[19] 中华人民共和国国家知识产权局

[51] Int. Cl.
G06F 21/00 (2006.01)

[12] 发明专利申请公布说明书

[21] 申请号 200780042490.X

[43] 公开日 2009 年 12 月 16 日

[11] 公开号 CN 101606160A

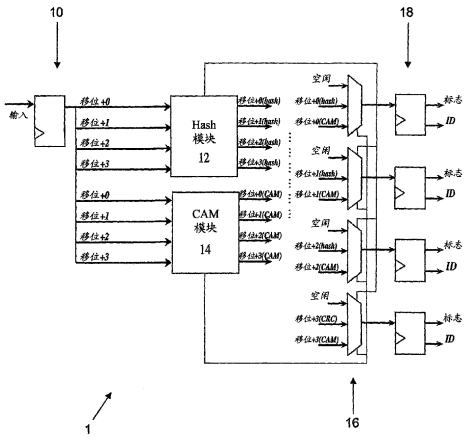
[22] 申请日 2007.10.10
[21] 申请号 200780042490.X
[30] 优先权
[32] 2006.10.10 [33] GB [31] 0620043.0
[86] 国际申请 PCT/GB2007/003833 2007.10.10
[87] 国际公布 WO2008/044004 英 2008.4.17
[85] 进入国家阶段日期 2009.5.15
[71] 申请人 英国贝尔法斯特女王大学
地址 英国贝尔法斯特
[72] 发明人 萨吉尔·塞泽尔
[74] 专利代理机构 北京安信方达知识产权代理有限公司
代理人 颜 涛 郑 霞
权利要求书 5 页 说明书 14 页 附图 4 页

[54] 发明名称

模式检测的相关改进

[57] 摘要

一种在多个数据块中检测模式的方法，包括生成包含一组被选模式中的模式的第一子集的第一数据库，生成包含所述一组被选模式中的剩余模式的第二子集的第二数据库，接收所述多个数据块，且对每个数据块，使用数据块和 Hash 函数来生成关键码，使用所述关键码来搜索所述第一数据库，定位第一数据库的相应于所述关键码的登记项，读取包括零或生成所述关键码的被选模式的登记项的内容，如果所述登记项的内容包括零，则确定所述数据块不包括被选模式，并输出指示所述数据块不包括被选模式的第一输出。



1. 一种方法，其用于在多个数据块中检测模式，所述方法包括：

生成包括一组被选模式中的模式的第一子集的第一数据库，

生成包括所述一组被选模式中的剩余模式的第二子集的第二数据库，

接收所述多个数据块，且对每个数据块，

使用所述数据块和Hash函数来生成关键码，

使用所述关键码来搜索所述第一数据库，

定位所述第一数据库的相应于所述关键码的登记项，

读取所述登记项的内容，所述登记项的内容包括零或生成所述关键码的被选模式，

如果所述登记项的内容包括零，则确定所述数据块不包括被选模式，并输出指示所述数据块不包括被选模式的第一输出，或者

如果所述登记项的内容包括被选模式，则确定所述数据块包括所述被选模式，并输出指示所述数据块包括所述被选模式的第一输出，或者

确定所述数据块不包括所述被选模式，并输出指示所述数据块不包括所述被选模式的第一输出，以及

使用内容可寻址存储器（CAM）比较所述数据块与所述第二数据库，

确定所述数据块匹配所述第二数据库中的被选模式，并输出指示所述数据块包括所述被选模式的第二输出，或者

确定所述数据块不匹配所述第二数据库中的被选模式，并输出指示所述数据块不包括被选模式的第二输出，

组合所述第一输出和所述第二输出，且如果任一输出指示所述数据块包括被选模式，则输出指示所述数据块包括所述被选模式的

标志。

2. 如权利要求 1 所述的方法，其中生成包括一组被选模式中的模式的第一子集的第一数据库的所述步骤包括:

确定每个可能的数据块，

使用每个可能的数据块和所述 Hash 函数来生成多个关键码，

比较生成关键码的数据块或每个数据块与所述一组被选模式，且

如果所述数据块或每个数据块不包括被选模式，则生成所述第一数据库的包括所述关键码和零的登记项，或者

如果所述数据块或任何的数据块包括被选模式，则生成所述第一数据库的包括所述关键码、包含被选模式的数据块或数据块中的一个数据块、以及数据块的标识符（ID）的登记项。

3. 如权利要求 1 或 2 所述的方法，其中生成包括所述一组被选模式中的剩余模式的第二子集的第二数据库的所述步骤包括生成所述第二数据库的登记项，所述第二数据库的所述登记项包括包含没有存储在所述第一数据库的登记项中的被选模式的每个数据块。

4. 如任一前述权利要求所述的方法，其中生成关键码的所述步骤包括生成相对于数据块被压缩的关键码。

5. 如任一前述权利要求所述的方法，其中确定所述数据块包括或不包括被选模式的所述步骤包括比较所述数据块与所述被选模式以确定它们之间的匹配是否出现。

6. 如任一前述权利要求所述的方法，其被用于检测在数据块的任何位置开始的模式。

7. 如任一前述权利要求所述的方法，其被用于检测具有不同长度的被选模式。

8. 如任一前述权利要求所述的方法，其中所述被选模式包括任何的全部或部分词语，或全部或部分串，或全部或部分 DNA 序列，或恶意内容的特征或特征段。

9. 一种模式检测电路，其用于在多个数据块中检测模式，所述电路包括:

多个 Hash 模块，每个 Hash 模块都包括包含一组被选模式中的模式的第一子集的第一数据库，其中每个 Hash 模块接收所述多个数据块，且对每个数据块，

使用所述数据块和 Hash 函数来生成关键码，

使用所述关键码来搜索所述第一数据库，

定位所述第一数据库的相应于所述关键码的登记项，

读取所述登记项的内容，所述登记项的内容包括零或生成所述关键码的被选模式，

如果所述登记项的内容包括零，则确定所述数据块不包括被选模式，并输出指示所述数据块不包括被选模式的第一输出，或者

如果所述登记项的内容包括被选模式，则确定所述数据块包括所述被选模式，并输出指示所述数据块包括所述被选模式的第一输出，或者

确定所述数据块不包括所述被选模式，并输出指示所述数据块不包括所述被选模式的第一输出，

多个 CAM 模块，每个 CAM 模块包括包含所述一组被选模式中的剩余模式的第二子集的第二数据库，其中每个 CAM 模块接收所述多个数据块，且对每个数据块，

比较所述数据块与所述第二数据库，

确定所述数据块匹配所述第二数据库中的被选模式，并输出指示所述数据块包括所述被选模式的第二输出，或者

确定所述数据块不匹配所述第二数据库中的被选模式，并输出指示所述数据块不包括被选模式的第二输出，

以及

组合器模块，其组合所述第一输出和所述第二输出，以及如果任一输

出指示所述数据块包括被选模式，则输出指示所述数据块包括所述被选模式的标志。

10. 如权利要求 9 所述的电路，其中每个 Hash 模块包括 RAM 装置。

11. 如权利要求 10 所述的电路，其中每个 RAM 装置存储所述第一数据库。

12. 如权利要求 11 所述的电路，其中关键码被通过以下方式使用来搜索 RAM 装置的所述第一数据库：将所述关键码用作地址来搜索被指派给所述 RAM 装置的多个存储单元的地址。

