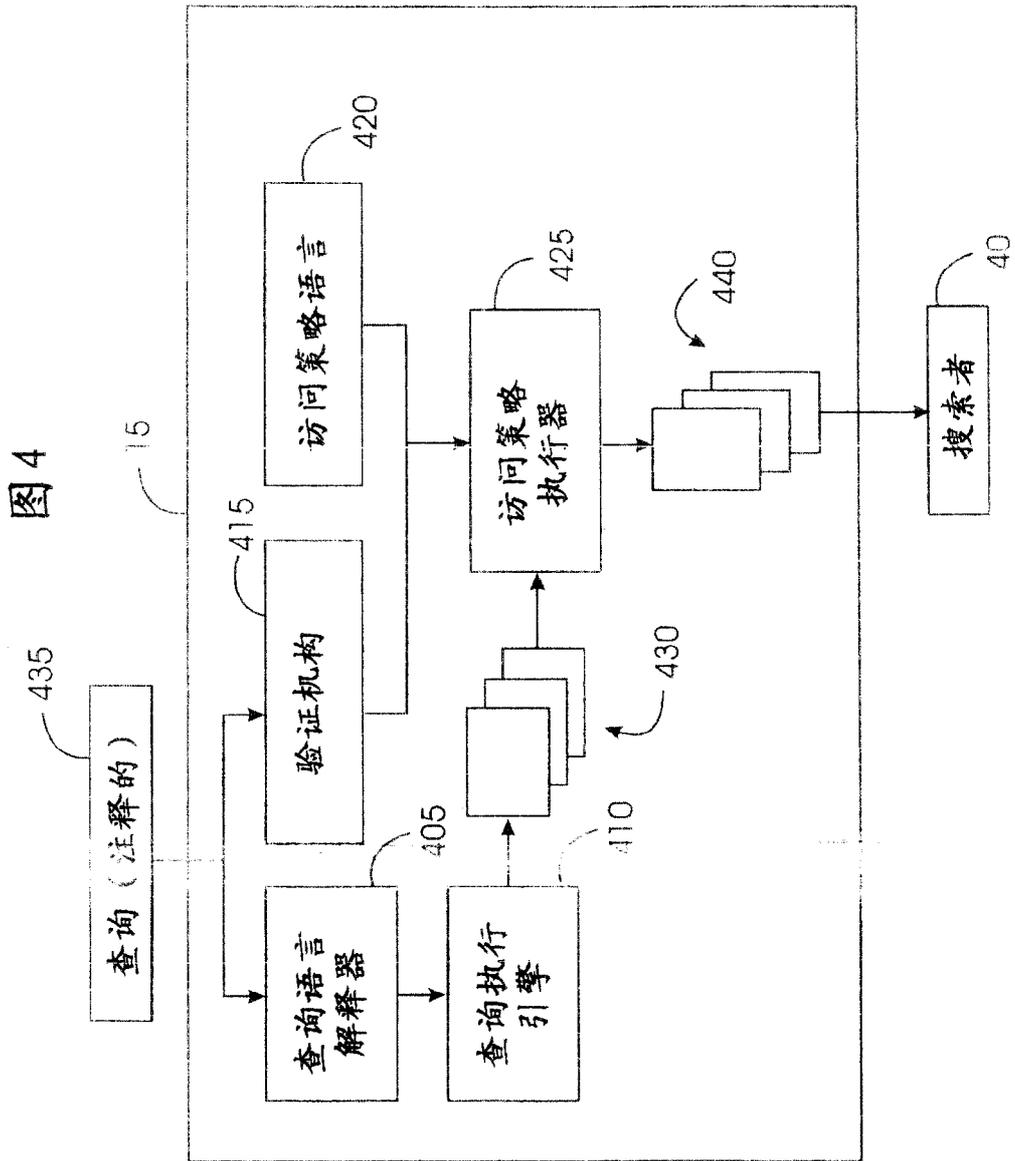


图2

图3