13. 如权利要求 9 至 12 中任一权利要求所述的电路，其中每个 Hash 模块包括多个 Hash 装置，所述多个 Hash 装置中的每一个 Hash 装置使用数据块和所述 Hash 函数来生成关键码。

14. 如权利要求 9 至 13 中任一权利要求所述的电路，其中每个 CAM 模块包括多个 CAM 单元。

15. 如权利要求 14 所述的电路，其中每个 CAM 单元存储包括所述第二数据库的模式的数据块。

16. 如权利要求 15 所述的电路，其中每个 CAM 单元包括多个比较器，所述多个比较器中的每一个比较器比较接收到的数据块与存储在所述 CAM 单元中的所述数据块。

17. 如权利要求 9 至 16 中任一权利要求所述的电路，其检测在数据块的任何位置开始的模式。

18. 如权利要求 17 所述的电路，其中所述模式检测电路包括多个 Hash 装置和多个 CAM 比较器，第一数据块输入到第一 Hash 装置中并输入到第一 CAM 比较器中，相对于所述第一数据块移位的第二数据块输入到第二 Hash 装置中并输入到第二 CAM 比较器中，等等。

19. 如权利要求 18 所述的电路，其中所述第二数据块相对于所述第一数据块移位所述块的一个或更多的位置。

20. 如权利要求 19 所述的电路，其中所述第一数据块和所述第二数

据块包括位或字节，且所述第二数据块相对于所述第一数据块移位包括所述块的一个或更多的位或字节的一个或更多的位置。

21. 如权利要求 9 至 20 中任一权利要求所述的电路，其包括多个部分，第一部分检测长度为 n 的模式，第二部分检测长度为 n-1 的模式，第三部分检测长度为 n-2 的模式，等等。

22. 如权利要求 9 至 21 中任一权利要求所述的电路，其中所述被选模式包括任何的全部或部分词语，或全部或部分串，或全部或部分 DNA 序列，或恶意内容的特征或特征段。

模式检测的相关改进

本发明涉及模式检测。

在许多应用中期望有检测信息中的模式的能力。这些应用包括串匹配，在串匹配中，选择特定的模式或串并在信息中搜索匹配的模式或串。这在许多领域都有应用，例如文件检索、记录检索、安全性（例如，可在数据或语音消息搜索包括特殊的词或词序列的模式的情况）。其他使用模式检测的应用包括生物学的应用，比如DNA测序，以及电信业中的各种应用，比如正则表达式处理、IP包分类和深度包检查（deep packet inspection）。在后一应用中，可在包中检查是否存在例如在恶意内容比如病毒或蠕虫中发现的模式。

模式检测应用得如此广泛，以至于一直在寻求检测的改进，例如检测的速度的改进。

根据本发明的第一方面，提供了一种在多个数据块中检测模式的方法，包括

生成包括一组被选模式中的模式的第一子集的第一数据库，

生成包括所述一组被选模式中的剩余模式的第二子集的第二数据库，

接收所述多个数据块，且对每个数据块，

使用数据块和Hash函数（散列函数）来生成关键码(key)，

使用所述关键码来搜索所述第一数据库，

定位第一数据库的相应于所述关键码的登记项（entry），

读取包括零或生成所述关键码的被选模式的登记项的内容，

如果所述登记项的内容包括零，则确定所述数据块不包括被选模式，并输出指示所述数据块不包括被选模式的第一输出，或者

如果所述登记项的内容包括被选模式，则确定所述数据块包括所述被选模式，并输出指示所述数据块包括所述被选模式的第一输出，或者

确定所述数据块不包括所述被选模式，并输出指示所述数据块不包括所述被选模式的第一输出，以及

使用内容可寻址存储器（CAM）比较所述数据块与所述第二数据库，

确定所述数据块匹配所述第二数据库中的被选模式，并输出指示所述数据块包括所述被选模式的第二输出，或者

确定所述数据块不匹配所述第二数据库中的被选模式，并输出指示所述数据块不包括被选模式的第二输出，

组合所述第一输出和所述第二输出，且如果任一输出指示所述数据块包括被选模式，则输出指示所述数据块包括所述被选模式的标志（flag）。

生成包括一组被选模式中的模式的第一子集的第一数据库，可包括确定每个可能的数据块，使用每个可能的数据块和 Hash 函数来生成多个关键码，比较生成关键码的数据块或每个数据块与所述一组被选模式，且如果所述数据块或每个数据块不包括被选模式，则生成所述第一数据库的包括所述关键码和零的登记项，或者如果数据块或任何的数据块包括被选模式，则生成所述第一数据库的包括关键码、包括被选模式的数据块或数据块中的一个数据块、以及数据块的标识符（ID）的登记项。

生成包括所述一组被选模式中的剩余模式的第二子集的第二数据库，可包括生成第二数据库的登记项，其包括包含没有存储在第一数据库的登记项中的被选模式的每一个数据块。

生成关键码可包括生成相对于数据块被压缩的关键码。生成压缩的关键码导致对内存的要求降低。

确定数据块包括或不包括被选模式的步骤，可包括比较所述数据块与所述被选模式，以确定它们之间的匹配是否出现。

组合第一输出和第二输出可包括复用输出。

本方法可用于检测在数据块的任何位置开始的模式。本方法可用于检

测具有不同长度的被选模式。

根据本发明的第二方面，提供了一种用于在多个数据块中检测模式的模式检测电路，包括

多个 Hash 模块（散列模块），每个 Hash 模块都包括包含一组被选模式中的模式的第一子集的第一数据库，其中每个 Hash 模块接收多个数据块，且对每个数据块，

使用数据块和 Hash 函数来生成关键码，

使用所述关键码来搜索第一数据库，

定位第一数据库的相应于所述关键码的登记项，

读取所述登记项的内容，其包括零或生成所述关键码的被选模式，

如果登记项的内容包括零，则确定数据块不包括被选模式，并输出指示数据块不包括被选模式的第一输出，或者

如果登记项的内容包括被选模式，则确定数据块包括被选模式，并输出指示数据块包括被选模式的第一输出，或者

确定数据块不包括被选模式，并输出指示数据块不包括被选模式的第一输出，以及

多个 CAM 模块，每个 CAM 模块包括包含所述一组被选模式中的剩余模式的第二子集的第二数据库，

其中每个 CAM 模块接收多个数据块，且对每个数据块，

比较数据块与所述第二数据库，

确定数据块匹配第二数据库中的被选模式，并输出指示数据块包括被选模式的第二输出，或者

确定数据块不匹配第二数据库中的被选模式，并输出指示数据块不包括被选模式的第二输出，以及

组合器模块，其组合第一输出和第二输出，以及如果任一输出指示所述数据块包括被选模式，则输出指示数据块包括被选模式的标志。

每个 Hash 模块可包括 RAM 装置。每个 RAM 装置可存储第一数据库。关键码可用于通过将所述关键码用作地址来搜索被指派给 RAM 装置的多个存储单元(memory location)的地址，来搜索 RAM 装置的第一数据库。

每个 Hash 模块可包括多个 Hash 装置，所述多个 Hash 装置中的每一个 Hash 装置使用数据块和 Hash 函数来生成关键码。

每个 CAM 模块可包括多个 CAM 单元。每个 CAM 单元可存储包括第二数据库的模式的数据块。每个 CAM 单元都可包括多个比较器，所述多个比较器中的每一个比较器比较接收到的数据块与存储在 CAM 单元中的数据块。

组合器模块可包括复用器。

所述模式检测电路可检测在数据块的任何位置开始的模式。所述模式检测电路可包括多个 Hash 装置和多个 CAM 比较器，第一数据块可输入到第一 Hash 装置中并输入到第一 CAM 比较器中，相对于所述第一数据块移位的第二数据块可输入到第二 Hash 装置中并输入到第二 CAM 比较器中，等等。第二数据块相对于第一数据块可移位块的一个或更多的位置。例如，第一数据块和第二数据块可包括位或字节，而第二数据块相对于第一数据块可移位包括块的一个或更多的位或字节的一个或更多的位置。这允许检测在数据块中任何位置开始的模式。

模式检测电路可包括多个部分，第一部分检测长度为 n 的模式，第二部分检测长度为 n-1 的模式，第三部分检测长度为 n-2 的模式，等等。

被选模式将包括多个模式，希望在数据块中检测到这些模式的存在。被选模式可包括任何的全部或部分词语，或全部或部分串，或全部或部分 DNA 序列，或恶意内容的特征（signature）或特征段（signature segment）。

应理解，术语“模式”用于描述任何字符或任何数量字符，且并不限于表示具有重复性的一定数量的字符。

现在将仅以举例的方式，参考附图，描述本发明的实施方式，其中：

图 1 是根据本发明的模式检测电路的图示，所述模式检测电路包括

Hash 模块和 CAM 模块，

图 2 是图 1 的 Hash 模块的部分的图示，

图 3 是图 1 的 CAM 模块的部分的图示，以及

图 4 是包括本发明的特征检测电路的深度包检查系统的图示。

在所描述的实施方式中，被选模式包括恶意内容的特征或特征段。然而，应认识到，这只是示例性的，且本发明可应用于许多模式类型的检测。

图 1 显示了模式或特征检测电路 1，其包括输入寄存器 10、Hash 模块 12、内容可寻址存储器（CAM）模块 14、多个复用器 16 以及多个输出寄存器 18。在此实施方式中，特征检测电路形成通信网络的深度包检查（DPI）系统的部分，并接收在多个实体间通信的数据。数据被格式化为包，每个包包括首部和有效负载。有可能任何包的有效负载可包含恶意内容，比如病毒或蠕虫。特征检测电路 1 检验数据，并向 DPI 系统标记在该数据中发现的任何恶意内容。诸如病毒的恶意内容一般每一个都包括独特的标识符或特征。到目前为止，具有相应的有限数目的特征的有限数目的病毒等是已知的。本发明的特征检测电路 1 通过寻找这些特征来检验网络中的数据是否有恶意内容。

在此实施方式中，网络数据被输入到特征检测电路 1 的输入寄存器 10 中，并由此输出且由该电路处理为一系列的 4 字节数据块。然而，应认识到，可以使用其他的数据块大小，例如 8 字节数据块或 16 字节数据块。恶意内容的每个特征通常将包括例如 1、2、3、4、6、8、12、14、16、24 个等一定数量的字节。因此，每个特征将散布在从输入寄存器 10 输出的一个或更多的数据块中。当将被检测的特征的长度小于数据块的长度（即在此实施方式中，小于 4 字节）时，特征检测电路将检查接收到的数据块的完整特征。当将被检测的特征的长度大于数据块的长度（即在此实施方式中，大于 4 字节）时，也是大多数的情况时，特征检测电路将检查接收到的数据块的特征段。关于被检测到的特征或特征段的信息从特征检测电路 1 输出，且在特征段的情况下，上述信息可被整理（collate）。

特征或第一特征段可在网络数据中的多个位置开始。这通过以下方式

而被考虑:配置特征检测电路 1 以处理相对于第一数据块而移位(或偏移)的例如移位一个或更多的字节的数据块。在此实施方式中,特征检测电路 1 通过将 4 字节的数据块例如 x1、x2、x3 和 x4(移位 = 0)输入到 Hash 模块 12 的第一 Hash 装置中,并还输入到 CAM 模块 14 的第一 CAM 比较器中而处理数据。移位 1 个字节即 x2,x3,x4 和 x5 的 4 字节下一数据块,被输入到 Hash 模块 12 的第二 Hash 装置中,并还输入到 CAM 模块 14 的第二 CAM 比较器中,对于 Hash 模块 12 的每个 Hash 装置和 CAM 模块 14 的 CAM 比较器,依此类推。Hash 模块 12 和 CAM 模块 14 均检验数据块的恶意内容。这些模块的输出由多个复用器 16 接收,而在数据块中发现的任何恶意内容的细节由复用器 16 输出到多个输出寄存器 18,并从这些输出寄存器输出到通信网络。

现将更详细地描述特征检测电路 1 的此实施方式的元件的功能。

图 2 详细示出了 Hash 模块 12 的一部分。这包括第一到第四 Hash 装置 20、第一到第四寄存器 22、复用器 24、RAM 装置 26、第一到第四寄存器 28 以及第一到第四比较器 30。将被检验恶意内容的网络数据以 4 字节的块由 Hash 装置 20 的每一个接收,如所示出的。

每个 Hash 装置以相同的方式工作,其基本的 Hash 函数是接收 4 字节(32 位)数据块,并生成关键码,所述关键码的值决定于数据块的值,且该关键码相对于数据块是被压缩的,即包括少于 32 位。在此实施方式中,由 Hash 装置生成的每个关键码具有 12 位的长度。然而,应认识到,可能生成不是 12(但小于 32)位大小的关键码。