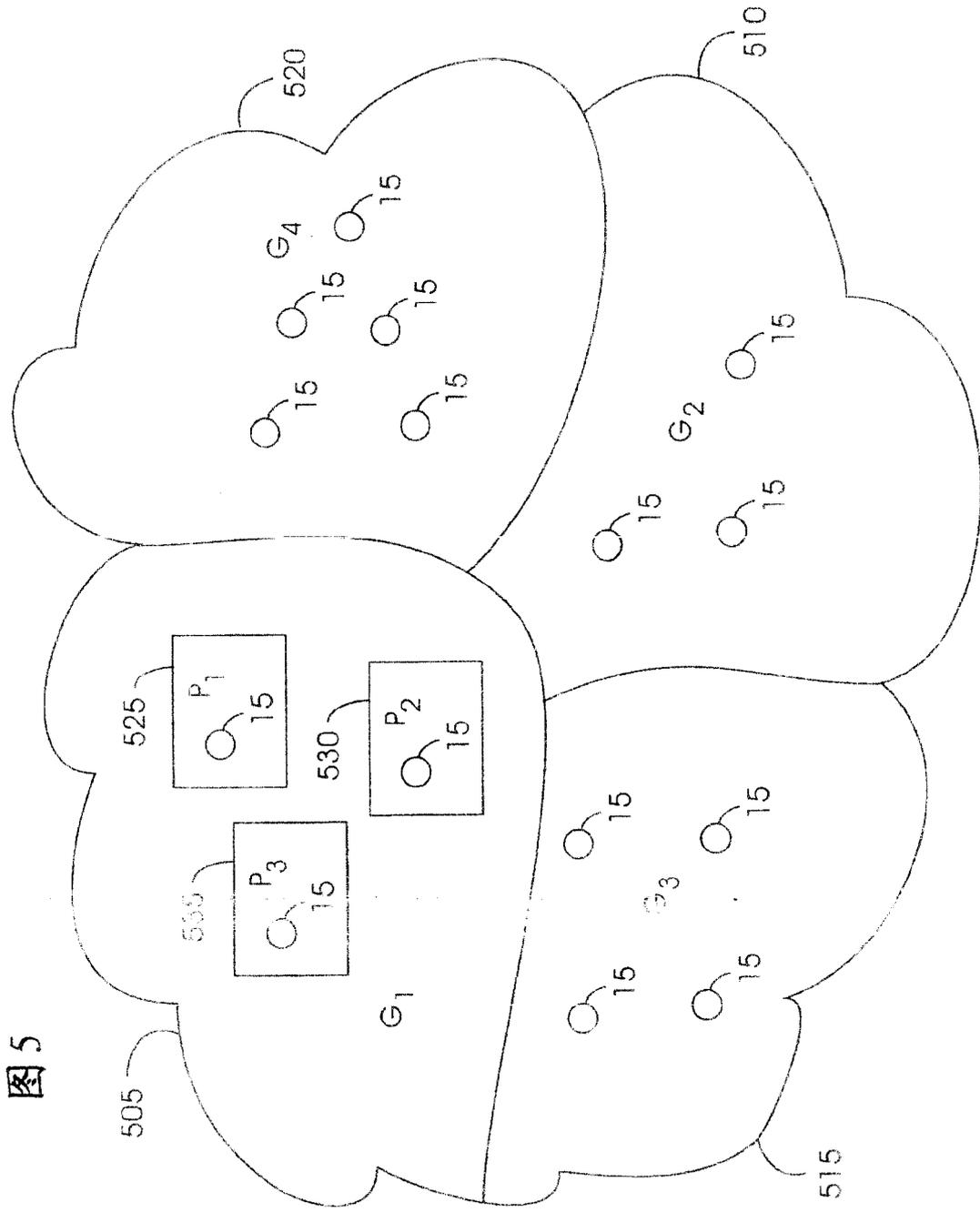


图 5

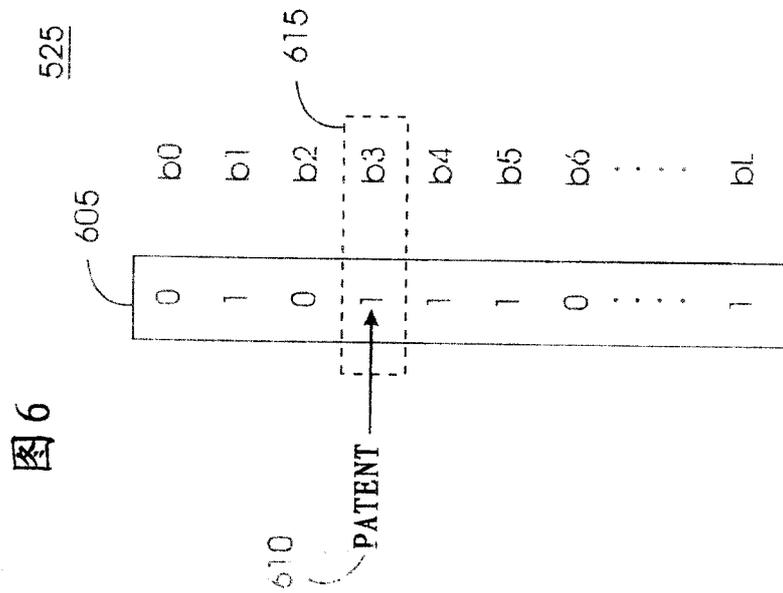