使用 Hash 函数很可能造成两个或更多不同的数据块将生成相同的关键码。例如,五个不同的数据块,其中三个包含恶意内容,而其中两个不包含恶意内容,可能生成相同的关键码。这种情况称为冲突(collision)。

当已经确定了将在 Hash 装置中使用的特定的 Hash 函数时,软件模块使用 Hash 函数生成关键码,用于每一可能的 32 位数据块。这允许绘制一表,各关键码都具有登记项,包括关键码的值以及零或生成关键码的数据块。如果关键码由一个或更多的数据块生成且每个数据块都不包含恶意内容,则关键码的登记项包括关键码值和零。如果关键码由一个或更多

的数据块生成，且每个数据块都由已知特征中的一个特征组成或包括已知特征中的一个特征的段(即包含恶意内容)，则关键码的登记项包括关键码值和数据块或数据块中的每一个，即特征或特征段，或特征或特征段中的每一个特征或特征段。数据块或数据块中的每一个的特征ID或特征段ID，无论哪个是合适的，也都添加到关键码的登记项，其使用将在下面进行描述。如果关键码由这样的数据块生成，即：数据块中的一个或更多个数据块由已知特征中的一个特征组成或包括已知特征中的一个特征的段，即包含恶意内容，而数据块中的一个或更多的数据块不包含恶意内容，则关键码的登记项包括关键码值和包含恶意内容的数据块或数据块中的每一个，以及数据块或这些数据块中的每一个的特征 ID 或特征段 ID，恶意内容即特征或特征段，或特征或特征段中的每一个特征或特征段。因此，由于使用 Hash 函数而造成的冲突是显著的。

所述表随后用于配置 Hash 模块 12 的 RAM 装置 26。RAM 装置 26 包括多个存储单元。每个存储单元都被指派地址和内容，所述地址具有等于关键码中的一个关键码的值，所述内容包括零或生成该关键码的一个数据块，如下所示。如果关键码由一个或更多的数据块生成且每个数据块都不包含恶意内容，则该关键码的存储单元的内容包括零。如果关键码由一个或更多的数据块生成，且每个数据块都由已知特征中的一个特征组成或包括已知特征中的一个特征的段（即包含恶意内容），则该关键码的存储单元的内容包括数据块或数据块中的一个数据块，即特征或特征段，或特征或特征段中的一个特征或特征段，以及特征 ID 或特征段 ID。如果关键码由这样的数据块生成，即：数据块中的一个或更多个数据块由已知特征中的一个特征组成或包括已知特征中的一个特征的段，即包含恶意内容，而数据块中的一个或更多的数据块不包含恶意内容，则该关键码的存储单元的内容包括包含恶意内容的数据块或数据块中的一个数据块，即特征或特征段，或特征或特征段中的一个特征或特征段，以及特征/特征段 ID。从后两种情况可注意到，当包含恶意内容的多个数据块生成相同的关键码时，数据块中的仅一个数据块，即特征/特征段，被选择用于 RAM 装置的存储单元的登记项。剩余的包含恶意内容的数据块被用来配置 CAM 模块 14，如下所述。

RAM 装置 26 具有用于每个不同的关键码值的存储单元，并因此包括等于可能的关键码的数量的一定数量的存储单元。每个关键码都包括 12 位。因此有 2^{12} 个可能的关键码值。因此 RAM 装置 26 包括 2^{12} 存储单元。每个关键码与生成其的数据块相比是压缩的，即与 32 位的数据块相对照，每个关键码只包括 12 位。与如果每个关键码包括 32 位则需要 2^{32} 个存储单元相对照，这导致只需要 2^{12} 个存储单元的 RAM 装置 26。因此，使用 Hash 装置压缩输入到特征检测电路 1 的数据允许极大地降低对 RAM 装置 26 的存储要求。

在操作中，Hash 装置每个都接收数据块，并生成关键码。每个 Hash 装置向寄存器 22 中的一个输出生成的关键码。每个寄存器随后向复用器 24 输出其关键码。复用器 24 接收地址输入（未显示），该地址输入将使复用器 24 依次接收其四个输入的每一个上的关键码，并依次向 RAM 装置 26 输出关键码。

RAM 装置 26 依次接收关键码。每个关键码被用作为存储单元地址，即关键码的值与 RAM 装置 26 的存储单元的地址比较，直到找到地址值匹配关键码值的存储单元为止。在找到 RAM 装置 26 的匹配的存储单元后，读取该匹配的存储单元的内容。存储单元的内容将包括零，或将包括包含恶意内容即特征或特征段的数据块以及特征/特征段 ID。在此实施方式中，因为数据块是 32 位长，因此，特征或特征段是 32 位长，特征/特征段 ID 的长度被选择为 12 位。

RAM 装置 26 依次向寄存器 28 的第一个寄存器，然后向第二个寄存器，然后向第三个寄存器，然后向第四个寄存器输出被寻址的存储单元的内容。寄存器 28 的每一个都向特征检测电路 1 的复用器 16（见图 1）输出其接收的存储单元内容的零或 12 位特征/特征段 ID 部分，用于与 CAM 模块 14 的输出比较，如下所述。

寄存器 28 的每一个也向比较器 30 中的一个比较器输出其接收的存储单元内容的 32 位特征/特征段部分，如所示。每个比较器接收两个输入，原始数据块（通过延迟提供，如所示出的）和由使用相同的数据块生成的关键码而产生的存储单元的内容的 32 位特征/特征段部分。每个比较器比

较原始数据块的值与存储单元内容的32位特征/特征段部分的值，并如果发现这些值是相同的，则输出匹配标志，该匹配标志指示已经找到恶意内容。

根据上述特征检测电路的操作，如果数据块不包含恶意内容，即不包括特征也不包含特征段，则数据块生成的关键码将产生零存储单元内容（当关键码由不包含恶意内容的一个或更多的数据块生成时），或产生包括32位特征/特征段的存储单元内容（当关键码由不包含恶意内容的一个或更多的数据块和包含恶意内容的一个或更多的数据块生成时）。在任一情况下，存储单元内容的32位特征/特征段与原始数据块的比较都将导致发现它们不是相同的，且不会生成匹配标志，即电路指示在该数据块中没有发现恶意内容。如果数据块包含恶意内容，即包括特征或包含特征段，则该数据块生成的关键码将产生包括等于该数据块的特征/特征段的32位特征/特征段的存储单元内容（当关键码由包含恶意内容的一个或更多的数据块生成，且此数据块被选作进入RAM装置的登记项时），或产生包括不等于该数据块的特征/特征段的32位特征/特征段的存储单元内容（当关键码由包含恶意内容的一个或更多的数据块生成，且此数据块未被选作进入RAM装置的登记项时）。在第一种情况下，存储单元内容的32位特征/特征段与原始数据块的比较将导致发现它们是相同的，且将生成匹配标志，即系统指示已经在数据块中发现恶意内容。在第二种情况下，存储单元内容的32位特征/特征段与原始数据块的比较将导致发现它们不是相同的，且不会生成匹配标志，即系统指示在该数据块中没有发现恶意内容。这不是正确的指示，但此情况通过使用CAM模块14而被考虑进来，如下所述。

图2所示的Hash装置等只包括特征检测电路1的实际Hash模块12的第一部分。Hash模块12的此第一部分能够检测长度为4字节的特征或特征段。Hash模块12进一步包括第二部分，该第二部分能够通过在三个最高有效字节中具有可能的特征数据而在剩余字节中具有‘通配符’数据的4字节数据块中寻找特征/特征段，检测长度为3字节的特征或特征段。Hash模块12进一步包括第三部分，该第三部分能够通过在两个最高有效

字节中具有可能的特征数据而在剩余字节中具有'通配符'数据的4字节数据块中寻找特征/特征段，检测长度为2字节的特征或特征段。Hash模块12进一步包括第四部分，该第四部分能够检测长度为1字节的特征或特征段。Hash模块的第二和第三部分包括与第一部分相同的元件，并以相同的方式起作用。Hash模块的第四部分包括简单的RAM装置，其能够提供足够的内存，以检测长度为1字节的特征或特征段，而没有过度的硬件要求。输入到如上所述的Hash模块的第一部分的数据块，也输入到Hash模块的第二部分、第三部分和第四部分。Hash模块12的这样的安排，允许其用于检测可变长度的特征或特征段。例如，如果将被检测的特征的长度为4字节，则这被提供给Hash模块的所有部分，且完整特征能够由Hash模块12的第一部分检测，而不被其他部分检测。如果将被检测的特征的长度为2字节，则这被提供给Hash模块的所有部分，且完整特征能够由Hash模块12的第三部分检测，而不被其他部分检测。如果将被检测的特征的长度为6字节，因为提供给Hash模块的部分的数据块的长度为4字节，包括所述特征的前4个最高有效字节的特征段被提供给Hash模块的所有部分，且此特征段能够由Hash模块12的第一部分检测，而不被其他部分检测，以及包括所述特征的剩余2字节的特征段和输入数据的下一个2字节被提供给Hash模块的所有部分，且此特征段能够由Hash模块12的第三部分检测，而不被其他部分检测。这样，两个特征段都可由Hash模块12检测，并从那里输出。这两个特征段可随后被整理，以允许产生指示已经检测到恶意内容的标志。

如上所述，由于Hash函数用于检测恶意内容，则很可能会出现冲突，即两个或更多不同的数据块生成相同的关键码。确定用于Hash装置的Hash函数中出现的冲突，并相应地配置RAM装置26。当不包含恶意内容的数据块和包含恶意内容的数据块每个都产生相同的关键码时，这对特征检测电路1检测恶意内容没有影响。这种情况下，RAM装置26将被配置成使得地址等于关键码的存储单元具有包括包含恶意内容的数据块的细节的内容，且将为包含恶意内容的数据块生成匹配标志。然而，当每个都包含恶意内容的两个或更多数据块每个都产生相同的关键码时，这可能影响特征检测电路1对恶意内容的检测。这种情况下，RAM装置26

将被配置成使得地址等于关键码的存储单元具有只包括包含恶意内容的数据块中的一个数据块的内容。如以上所详述的，这可导致对事实上包含恶意内容的数据块并不生成匹配标志。通过 CAM 模块 14 将这种情况考虑进来。

图 3 中示出了 CAM 模块 14 的部分。其包括多个 CAM 单元（cell）40、多个解码器 42、多个寄存器 44、复用器 46、RAM 装置 48 和多个寄存器 50。每个 CAM 单元包括内容寄存器和多个比较器。

CAM 单元被定制成处理包含恶意内容的两个或更多数据块的冲突。如上所述，由软件模块确定造成这种冲突的数据块。数据块之一被选择用于模块 12 的 RAM 装置 26 的存储单元中的登记项（且因而如果等于此被选择的数据块的数据块输入到特征检测电路，则将检测到它的恶意内容）。剩余的数据块通过使用一个或更多的 CAM 单元而被考虑进来。CAM 单元被定制成通过将数据块存储在 CAM 单元的内容寄存器中而将一个这样的数据块考虑进来。因此 CAM 模块 14 将包括 k 个 CAM 单元，其中 k 等于没有被选择用于储存在 Hash 模块 12 的 RAM 装置 26 的包含恶意内容的数据块的数量。

每个 CAM 单元包括四个比较器。对于每个 CAM 单元，每个比较器都接收网络数据的输入数据块，所述输入数据块如之前所详述的已相对于第一数据块移位。对于每个 CAM 单元，每个比较器还接收存储在 CAM 单元的内容寄存器中的数据块。每个比较器比较输入数据块与内容寄存器数据块，并在它们不相同的情况下，输出等于 0 的匹配，或在它们相同的情况下，输出等于 1 的匹配。在后一情况下，这意味着输入数据块包含恶意内容（即特征或特征段），其与导致冲突的包含恶意内容的数据块中的一个数据块相同。每个 CAM 单元的第一比较器的输出被输入到第一解码器，每个 CAM 单元的第二比较器的输出被输入到第二解码器，等等，如所显示的。对解码器接收的每个等于 1 的匹配，解码器确定 CAM 单元的标识（identity）以及确定输出该匹配的 CAM 单元的比较器的标识，并输出指示匹配的起源位置的二进制值。对解码器接收的每个等于 0 的匹配，解码器输出二进制值零。

每个解码器向寄存器 42 的一个寄存器输出二进制位置值和零值，如所显示的。每个寄存器随后向复用器 44 输出其二进制位置值和零值。复用器 44 接收地址输入（未显示），该地址输入使复用器依次在其四个输入的每个输入上接收二进制位置值或零值，并向 RAM 装置 48 依次输出二进制位值和零值。

RAM 装置 48 依次接收二进制位置值和零值。每个二进制位置值和零值用作为存储单元地址。当接收到零值时，这映射到 RAM48 的地址等于零的存储单元，而等于零的此存储单元的内容被输出到寄存器 50 中的一个寄存器。当接收到二进制位置值时，这与 RAM 装置 48 的存储单元的地址比较，直到发现地址匹配二进制位置值的存储单元为止。在找到 RAM 装置 48 的匹配存储单元之后，匹配存储单元的内容被输入到寄存器 50 的一个寄存器。存储单元的内容将包括生成那个产生了该二进制位置值的匹配的数据块的 12 位的特征/特征段 ID。