图7

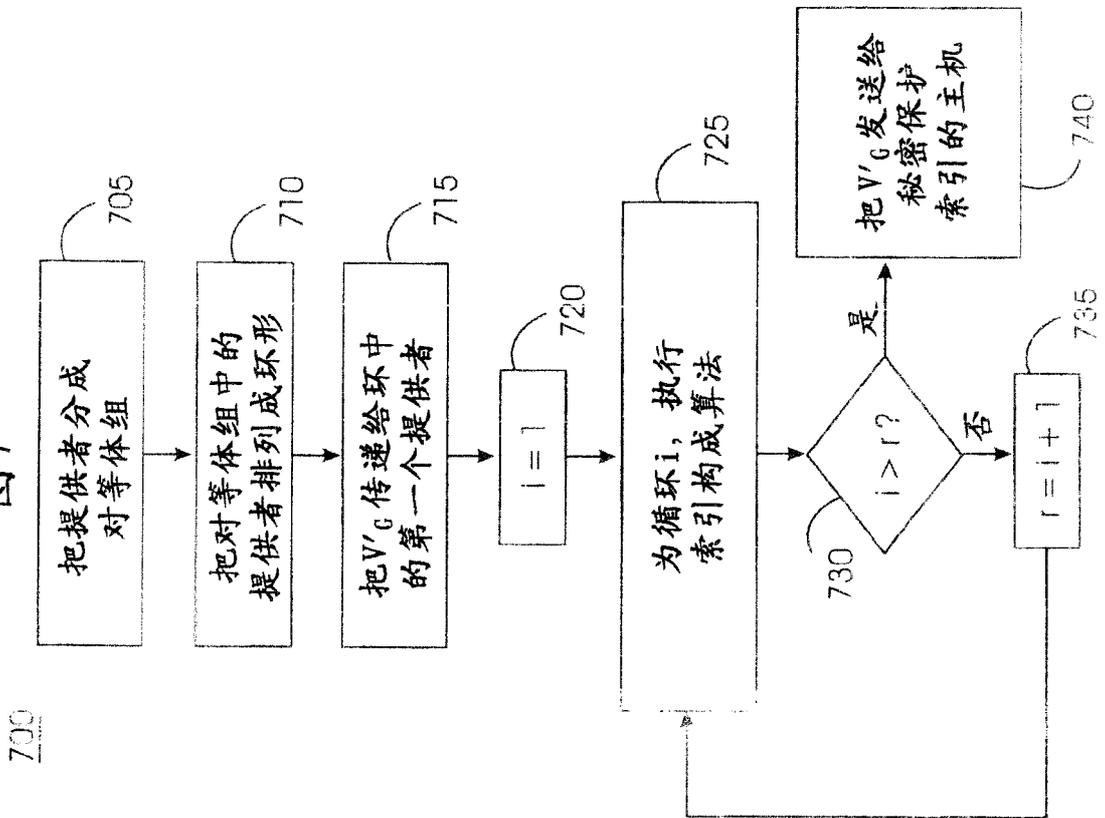


图 8