RAM 装置 48 向寄存器 50 的第一个寄存器、随后向第二个寄存器、随后向第三个寄存器、随后向第四个寄存器依次输出零值和 12 位特征/特征段 ID。寄存器 50 的每个寄存器都向复用器 16（见图 1）输出零值和 12 位特征/特征段 ID，用于与 Hash 模块 12 的输出相比较，如下所述。

与 Hash 模块 12 一样，图 3 示出的 CAM 装置等也只包括特征检测电路 1 的实际 CAM 模块 14 的第一部分。CAM 模块 14 的此第一部分能够检测长度为 4 字节的特征或特征段。CAM 模块 14 进一步包括第二部分，该第二部分能够通过在三个最高有效字节中具有可能的特征数据而在剩余字节中具有‘通配符’数据的 4 字节数据块中寻找特征/特征段，检测长度为 3 字节的特征或特征段。CAM 模块 14 进一步包括第三部分，该第三部分能够通过在两个最高有效字节中具有可能的特征数据而在剩余字节中具有‘通配符’数据的 4 字节数据块中寻找特征/特征段，检测长度为 2 字节的特征或特征段。CAM 模块 14 进一步包括第四部分，该第四部分能够检测长度为 1 字节的特征或特征段。CAM 模块的第二和第三部分包括与第一部分相同的元件，并以相同的方式起作用。CAM 模块的第四部分包括简单的 RAM 装置，其能够提供足够的内存，以检测长度为 1 字节

的特征或特征段，而没有过度的硬件要求。输入到如上所述的CAM模块的第一部分的数据块，也输入到CAM模块的第二部分、第三部分和第四部分。CAM 模块 14 的这样的安排，允许其用于检测可变长度的特征或特征段。例如，如果将被检测的特征的长度为3字节，则这被提供给CAM模块的所有部分，且完整特征能够由 CAM 模块 14 的第二部分检测，而不被其他部分检测。如果将被检测的特征的长度为1字节，则这被提供给CAM模块的所有部分，且完整特征能够由CAM模块14的第四部分检测，而不被其他部分检测。如果将被检测的特征的长度为7字节，因为提供给CAM模块的部分的数据块的长度为4字节，包括所述特征的前4个最高有效字节的特征段被提供给 CAM 模块的所有部分，且此特征段能够由 CAM 模块 14 的第一部分检测，而不被其他部分检测，以及包括所述特征的剩余3字节的特征段和输入数据的下个字节被提供给CAM模块的所有部分，且此特征段能够由 CAM 模块 14 的第二部分检测，而不被其他部分检测。这样，两个特征段都可由CAM模块14检测，并从那里输出。这两个特征段可随后被整理，以允许产生指示已经检测到恶意内容的标志。

对输入到特征检测电路 1 中的每个数据块，电路的复用器 16 的每个复用器都从 Hash 模块 12 接收零值或 12 位特征/特征段 ID，从 CAM 模块 14 接收零值或 12 位特征/特征段 ID 以及接收空闲信号，如所显示的。每个复用器 16 如果接收到 Hash12 位特征/特征段 ID，则输出 Hash12 位特征/特征段 ID，或如果接收到 CAM12 位特征/特征段 ID，则输出 CAM12 位特征/特征段 ID，或如果从 Hash 模块 12 和 CAM 模块 14 两者都接收到零值，则输出空闲信号。复用器 16 的输出由寄存器 18 接收。寄存器的每一个都输出 Hash12 位特征/特征段 ID 或 CAM12 位特征/特征段 ID，连同输出指示在网络数据的数据块中发现恶意内容的标志，或输出空闲值。把这些从特征检测电路 1 输出到 DPI 系统，以便在那里使用。因为特征/特征段 ID 只有 12 位，因此，与 32 位特征/特征段相对照，例如就存储它们所需的内存方面而言，ID 比特征/特征段可更容易地使用。

在此实施方式中，特征检测电路 1 构成 DPI 系统的部分，如图 4 所示。

DPI 系统接收 IP 包，如在附图的下部所显示的。DPI 系统处理 IP 包以从其提取有效负载，如在附图的中部所显示的。特征检测电路被用来检测有效负载中的特征，如在附图的上部所显示的。这示出了，在检测到特征段时，整理这些特征段来形成完整特征，以便确定有效负载中恶意内容的存在。

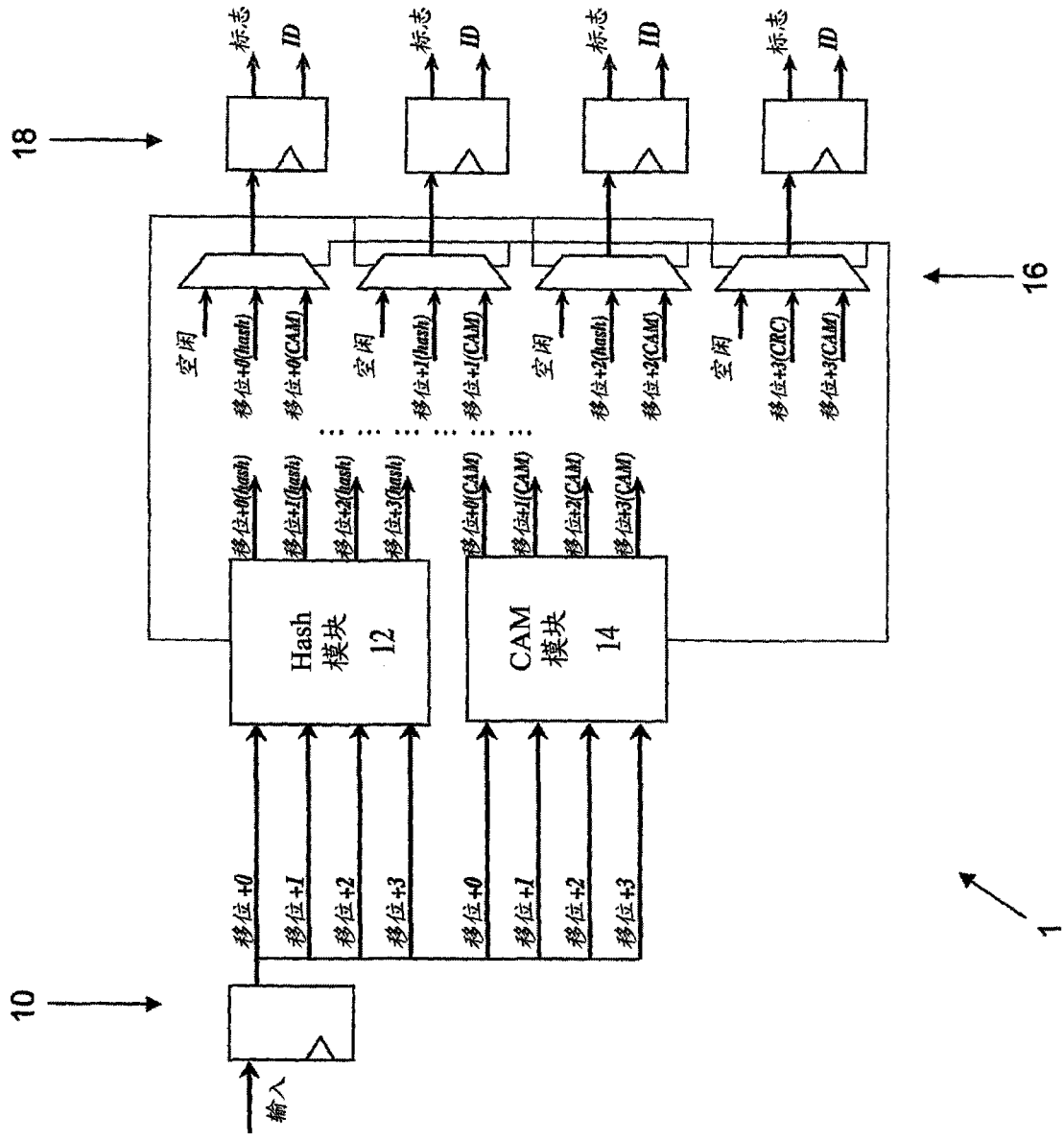

18
标志
ID
16
空闲
移位+0(hash)
移位+0(CAM)
移位+1(hash)
移位+1(CAM)
移位+2(hash)
移位+2(CAM)
移位+3(CRC)
移位+3(CAM)
Hash 模块 12
CAM 模块 14
移位+0
移位+1
移位+2
移位+3
10
输入
1

图1

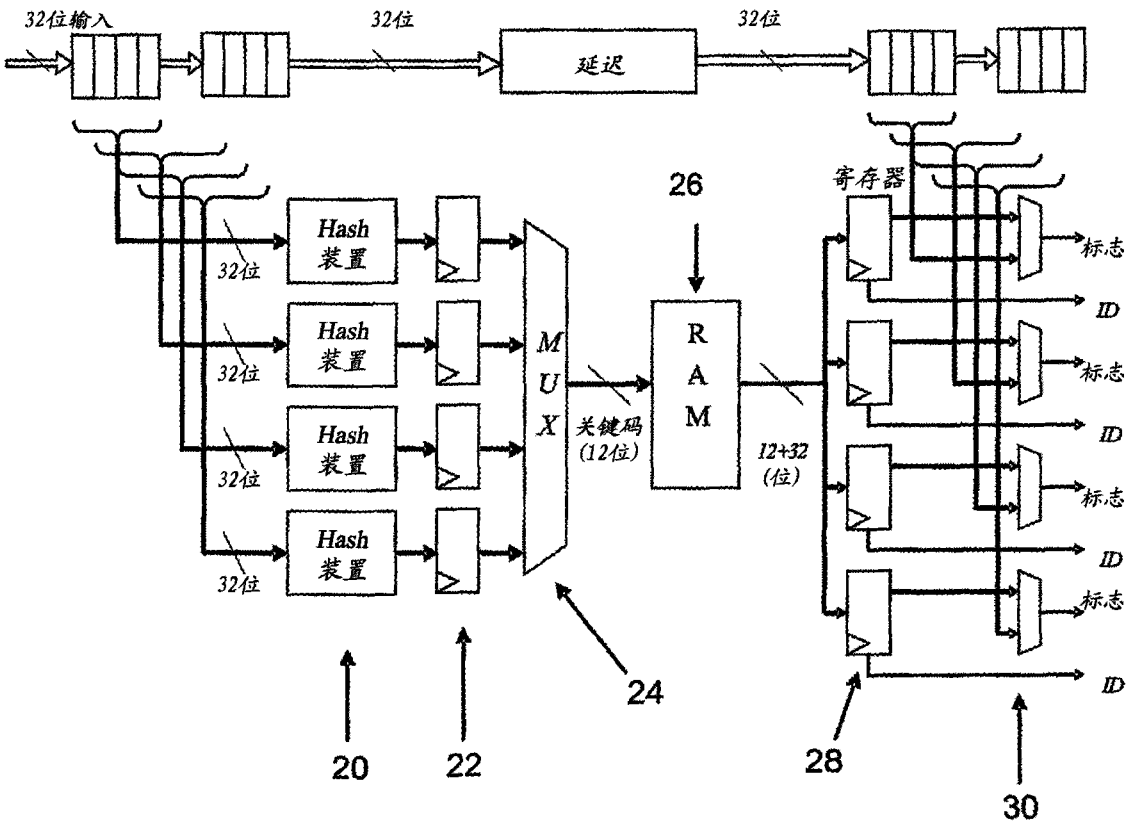

32位输入
32位
延迟
32位
寄存器
26
Hash
装置
32位
MUX
关键码
(12位)
RAM
12+32
(位)
标志
ID
20
22
24
28
30

图2

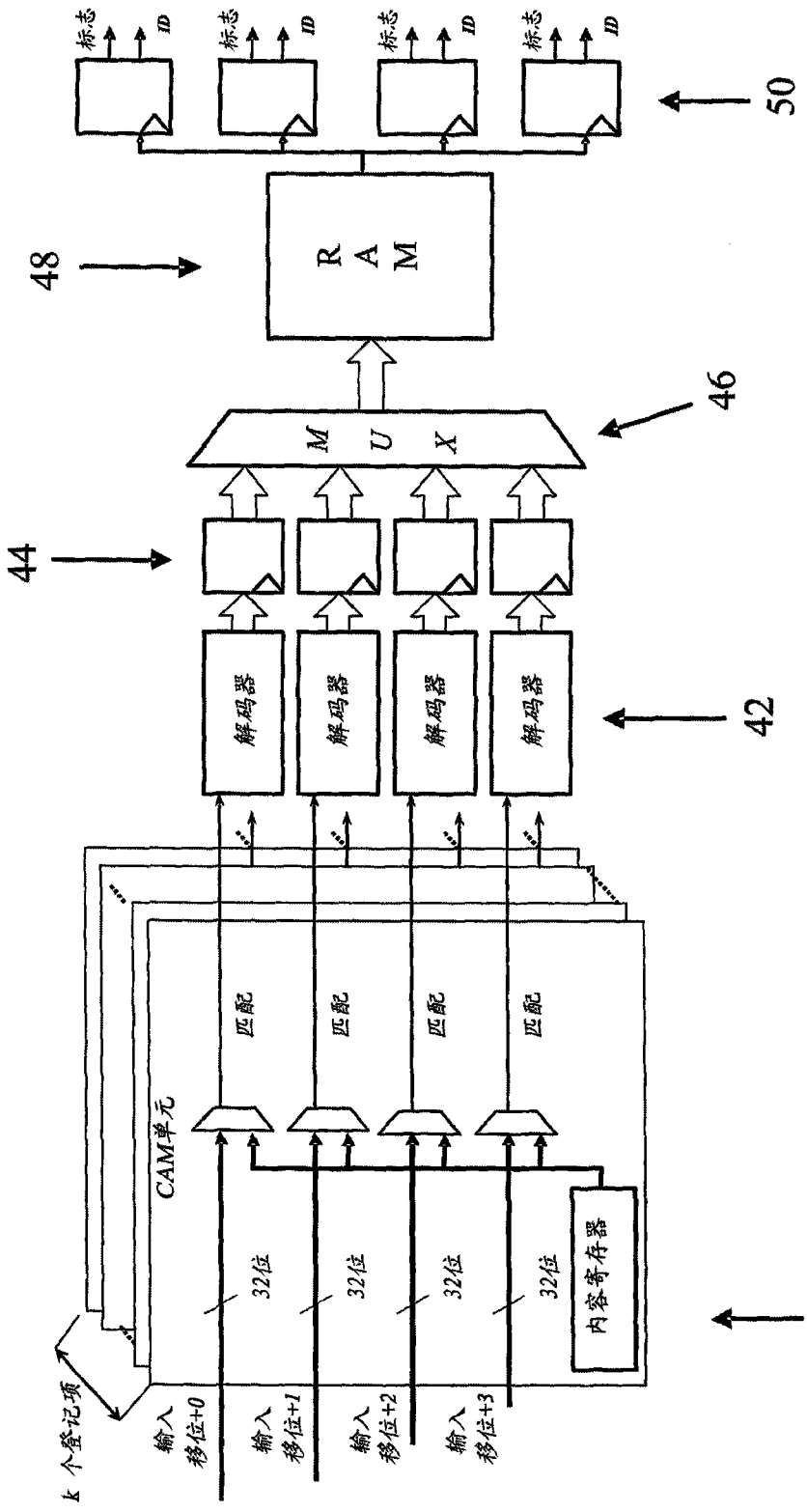

标志
ID
50
48
RAM
46
MUX
44
解码器
42
匹配
CAM单元
32位
内容寄存器
40
k个登记项
输入移位+0
输入移位+1
输入移位+2
输入移位+3

图3

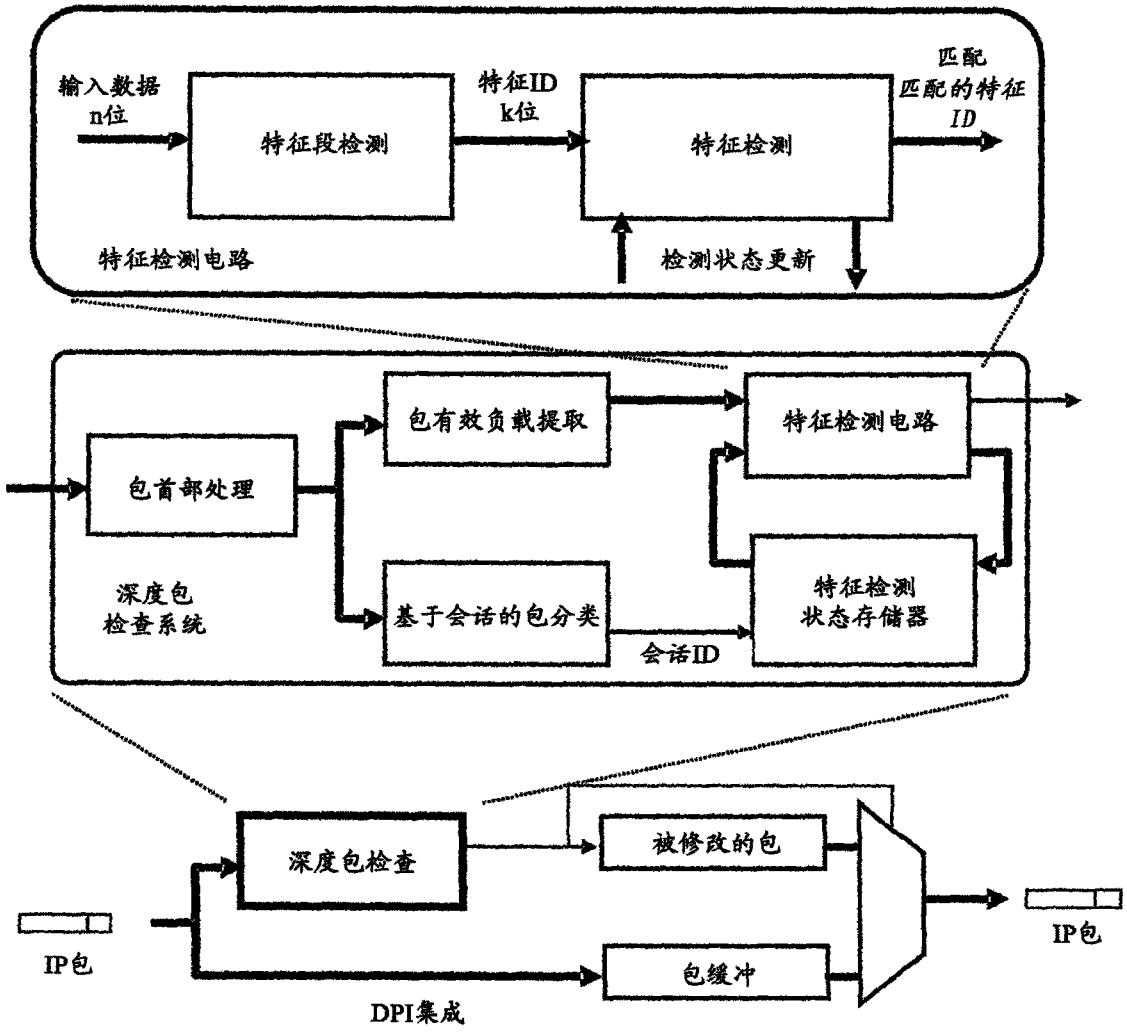

输入数据
n位
特征段检测
特征ID
k位
特征检测
匹配
匹配的特征
ID
特征检测电路
检测状态更新
包首部处理
包有效负载提取
特征检测电路
深度包
检查系统
基于会话的包分类
特征检测
状态存储器
会话ID
深度包检查
被修改的包
IP包
包缓冲
IP包
DPI集成

图4